UMBC

DISSERTATION APPROVAL SHEET

Title of Dissertation: Learning Natural Language from Probabilistic Perceptual Representations with Limited Resources

Name of Candidate: Nisha Pillai

Doctor of Philosophy, 2021

Graduate Program: Computer Science

Dissertation and Abstract Approved:

Cynthia Matuszek

Cynthia Matuszek

Assistant Professor

Computer Science and Electrical Engineering

7/24/2021 | 11:31:31 AM EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

NISHA PILLAI

nmbpillai@gmail.com www.github.com/nispillai
+17188449866 npillai.com www.linkedin.com/in/nispillai

A Ph.D. graduate in Computer Science from the University of Maryland, Baltimore County, who worked with Dr. Cynthia Matuszek and Dr. Francis Ferraro. My research interests are in machine learning, natural language processing, or grounded language learning.

SKILLS

Languages: Python, Java, and Perl.
Tools: Scikit-learn, Keras, Tensor-flow, Pandas, Gensim, NLTK, Weka, SciPy.
Concepts: Data Science, Design Patterns, Data structure and algorithms.
Courses: Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, Robotics, Human Robot Interaction.

EXPERIENCE

Research Assistant (Machine Learning, Natural Language Processing, Robotics)
UMBC May 2016 - August 2021

- ❏ Designed and developed language learning systems that jointly learn the linguistic concepts and visual percepts using various probabilistic supervised and unsupervised algorithms and advanced robot learning of generic novel language from untrained users in the real world in limited resource settings.
- ❏ Thorough exploration of best of active learning approaches using point process modelling and probabilistic clustering in linguistically complex and visually differentiated real world perception and presented suggestions for what approach may be suitable given the complexity of a problem. And the proposed unsupervised probabilistic gaussian mixture model based active learning approach *reduced the annotation data by at least 5% in all complex settings*.
- ❏ Developed a generic category-free unsupervised techniques using deep generative variational autoencoder and *guaranteed a minimum linguistic grounding score of 0.45* compared to minimum baseline performance of less than 0.1.
- ❏ Developed statistical tools to *quantify the complexity of linguistic and perceptual sensor multi-lingual data* in order to aid grounded language learning to apply differentiated learning approaches in the cross-modal grounding.
- ❏ Conducted human-robot interaction studies on interactive robot learning and presented the observations to increase efficacy and ease in human-robot communication.

Teaching Assistant (Advanced Robotics, Human Robot Interaction)
UMBC Fall 2015 - Spring 2016

- ❏ Developed robotics and machine learning project plans and trained students on object grasping, object recognition, and active learning based human-robot interaction projects.

Teaching Assistant (Data Structures)
UMBC Fall 2014 - Spring 2015

- ❏ Taught students who needed extra support in data structure lessons and projects.
- ❏ Evaluated data structure assignments and projects that implement complex applications and analyze their asymptotic performance.

EDUCATION

- ❏ **Ph.D. in Computer Science** 2014 - 2021
  Advisors: Dr. Cynthia Matuszek, Dr. Francis Ferraro
  University of Maryland, Baltimore County
- ❏ **Chartered Engineer (India) in Computer Engineering** June 2020
  Institution of Engineers, India

RESEARCH INTERESTS

Machine Learning, Active Learning, Deep learning, Unsupervised Learning, Robotics, Natural Language Processing, Human-Robot Interaction, Artificial Intelligence.

CERTIFICATIONS

[1] Deep Learning Specialization Coursera
[2] Mathematics for Machine Learning Specialization Coursera

RESEARCH & PUBLICATIONS

[1] Neural Variational Learning for Grounded Language Acquisition. **Nisha Pillai,** Cynthia Matuszek, and Francis Ferraro. *IEEE International Conference on Robot and Human Interactive Communication (Ro-Man).* 2021

[2] Measuring Perceptual and Linguistic Complexity in Multilingual Grounded Language Data. Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. *34th International FLAIRS Conference(FLAIRS).* 2021

[3] Sampling Approach Matters: Active Learning for Robotic Language Acquisition. Nisha Pillai, Edward Raff, Francis Ferraro, and Cynthia Matuszek. *IEEE BigData (special session on machine learning in big data).* 2020

[4] Building Language-Agnostic Grounded Language Learning Systems. Caroline Kerry, Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. *IEEE International Conference on Robot and Human Interactive Communication (Ro-Man).* 2019

[5] Deep Learning for Category-Free Grounded Language Acquisition, Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. *NAACL Workshop on Spatial Language Understanding and Grounded Communication for Robotics (NAACL-SpLU-RoboNLP).* 2019

[6] Optimal Semantic Distance for Negative Example Selection in Grounded Language Acquisition. Nisha Pillai, Francis Ferraro, and Cynthia Matuszek. *Robotics: Science and Systems (R:SS) Workshop on Models and Representations for Natural Human-Robot Communication.* 2018

[7] Unsupervised Selection of Negative Examples for Grounded Language Learning. Nisha Pillai and Cynthia Matuszek. *32nd Conference on Artificial Intelligence (AAAI).* 2018

[8] Identifying Negative Exemplars in Grounded Language Data Sets. Nisha Pillai and Cynthia Matuszek. Robotics: Science and Systems Workshop on Spatial-Semantic Representations in Robotics. 2017

[9] Improving Grounded Language Acquisition Efficiency Using Interactive Labeling. Nisha Pillai, Karan. K. Budhraja, and Cynthia Matuszek. *Robotics: Science and Systems (R:SS) Workshop on Model Learning for Human-Robot Communication.* 2016

POSTERS

[1] Generic Purpose Visual classifiers - Learning from Fixed category to General category. Nisha Pillai and Cynthia Matuszek. *Women in Robotics III, Robotics : Science and Systems.* 2017

[2] Robot language Acquisition and Active learning approaches. Nisha Pillai and Cynthia Matuszek. *Grad Cohort, Computing Research Association - Women.* 2016

[3] Active Learning Strategies for Efficient Grounded Language Acquisition. Nisha Pillai and Cynthia Matuszek. *Becoming a Robot Guru, ICRA.* 2015

ACADEMIC PAPER REVIEWS

[1] Conference on Robot Learning (CoRL) 2019
[2] Conference Empirical Methods in Natural Language Processing (EMNLP) 2019
[3] Conference on Human-Robot Interaction (HRI) 2020
[4] ACM Transactions on Human-Robot Interaction (THRI) 2020
[5] SpLU-RoboNLP workshop at ACL 2021
[6] Conference Empirical Methods in Natural Language Processing (EMNLP) 2021

AWARDS AND GRANTS

- ❏ National Science Foundation (NSF) fund for attending CRA-W 2016
- ❏ National Science Foundation (NSF) fund for attending ICRA 2015
- ❏ UMBC Graduate Student Association fund for attending conferences 2015 - 2018
- ❏ Individual Excellence Award, VMware Software India Pvt Ltd. 2011

CO-CURRICULAR

President of UMBC ACM Student Chapter 2017 - 2018
UMBC Computer Science Senator 2016 - 2018
Organizer of UMBC ACM Research Symposium 2018
Program Committee member of SpLU-RoboNLP workshop at ACL 2021

ABSTRACT

Title of dissertation: Learning Natural Language from Probabilistic Perceptual Representations with Limited Resources

Nisha Mannukkunnel Balan Pillai
Doctor of Philosophy, 2021

Dissertation directed by: Assistant Professor Cynthia Matuszek
Department of Computer Science
and Electrical Engineering

The advent of artificially intelligent technologies has generated an explicit requirement to study the semantic comprehension of perceptual and semantic world experiences. My thesis focuses on designing an integrated grounded language acquisition system composed of linguistic and visual symbols generated by obtaining meaningful perceptual representations from a physically grounded world. Specifically, my research presents semantic models that holistically enhance language acquisition by enabling learning systems to construct concise, category-free language from visual content.

Definitive knowledge of a visual concept requires not only a precise understanding of its positive information (information that provides a valid inference about a subject) but also of its negative information (information that provides information about what a subject is not). Obtaining negative examples of language referents is a challenging problem; people tend to describe things that are true of a particular situation, rather than negatives about it [199, 64, 56]. To address this problem in information acquisition, in the first work I employed semantically inferred linguistic information to overcome the difficulty of naturally finding negative perceptual data. My experiments show that such semantic measures are effective in choosing positive and negative samples for perceptual

learning, thus reducing the need for explicit data collection.

My research also explores the complexities involved in multimodal language–visual grounding tasks. In the second work presented in this thesis, I quantify the complexity of linguistic and visual observations associated with multi-modal language acquisition to help researchers make informed design decisions that grounded language learning performance. I employ entropy-based and compression error-based metrics to quantify the diversity in visuo-linguistic grounding inputs. The results formalize the linguistic and visual complexity present in language acquisition tasks and provide insight into the cross-modal grounding performances to keep task success consistent in the following works.

Subsequently, in the third work, I present and explain how a correctly presented order of visual content accelerates language acquisition and makes it more efficient. I demonstrate the benefits of careful selection of representative and diverse samples from a pool of unlabeled visual representations using active learning techniques and advanced language acquisition. For this purpose, I utilize probabilistic clustering characteristics and point process modeling as active learning strategies. My research also explores the user experience side of interactive learning in grounded language acquisition using a joint model of vision and language.

Finally, this research presents a unified generative method that infers meaningful, representational, and latent visual embedding for generalizing language acquisition. Such a generative approach helps grounded language acquisition to move away from learning predefined categories and toward category-free learning. I tackle the problem of category-free visual language learning using unsupervised approaches. Experimental results indicate that the methods suggested are competent in building semantic, linguistic, and visual models and make grounded language acquisition more efficient.

Learning Natural Language from Probabilistic Perceptual Representations with Limited Resources

by

Nisha Mannukkunnel Balan Pillai

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Dr. Cynthia Matuszek, Chair/Advisor
Dr. Frank Ferraro
Dr. Anupam Joshi
Dr. Tim Finin
Dr. Stefanie Tellex

Dedication

To Brahman, who is being and non-being.

Acknowledgments

I owe all the people who have supported me a debt of gratitude to make this thesis possible. I am grateful to all my IRAL lab mates at the University of Maryland, Baltimore County, especially Carolin Kerry, Luke E. Richards, Kasra Darvish, John Winder, and Edward Raff, for their immense support during my research life.

I am beyond grateful to Dr. Cynthia Matuszek for standing firm with me with unending support, motivation, and guidance. I also sincerely thank Dr. Francis Ferraro for his support, understanding, and guidance. Without both of you, I could have lost my way. I thank both of you from the bottom of my heart for leading and inspiring me in times of failure and success. I would not be where I am today without your trust and care. I also thank my committee members, Dr. Tim Finin, Dr. Anupam Joshi, and Dr. Stefanie Tellex for your kindness and assistance.

Finally, I thank my grandmother, late Mrs. Kamalakshi, for the inspiration and emotional support. I express my gratitude to my family, Mr. Balan Pillai, Mrs. Devaki Amma, Dr. Nitheesh, and Ms. Neethu, for your kindness. I am indebted to my friends, Ms. Atulya, Mr. Sibi, Dr. Sandeep Nair, Dr. Sudip Mittal, Dr. Karan Budhraja, Dr. Ashwinkumar, Dr. Nilavra Pathak, Mr. Ajinkya Borle, Ms. Sai Sree Laya Chukkapalli, and Mr. Vishnu for being there to guide me in all my decisions.

I appreciate all who helped me these years. Thank you all!

Table of Contents

List of Tables vi

List of Figures viii

1 Introduction 1
- 1.1 Motivation and Goals . . . 1
- 1.2 Roadmap and Contributions . . . 2
  - 1.2.1 Deriving a Counter-Perspective for Visual Experiences . . . 3
  - 1.2.2 Estimating the Complexities in Visual and Linguistic Experiences . . . 3
  - 1.2.3 Optimized Active Learning for Language Acquisition . . . 4
  - 1.2.4 Comprehensive Representation of Visual Experiences . . . 5
- 1.3 Organization of Dissertation . . . 6

2 Related Work, Background, and Data Corpus 7
- 2.1 Data: Corpora and Features . . . 7
- 2.2 Grounded Language Acquisition . . . 10
  - 2.2.1 Joint Model of Language and Vision . . . 15
- 2.3 Relevant Linguistic Concepts . . . 16
  - 2.3.1 TF*IDF . . . 17
- 2.4 Counter-Perspectives for Visual Samples . . . 18
  - 2.4.1 Paragraph Vector . . . 19
- 2.5 Linguistic and Perceptual Complexity Measures . . . 21
- 2.6 Active Learning to Improve Category-Based Learning . . . 22
- 2.7 Generalized Language Acquisition . . . 25

3 Unsupervised Selection of Negative Examples for Grounded Language Learning 27
- 3.1 Approach . . . 28
  - 3.1.1 Selecting Relevant Terms . . . 28
  - 3.1.2 Finding Negative Examples for Concepts . . . 29
  - 3.1.3 Classifier Learning . . . 31
- 3.2 Experimental Results . . . 31
  - 3.2.1 Selecting Terms . . . 31
  - 3.2.2 Negative Example Selection . . . 32
  - 3.2.3 End-to-End Quality of Trained Classifiers . . . 33
- 3.3 Discussion . . . 37

4 Measuring Complexities in Perceptual and Multilingual Data 39
- 4.1 Approach: Measuring Complexity . . . 40
- 4.2 Analysis: Linguistic Complexity . . . 43
- 4.3 Analysis: Visual Complexity . . . 46
- 4.4 Discussion . . . 47

5 Active Learning for Enhanced Grounded Language Acquisition 49
5.1 Fewer Descriptions and Better Learning . . . 50
5.2 Approach . . . 52
5.2.1 Learning Concept Classifiers . . . 52
5.2.2 Core Sampling Methods . . . 53
5.3 Experimental Setup . . . 56
5.4 Results and Per-Characteristic Analysis . . . 56
5.4.1 In-Depth Analysis of Active Learning Performance . . . 57
5.4.2 Pool vs. Uncertainty-Based Active Learning Methods . . . 59
5.4.3 The Impact of Visual Features . . . 60
5.4.4 Analysis with Different Classifiers . . . 61
5.4.5 Analysis with Different Datasets . . . 62
5.4.6 The Impact of Seed Language . . . 62
5.4.7 Performance with Varying Data Size . . . 64
5.5 Analysis of Results . . . 65
5.5.1 Method-Specific Findings . . . 65
5.5.2 General Considerations . . . 67
5.6 Discussion . . . 69

6 Generalized Category-free Grounded Language Learning 70
6.1 Learning Beyond Constraints . . . 71
6.2 Approach . . . 72
6.2.1 Unified Discriminative Learning Model . . . 73
6.3 Experimental Results . . . 75
6.3.1 UDM Specification . . . 75
6.3.2 Experimental Setup . . . 76
6.3.3 Limited Resource Classifications . . . 77
6.3.4 Multilingual Verification . . . 87
6.3.5 Highly Complex, Multi-Colored Resource Verification . . . 88
6.4 Discussion . . . 89

7 Conclusion 90
7.1 Future Work . . . 90
7.2 Synopsis . . . 92

List of Tables

4.1 Kolmogorov–Smirnov test results for each dataset and language, comparing trait *vs.* not-trait. D represents the maximum distance between the two samples' empirical CDF, i.e., the trait and non-trait cumulative distributions. All results are significant to at least p=0.013, with p-values provided in parentheses. This table shows that the UMBC dataset has fairly consistent color descriptions (larger K-S distances), but the UW-RGBD dataset, which contains more complex, multicolored objects, is less consistent (smaller K-S distances). The K-S distances for shape and object traits are smaller, indicating complex, varied descriptions. 42

4.2 A qualitative summary of typical complexity of sentences describing images, separated out by datasets (UMBC vs. UW RGB-D). 43

4.3 The average of the linguistic complexity comparisons between trait vs. non-trait for each dataset and language. Higher differences between average values indicate conciseness in the description; color descriptions were more concise than shape and object descriptions. 46

4.4 The average value of the visual complexity measures of color and shape distributions for each dataset. The smaller mean for my color complexity metric indicates a lack of variety in color features, whereas larger values for shape complexity are a result of the complicated edges and shapes in the feature set. 47

5.1 AUC summaries for each method's F_1 performance, grouped by the characteristic learned. All Active Learning (AL) techniques performed better in characteristic grounding by selecting significant points from the pool. 57

5.2 AUC summaries of F_1 performance for Pool and Uncertainty sampling performance, grouped by the characteristic learned. Uncertainty sampling (which depends on feature variability) does not perform well in object grounding, which has a noisy, highly varied data pool. 59

5.3 AUC summary results for each visual feature's F_1 performance for "object" characteristics. The DPP and GMM pools consistently outperform the baseline for both types of visual features (non-neural kernel descriptors and CNN features). 60

5.4 AUC summary results for each classifier's F_1 performance for "color" characteristics. Logistic regression can effectively classify types with diverse and meaningful points. 61

5.5 AUC summary results for each dataset's F_1 performance for COLOR. The GMM pool and GMM-DPP were able to consistently outperform the baseline, even with a multi-colored UW RGBD+ dataset. 62

5.6 AUC summaries for each method's F_1 performance, grouped by the characteristic learned. Both AL techniques performed better in characteristic grounding by selecting significant points from the pool. 63

6.1 Overall summary of the F1-score distribution comparisons of all concepts. The minimum, mean and the maximum of our method are higher than all baselines, with the UDM with 50 latent dimensions showing better learning especially for difficult categories. 77

6.2 F1-score performance of UDM in multilingual classification with less training data. UDM provided a consistent improvement compared to the category-free logistic regression baseline with both Spanish and Hindi training data. 88

List of Figures

2.1 Sample RGB images for each class in the UMBC dataset, as taken with a Kinect v2 camera and presented to Mechanical Turk annotators. 8

2.2 Sample RGB images in the UMBC dataset, as taken with a Kinect2 camera and shown to annotators [153]. In this visually varied dataset, shape and object classification are nontrivial. 9

2.3 RGB-D sensor data and descriptions [153]. Each concept was used by multiple annotators to describe each of the corresponding images, showing the noise and variability of human descriptions. 10

2.4 F1-Score results of Language acquisition with varying paragraph vector size. The results suggest that 1000 dimension give the best results. 21

3.1 Automatically selected terms and training data for grounded language learning. . . 27

3.2 Selected and discarded terms after tf*idf. Terms above the threshold (green) name a classifier that uses this object as a training example; terms below the threshold (red) do not. 29

3.3 Cosine similarity of the paragraph vectors of descriptive documents for a single banana in the UMBC dataset vs. selected other objects. Each PV represents an individual object in the dataset. 30

3.4 Precision (blue) and recall (orange) of term selection as the tf*idf threshold is varied. 31

3.5 Examples of AMT similarity results. Five participants selected which of the two choices was more similar to a target object. In the first row, most users selected the green arch, whereas the second row shows a less clear preference. 32

3.6 Performance of color classifiers for words (y-axis) versus the ground truth (x-axis). Only a small subset of representative classifiers is shown, as one is created for each keyword in the corpus. This confusion matrix shows the confidence of trained classifiers when run against objects of each type; for example, the trained model for the word "yellow" classifies the first object as positive with 93 % confidence but is only 20 % confident that the second object matches. Classifiers associated with color words have strong predictive power, as does the color classifier associated with the token "tomato." In contrast, the visually uninformative word "building" was not strongly associated with a classifier. 34

3.7 Performance of selected shape classifiers (x-axis) against objects (y-axis). The confusion between rectangles and arches is a product of ambiguity present in the data, as the blocks usually described as arch-shaped have a rectangular top. This confusion matrix shows the confidence of trained classifiers when run against sample objects of each type. 34

3.8 Performance of selected object classifiers (x-axis) against objects (y-axis). This confusion matrix shows the confidence of trained classifiers when run against sample objects of each type. 35

3.9 Average performance of color, shape, and object classifiers. Negative data were selected randomly (gray), using all non-overlapping objects (red) and using my dissimilarity measure. Incorporating meaningful negative examples improved performance in every category. 36

3.10 Average cross-validation performance of classifiers for words. In general, color classifiers (top left) performed best, although the outlier, purple, reflected the color differences between the objects described as purple (typically eggplants, red cabbage, and plums). Classifiers for object types (bottom left and right) generally performed well. Shape classifiers (top right) performed worst, resulting from the fact that people do not provide a shape description as often as they do in the other classes. 36

4.1 Comparison of traits "Color," "Shape," and "Object" via lexical entropy for the UMBC and UW RGB-D+ datasets. The K-S statistics quantify the divergence within each facet (subplot). Note that the entropy for color concepts is lower than for non-color concepts, indicating the concise, less varied vocabulary used to describe colors. The object trait entropy was higher, indicating linguistic variability. Only 9.5% of the UW dataset instances had shape concepts in the description at least once. Spanish descriptions contained varied but semantically similar shape/object tokens in their vocabulary. 44

4.2 Visual complexity of "color" & "shape" for both datasets. Lower standard deviations are a good indication of greater visual color consistency. The left-skew of the compression errors illustrates the high variations of the "shape" concept. 46

5.1 Performance of **visual classifiers** for Object type as the learning progressed with varying data sizes. In total, 216 distinct object images and their annotations were used in training. The F_1-score is shown on the y-axis, and the number of data samples is shown on the x-axis. The VL-GMM approach showed promising performance for more complex shape and object classification problems. However, the addition of noisy, highly varied descriptions during training affected the consistency in classification. Linguistic variability within the description caused the VL-GMM performance to oscillate as it learned the language during training. 64

6.1 Design diagram of the unsupervised concept grounding using the latent feature discriminative method. For every object, I extracted visual features and trained a representative feature embedding by applying a latent feature discriminative model. The visual variation encoder (v_{enc}) embeds the cumulative visual features to a low-dimensional feature representation, and the visual variational decoder (v_{dec}) decodes the embedding of the visual features. The extracted low-dimensional feature embeddings are then used to create a concept classifier ($c_{concept}$) for language grounding. 71

6.2 Comparison of the F1-score results with hidden dimensions ranging from 100 to 700. Hidden Dimension 100 and 700 with a latent dimension 50 shows better performance compared to other dimensions. 76

6.3 The comparison of the F1-score distribution of all concepts of the unified discriminative method vs. category free logistic regression. The goal is a high F1 score with a smaller number of occurrences, so the upper left quadrant (shaded) is the target. F1-score performance of UDM is both high and consistent with limited annotation. 78

6.4 The comparison of the F1-score distribution of all concepts of the discriminate method vs. convolutional neural network baselines (leftmost two bars). Though the averaged F1-score of SmallerVGGNet and UDM + SmallerVGGNet is 0.81, SmallerVGGNet scores are as low as 0.0 for some concepts. But the minimum of UDM + SmallerVGGNet is 0.37. F1-score of UDM with NASNetLarge is 0.73 which performs better than NASNetLarge where F1-score is 0.71. 79
6.5 Classification performance of UDM in different architecture variations with less training data. This thorough analysis considers two negative sample varieties (semantically dissimilar patterns as negative examples versus all except positive samples as negative examples), and feature input combinations (CNN features with and without RGB-D features). 80
6.6 F1-score distribution comparison of a CNN variant (SmallVGGNet) vs. UDM, for all concepts with varying annotation frequency (horizontal axis). I operationally defined setting using either 10% or 20% of the labeled data. The performance of the UDM was high and consistent, even with limited annotations. 80
6.7 Averaged macro F1-score comparison of the unified discriminative method against other approaches for every concept with RGB-D features. I segmented the classifiers by category here for ease of analysis as my UDM models do not consider category types. UDM with a latent dimension of 50 can provide promising performance in grounded language acquisition for all categories. Color-specific visual classifiers performed better than the category-free logistic regression baseline. Object and shape classifiers performed well with my method (UDM) with latent dimension 50 compared to other approaches. 82
6.8 Prediction probabilities of selected visual classifiers (x-axis) against ground truth objects (y-axis) selected from a held-out test set with RGB-D features. This confusion matrix exhibits the prediction confidence of the unified discriminative method (UDM) run against real-world objects. Color, shape, and object variations added complexity to performance. 83
6.9 The comparison of the F1-score distribution of all concepts of the unified discriminative method vs. baselines (leftmost two bars). Minimum, mean, and maximum F1-score performance of UDM using 50 latent dimensions is both high and consistent compared to both baselines and other latent dimension variants. 84
6.10 Averaged micro F1-score performance of visual classifiers. The unified discriminative method (UDM) shows improved performance than predefined category classifier where classifiers are learned per category and the category-free logistic regression where the concatenated feature set is learned per concept. 85

Chapter 1

Introduction

Recent advances in artificially intelligent systems shape human life and reduce the need for human effort in every facet of life, including education, manufacturing, and medicine. However, developing competent intelligent systems is difficult, as a myriad of factors contribute to how people understand and reason about their surroundings and experiences. Language and visual perception play a vital role in forming these mental models to think, express, communicate, and coordinate.

An intelligent system should therefore be able to learn from its surroundings to conceptualize, ground, and converse about its experiences, while interacting with untrained, naturally behaving humans. My thesis is centered around grounded language learning in which intelligent systems efficiently learn how linguistic constructs are grounded in the underlying perceived world.

1.1 Motivation and Goals

As Andrew Ng, the former director of Google's AI team, stated, "We may be in the eternal spring of AI [221]." The proliferation of intelligent systems and modern technologies has increased the popularity of AI-based approaches in the present-day digital world. As these technologies become more capable and affordable, the potential to deploy them in human-centric environments becomes more realistic. For these intelligent systems to be beneficial to society, they need to be able to take instructions from people in a natural, intuitive way [189], including instructions pertaining to—that is, grounded in—the specific environment in which they are operating.

An intelligent system internalize their experiences by analyzing the complexities of their environment, building semantic models based on those intricacies, and generalizing their conceptual learning to execute instructions from humans naturally in the real world. When operating in novel environments, such intelligent systems need to learn the language quickly and accurately from limited data. By enhancing learning, the intelligent systems provide innovative yet light-weighted services to humanity.

Semantic representations of real-world environments are powerful tools for supporting user interactions and action planning. My goal is to obtain such representations from conversations with users, allowing physically situated agents to learn appropriate world models "on the fly" for a wide range of situations. Learning these models from natural language provides a framework for learning semantics at the appropriate level of granularity in an intuitive and natural way. In this thesis, I intend to address the obstacles and challenges of interactive perceptual language learning.

1.2 Roadmap and Contributions

The primary objective of my thesis is to advance the learning of noisy and natural language associated with visual percepts, which robots encounter during real-world interactions with naturally behaving humans. My contributions to an end-to-end natural language understanding are in four areas: deriving negative perspectives of the world, estimating the complexities of visual and linguistic experiences, optimizing perceptual learning using active learning techniques, and deriving a generic semantic model of the perceived world. Taken together, these four areas of contribution demonstrate that an intelligent system with no prior knowledge can construct generic linguistic semantic models from fewer annotations and more efficient learning than existing approaches.

1.2.1 Deriving a Counter-Perspective for Visual Experiences

Building semantic models from natural language is challenging. The way in which people use language is frequently not a good basis for statistical learning systems. For example, descriptions of physical things rarely contain negative data; it is unusual for people to provide negative examples without prompting (e.g., objects are rarely described as "not yellow."). A lack of a positive label does not imply a negative grounding. Something described as "an apple" is not a good negative grounding for a "red" classifier. This problem affects parser learning [70], lexical acquisition [167], and human grammar acquisition [21, 108].

As an effective solution to this problem, I generated counter-visual perspectives using semantic similarity measurements to learn linguistic concepts. I also measured the effectiveness of these similarity metrics through human experimental evaluation. Additionally, I calibrated the differences in grounded language acquisition performance using various thresholds for the similarity metrics.

1.2.2 Estimating the Complexities in Visual and Linguistic Experiences

In *grounded language acquisition*, visual samples acquired from a physical or simulated context are used to drive language learning. Handling grounded language is a challenging problem for many reasons. When language is paired with real, physical data from robot sensors, groundings must be learned from often noisy, ambiguous, and complex channels. Although significant recent efforts have been made on grounded language learning [81, 198, 208, 225], there has been little emphasis on understanding the traits of the inputs involved. In this work, I aim to correct this by analyzing the visual and linguistic complexity of real, physical data.

In this work, I provide analytical, quantifiable statistical tools to represent the complexities

in visual and linguistic features. Entropy-based statistical measurements indicated the variations present in natural language descriptions of different traits of objects in the environment. I use these results to outline the reasons behind the difficulties involved in learning different traits from natural language. In addition, I explain how a compression technique can be used to measure the variance in visual images. Finally, I quantify the challenges of learning different visible traits from the environment.

1.2.3 Optimized Active Learning for Language Acquisition

Machine learning of grounded language often demands large-scale natural language annotations of things in the world, which can be expensive and impractical to obtain. It is not feasible to build a dataset that encompasses every object and every possible linguistic description of those objects. Novel environments require symbol grounding to occur in real time, based on inputs from a human interactor. Learning the meanings of language from unstructured communication with people is an attractive approach, but requires fast, accurate learning of novel concepts. People are unlikely to spend hours manually annotating even a few hundred samples, let alone the thousands or millions commonly required for machine learning.

In this work, I studied *active learning*, in which a system deliberately seeks information that will lead to an improved understanding with less data than more passive approaches, as a means of minimizing the number of samples/human interactions required to learn about a topic. This research compares pool-based and uncertainty-based active learning approaches in classifying distinct traits of real-world percepts. Additionally, I verified the performance differences with respect to different machine learning algorithms for pool-based active learning methods. For this,

variants of probabilistic-based Gaussian mixture models (GMMs) and determinantal point processes (DPPs) were considered. I studied which machine learning algorithms and which active learning approaches were able to achieve better performance with limited training data. Furthermore, I evaluated the effectiveness of neural-based features relative to traditional non-neural features in active learning. In summary, this research explores a principled study of active learning approaches in unsupervised data sampling techniques.

1.2.4 Comprehensive Representation of Visual Experiences

The joint modeling of language and vision [127, 153], wherein natural language is paired with sensor information to train visual classifiers, allows learning when both the language space and the perceptual space are novel. That is, such systems can learn novel words describing objects, attributes, or actions that do not already exist in the formal representation language. Although this method learns language groundings from visual features for multiple attribute classes, in previous attribute learning work classifiers were still trained for specific domains, such as object type or color. However, modeling semantics that are specific to particular attribute types still constrains language acquisition.

In this work, I present general visual classifiers that learn language without relying on predefined visual categories. My method generalizes language acquisition by using novel, generally applicable visual percepts from natural descriptions of real-world objects. I evaluate its efficacy by predicting the visual semantics of ground truth objects and comparing the performance with neural and non-neural baselines. I use a latent discriminative model to learn the low-dimensional representative features for category-free language learning. The performance of this approach

is assessed with a limited data training set and high resource training data. A detailed analysis of concept-wise performance elucidates additional information on the intricacies involved in the learning problem.

1.3 Organization of Dissertation

Chapter 2 provides details of the background and related work involved in my research. It describes grounded language acquisition, active learning approaches, paragraph vector (PV) models, and variational autoencoders (VAEs). It also lists existing past research works on grounded language acquisition and active learning approaches. Chapter 3 describes an effective unsupervised approach to obtain counter-visual exemplars from positive language descriptions. Chapter 4 is dedicated to measuring the complexities involved in grounded language acquisition in multilingual multimodal data. In Chapter 5, I analyze how different active learning sampling approaches influence grounded language acquisition. Chapter 6 presents an effective unified generative method to acquire shared semantic/visual embedding for learning natural language in a category-free manner. Finally, Chapter 7 concludes the thesis with the major contributions of this research and presents a description of ongoing work.

Chapter 2

Related Work, Background, and Data Corpus

As artificially intelligent devices such as robots have become safer and more capable, the idea of deploying them in situations where they interact with non-specialists (e.g., in homes, hospitals, or schools) has become more realistic. However, in order for non-specialists to interact with these robots, they need a means to communicate with them. Natural language is an intuitive and widely understood way of conveying instructions and information. However, building appropriate language models for a wide range of real-world situations and users is an enormous challenge, particularly in the area of *grounded language*, wherein language refers to objects and actions in a particular robot's perceptual world. To address this, I aimed to obtain representations from conversations with users, allowing systems to naturally learn world models applicable to a wide range of situations. Learning these models from natural language provides a framework for understanding such semantics at the right granularity in an intuitive, natural way.

In this chapter, I will cover the background, related work, and the datasets used in this thesis. In particular, background topics for my thesis include natural language processing, and statistical and machine-learning models.

2.1 Data: Corpora and Features

Throughout this thesis, I use two existing datasets for learning from descriptions: the UMBC dataset [152] which contains 72 objects (see Fig. 2.1), and the UW RGBD+ dataset [107], which

contains 300 objects.

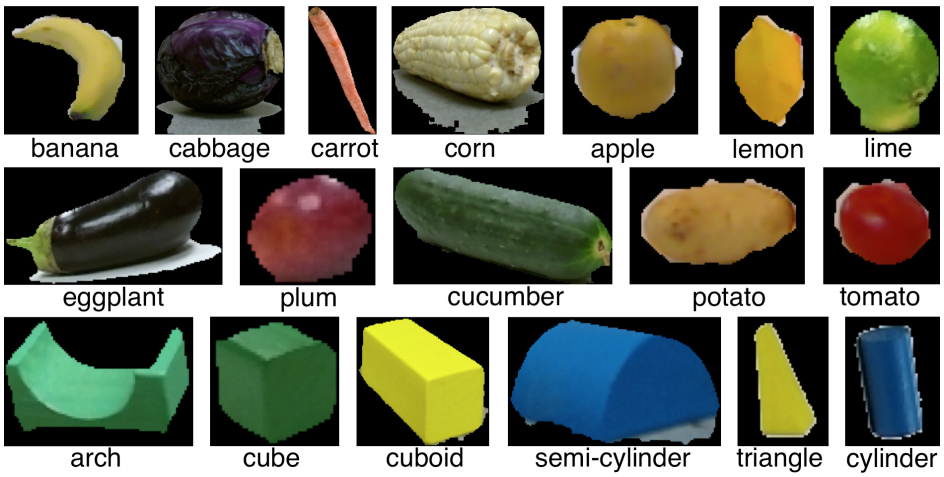

Figure 2.1. Sample RGB images for each class in the UMBC dataset, as taken with a Kinect v2 camera and presented to Mechanical Turk annotators.

Language. The corpora consists of Kinect2 depth images of objects paired with human descriptions. Each object instance has multiple associated language descriptions. The RGB images were posted on Amazon Mechanical Turk to obtain descriptive language, and users provided short descriptions. I use three different language descriptions for UMBC dataset: English, Spanish, and Hindi. While Kery [88] collected 5,100 Spanish and 5,700 Hindi descriptions for the UMBC dataset, I collected approximately 6000 natural language English descriptions for the same dataset(see Fig. 2.2). Similar to Richards and Matuszek [164], Kery et al. [89], I worked on learning to understand language referring to different types of characteristics: COLOR, SHAPE, and OBJECT TYPE (see Fig. 2.3). The UMBC object dataset contains 8 color, 9 shape, and 18 object characteristics. While the Spanish language dataset contains 35 color, 51 shape, and 138 object concepts, the Hindi descriptions include 25 color, 34 shape, and 135 object concepts. Shape and object concepts in both the datasets are highly varied and diverse, causing the classification to be difficult. Color concepts in the Spanish set seem concise and less varied, but gender-based inflectional differences in the

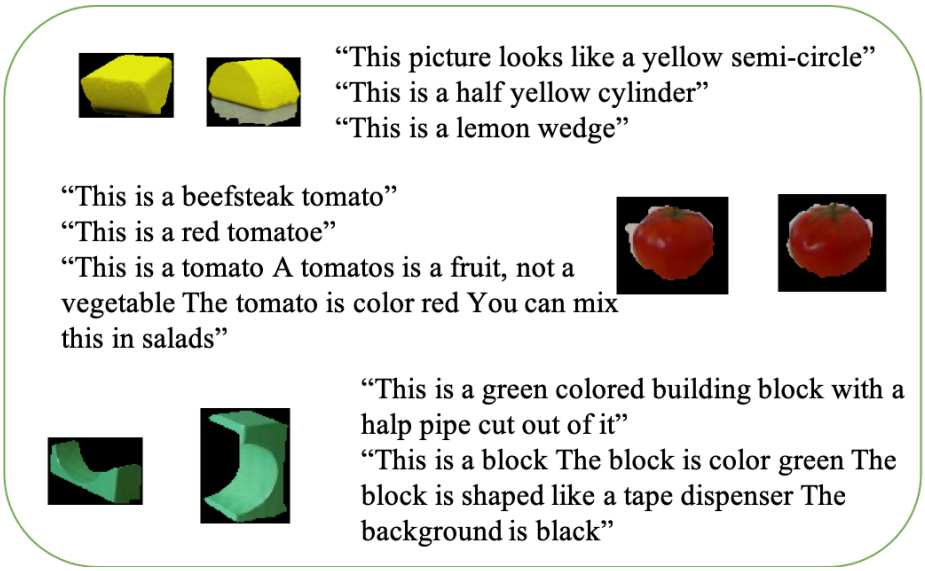

Figure 2.2. Sample RGB images in the UMBC dataset, as taken with a Kinect2 camera and shown to annotators [153]. In this visually varied dataset, shape and object classification are nontrivial.

description cause the color concepts diverse in the Hindi set. For the second dataset, I used the 1500 English descriptions collected by Richards and Matuszek [164] for UW RGBD+ dataset. The UW RGBD+ dataset includes 14 color, 13 shape, and 51 object characteristics. Shape concepts are reported only in 9.5% descriptions of all 300 object UW RGBD+ dataset annotations. In the UMBC dataset, 53%, 14%, and 73% of annotations reported color, shape, and object concepts respectively. Across all combinations I found that COLOR is relatively easy to learn; SHAPE, which depends in part on camera angle and is less likely to be mentioned, is more difficult; and OBJECT TYPE is the finest grained, with the highest visual complexity.

Visual Perception. The physical context for language grounding is provided by depth and color images of each object, taken with an RGB-D camera mounted on a typical of a robot sensor platform. From each RGB-D image, I extracted perceptual features η_{TRAIT} for each type of characteristic. I used three different types of visual features for learning. First, for a kernel descriptor-based approach, I used the average RGB values for color, HMP-extracted kernel descriptors [107, 19]

Type	Image	English annotation
color		This is an **orange** object.
shape		This looks like a green **upside down C shape**.
object type		This is an **Italian Eggplant**. It is firm and dark purple when ripe.

Figure 2.3. RGB-D sensor data and descriptions [153]. Each concept was used by multiple annotators to describe each of the corresponding images, showing the noise and variability of human descriptions.

for shape, and a combination of the two for objects. Kernel descriptors were model size, 3D shape, and depth edge from the RGB-D depth channel, and were efficient in shape and object classification. While this set of features does not use a neural network, I experiment with them because they are a proven and reliable set of features within the robotic-based vision and language processing [127]. In the second approach, I used a convolutional neural network, Neural Architecture Search network (NASNetLarge) [227], with pretrained ImageNet [46] weights for extracting a 1024-dimensional feature vector from the RGB images. It is proven effective in visual classification as Neural Architecture Search Network (NASNetLarge) obtained better top-1 and top-5 accuracy on multiple datasets compared to popular architectures like ResNet, Inception, and VggNet. In the third case, I extraced a 1024 dimension feature vector from SmallerVGGNet, a variant of the popular and strong object classifier VGGNet neural architecture Simonyan and Zisserman [183].

2.2 Grounded Language Acquisition

Describing human thinking as a symbolic system resulted in a myriad of ways for researchers to envision the future of artificial machines that think, “as we may think [24].” Five years after

that initial revelation, Turing [201] proposed the notion of intelligent machines in his famous *Mind* paper in 1950. Since then, to fit a wide spectrum, research in the field of artificial intelligence has widened its definition of what thinking is. The core element of this thesis fits within that widened spectrum, specifically into the category of *grounded language acquisition*, which is the integrated learning of language and environment [138, 69, 137, 142, 174]. The core question that I intend to answer through my research is, "How can an intelligent system effectively learn the semantics of language associated with its perception from very low resources?"

In grounded language theory, the semantics of language are based on how symbols connect to the underlying real world through the so-called "symbol grounding problem [69]." Recent researchers have addressed two main problems in this domain: language understanding and language generation [193]. Recent research, such as Song et al. [186], has proposed a model that integrates language grounding and language generation into one framework. My thesis contributes to this body of work, specifically to the effective language grounding of real-world percepts in low-resource settings. My work falls under three main areas: learning mappings between language and predefined object categories, learning mappings between language and generalized object perspectives, and improving learning from reduced resources in a robot-driven setting. My primary objective is an intelligent learning system that grounds the recognition object (defined by a set of visual percepts from the real world) to a canonical symbol, for example, producing the symbol 'eggplant' when it sees an eggplant. When a user asks, "Please grab me the eggplant," the robot should ground the natural language *word* "eggplant" to the same *symbol* that denotes the relevant visual percepts. Once both language and vision are successfully grounded to the same symbol, it becomes possible for the robot to complete the task. I teach the robot this connection by using physical sensors in conjunction with language learning; paired language and perceptual data are used to train a joint model of how

linguistic constructs apply to the perceivable world.

Grounded language acquisition is generally relevant for a wide range of applications. Similar to language generation, scene or image generation has also attracted attention in this domain. Cheng et al. [35] generated realistic scenes conditioned on natural language descriptions. They used object instances from the Internet to synthesize scenes. In another study, Yan et al. [212] generated an image from visual attributes (age, hair, gender, expression, etc.) using a VAE framework. In contrast, my efforts are in the opposite direction, namely to learn the grounding of language from a generalized perspective. A wealth of research has contributed toward the field of language generation, including on caption generation for news images [15, 53, 78, 102, 60, 76, 54, 55] and description generation from images of videos [202, 218, 14, 96, 97, 33, 215, 113, 147]. Barbu et al. [11] presented sentence descriptions that included properties of objects, such as color, shape, and size, from videos using natural language collected from AMT. My work also associates objects' characteristics with natural language from AMT, but my application focuses on robots learning the language in real-world settings. Higgins et al. [77] train a robot in combining fully simulated sensing and actuation with human interaction. Guadarrama et al. [66] developed a technique to describe the main activity taking place in a sample video by predicting approximate action words using a hierarchical semantic approach learned from the data. In a similar application, Huang et al. [83] introduced a visual language for visual storytelling. My work is related to this wide range of vision–language problems, but I focus specifically on learning the groundings of language.

Visual grounding, which involves relating expressions to images [117, 45, 213] and videos [223, 168, 214], has been effectively utilized in many different ways. The most popular 3D visual groundings [44] associate objects from segmented scenes with expressions. A recent approach [119] fuses the language and visual features to localize the relevant regions from the unsegmented, complete

RGBD image. However, similar to Jenkins et al. [85], I used depth images taken with an RBB-D camera mounted on a robot platform to map them with language attributes for my experiments. I later segmented the objects and extracted the visual features to learn the association between visual attributes and the related linguistic concepts. The problem space considered in this thesis assumes that there are no pre-existing models of language or objects in the world. That is, an agent learns from novel, previously unencountered language about previously unseen objects [127, 200, 126, 193], making the evaluation more broadly applicable.

This thesis primarily focuses on learning the association between real-world objects and their natural language descriptions. In similar research, Cohen et al. [41] demonstrated a method to learn to disambiguate object instances within the same class. Krishnamurthy and Krishnamurthy and Kollar [103] developed a method that learned the mappings of categorical concepts and their relations from the statements with objects in RGB images. Chen et al. [32] developed a method that learned the joint embedding between freeform text descriptions and colored 3D shapes, and Nguyen et al. [146] mapped the object-based descriptions of their usage. Nguyen et al. [143] learned a visual, linguistic association by sampling triples of anchor, positive, and negative data points from RGB-depth images and their natural language descriptions. Al-Omari et al. [4] grounded color, shape, location, relative direction, and relative distance as seen in video clips. These works are related to my research, although I instead focus on grounding natural language descriptions with object characteristics based on categories using depth and color images. Misra et al. [136], similarly, presented a model that was capable of handling the ambiguities present in grounding natural language instructions. Like the work composing this thesis, that research grounded the often incomplete and noisy robotic instructions. In another similar work, Duvallet et al. [50] treated language as a type of input sensor to formulate a prior distribution over the unknown parts of

the environment to ground natural language instructions in unknown environments. In a similar study, Chai et al. [29] teach a robot to ground actions by communication and action demonstrations involved in interactive task learning using action demonstrations.

Vision- and language-based navigation is another popular research domain that generates associations between visual and linguistic stimuli. Nguyen et al. [145] presented a grounded vision–language task for finding objects in indoor environments with human assistance, with subtasks such as "Go forward three steps, turn left." In a similar study, Nguyen and Daumé [144] used a simulated human assistant that provided natural language and visual instructions to direct the robotic agent. Anderson et al. [6] provided the Room-to-Room dataset for navigation based on crowd-based navigation instructions. Finally, Jain et al. [84] provided a new metric with which to evaluate the end goals of a navigation sequence by including an evaluation of the sub-goals related to the success. While these researchers aim to execute the commands, my work focuses on grounding real-world objects with natural language.

Deep networking models have been successfully used in visual–linguistic mapping, including bilinear attention networks [91], dual-attention networks [139], co-attention networks [220], faster R-CNN [219], and reinforced encoder–decoder networks [224]. I have also used VAEs and other popular deep networking models to retrieve object features and obtain meaningful visual embeddings for building a semantic language association [165]. The semantic embeddings of language descriptions can also be obtained using the PV deep network model [109] to obtain similarity metrics of visual perspectives. Other deep network models have been widely used in the natural language processing domain. Multilingual BERT [47], RoBERTa [120], XLM-R [43], cross-lingual XGLUE [116], among others, generate pretrained language embeddings for multiple vision–language tasks. On the contrary, for this thesis I generated language embeddings from the

descriptions provided in real time for my semantic embedding models.

Although most studies in grounded language have focused on English [73, 39, 10], there is a growing interest in multilingual grounded language. These efforts encompass both image captioning [15, 53, 76, i.a.] and learning spatial relations [13, 52] domains. Kery et al. [89]'s work on multilingual grounded object description is most relevant in this regard. Similar to Kery et al. [89], my work also primarily focuses on visio-language grounding, but I analyze the linguistic intricacies involved in multi-lingual visual groundings.

Thus, although significant recent efforts have been made on grounded language learning [81, 198, 208, 225], handling grounded language remains a challenging problem, in part because groundings are learned from noisy, ambiguous, and complex channels.

2.2.1 Joint Model of Language and Vision

Throughout this thesis, I used the *joint model of vision and language* [127, 154], in which a joint model of language and perception is used to acquire groundings for language that describes the perceived characteristics of objects in the environment. This model allows for learning words that have no pre-existing counterparts in the underlying formal grammar. In this approach, descriptions are treated as labels for visual percepts, making it possible to learn novel languages that describe entirely novel visual concepts.

Upon encountering an unfamiliar language token, the system creates visual classifiers trained on percepts (for the vision system) and associated with tokens or keywords (for the language system). Visual classifiers and language learning are treated as a single joint model with a shared learning objective. The result is a set of visual attribute classifiers that identify objects in a scene as they are

referred to in language.

As training data is added, the joint model obtains an expanding group of classifiers, and repetition of object attributes reinforces classifiers related to the corresponding keyword. In this way, keywords that refer to attributes of an object are likely to collect positive examples consisting of similar percepts (similar colors, in this example). Such classifiers (*e.g.*, associated with the word "red") will therefore gain predictive power. Classifiers with good predictive power then collectively form the grounded language model, whereas classifiers with reduced predictive power (*e.g.*, associated with "its.") can eventually be pruned.

The model is *joint* in the sense that each classifier, when created, is associated with a language token, for example, `new-classifier-called-``red.''` This is a deliberately simplified classification problem, as the goal of this work is to enhance multimodal robotic learning, rather than to solve a novel vision problem.

2.3 Relevant Linguistic Concepts

To choose suitable language concepts for training visual classifiers, I relied on the well-known tf*idf algorithm [170], which can be used to determine the descriptive power of terms [170], their relevance to particular documents [226], or as a document similarity metric [169]. Previous research described a method to select positive language samples when generating advanced sentences to describe images by predicting the most likely nouns, verbs, scenes, and prepositions [215]. Similarly, Cheng et al. [35] described a method that processed screen descriptions through a pipeline of natural language components to identify verbs and their arguments.

In Chapter 3, I describe a unigram language model that I applied to learning visual classifiers

for grounded language acquisition. I first converted descriptions of images into language *concepts*, removing common stop words and lemmatizing the remainder. I then identified meaningful, relevant, and representative concepts by applying tf*idf [170], which yielded concepts such as "banana" and "yellow," while rejecting less predictive words, such as "object" and "look." Empirically, this helps identify domain-meaningful words for which classifiers can be learned. Using this metric, the score of a word decreases with the number of documents it appears in and increases with the number of times it appears in a document. Terms such as "image," "picture," and "object" appear in an overwhelming number of documents, but relevant terms such as "carrot" or "banana" appear disproportionately in fewer documents. For example, for the image instance of a tomato, the description could be "This is an image of a red tomato," resulting in "image," "red" and "tomato" as focal tokens.

2.3.1 TF*IDF

To select relevant terms whose meaning needs to be learned, I used tf*idf, which stands for *term frequency-inverse document frequency*, a well-studied metric reflecting how important a word is to a document in a corpus. The tf*idf value of a term *increases* proportionally to the number of times that term appears in the document, reflecting the relevance of the term to that document, and the value *decreases* with the number of documents containing that term, reflecting its discriminative power. Intuitively, if a term such as "cabbage" appears frequently in a document, it is important to that document, but common words that appear in many documents, such as "very," have less discriminative power.

In this work, I use the simplest definition of term frequency: $tf(t, d)$ is the raw count of

the number of times a term t appears in the document d. The inverse document frequency is the inverse logarithmic fraction of the number of documents that contain the term t from the set of all documents D. This gives the tf*idf value of t for a particular descriptive document d:

$$tf*idf(t, d, D) = tf(t, d) \cdot \log \frac{N}{|\{d \in D : t \in d\}|}$$

where $tf(t, d)$ is the number of times a term t appears in document d, N is the size of the set of documents $N = |D|$, and $|\{d \in D : t \in d\}|$ is the number of documents in which the term t appears. I use tf*idf to separate relevant concepts from the descriptions, which returns concepts such as "cucumber" and "green," while discarding unhelpful concepts such as "picture" and "thing." This statistical technique was used to filter important concepts throughout the research presented in this thesis.

2.4 Counter-Perspectives for Visual Samples

In my thesis, the next focus is to automatically generate negative perspectives in limited-resource settings. In Chapter 3, I describe a way to automatically find negatives for visual percepts in grounded language learning. The most straightforward approach is to explicitly collect negative labels [191, 48], possibly through crowdsourcing [192, 94] or gameplaying [194]. However, this may not be applicable for all methods of gathering language. Another possibility is to associate randomly chosen groundings with terms that are not used to describe those images [182, 38, 98]. Goyal et al. [65] used negative samples in their experiments by randomly choosing alternative samples from the dataset, including or excluding positives. However, because the language used in these descriptions is not exhaustive, this approach is noisy and may require manual cleanup [189]. Another practical

technique is to design language collection trials that either use objects that have no shared visual characteristics [127] or explicitly design trials that exhibit negative characteristics [172]. My work is most similar to the fully unsupervised label identification of Roy [167], but it uses document similarity metrics, rather than term clustering.

To find negative examples, I used a similarity metric to maximize the semantic distance between object descriptions. My selection of negative labels used the PV algorithm (see section 2.4.1), which learns representations of features from documents of varying lengths [133, 134]. I employed the Distributed Memory Model of Paragraph Vectors (PV-DM) for this study [109]. Previous language representations have used vector models and multimodal topic models for image retrieval [185, 118], whereas I used a vector model of language to measure the similarity between descriptions of images [155]. There exist several popular approaches for representing language concepts in vector embeddings, including GloVe [151], skip-thought [93], and FastText [20]. While Hu et al. [80], Jain et al. [84], Shi et al. [177] initialized language descriptions using GloVe word embeddings, Liang et al. [115] used FastText, and He et al. [71] used skip-thought for grounding natural language descriptions for videos. Goyal et al. [65], alternatively, embedded natural language instruction with a combination of pretrained GloVe word embeddings and a two-layer GRU [37]. Although these approaches are effective, the PV model is more memory efficient and is most suitable for "on-the-fly" semantic model representation.

2.4.1 Paragraph Vector

Paragraph Vector is an unsupervised learning algorithm that maps documents into a fixed-length feature vector that is robust against varying document sizes [109]. A neural network with one

hidden layer is used to derive the error gradients from the loss function, which is calculated using the probability of words in a visual context given the input terms. This model is used to measure the dissimilarity between descriptions. In the PV model, paragraphs and words within these paragraphs are mapped to vectors P and W, respectively. The non-normalized log-probability vector of P is calculated as follows:

$$y = b + Uh$$

Here, y_i is the non-normalized log-probability of a word in the vector, U and b are softmax parameters, and h is a vector formed by the concatenation of word vectors W and the PV P. Prediction of the "next word" in the context or "topic" of the paragraph is achieved using a softmax classifier. A fixed-length sliding window is applied to choose contexts. Here, w_1,w_2, ..., w_T denotes the sequence of words being trained on:

$$p(w_t|w_{t-k}, ...w_{t+k}) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}}$$

The average log probability is then maximized:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k}, ...w_{t+k})$$

Training is performed using gradient descent with backpropagation. The output is a fixed-length dense vector, as in a bag-of-words model, but retains the predictive power of a more semantically informed model. The trained PV represents the "topic" of a document and has shown good performance in predicting other terms that may be found in that document. PV maps every document to a

point in a fixed-dimensional space, irrespective of their varying description size; from my empirical analysis, 2000 dimensions provided sufficient representative power. Recent experiments (See figure Fig. 2.4) show that the paragraph vector with 1000 dimension performs the best.

2.5 Linguistic and Perceptual Complexity Measures

The success of grounded language acquisition using perceptual data (*e.g.*, in robotics) is affected by the complexity of both the perceptual concepts being learned and the language describing those concepts. Chapter 4 presents methods for analyzing this complexity using both visual features and an entropy-based evaluation of sentences [159].

Linguistic complexity has been investigated through numerous psycholinguistic approaches, including concreteness and imageability [122, 176, 74, 30, 75], cost of learning [12, 86], and the length of words in the text [112]. While Ferraro et al. [56] presented multiple syntactic-, concreteness-, and language modeling-based approaches for quantifying the complexity of vision-

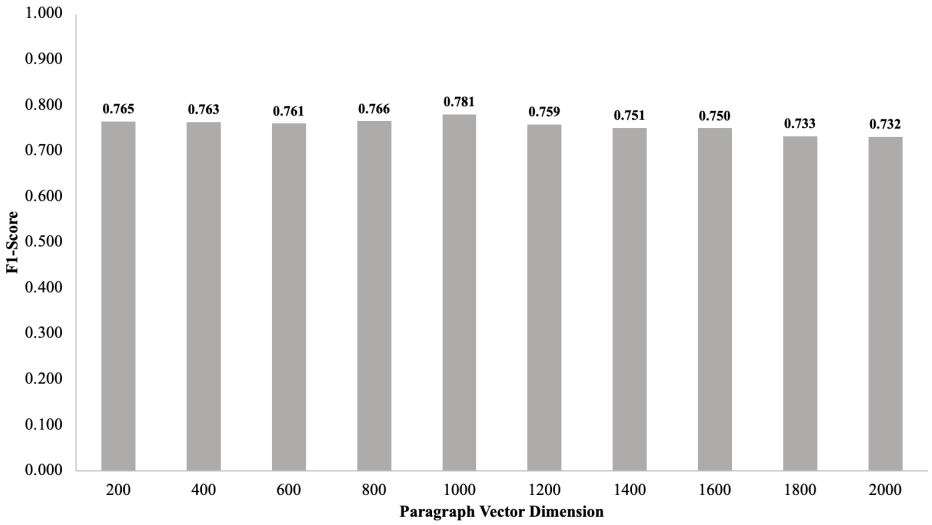

Figure 2.4. F1-Score results of Language acquisition with varying paragraph vector size. The results suggest that 1000 dimension give the best results.

.

and language-based *datasets*, I am interested in examining the complexity of semantic *traits* (i.e., categories of concepts) encompassed by those datasets. Therefore, many of these previous approaches are less relevant to the problems I study, in that I intend to quantify the differences in complexity, rather than discover the cognitive sources of those differences. Further, as my data are drawn from robotics, almost all of the concepts being learned are similarly concrete, and not subject to differences in human psychology.

Computationally measuring visual complexity in accordance with human perception is challenging. Human reactions can be influenced by familiarity, style, and other perceived factors, which are challenging to evaluate and operationally define [124]. However, I intended to find an automated measure for the *concept complexity* of an image [135]. As shape complexity varies widely [9], compression techniques were used for this category [57, 49].

2.6 Active Learning to Improve Category-Based Learning

It has become increasingly apparent that an interactive learning system in which robots learn from unstructured communication with untrained users is a necessary requirement in many contexts of human—robot interaction. Chapter 5 aims to demonstrate a technique to learn widely diverse, real-world objects with minimal labeling efforts, and to develop an approach wherein robots query unclear information from users and can learn from their responses about the meanings of words used to describe objects.

Active learning has been applied successfully to a variety of problems [95, 179, 23], providing performance improvements in areas as diverse as learning from demonstration [27, 22], following directions [72], and learning about object characteristics [196]. Although a well-chosen active

learning approach can reduce the number of labels required for grounded language learning [5, 154, 114, 68], such an approach also raises questions about what queries to ask, and when to ask them [26, 190, 184, 110, 63, 162].

Advances in active learning techniques have improved the ability to identify the most useful data points. Unsupervised learning techniques, such as subspace clustering, have been used to identify influential points from a cluster [150]. A hybrid method that connects active learning and data programming [140] has shown improvements in the reduction of noisy data in large-scale workspaces [34]. Similar to my work, active learning approaches [67, 211, 178] have been effective in training biased and highly varied datasets. Additionally, researchers have put effort into utilizing different active learning methods, depending on the complexity of the problem [59]. Traditional active learning methods have also helped to improve performance in other tasks, such as finding data faults or in fake news detection [79, 16]. Although in this work I consider efficiency as being more important than time complexity, researchers have studied methods that are time efficient, especially in large-scale applications [82]. Similar to my research, Chhatwal et al. [36] compared two traditional active learning algorithms for selecting important points from a pool of training data. Although similar, I also consider distinct machine-learning approaches with small-scale and large-scale datasets in my comparisons. Various Bayesian techniques have also seen popular use in selecting diverse points as the most influential from a dataset [160], and I use different variants of determinantal point processes (DPP) to select distinct data points as the active learning technique in batch sample selection [157].

Melville and Mooney [129, 130] developed a query-based committee model that used diversity as a factor to build different feature ensembles. In addition, that work demonstrated that this technique produced better results than traditional query-by-boosting and query-by-bagging

models. The same method can also be effectively used to choose diverse feature selections, and is used to enhance the quality of active learning by selecting appropriately diverse data points. In addition, Melville et al. [131] used uncertainty sampling as a technique for effective feature value acquisition. The selection of quality samples is beneficial in reducing costs associated with misclassification. Finally, Melville et al. [132] used similarity measures on probability distributions to estimate good class probability estimates.

In this study, my goal was to perform a principled exploration of selecting what data to query for labeling [121], and basing that decision on informativeness and uncertainty metrics [204, 23] from grounded language problems of varying complexity. I draw on existing techniques, particularly pool-based learning [222, 100, 188], uncertainty sampling [111, 99], and probabilistic sample selection [171]. I took inspiration from this body of research to select my set of experimental approaches, which included sample selection via GMMs [42, 90] and DPPs [106], which have proven effective in modeling diversity [61, 207]. Using supervised learners as active learning techniques [17, 188] is not suitable for our current study because I concentrated on building a language model capable of operating without prior knowledge [101].

My work is most closely related to that of Thomason et al. [197], who incorporated "opportunistic" active learning in a system that learned language in an unstructured environment [196, 148]. However, that work focused on opportunistically querying for labels whenever annotators were present; in contrast, this work focuses on exploring the best way of selecting choices from a large range of possible queries, reflecting the assumption that opportunities to query users for clarifying details will often be severely limited.

2.7 Generalized Language Acquisition

In the previous chapter, I described how visual classifiers used for language learning are often still trained for specific domains, such as object type or color. However, modeling semantics specific to particular attribute types constrains language acquisition. Concepts denote visual classifiers that are then trained with visual features extracted from a fixed set of semantic categories. Previous approaches are therefore limited to learning predefined visual categories, such as color, shape, and object words [153]. Chapter 6 presents general visual classifiers that learn language without relying on predefined visual categories. My method generalized language acquisition using novel, generally applicable visual percepts generated from natural descriptions of real-world objects. I propose a latent discrimination model to generalize the visual grounding using a VAE [158, 156]. My architecture predicts visual percepts associated with language by training the representative latent probability distribution generated from cumulative visual features using a deep generative model.

Representative embedding can be learned in several ways. Yang et al. [216] used VGGNet [183] to produce a representative embedding for images, whereas I used a single hidden layer, deep generative VAE model to generate latent embedding from visual features. Shridhar and Hsu [180] used a two-stage neural network model and the Visual Genome dataset to ground expressions, such as object name, color, and shape. In contrast, I focused on grounding expressions from very limited and sparse resources.

Autoencoding has proven useful for a variety of tasks, including image captioning [186], image-to-image translations [205], sign language translation [40], machine translation [217, 123], 3D shape analysis [187], hand pose estimation [203], sentence annotations [3], denoising [125],

and scene understanding [25, 161]. Silberer and Lapata [181] demonstrated a technique that uses a stacked autoencoder that grounded semantic representations of concepts by mapping language and vision into a joint embedding space. Although my objective is similar to theirs, they trained stacked autoencoders for every modality by treating them separately and fusing them in the last layer to obtain meaningful representations, whereas I combine all available raw visual features before feeding them into a deep network with no differentiation among the attribute types. Therefore, my technique has the benefit of requiring less supervised handling of the data. Nguyen et al. [145] presented a long short-term memory (LSTM) based encoder–decoder model for language assistance in vision-based navigation. Similarly, [80] encoded an instruction to actions in context with an LSTM and decoded using an LSTM decoder that conditioned the encoded instruction and visual features into output actions. Rohrbach et al. [166] also employed a deep network with LSTM to ground textual phrases in images with no, few, or all grounding annotations available. By contrast, in this work, I grounded the semantics of the learned concepts without specifying their attribute types, using the annotations from natural descriptions provided by AMT users.

In summary, in this chapter I explained the background, related work, and datasets used in this research. In the following chapters, I explain the set of research work that I conducted to enhance interactive robot learning performance in the physical world. In the next chapter, I present an effective approach to automatically select negative examples of visual concepts from descriptions using language processing techniques.

Chapter 3

Unsupervised Selection of Negative Examples for Grounded Language Learning

Modern applications of artificial intelligence have become increasingly dependent on their ability to learn from people. In addition, the ability to understand and predict *concepts* related to various objects based on human input is essential. My focus in this dissertation is to strengthen grounded language learning by building generic, efficient semantic models to facilitate language learning "on the fly." To this end, for optimized perceptual learning, a learner needs to have both positive and negative perspectives when operating in limited-resource settings. However, grounding language with perceived concepts is frequently hindered by a lack of negative descriptions of concepts. This chapter proposes an unsupervised system that learns visual classifiers associated with words, using semantic similarity to automatically choose negative examples from a corpus of perceptual and linguistic data.

I used statistical language processing tools to address two outstanding problems in grounded language learning. First, I automatically selected which terms to consider as candidate labels for

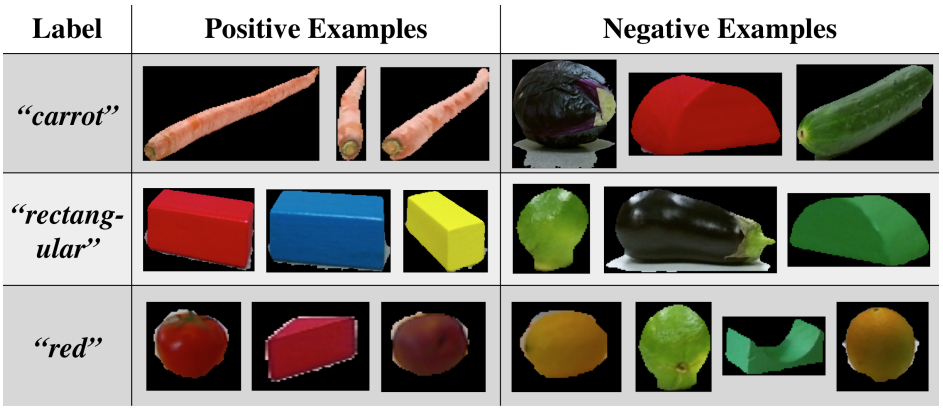

Figure 3.1. Automatically selected terms and training data for grounded language learning.

visual classifiers; and second, I used document similarity metrics to select appropriate negative examples from a corpus of training data (see Figure 3.1). I evaluated my approach with a novel dataset of objects and descriptions (the UMBC dataset), and the initial results support the idea that purely linguistic tools can be used to overcome weaknesses in the corpora of perceptual training data.

3.1 Approach

I build on previous work that treats the grounding problem as one in which words are associated with classifiers, jointly training classifiers and descriptive language to develop a semantic understanding of the visual characteristics of objects [127, 154]. I used a two-step approach: first, choosing relevant terms for which to train visual classifiers; and second, using semantic dissimilarity between descriptions of objects to find negative examples of those terms.

Specifically, I concatenated all of the descriptions of a particular object, and treated that concatenation as a "document" associated with that object. I then used tf*idf to find the most discriminative terms for a particular document, and used all the objects that people described using that term as positive examples for a classifier. I then chose negative examples by learning a PV for each document and using cosine similarity to find the most distant PVs.

3.1.1 Selecting Relevant Terms

To select words to learn,I employed tf*idf to find discriminative terms from the set of descriptive documents and passed it through an activation function to learn the importance of the term to that document. Currently, the activation function used here is thresholding, however in

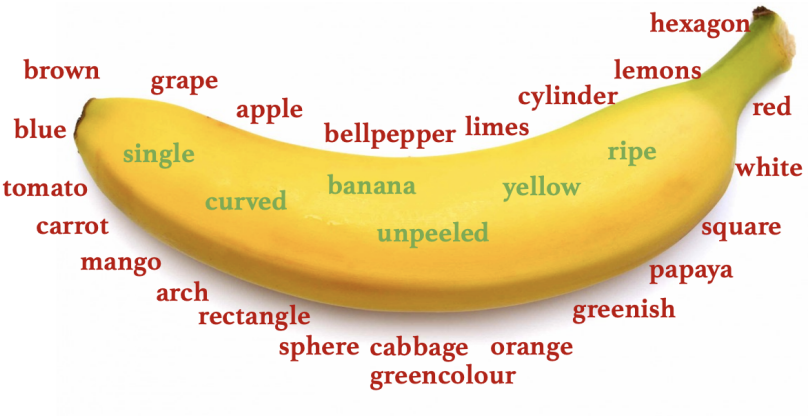

*Figure 3.2. Selected and discarded terms after tf*idf. Terms above the threshold (green) name a classifier that uses this object as a training example; terms below the threshold (red) do not.*

the future, it would be beneficial to experiment with more sophisticated, context-aware functions. Important terms are then used as labels for visual classifiers (see Figure 3.2 for examples). Varying this threshold affects the precision of the selection process.

For each term, all images that have been described using that term become positive examples for training a classifier. From the original 19,947 words used to describe 72 objects, 230 words were selected as tokens for classifier training. This process successfully screened out words that are used frequently when people are asked to describe objects, but that have poor discriminative or semantic power (such as "picture," "look," or "image").

3.1.2 Finding Negative Examples for Concepts

Based on prior research, I built a world model in which both the words being used and the concepts they describe are initially unknown. Once a set of images has been selected as positive training examples, the next step is to find dissimilar objects in the corpus to serve as negative examples. This presents a bootstrapping problem; counterexamples are critical to the efficient learning of word meanings [51] for a new term, but no classifier has yet been trained to

automatically select negative examples. However, semantic data are available in the descriptions of objects, and I expect that the descriptions of similar objects would be semantically similar.

A PV model was used to find the semantic distance between descriptions in the vector space, which can then be treated as reflective of the dissimilarity between objects in the world. All descriptions of each object were concatenated into an unordered "document," from which a PV was generated. The cosine similarity of these PVs then served as a distance metric (Figure 3.3). From a matrix of all cosine similarities, I chose objects with the most semantically dissimilar descriptions as negative training data. The experimental results validated this approach.

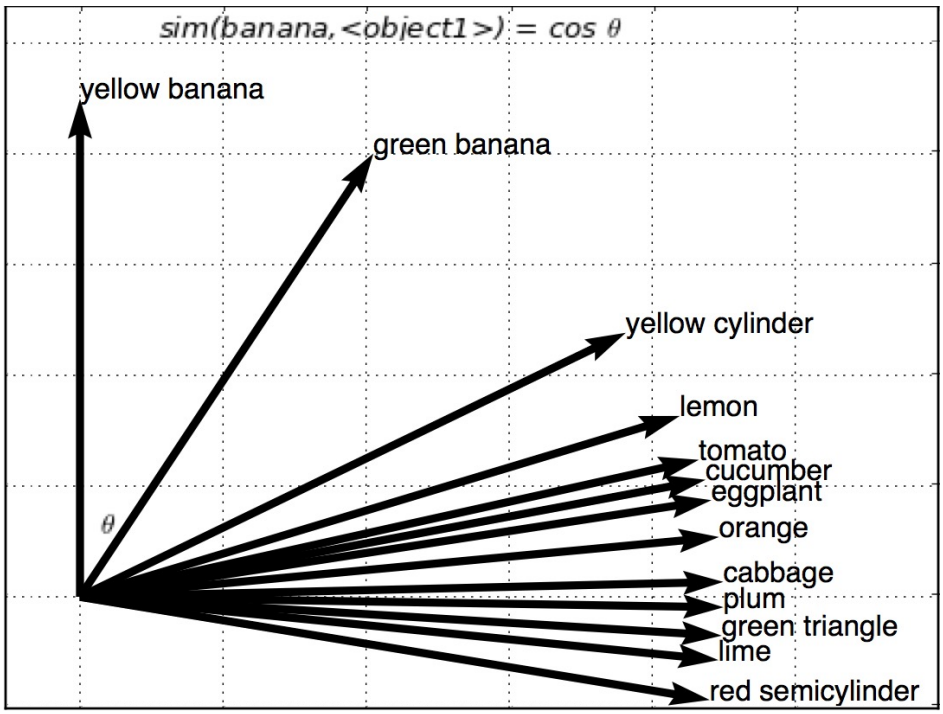

Figure 3.3. Cosine similarity of the paragraph vectors of descriptive documents for a single banana in the UMBC dataset vs. selected other objects. Each PV represents an individual object in the dataset.

3.1.3 Classifier Learning

First, I selected terms for which to create classifiers, as described above. For this perceptual learning problem, I use kernel descriptors extracted from RGB and RGB-D images of UMBC dataset objects. To test the effectiveness of the approach, I used three different types of classifiers: color, shape, and type of object.

3.2 Experimental Results

In this section, I present experiments that test each stage of the learning pipeline: selecting semantically meaningful words, finding negative training data, and assessing the quality of the final trained classifiers.

3.2.1 Selecting Terms

To evaluate the effectiveness of this approach for finding semantically meaningful words, I compared the results to the ground truth provided by human annotators. All unique words in the data set were assigned to two annotators to categorize them as “Visually meaningful” or “Not visually

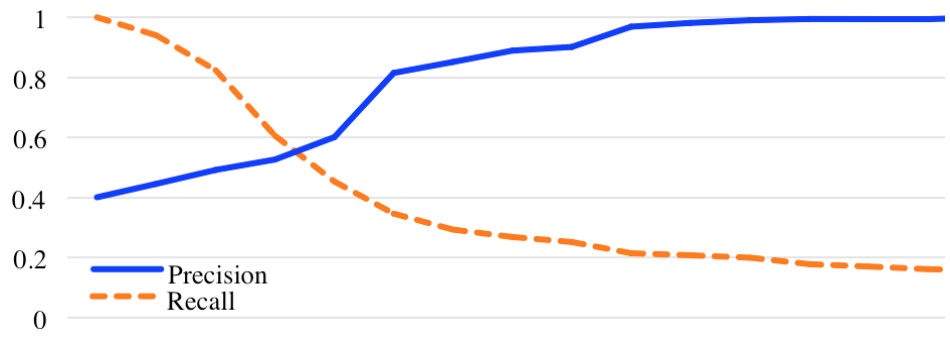

*Figure 3.4. Precision (blue) and recall (orange) of term selection as the tf*idf threshold is varied.*

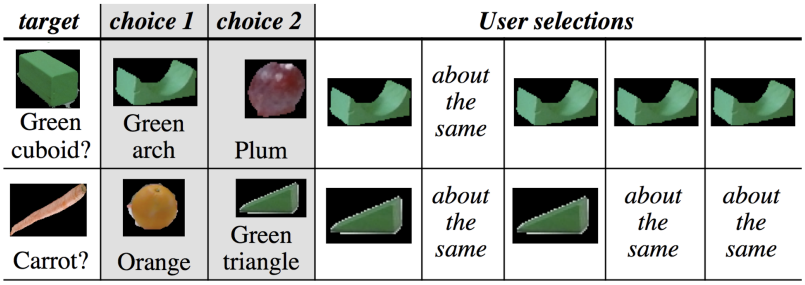

Figure 3.5. Examples of AMT similarity results. Five participants selected which of the two choices was more similar to a target object. In the first row, most users selected the green arch, whereas the second row shows a less clear preference.

meaningful."[1] Figure 3.4 shows the fluctuation in precision and recall as the tf*idf threshold used for term selection was adjusted. The proposed method provides promising results for determining the significance of words for which to learn visual groundings.

Discussion: As presented, this method selects preferentially for precision, i.e., reliably returns semantically meaningful terms at the cost of thoroughness. This is appropriate, as classifiers trained on visually uninformative words will result in poor predictive power and can be screened later. The purpose of term selection is to focus the learning effort on the most promising terms.

3.2.2 Negative Example Selection

One of the primary contributions of this study is a distance metric for perceptual training data that is based entirely on a paired, novel language. Using the PV model addresses a major failure in the simpler bag-of-words model, in that the PV model considers the ordering and semantics of words, but also still allows vector space–based comparisons. I treated the cosine distance between the PVs as an implicit distance in the grounding space (see Figure 3.3). Images of the most distant objects could then be used as negative samples for training the visual classifier (see Figure 3.1 for examples).

[1]For ease of annotation, the choice "error" was also provided, but only 27 words appeared to be an error.

As the "similarity" of objects is highly contextual, the ground truth for this distance metric is not clearly defined. I approximated the ground truth by using AMT infrastructure to ask people to evaluate object similarity. Because asking for a complete ordering of objects in the dataset was impractical, I tested a subset of cases, asking five annotators to decide which of the two objects was most similar to another. I presented 360 comparisons of the 72 objects in the dataset to five different evaluators for a total of 1800 comparisons. A simple majority of annotators agreed with my similarity metric in 84% of cases. Figure 3.5 shows examples of the results.

Discussion: The PV model was generally able to select good negative samples from the corpus, as evidenced by comparisons with human evaluators. Visual classifiers trained using these negative samples outperformed the baseline classifiers trained using random sampling from the dataset. A more complex evaluation of similarity with better-defined parameters might be appropriate in the future; for example, some users never considered color when designating similarity, whereas others clearly based their decisions on whether something was food or not. These are informed and reasonable aspects of similarity, but do not always align with the visual classifier training problem.

3.2.3 End-to-End Quality of Trained Classifiers

The quality of the grounded language model, that is, the trained model of the relationship between language and percepts, is a product of the association between the language tokens and the trained visual classifiers. Ideally, attribute descriptions should be associated primarily with a single classifier with good predictive power.

As a baseline, I compared the classification accuracy of the end-to-end system described in this paper with a model that chooses random negative samples and all non-overlapping samples

Color classifier denoted by "term"	Ground truth: yellow	red	green	white	orange
"yellow"	0.93	0.20	0.37	0.05	0.02
"building"	0.09	0.11	0.00	0.00	0.17
"red"	0.00	0.89	0.05	0.16	0.35
"green"	0.27	0.00	0.89	0.02	0.00
"tomato"	0.24	0.94	0.00	0.00	0.00
"white"	0.06	0.68	0.55	0.85	0.73
"orange"	0.50	0.93	0.21	0.26	0.66

Figure 3.6. Performance of color classifiers for words (y-axis) versus the ground truth (x-axis). Only a small subset of representative classifiers is shown, as one is created for each keyword in the corpus. This confusion matrix shows the confidence of trained classifiers when run against objects of each type; for example, the trained model for the word "yellow" classifies the first object as positive with 93 % confidence but is only 20 % confident that the second object matches. Classifiers associated with color words have strong predictive power, as does the color classifier associated with the token "tomato." In contrast, the visually uninformative word "building" was not strongly associated with a classifier.

Shape classifier denoted by "term"	Ground truth: cube	cylinder	sphere	arch	triangle
"cylinder"	0.32	0.87	0.06	0.29	0.29
"rectangular"	0.82	0.43	0.51	0.78	0.30
"circle"	0.25	0.25	0.75	0.26	0.21
"archshaped"	0.29	0.27	0.12	0.82	0.33
"triangle"	0.54	0.60	0.52	0.31	0.82

Figure 3.7. Performance of selected shape classifiers (x-axis) against objects (y-axis). The confusion between rectangles and arches is a product of ambiguity present in the data, as the blocks usually described as arch-shaped have a rectangular top. This confusion matrix shows the confidence of trained classifiers when run against sample objects of each type.

from the dataset. I used the same dataset to evaluate my method, the random selection method, and all other sampling methods. The evaluation was conducted using my corpus of images and descriptions. Cross-validation was performed for testing. As described above, I trained the color, shape, and object classifiers for all selected terms.

Color: The color classification results demonstrate good results for color labels (see Figure 3.6). However, there was some overfitting resulting from the relatively small set of objects. For example, objects were frequently described as being on a white background, leading to conflation in the

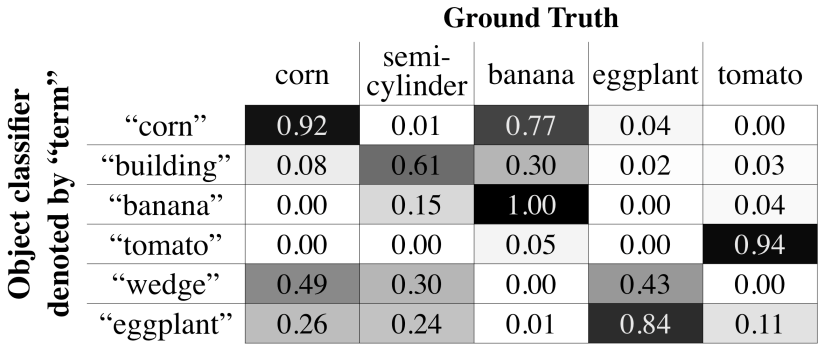

Figure 3.8. Performance of selected object classifiers (x-axis) against objects (y-axis). This confusion matrix shows the confidence of trained classifiers when run against sample objects of each type.

use of the word "white." The "orange" and "red" classifiers overlapped, in part because users described both tomatoes and carrots using both terms. In addition, polysemy had a negative impact, for example because the term "orange" can refer to the color or the object. Underfitting played a key role in the classification using the words "half" and "white," as there were fewer than 100 descriptions using these labels, when other labels received an average of 300–450 descriptions. One possible solution to the need for extensive annotation is the use of efficient active learning techniques. Previous grounded language acquisition experiments that utilized active learning techniques [154] have shown promising outcomes in reducing annotation efforts without compromising classification accuracy.

Shape: Figure 3.7 shows the results of some selected shape classifiers. Training shape classifiers on small RGB-D images was significantly more difficult than for color, in part because the shape of an object can vary considerably when viewed from different angles. Although the shape classifiers still performed well, the quality of the results was somewhat lower than for color. A few sources of complications included the tendency of annotators to not describe the shape of common objects; for example, cucumbers were frequently referred to as green, but never as cylindrical. In addition, certain terms, such as "rectangular" were overused, which influenced their classification success.

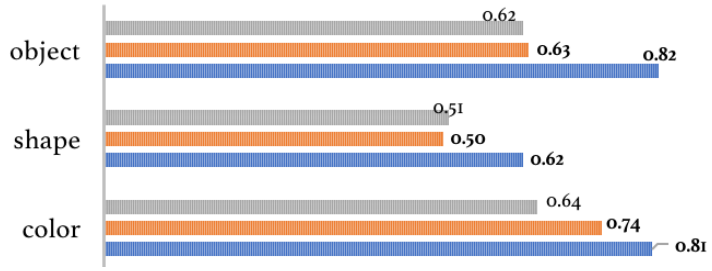

Figure 3.9. Average performance of color, shape, and object classifiers. Negative data were selected randomly (gray), using all non-overlapping objects (red) and using my dissimilarity measure. Incorporating meaningful negative examples improved performance in every category.

Object class: Object classifiers, which are intended to determine the class an object belongs to, are trained using a combination of color and shape features. Although the object classification demonstrated positive results with my dataset, this is partly due to the strong influence of color in classification; both the toys and the food objects in my dataset tended to be primarily a single strong color.

Overall: The proposed system convincingly outperformed two baseline models (see Figure 3.9), one that randomly selected objects to serve as negative examples, and one that used all other objects as negative examples, demonstrating that my method improves on the state of the art for unsupervised grounded language acquisition. A classifier trained with all other samples as

blue:	0.995	arch:	0.532	banana:	0.942	lemon:	0.777
green:	0.947	cube:	0.590	cabbage:	0.879	lime:	0.936
orange:	0.720	cylinder:	0.725	carrot:	0.887	orange:	0.921
purple:	0.499	rectangle:	0.621	corn:	0.922	potato:	0.715
red:	0.844	triangle:	0.649	cucumber:	0.615	tomato:	0.926
white:	0.772			eggplant:	0.646		
yellow:	0.918						

Figure 3.10. Average cross-validation performance of classifiers for words. In general, color classifiers (top left) performed best, although the outlier, purple, reflected the color differences between the objects described as purple (typically eggplants, red cabbage, and plums). Classifiers for object types (bottom left and right) generally performed well. Shape classifiers (top right) performed worst, resulting from the fact that people do not provide a shape description as often as they do in the other classes.

negative data performed well, whereas random sampling performed almost as well in most cases, but represented a fairer comparison in terms of training time and resources.

The overall goal of this work is to allow agents to improve their ability to learn semantic representations of their perceived environments, using natural language as the training signal. Although not a complete metric, one way of considering whether this work makes progress toward that goal is to verify that the most obvious terms for the intended ground truth have been identified as having important semantic relevance, and assess how accurately the classifiers associated with those terms perform on the complete dataset. Using this metric, I found that all of the ground truth labels were discovered using this technique. The classifier performance is shown in Figure 3.10.

3.3 Discussion

Although a number of different approaches have explored how to acquire semantic representations of perceptual data, the need for automated selection of learning targets and, especially, negative natural language exemplars is a frequent concern throughout the literature. The results presented here demonstrate that statistical tools from natural language can be applied to corpora of mixed language and perceptual data, automatically identifying terms that should be considered as candidates for learning groundings and automatically selecting negative examples for training classifiers. This reduces the need for human supervision, allowing language-learning agents to learn end-to-end in an unsupervised fashion, from collecting data to fully trained grounded language models.

An evaluation of this process for finding meaningful words and selecting negative examples suggests that these approaches are effective. These results illustrate the performance and effectiveness of the classification model by comparing it with two baseline models, one that randomly

selects negative samples, and one that uses all non-positive examples as negatives. I used the word-as-classifier approach because, although it is a simplification of the language problem, it is an applicable starting point for the robotic language understanding task to be applied to noisy perceptual data. This language model is preliminary, and I intend to extend this to a more semantically driven and context-sensitive model in the future. I also hope to see this research used in a conversational agent. In a conversation-based interaction, the system will have the opportunity to explicitly ask for negative examples, which I hope will improve the results. The approach in this chapter would then be useful in reducing the number of (possibly repetitive) questions and enhancing the quality of the dialog.

In the subsequent chapters, I effectively use this negative sampling approach with a more varied set of objects, additional classifiers, multi-color, multilingual, and more complex visual classification tasks. The automatic selection of meaningful counterexamples for the perceived world serves as a foundation for perceptual grounding in this dissertation's remaining research works.

Although the selection of suitable input samples supports grounded learning, analyzing the complexities involved in the visual and linguistic sensor information is necessary to select the appropriate techniques to enhance the learning quality. In the next chapter, I demonstrate the use of well-known statistical techniques to measure the variability and complexity of linguistic and visual data to examine the complexities involved in language learning tasks.

Chapter 4

Measuring Complexities in Perceptual and Multilingual Data

There has been significant recent research on grounded language learning [81, 198, 208, 225], but little emphasis has been placed on understanding the complexities of the inputs involved. Although having appropriate positive and negative samples is vital in language acquisition, knowing the intricacies of the language further assists in designing suitable learning approaches. This chapter presents methods for analyzing these complexities using statistical language- and image-processing approaches. These methods illuminate the core, quantifiable statistical differences in how language is used to describe different traits of objects and the visual representation of those objects. These methods provide an additional analytical tool for research on perceptual language learning.

Previous studies on language grounding [153, 89] have demonstrated that the amount of data required to learn about different traits of an item, such as its color, shape, or overall object type, varies significantly. There has been speculation that the "complexity," broadly defined, of the trait being linguistically described (or visually represented) is a key correlate to this varied performance. This is generally intuitive, although the lack of a clear quantitative measure limits the conclusions that can be drawn. Therefore, it is important to measure the complexities of learning tasks to find suitable techniques for improving language learning performance.

My primary contributions are: (1) the introduction of "trait-based" complexity to the AI and grounded language learning communities, and (2) the identification of appropriate metrics and statistical tools to measure the complexity of perceptual data and linguistic variability as a means of

predicting efficiency in grounded language learning. Language is measured using sentence-based entropy analysis and visual complexity is measured by examining visual features. I argue that this straightforward approach is beneficial, since it is simple to compute yet effective at discerning key differences in grounded language. I further argue that the complexity measures provide grounded language learning researchers with an additional tool for analyzing and understanding their data and underlying learning problems.

4.1 Approach: Measuring Complexity

Although perceptual and linguistic complexity are intuitive concepts for many people, they are difficult to verbalize or define. In general, humans are poor at providing numerical priors or rankings of subtle concepts, particularly over a very large dataset [7]. This study attempts to clarify the concept of visual and linguistic complexity individually. Accordingly, I introduce automated metrics here. Note that in identifying these metrics, I do *not* claim that they are the only possible metrics. Indeed, I argue that these complexity measures provide grounded language learning researchers with an additional tool for analyzing and understanding their data and underlying learning problems, and augment the tools already used in this domain of research. I hope that my work will encourage the community to begin examining these notions of complexity in their other efforts and across other grounded language tasks. Approximating the combined visual–linguistic complexity would be an exciting topic for future research.

Linguistic Complexity: This work calculates the linguistic complexity by computing lexical entropy, extracted for each concept from the descriptions. For every object instance i, I combined all the descriptions into a pseudo-document d_i. I calculated the frequency for every descriptive concept v in

d_i as $p_{i,v} \propto \text{count}(v \in d_i)$. I then computed the entropy h_i of the instance i: $h_i = -\sum_u p_{i,u} \log p_{i,u}$. Descriptions and entropies can be separated at the pseudo-document level depending on whether a characteristic word, e.g., "red," was used or not. As entropy reflects the diversity of language used to describe an instance, examining its variability helps explain the linguistic complexity of a trait. Although straightforward, this approach is consistent with the traditional concepts-as-classifier grounding approach used in previous work [173, 195, 1]. I then calculated the density estimates of entropy for the distribution of language describing a trait such as color vs. the distribution of language that does not describe that trait. Descriptions of one instance may include concepts of all three traits (e.g., "a round, purple eggplant"). I then calculated the linguistic complexity of a trait by combining the entropies of descriptions that referenced a concept associated with that trait. I compared them with the cumulative entropy calculated from all other concepts that were not related to that trait. For example, I combined all the descriptions of eggplant instances and calculated the entropy using every concept count. For color, I selected all the concepts associated with color (e.g., "purple"), and added the entropies of all the instances described by that concept. To calculate the distribution of non-color traits, I added all entropies that were not related to color. I categorized the concepts corresponding to each trait using Google Translate [209].[1] I performed a Kolmogorov–Smirnov (K-S) test to quantify these distributions. K-S tests are an efficient way of comparing two distributions (or samples from unknown distributions) with the null hypothesis that they do not differ. The K-S test returns the maximum distance *D* between the two curves, with *D* bounded by 0 (for identical distributions) and 1. The results are shown in Tab. 4.1. Although measures such as the KL and Jensen–Shannon divergence quantify the "difference" between two distributions, I use the K-S test because it not only provides a similar "difference" score, but also

[1] Available with code at `https://github.com/iral-lab/MultiModalComplexityEval`

Dataset	Language	*D*, Color Concepts	*D*, Shape Concepts	*D*, Object Concepts
UMBC	**English**	0.41 (1.63E-7)	0.28 (1.163E-3)	0.36 (9.0E-6)
	Spanish	0.58 (1.3E-14)	0.23 (1.3E-2)	0.39 (9.9E-7)
	Hindi	0.48 (5.4E-10)	0.41 (2.0E-7)	0.58 (2.8E-14)
UW	**English**	0.20 (2.2E-16)	0.56 (2.2E-16)	0.56 (2.2E-16)

Table 4.1. Kolmogorov–Smirnov test results for each dataset and language, comparing trait vs. not-trait. D represents the maximum distance between the two samples' empirical CDF, i.e., the trait and non-trait cumulative distributions. All results are significant to at least p=0.013, with p-values provided in parentheses. This table shows that the UMBC dataset has fairly consistent color descriptions (larger K-S distances), but the UW-RGBD dataset, which contains more complex, multicolored objects, is less consistent (smaller K-S distances). The K-S distances for shape and object traits are smaller, indicating complex, varied descriptions.

provides an efficient way to reject a null hypothesis (that the two distributions do not differ).

Visual Complexity: I used the methods proposed and *validated* against the results of users reported by Machado et al. [124] to estimate the complexity and variability of visual traits, using edge density features and compression errors. I considered different categories in concept-specific ways. For color, the approach was simple: I used the empirically validated approach of computing the standard deviation of raw RGB values. To measure the shape complexity, I computed the compression loss of the detected edges. I extracted HSV values from the RGB images, computed edge densities over these using standard edge detection algorithms ([28, 87]), and estimated the compression errors using JPEG compression (see Fig. 4.2). Machado et al. [124] presented user studies that validated this approach; other compression techniques would need to be validated in a similar fashion, which is an effort deserving of a dedicated study.[2]

Previous work that predicted trait-based concepts used a combination of color and shape features to predict the object trait [153, 89], so I do not directly address the visual complexity of object types. To meaningfully analyze object type as distinct from color and shape, I would

[2]Canny edge detection coupled with JPEG compression provides one of the highest correlations between human and computational estimates of visual complexity. This implies that edge density and compression error are reliable predictors of people's perception of visual complexity.

Dataset	Lang.	Color	Shape	Object Type
UMBC	**English**	Concise, less varied concept vocabulary	Contrasting description; 83.3 % instances were described with shape concepts	Varied and diverse description; all instances are described with object concepts
	Spanish	Concise and less varied concepts	Various concepts of similar meaning; all instances are described with shape concepts at least once	Varied and diverse description; all instances are described with object concepts
	Hindi	Semantically similar, but gender-based inflectional differences present	Highly varied and diverse concepts; all instances are described with shape concepts	Varied and diverse description; all instances are described with object concepts
UW	**English**	Multicolor objects, medium consistency in description; Not all descriptions have color concepts	Various concepts of similar meaning; only 9.5% of all instances have shape concepts in their description at least once	Varied and diverse description; 98.9% of instances have object concepts in description

Table 4.2. A qualitative summary of typical complexity of sentences describing images, separated out by datasets (UMBC vs. UW RGB-D).

need to consider a different featurization that captures more of the semantics of "object-ness," which remains a topic for future work. Nevertheless, I expect my approach to generalize to other language-grounding problems that are currently of significant interest to the field [65, 6]. I focused on RGB-D data, but variations of my measures apply to most data with a visual component, and the language analyses will be directly applicable.

4.2 Analysis: Linguistic Complexity

English: I considered both the UMBC and UW datasets for linguistic complexity evaluations in English. Fig. 4.1 shows the density computed from the entropies of the UMBC and UW datasets (results are broadly consistent between them). The variability of the traits can be seen in the entropy results in the figure. We can see that color entropies are concentrated toward "zero" compared to non-color entropies, indicating the concise, less diverse vocabulary used to describe "color." For example, the concept BLUE is described using exactly the term "blue" in 95% of the descriptions.

Non-color entropies were more diverse, indicating the variance in the descriptions using these

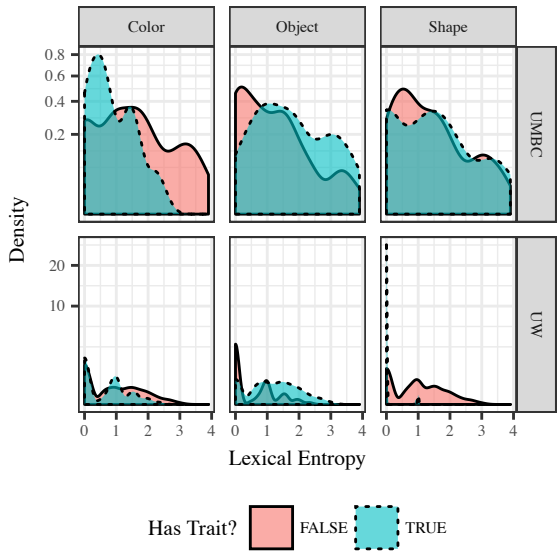

Figure 4.1. Comparison of traits "Color," "Shape," and "Object" via lexical entropy for the UMBC and UW RGB-D+ datasets. The K-S statistics quantify the divergence within each facet (subplot). Note that the entropy for color concepts is lower than for non-color concepts, indicating the concise, less varied vocabulary used to describe colors. The object trait entropy was higher, indicating linguistic variability. Only 9.5% of the UW dataset instances had shape concepts in the description at least once. Spanish descriptions contained varied but semantically similar shape/object tokens in their vocabulary.

concepts, and demonstrating that "color" was linguistically simpler in these datasets than other traits. "Shape" was the most varied trait, with high variance in the annotations, both according to my metrics and in practice. "Object" annotations were more consistent than "shape," as users were mostly consistent in describing vegetables (e.g., "corn," "cabbage") but less consistent in annotating children's toys (e.g., "arch," "cube").

There were differences between datasets. In the UW dataset, not every instance was described with a color, which is reflected in the lower K-S distances. "Object" descriptions in the UW dataset were also more diverse compared to the UMBC dataset. The atypical "shape" behavior indicates the lack of "shape" words in the description. Additional analysis revealed that only 9.5% of instance descriptions had shape descriptors.

Hindi: Figure 4.1 shows the densities calculated from the UMBC Hindi dataset. Color complexity (i.e., diversity of language describing color) was much smaller than that of shape and object. From the annotations, I found that different forms of the same words were used to describe the object: For example, the "color" concepts were semantically similar but exhibited noun inflection based on gender. Such discrepancies affect language acquisition performance. Diverse words are used for shapes, particularly to describe cylindrical objects, making the downstream language learning problem more complex. A high entropy implies a weak agreement between the annotators. The trait complexity patterns in Hindi are nonetheless approximately analogous to those in English.

Spanish: The diversity of terms used in Spanish across the three traits was similar to that of English and Hindi; language about colors was consistent and straightforward, but became more complex for shape and object. Figure 4.1 shows the densities calculated for the UMBC Spanish data. "Color" showed the least variation of the three traits, although there was more variance in color descriptions for concepts with similar meanings, such as the very similar terms *morado*, *púrpura*, and *violeta* for purple.

The vocabulary used for shape features varied and was inconsistent. Of the words *rectangulo*, *poliedro*, and *paralelepípedo*, all appeared when describing rectangular solids. Similarly, object terms varied widely, possibly due to a difference in which objects are routinely found and discussed in day-to-day life. For example, a cucumber was described as a *pepino* (cucumber) and a *pepinillo* (pickle), but also several times as "looking like a small *sandía* (watermelon)," as well as by the category hypernyms *vegetal*, and *fruta* (vegetable and fruit).

Overall, the relative linguistic complexity of traits was comparable to that of English and Hindi. Therefore, all three languages have a consistent and straightforward vocabulary for the "color"

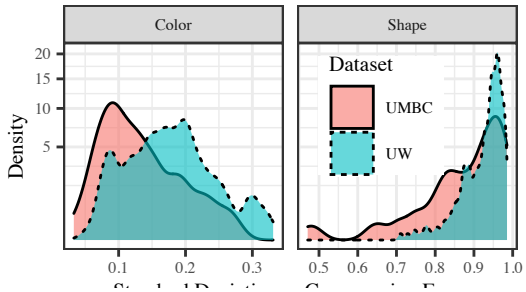

Figure 4.2. Visual complexity of "color" & "shape" for both datasets. Lower standard deviations are a good indication of greater visual color consistency. The left-skew of the compression errors illustrates the high variations of the "shape" concept.

Dataset	Language	Color		Shape		Object	
		Yes	No	Yes	No	Yes	No
UMBC	**English**	0.71	1.45	1.20	1.18	1.67	0.95
	Spanish	1.17	2.23	2.07	1.78	2.38	1.63
	Hindi	1.09	1.61	0.95	1.68	2.27	1.03
UW RGB-D+	**English**	0.39	0.58	0.01	0.71	1.03	0.22

Table 4.3. The average of the linguistic complexity comparisons between trait vs. non-trait for each dataset and language. Higher differences between average values indicate conciseness in the description; color descriptions were more concise than shape and object descriptions.

concept, but a varied and complex vocabulary for "shape" and "object" concepts.

4.3 Analysis: Visual Complexity

In modeling visual complexity, I considered shape and color differences between the two datasets, omitting object type for the reasons described above. The results are shown in Fig. 4.2.

In the smaller UMBC dataset, the standard deviations of the RGB values were a good indication of greater visual consistency, whereas lower compression rates were a good indication of reduced complexity. From these results, I can conclude that the overall color deviation was small, which is accurate for the dataset being measured. The compression rates of the shape concepts were more varied, which is indicative of greater visual variety.

In the UW dataset, the results were similar. Although there were subtle differences, the overall

Dataset	Color	Shape
UMBC	0.120	0.910
UW RGB-D+	0.171	0.942

Table 4.4. The average value of the visual complexity measures of color and shape distributions for each dataset. The smaller mean for my color complexity metric indicates a lack of variety in color features, whereas larger values for shape complexity are a result of the complicated edges and shapes in the feature set.

complexity profiles between the two datasets were similar, with perhaps slight diversity in the UW dataset color standard deviation, presumably due to the fewer monochromatic objects in this dataset.

Previous chapter reported large performance drops in classification surrounding "color" concepts vs. "shape"-based concepts [153], wherein "color" yielded an accuracy of 0.81, but "shape" was much lower at an accuracy of 0.62. This roughly tracks with this complexity measure; both linguistic and visual complexity measures for the "color" trait are lower (indicating lower complexity, and more successful classification), whereas the complexity measures for the "shape" trait are higher (indicating higher complexity, more complicated descriptions/visuals, and more difficult classification). Additionally, I found that the level of ambiguity in learning varies with multi-sense concepts in the context of dealing with concrete objects. For example, "orange" is both a color and an object. Learning the meaning of "orange" as a color is simpler than "orange" as an object, and the complexity measures reflect this difference.

4.4 Discussion

In this work, I analyzed multilingual grounded language data modalities and presented models that allow automated analysis of the complexity of descriptions and visual inputs. I verified that there is a consistent, statistically verifiable pattern of complexity across the traits I considered, making it possible to consider differentiated learning approaches in future cross-modal grounding

tasks. I anticipate that this will help grounded language learning researchers better understand the data they are working with and therefore yield and aid improved design decisions, such as more appropriate feature selection and classification models.

In this dissertation, the first two chapters concentrated on building and analyzing the necessary building blocks in perceptual learning. However, a thorough exploration of techniques to enhance learning performance is still required. Identifying appropriate technologies to optimize and generalize language grounding from complex language signals with fewer annotations is a significant contribution. The following two chapters further explore effective ways to optimize semantic language models for interactive robots.

In this chapter, I have demonstrated the efficacy of my approaches to quantify the variability and complexity of three characteristics of real-world objects. In the next chapter, I present the methods that are suitable for improving data efficiency from fewer annotations in learning these characteristics.

Chapter 5

Active Learning for Enhanced Grounded Language Acquisition

Learning unconstrained language paired with sensor and actuator data about novel object attributes (e.g., attributes that have no representation in the underlying language model until encountered) [127, 128] is a critical task for social systems. Chapter 3 demonstrated an approach for finding essential negative percepts for learning the language models, and Chapter 4 provided appropriate techniques for calibrating the complexities present in the learning problem. All these methods are useful in designing visual–linguistic learning models. However, to accomplish the longer-term goal of learning groundings for descriptions of objects from end-users, it is necessary to reduce the amount of annotation needed to learn about these objects. In this Chapter, I present an exploration of machine learning methods to improve the efficiency of grounded language learning with fewer annotations.

Machine learning of grounded language often demands large-scale natural language annotations of things in the world, which can be expensive and impractical to obtain. It is not feasible to build a dataset that encompasses every object and its possible linguistic descriptions. Novel environments require symbol grounding to occur in real time, based on inputs from a human interactor. In this work, I present an exploration of active learning approaches applied to three grounded language problems of varying complexity to analyze which methods are suitable for improving data efficiency in learning. In addition, I report on how characteristics of the underlying task, along with design decisions, such as feature selection and classification models, drive the results. I observe that

representativeness, along with diversity, is crucial in selecting data samples.

5.1 Fewer Descriptions and Better Learning

Learning the meanings of language from unstructured communication with people is an attractive approach, but it requires fast, accurate learning of new concepts, as people are unlikely to spend hours manually annotating even a few hundred samples, let alone the thousands or millions commonly required for machine learning. In this work, I study *active learning*, in which a system deliberately seeks information that will lead to improved understanding with less data, to minimize the number of samples/human interactions required. The field of active learning typically assumes that a pool of unlabeled samples is available, and the model can request specific examples for which it would like to obtain a label. By having the model select the most informative data points for labeling, the number of samples that need to be labeled is reduced. This aligns with the goal of human–robot learning for minimal training data to be provided by humans. Furthermore, active learning can be incorporated as part of a pipeline with other few-shot learning methods [58].

However, active learning is not a magic bullet. When not carefully applied, active learning does not outperform sequential or random sampling baselines [163]. A thorough selection of suitable approaches for these problems is required. While active learning has been used for language grounding [154, 8, 148], I present the **first broad exploration of the best methods for active learning for grounding vision–language pairs**. In this chapter, I focus on developing guidelines by which active learning methods might be appropriately selected and applied to vision–language grounding problems. I test different active learning approaches on grounded language problems of varying linguistic and sensory complexity, and use the results to drive a discussion of how to select

active learning methods for different grounded language data acquisition problems in an informed manner.

Here I consider the grounded language task of learning novel language about previously unseen object types and characteristics. My emphasis is on **determining which methods can reduce the amount of training data** required to achieve performance that is consistent with human evaluation. Primarily, I address five relevant questions concerning characteristic-based grounded language learning: (1) How much do active learning techniques help when learning with limited data? (2) Do different active learning techniques, e.g., pool-based vs. uncertainty-based approaches, lead to noticeable differences in performance? (3) Are the methods robust across both neural and non-neural features and classifiers? (4) How important are the characteristics of the dataset? (5) How much does incorporating some seed language affect the performance? I make conclusions with respect to these questions in §5.3. In addition to addressing the above research questions, I verify how generalizable these learning techniques are beyond characteristic-based grounding.

I found that a correct ordering of training data makes it possible to successfully learn from significantly fewer descriptions than other models for most cases, but also that the active learning methodology chosen is specific to the nature of the learning problem. My main contribution is a **principled analysis of using active learning methods as unsupervised data sampling techniques** in language grounding, with a discussion of what aspects of those problems are relevant to approach selection. Although my contributions are primarily analytic rather than algorithmic, I argue that they address a critical need in language understanding, a research area in which questions of efficiency and data collection are widespread.

5.2 Approach

For different active learning methods, I trained associations between RGB-D images (color + depth) of objects in a dataset and the language that describes them (Sec. 2.1). I note that the evaluations performed in this work are intended to *compare* the success of different active learning approaches for the same problem.

I limited the training data to a single description of each object to mimic the limited training available from human interactions. To perform replicable experiments, I used active learning approaches in which objects (and associated training and evaluation information, such as descriptions and identified concepts) were drawn from a pre-existing pool of data, rather than obtained *de novo* through human interaction. In my primary experiments, I varied the active learning approach used to select new descriptions of objects to add to the training pool. In addition, I experimented with different features and classification techniques. Because my problem focuses on choosing objects to obtain labels for, this is consistent with the task of asking a person to describe a particular object, but also allows me to perform larger-scale and more replicable experiments. My goal was to explore data selection decisions in limited settings to improve performance at the early stages of learning. The goal was thus not to improve absolute learning performance, as using a novel or complex approach runs the risk of introducing poorly understood confounding factors.

5.2.1 Learning Concept Classifiers

Similar to previous tasks, I learn the associations between perceptual inputs and descriptive concepts to test active learning approaches. Once perceptual features were extracted from the images, a visual classifier for each characteristic was learned. These classifiers were trained using

every image that had been described using the concept *and* selected by an active learning method.

Given an instance x_i and a characteristic-specific perceptual representation $\eta_{\text{TRAIT}}(x_i)$, I trained characteristic-specific probabilistic binary classifiers for each concept, $p_{\text{TRAIT}}(w_{concept} \mid \eta_{\text{TRAIT}}(x_i))$, where $w_{concept} \in \{0, 1\}$ represents the probability of x_i's characteristic TRAIT being described as *concept*. Note that this problem is two-fold: the system must learn how to both describe objects properly, and how to *avoid* characterizing objects in a way that does not make sense. I used LR as my primary classifier type p_{TRAIT} (see §5.4.4 for the impact of this decision) and extract characteristic-specific features η_{TRAIT}.

5.2.2 Core Sampling Methods

Intuitively, I wanted my algorithms to preferentially select the most informative and diverse objects for labeling from the pool of unlabeled objects. Driven by both long-standing and recent findings in active learning [42, 61, 175, 206], I used probabilistic clustering, and point process modeling in particular, as active learning strategies. Because the data are inherently noisy, I found in my early experiments that variations in GMMs and DPPs were robust selection algorithms. GMMs accommodate mixed membership, and soft cluster assignments allows one to model uncertainty. Parametric methods were selected as my learning techniques, as they are statistically stable [149] compared to nonparametric models. Therefore, I focus on GMM- and DPP-based approaches, applied to visually grounded object features, to select the most informative points from a set of unlabeled instances.

As I focused on learning from limited data, I did not consider deep learning approaches, which generally operate best over large datasets. Across all of my experiments, I examined five

different active learning models: three pool-based methods (GMM Max Log Density Based, VL-GMM,[1] and DPP) and one uncertainty-based (GMM log density) method. I introduced a structured DPP (GMM-DPP)-based active learning technique, which is a novel approach for the grounded language problem. I compared these variants of active learning strategies with a random sampling baseline across the three characteristics (color, shape, and object). Although the initial experiments considered entropy-based sampling methods (computed by my GMM's posterior entropy), they were found to perform substantially worse than those listed, and subsequent experiments did not include them. For all GMM approaches, I selected the number of components C empirically using four-fold cross-validation. In GMM-based methods, I compared the test performance with the number of components ranging from 5 to 35, and received the best results with 15 components. In GMM-based pool sampling experiments, I clustered instances using their informativeness and ranked the instances according to their learned conditional densities.

My methods selected instances that are informative and diverse by querying from all N items at once. This is also called querying in "batch mode," and has been applied successfully in prior research [171, 31]. I drew from an existing pool of human-provided descriptions, rather than explicitly seeking new labels via interactions, to enable broader and more repeatable experiments.

Max Log-Density-Based GMM Sampling: This model uses a C-component GMM to cluster unique image features and rank them according to their maximum multivariate densities from the unlabeled data pool. Those with greater density are selected as they are potentially more informative. I used 15 Gaussian components (selected empirically as a hyperparameter), initialized the mixing weights and Gaussian parameters using k-means, and fit the GMM with the standard expectation

[1] VL-GMM is included to show the difference between vision-only *vs.* vision–language clustering-based learning, and so does not occur in other reported results.

maximization algorithm to learn the parameters.

DPP Sampling: DPPs have proven to be effective in modeling diversity [62]. I used DPPs as a technique to find the most representative and diverse data points from the pool of data instances. This method uses the pool of all unlabeled image samples to find the most diverse data points by using a radial basis function (RBF) kernel with carefully selected parameters. In this setting, DPPs define a discrete probability distribution of all subsets of the image data samples. If $\mathbf{X}$ is the random variable for selecting a subset of images X from a larger set $\mathcal{X}$, then $P(\mathbf{X} = X) = \det(K_X^{(0)})/\det(K_{\mathcal{X}}^{(0)} + I)$, where I represents the identity matrix. Applied to all pairwise elements of X, the *kernel* $K_X^{(0)}$ is a positive semi-definite matrix, where the (i, j) element of the matrix is the value of the kernel applied to items x_i and x_j. I used the RBF kernel, $K^{(0)}(x_i, x_j) = \exp\left(-h\|x_i - x_j\|_2^2\right)$, by cross-validating with $h \in \{100, 25, 4\}$.

GMM-DPP: I combined the DPP kernel with the GMM marginal probability derived from the image samples to rank input samples based on diversity. Following Kulesza and Taskar [104] and Affandi et al. [2], I combined a DPP kernel $K^{(0)}(x_i, x_j)$ defined on images x_i and x_j with individual "quality" scores for each image. I used $P_{\text{GMM}}(x)$, the marginal probability of image x according to the GMM, as the quality scores, and defined a new kernel as:

$$K^{(1)}(x_i, x_j) = P_{\text{GMM}}(x_i) K^{(0)}(x_i, x_j) P_{\text{GMM}}(x_j)$$

The marginal probability modulates the diversity of the data. It allows a separate model, with its own assumptions, to help designate which data is and is not diverse. To the best of my knowledge, this is a novel kernel for grounded language learning. Similar to the GMM-based sampling approach, I used 15 Gaussian components in the GMMs and initialized the mixing weights and Gaussian

parameters using k-means.

5.3 Experimental Setup

I estimated the quality of grounded language acquisition using the predictive power of learned concept classifiers against test objects. In Tab. 5.1 I calculated the area under the curve (AUC) from the F_1-score performance of the concept classifiers. My baseline randomly selected images to train visual classifiers, whereas active learning approaches sample data points as described above. This is meant to mimic the performance of a robot asking random questions about objects in the environment.

The baseline and active learning methods only observe concept words from a single-text description for each image. Images that were described by these words were selected as positive instances. Similarity metrics were used to find negative examples of these words [155]. All results were averaged over 4–12 runs for each characteristic: object, shape, and color. I selected hyperparameters, such as the number of components of my GMM model, empirically via cross-validation. I also empirically selected the query size for each experiment.

5.4 Results and Per-Characteristic Analysis

The overall performance of each approach during the language learning task is shown in Tab. 5.1, divided into the three characteristic learning problems addressed: color, shape, and object.

Sampling	**Color**	**Shape**	**Object**
Baseline: Random	0.75	0.19	0.49
Max-Log-Density-Based GMM Pool	0.82	0.25	0.62
DPP Sampling	0.8	0.22	0.59
GMM - DPP Sampling	0.78	0.27	0.58

Table 5.1. AUC summaries for each method's F_1 performance, grouped by the characteristic learned. All Active Learning (AL) techniques performed better in characteristic grounding by selecting significant points from the pool.

5.4.1 In-Depth Analysis of Active Learning Performance

The effect of active learning techniques in grounded characteristics learning was measured by comparing the three pool-based active learning techniques described previously, with the random sampling baseline for color, shape, and object characteristics (Tab. 5.1). Below, I will present an analysis of the results with respect to color, shape, and object grounding.

Color COLOR is the simplest of the three categories of characteristics learned. This observation is, in part, a result of the dataset, in which the objects were primarily all of one color, but color learning is also a simpler vision problem overall. As such, there was little variation in the color descriptions provided. Most annotators used simple color names (e.g., "red") rather than the full range of available English terms (e.g., "crimson"). However, occasional noisy annotations such as a carrot being described as "purple" and "rose" made the learning problem more difficult. Here, RGB features extracted from the segmented objects define the η_{COLOR} and were shared across all approaches.

All active learning techniques outperformed a random baseline in learning the groundings for color concepts. When neither visual percepts nor descriptive language varies widely, the primary consideration is to choose representative data quickly. DPP-based sampling methods, which by design select diverse points, also learn effective classifiers with limited data.

Shape The second category of results, SHAPE, is the most visually complex, and contains the most extreme linguistic difficulty due to the limited set of annotations. Training shape classifiers is a comparatively complex problem, as the shape of an object varies with the viewing angle. A wider variety of words were used to describe shapes; however, unlike describing colors, users tended to not explicitly specify object shapes. For example, when asked to describe a lemon, most people responded with "yellow," but relatively few responded with "round." Kernel descriptors of the segmented object define η_{SHAPE} and were shared across all approaches.

The random sampling baseline was affected by the lack of shape tokens in the description, requiring nearly 30 descriptions to learn the first few shape words. The GMM-based DPP showed a noticeable improvement in the speed of learning and also, on inspection, found distinct shape words faster than did the random sampling approach. All active learning approaches that found diverse points at earlier stages also outperformed the random baseline.

Object Type The next challenging grounding task considered in this work was the OBJECT: learning language that describes membership in an object class, that is, object recognition. To train object classifiers, I extracted both RGB and kernel descriptors [18], which define the η_{OBJECT}, meaning that object recognition was treated approximately as a superset of color and shape learning.

The performance of the Max-Log-Density-Based GMM Pool sampling approach was significantly better than that of the random baseline. I believe this result is because the number of classes for objects was larger (and membership is therefore sparser) than for color and shape characteristics, reflecting the complexity of real-world sensor data. This sparsity made careful selection of samples particularly critical.

Sampling	**Color**	**Shape**	**Object**
Baseline: Random	0.75	0.19	0.49
Max-Log-Density-Based GMM Pool	0.82	0.25	0.62
Log Density Based GMM Uncertainty	0.83	0.23	0.44

Table 5.2. AUC summaries of F_1 performance for Pool and Uncertainty sampling performance, grouped by the characteristic learned. Uncertainty sampling (which depends on feature variability) does not perform well in object grounding, which has a noisy, highly varied data pool.

5.4.2 Pool vs. Uncertainty-Based Active Learning Methods

Uncertainty sampling methods use learned probability models to measure the uncertainty in unlabeled data points. *Log-Density-Based GMM Uncertainty Sampling:* uses a learned GMM to select outliers. I selected these by finding the images that had the *lowest* log-density of any GMM component. I aimed to select the most uncertain data points to obtain a diverse dataset.

Max-log-density-based GMM pool-based sampling (Tab. 5.2) selects representative data points from the unlabeled pool of objects, whereas uncertainty sampling selects diverse points by considering outliers as useful points. This selection depends on the variability of the features. For learning color and shape concepts, both pool- and uncertainty-based sampling performed better than the baseline. However, while learning object types, uncertainty sampling could not obtain the required concepts from the most varied visual set and the limited annotation dataset.

I hypothesize that the deterioration of uncertainty pooling on the object task relates to the nature of the information's utility in an active learning context. As more information and descriptors become available in the object scenario, it becomes easier for outliers to occur; points with unusual shapes and color combinations that are not well described will increase model uncertainty. Obtaining a label for an outlier may have limited utility for future data because of the inherent quality of being an outlier; its behavior is inconsistent with the rest of the data. This may make uncertainty-based approaches less attractive as more complex grounded language datasets become available, or it may

Sampling	**CNN**	**KernelDesc**
Baseline: Random	0.53	0.49
Max-Log-Density-Based GMM Pool	0.66	0.62
DPP Sampling	0.55	0.59
GMM - DPP Sampling	0.49	0.58

Table 5.3. AUC summary results for each visual feature's F_1 performance for "object" characteristics. The DPP and GMM pools consistently outperform the baseline for both types of visual features (non-neural kernel descriptors and CNN features).

indicate a need for refinement to uncertainty-based approaches.

5.4.3 The Impact of Visual Features

Convolutional neural network (CNN) features have been shown to be effective in learning characteristic types [210]. In this section, I examine the robustness of my active learning methods across both neural and non-neural features. In contrast to the "kernel descriptors" (the RGB and HMP features used in the previous section), I extracted 1024-dimensional features from the Neural Architecture Search Network (NASNetLarge), which was pretrained on ImageNet. The NASNetLarge features are henceforth referred to as the "CNN" features in this section.

Table 5.3 shows that, similar to grounded learning with kernel descriptors, most of the active learning techniques outperformed the random baseline on CNN features. DPP and max-log-density-based GMM pool active learning techniques could select diverse and representative points at earlier stages than could the random baseline. The characteristic learning example above shows that active learning was effective in selecting meaningful and diverse points faster, irrespective of the underlying visual features. These results also demonstrate that in a low-data setting, using a CNN over kernel descriptors without first considering the specific method of active learning can lead to inferior results. Using CNN features with both DPP sampling approaches yields a lower AUC than

Sampling	LR	SVM	MLP
Baseline: Random	0.75	0.72	0.62
Max-Log-Density-Based GMM Pool	0.82	0.66	0.54
DPP Sampling	0.8	0.66	0.6
GMM - DPP Sampling	0.78	0.63	0.5

Table 5.4. AUC summary results for each classifier's F_1 performance for "color" characteristics. Logistic regression can effectively classify types with diverse and meaningful points.

does kernel descriptors. Although the max-log-density approach dominated in this setting, these results showcase why the study of the impact of features in combination with active learning is important.

5.4.4 Analysis with Different Classifiers

In this section, I revisit my choice to use an LR classifier for $p_{\text{TRAIT}}(w_{concept} \mid \eta_{\text{TRAIT}}(x_i))$, and examine how robust the active learning methods were across different classifiers (Tab. 5.4). I consider a support vector machine (SVM) and a multilayer perception (MLP). SVMs are well-known linear models that find the maximum-margin hyperplane, which distinctly classifies data samples. An MLP, alternatively, is a feedforward artificial neural network that uses nonlinear activation functions. Both have been widely used for classification purposes.

In this experiment, I examined the "color" characteristic learned using the three classifiers (LR, SVM, and MLP). In Tab. 5.4 I demonstrate that, across active learning methods, LR classifiers were better able to classify colors than could the random sampling baseline. In contrast, neither the SVM nor the MLP resulted in effective classification models when paired with active learning approaches. These results suggest that complex classification methods may not yield improved performance and show the need to jointly consider the selection of active sampling methods and downstream classifiers.

Sampling	**UMBC**	**UW RGBD+**
Baseline: Random	0.75	0.53
Max-Log-Density-Based GMM Pool	0.82	0.58
DPP Sampling	0.8	0.51
GMM - DPP Sampling	0.78	0.64

Table 5.5. AUC summary results for each dataset's F_1 performance for COLOR. *The GMM pool and GMM-DPP were able to consistently outperform the baseline, even with a multi-colored UW RGBD+ dataset.*

5.4.5 Analysis with Different Datasets

In this section I examine if my techniques were effective for a large dataset (Tab. 5.5) that is visually and linguistically noisy and diverse. In addition to the limited features dataset, I tested the active learning techniques over a 300-object UW RGDB+ multi-colored dataset (Tab. 5.5), for just "color" characteristics due to space constraints. It contained 51 objects and 1500 annotations (Sec. 2.1). In the UW RGBD+ dataset, not every description contains color information. Additionally, the words used to describe the color concepts are inconsistent. Because the dataset contains fewer monochromatic objects, the visual variation is also high, making vision–language grounding a challenging task. Even in these experiments, most of the learning techniques that selected diverse and representative points were able to perform better than a random baseline. The DPP failed to rank in order of importance when linguistic and visual data were inconsistent. Taken together, these results indicate that my active learning techniques are generalizable and equally beneficial to datasets on different scales.

5.4.6 The Impact of Seed Language

Thus far in my analyses, my proposed methods have selected images without considering the concepts that the objects represented. In this section, I revisit that restriction and examine whether

active learning methods can benefit from considering both the image and language description together. To achieve this, I defined a joint vision–language pool-based model that uses a combination of language informativeness and visual features to select sample points from the data pool. I refer to this method as *VL-GMM sampling* . I used PVs [109] to semantically represent a language description associated with the image data point in the vector space. I used C-component GMMs to cluster the feature vectors (combined image features and PVs) and rank them. I considered the features that were closest to the center of the cluster points to be the most informative data points and selected them for training.

Sampling	**Color**	**Shape**	**Object**
Baseline: Random	0.75	0.19	0.49
Max-Log-Density-Based GMM Pool	0.82	0.25	0.62
VL-GMM Sampling	0.8	0.22	0.57

Table 5.6. AUC summaries for each method's F_1 performance, grouped by the characteristic learned. Both AL techniques performed better in characteristic grounding by selecting significant points from the pool.

VL-GMM sampling (Tab. 5.6) outperformed a random baseline in learning groundings for color, shape, and object concepts, selecting the most diverse and informative data points at the earlier stages. VL-GMM consistently exhibited better performance, which makes intuitive sense, as this method uses language in addition to image characteristics to select training data, and therefore has more information on which to decide. While learning object types, VL-GMM selected only informative points at the initial stages, and the initial performance was comparable to the baseline. After 50 data samples, diverse and representative data samples were found, which ultimately outperformed all other sampling strategies.

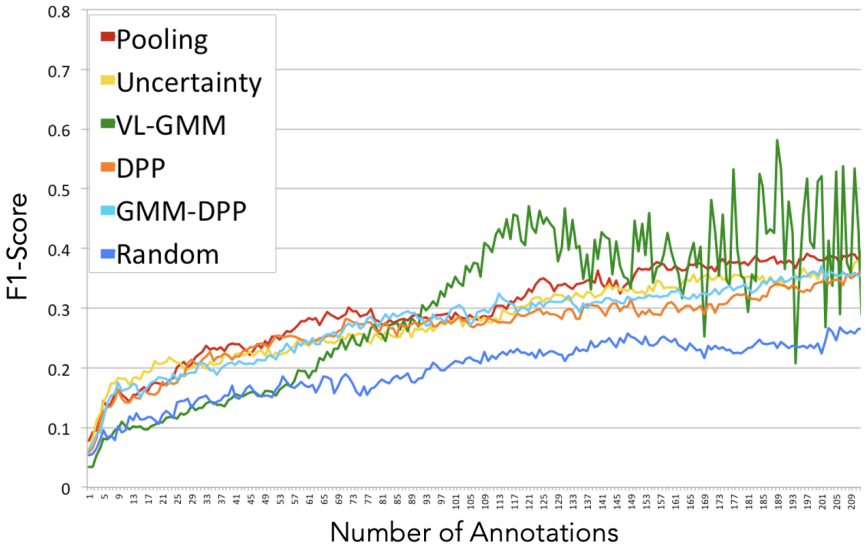

*Figure 5.1. Performance of **visual classifiers** for Object type as the learning progressed with varying data sizes. In total, 216 distinct object images and their annotations were used in training. The F_1-score is shown on the y-axis, and the number of data samples is shown on the x-axis. The VL-GMM approach showed promising performance for more complex shape and object classification problems. However, the addition of noisy, highly varied descriptions during training affected the consistency in classification. Linguistic variability within the description caused the VL-GMM performance to oscillate as it learned the language during training.*

5.4.7 Performance with Varying Data Size

In this experiment (Fig. 5.1), I attempted to mimic real-world human–robot learning that used noisy, inconsistent, and limited data resources. For training, I used 216 distinct depth images and each image's description for training. The remaining 72 images were used for testing. The descriptions were highly noisy and varied. Most of these methods did not provide shape or object information. My objective here was to understand how the active learning methods performed across varying amounts of available training data. Owing to space constraints, and to examine the performance of my methods under a "harder" setting wherein concepts are not frequently described, I display the results for "object" classification in Fig. 5.1. With highly varied and noisy features, all active learning algorithms could select diverse and important points from the pool using image features and performed better than the baseline for shape-and object-type words. The results show

that Max-Log-Density-Based pool sampling was consistently effective for all cases. This experiment also suggests that active learning algorithms that select informative and diverse points increase language acquisition quality, especially when the training data are diverse and noisy.

5.5 Analysis of Results

The main conclusion of this study is that the selection of an appropriate active learning method depends on the difficulty of the problem with respect to perceptual complexity, linguistical complexity and coverage, and sparsity of objects in each class. However, I found that incorporating different active learning methods can improve learning speed, overall performance, or both across all cases. Overall, I found GMMs to be a reliable choice for enhancing the overall learning performance. These results are discussed in more detail further in this section.

5.5.1 Method-Specific Findings

GMM clustering with image features recovers the selection of data with both informative and diverse representations. This approach probabilistically clusters similar features from the same component. However, the uncertainty-based GMM was unable to effectively find patterns faster at the initial stages than at the end stages in the dataset when object classes were scattered in the visual space. Uncertainty sampling depends on the feature variability for finding uncertain points in GMM clustering, and the sampling selects noisy outliers when the variability is greater. This finding echoes the performance reduction of uncertainty-based sampling in object feature spaces when compared to pool-based approaches.

DPP variants of active learning methods with careful parameter tuning are well suited for

selecting the most diverse points in the early stages of learning, which is appropriate when highly varied perceptual features make sample diversity important. Coverage of the more complex SHAPE and OBJECT attributes was attained significantly faster through these methods than through random sampling. Visually varied datasets require more examples of concepts, in addition to requiring diversity in labeling. k-DPP sampling provides diverse samples from the dataset and has proven to be sufficient for faster convergence of characteristic concepts. The DPP-based method is able to find diverse data samples during the initial stages and provide faster convergence to the classification tasks with kernel descriptors as well as with CNN features. However, *representativeness* and *diversity* are necessary to ensure a consistent improvement in performance. DPP sampling does not ensure representativeness and is not effective in the case of multi-colored, or visually-confusing samples. GMM-based structured DPPs provide breadth as well as diversity and perform well for both simple and complex kernel descriptor data. However, this approach is weaker for CNN-based object classification, possibly because the process of selecting representative data adds unnecessary constraints which limit performance.

Although the requirement for large and diverse quantities of language in selecting data samples would be a limitation for large datasets, I found that sampling methods that could consistently augment the visual features with a small amount of language yielded improved grounded language systems.

Time Considerations. For a dataset with N number of training data with D dimensions, the DPP computation requires $O(Nk + k^2)$ [105] if the eigen decomposition of the positive semi-definite kernel $K^{(0)}$ is available. In addition, the eigendecomposition takes approximately $O(D^3)$ time complexity. Here, k denotes the size of the subsets considered in DPP sampling. Similarly, the

GMM requires $O(D^3)$ to calculate weights that involve finding the inverse and determinants. Because I calculated weights for every component and every data point, the overall time complexity of the Max-Log-Density-based approaches is $O(C * N * D^3)$, where N is the number of data points, and C is the number of components. Structured DPP calculation involves GMM and DPP, so it requires $O(((N + k) * k + C) * N * D^3)$ operations in total. After comparing the time complexities, Max-Log-density-based pool sampling seems suitable for large-scale datasets.

5.5.2 General Considerations

In all but the most trivial cases, random sampling from a dataset outperformed a sequential baseline. Since describing objects in order is a normal human behavior, this suggests that, lacking any other change, having an agent ask widely ranging questions in a varying order may improve learning efficiency compared to passive learning.

For cases in which neither visual percepts nor descriptive language varies widely, such as COLOR, all active learning techniques are appropriate. I show that careful selection of *informative* points is most critical under these circumstances. Because the features used are simple, the main consideration here is to select representative data quickly, assuming that learning groundings (i.e., training visual classifiers) will also proceed quickly.

For visually differentiated, linguistically complex datasets, the importance of having a wide *variety* of samples increases. DPPs [106] are a class of 'repulsive' processes designed to increase diversity (see the discussion of k-DPP, above). Tuning with GMM parameters allows the DPP method to choose distinct, representative, and salient points in the dataset during very early stages of learning. Uncertainty-based max posterior GMM sampling performs well on complex data but

does not perform as strongly for sparsely populated features.

I have demonstrated that active learning techniques with carefully selected points reduce the amount of training data needed (see Tab. 5.1). When dealing with more complex datasets, selecting diverse and meaningful points increases performance as compared to selecting outliers. My experiments have also demonstrated that active learning helps establish the correct order of data points that best improve learning efficiency for both neural and non-neural visual features, and show that the addition of language features is not necessary for pool-based learning techniques to reduce the label cost.

To summarize, GMM pool sampling, which determines certainty based on the density of the clustered data points, is the most reliable active learning choice for simple, complex, noisy, multi-colored, and highly varied datasets. It is consistently able to outperform random selection with at least a 5% increase in predictive power. GMM uncertainty sampling is not a reliable choice in the case of visual data with extremely noisy outliers. LR is the most robust classification model for modeling diverse limited data when compared with SVM and MLP. DPP-based and GMM-based pool sampling produce good results in the case of neural and non-neural visual features. I observed that feature variability affected the selection techniques more than the characteristics of the dataset did. I believe that the vision-and-language sampling method considers the complexity and variance in visual features as well as language features, and as a result it aids in selecting the most diverse samples.

5.6 Discussion

In this chapter, I have presented a thorough exploration of different active learning approaches to ground unconstrained natural language using real-world sensor data. I demonstrated that active learning has the potential to reduce the amount of data necessary to ground language about objects, an active area of research in both natural language processing (NLP) and robotics, as well as machine learning from sparse data generally. I also provided suggestions for what approach may be most suitable given the perceptual and linguistic complexity of a particular problem. Given my analysis of the causes of variability in performance for different algorithms and cases, I believe these results can generalize beyond the relatively simple data seen here, making it possible for these guidelines to apply to more complicated language grounding tasks in the future.

However, previous the optimization approaches in previous chapters were limited to learning predefined visual categories, such as color, shape, and object words. The next chapter advances perceptual learning optimization by presenting a generalized category-free language model of the perceived world by omitting the need for learning visual classifiers for each category.

Chapter 6

Generalized Category-free Grounded Language Learning

Although tasks that require language, such as navigation assistance applications, are important, grounded language learning has traditionally depended on a very fine-grained understanding of individual objects and their properties. This dissertation has thus far focused on optimizing the grounded language learning problem of salient traits observed in real-world perception. The overall objective is to develop a generalized and optimized semantic model of the perceived world from noisy and complex language signals. Previous chapters have explored and demonstrated effective approaches to conceptualize various categories of perception with reduced annotation costs. In this Chapter, I focus on the category-free concept-learning problem in low-resource settings, and specifically on removing the assumption that only concepts in pre-defined categories are to be learned.

For an end-to-end efficient learning system, a generalized understanding of the underlying linguistic concepts is necessary. In this work, I analyzed methods designed to solve the symbol grounding problem. I present a *computationally reasonable* visual pretraining approach that improves on existing concept learning systems, is robust to modality featurization/embedding, and performs well in low-data settings.

This work demonstrates a semantic model in which language is grounded in visual percepts without specific predefined categories of terms. I present a unified generative method to acquire a shared semantic/visual embedding that enables the learning of language about a wide range of

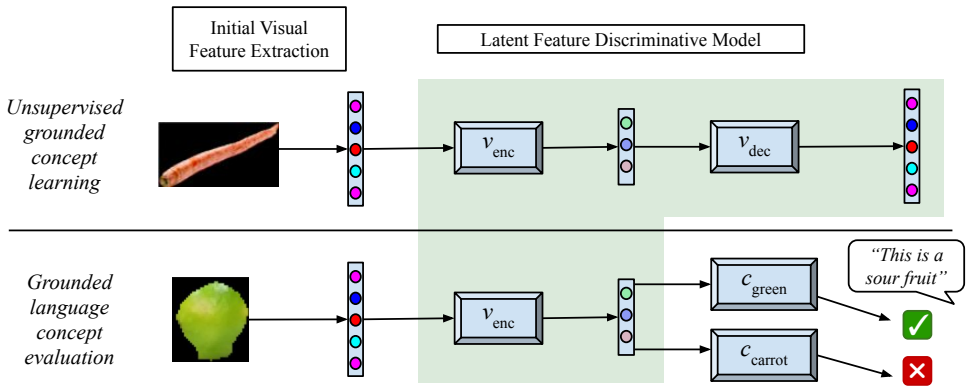

Figure 6.1. Design diagram of the unsupervised concept grounding using the latent feature discriminative method. For every object, I extracted visual features and trained a representative feature embedding by applying a latent feature discriminative model. The visual variation encoder (v_{enc}) embeds the cumulative visual features to a low-dimensional feature representation, and the visual variational decoder (v_{dec}) decodes the embedding of the visual features. The extracted low-dimensional feature embeddings are then used to create a concept classifier ($c_{concept}$) for language grounding.

real-world objects. In addition, I evaluate the efficacy of this learning by predicting the semantics of ground truth objects and comparing their performance. I demonstrate that this generative approach exhibits promising results in language grounding without pre-specifying visual categories under low-resource settings. The experiments demonstrate that this approach is generalizable to multilingual and highly-varied datasets.

6.1 Learning Beyond Constraints

In this work, I present general visual classifiers (see Fig. 6.1) that learn language without relying on predefined visual categories such as color and shape. My method generalizes language acquisition by using novel, generally applicable visual percepts and natural descriptions of real-world objects. Instead of creating classifiers for a fixed set of high-level object attributes, I used a combination of features to create a general classifier for terms that I observe in language used to describe real-world objects. I used deep generative models to obtain a representative unified visual

embedding from the combination of visual features to move away from category-specific language learning constraints.

My core contribution is a mechanism for generalizing language acquisition with an unsupervised neural variational autoencoder, which relies only on small amounts of data and requires no pre-trained image models. To compare to existing work, I evaluate the performance of my proposed method against learning concepts both with predefined categories as well as without; however, importantly the work presented here does not rely on existing categories. I also demonstrate consistent improvements over the ability of previous methods to understand Spanish and Hindi. The VAE approach I explore in this paper has the benefits of simplicity and approachability, while still demonstrating effectiveness in the low-resource settings, without the high overhead of large pretrained transformer models.

6.2 Approach

I suggest an effective generic visual classifier for training real-world object features with noisy natural human descriptions. To learn the language and its association with visual perception, I extracted a latent semantic embedding from the cumulative visual data and joined it with linguistic concepts. A high-level view of my approach can be formulated as follows: 1) Extract visual features that are associated with perception; 2) Join all extracted visual features; 3) Use the latent feature discrimination method [141, 92] based on an unsupervised neural VAE to extract meaningful, representative latent embedding from the cumulative feature set; and 4) Learn a general visual classifier using the latent embedding created from the cumulative feature set (see Fig. 6.1). Here, I intend to demonstrate how this simple discriminative method is effective for generalizing visual

classifiers. I describe the data corpus and model in sections §2.1 and §6.2.1.

6.2.1 Unified Discriminative Learning Model

My objective is to associate linguistic concepts W with a set of real-world objects, O, in limited data settings. To learn this grounded association, I created a generalized visual feature embedding out of the features extracted from the object instances and used it to train a general classifier. The components of the unified discriminative model (UDM) are detailed below.

Latent Feature Discriminative Model I used a deep generative autoencoder [92] that provides latent feature embedding for training grounded concept classifiers. Although currently VAEs may be considered a "standard" part of one's language learning toolbox, I argue that this a core strength of this paper's approach. Nevertheless, I review the use of VAEs.

This VAE consists of an encoder, a decoder, and a loss function. The encoder is a neural network that translates input data X into latent (hidden) variables Z. I represent this as $P(Z|X)$. I used a variational distribution $q_\theta(Z)$, which can be viewed as the *encoder* of X into latent features, to approximate $P(Z|X)$. The decoder is also a neural network that attempts to reconstruct X from the latent variables Z. It is modeled as $P_\phi(X|Z)$, with learnable parameters ϕ.

My objective was to learn useful and meaningful latent representations (Z) from the input data (inference network/encoder network) for use in my classification. The posterior probability was approximated using a Gaussian function $q_\theta(z|x) = \mathcal{N}(z|\mu_\theta(x), \text{diag}(\sigma^2_\theta(x))$ where $\sigma^2(x)$ is a vector of standard deviations, and $\mu(x)$ is a vector of means that are learned via multilayer perceptrons (MLPs). These are learned by minimizing

$$L = -\mathbb{E}\left[\log p(x|z)\right] + \text{KL}(q(z|x)||p(z)). \tag{6.1}$$

This is the standard VAE loss: the sum of the reconstruction error (expectation of negative log-likelihood) and the KL divergence of the approximation function and prior distribution ($KL(q(z|x)||p(z))$.

Straightforward Pretraining for a VAE As described above, my goal here was to form a unified probability distribution, $P(z|x)$ of latent variables (z) out of extracted feature variables and use it as the general embedding needed to learn the visual classifier. Here, X is defined as the feature vector extracted from the object o. In the experiments where attribute-based visual features were used, X is $\langle f_1, f_2 ... f_n \rangle$, and f_i is a type of visual feature extracted from the object, o. In this research, the challenge was to find an efficient representation of the feature space $P(Z|X)$ for my grounded learning tasks based on limited data.

Inspired by the discriminative VAE [92], I constructed a representative, meaningful low-dimensional embedding, accepting the cumulative feature vector X as the input. My model uses a neural network, $q_\theta(Z)$, to approximate $P(Z|X)$, which is a generic latent representation. Another network $P_\phi(X|Z)$, which is considered as the decoder, is used to reconstruct X from the latent variables Z. Employing an encoder function represented by a neural network (see section §6.2.1), my approach learned the encoder weights of the UDM by applying all the training data as input.

Category-free Visual Classifiers Similar to previous chapters, the system learns a binary classifier $P(y_w = 1|Z)$ for the positive items and $P(y_w = 0|Z)$ for the negative items for every relevant concept w (see sample selection from section §6.3.1). Z was defined as the visual ground generated from cumulative features. Vectors of the mean $\mu(x)$ and the standard deviation $\sigma^2(x)$ that were extracted from the generator network defined the latent embedding Z. Unlike previous approaches, instead of creating a concept-per-attribute classifiers, I learn a single concept classifier. For example, instead of learning a "red-as-color" classifier by training on color features (alone), I created a unified

general classifier for the concept “red” by associating a generalized probability distribution based on the visual features extracted from the perceived objects. I used a binary logistic regression classifier to learn the concept classifiers.

6.3 Experimental Results

In section §6.3.1, I detail the preprocessing steps used in the training data and the instantiation of the UDM model. In section §6.3.2, I describe the baselines, evaluation metrics, and cross-fold setup.

6.3.1 UDM Specification

Initial Visual Features The UDM VAE learns and computes refined embeddings, and I experimented providing the VAE with three different initial visual embeddings. In the first case, I used averaged RGB values and kernel descriptors from associated depth images [19]. I also examined neural image processing approaches (with pretrained ImageNet [46] weights) to demonstrate the generalizability and flexibility of the VAE. In the second case, I used SmallerVGGNet features and as the third case I used Neural Architecture Search Network (NASNetLarge) (See section §2.1).

Sample Selection Positive object instances were selected for every meaningful concept identified. I considered an object instance a ‘positive’ example if the object was described by the concept’s corresponding lexical form in any of that object’s descriptions. If an instance used a novel concept, a new visual classifier was created. To examine the significance of negative samples in the UDM model, I considered two different types of negative samples during learning. In the first approach, all samples except the positive samples were considered negative [182]. In the second case, I utilized

semantic similarity measures over the descriptions [155, 153].

UDM Structure I experimented with latent embedding lengths (size of Z) ranging from 12 to 100. During early development, I found an embedding length of 50 to yield the best results (See Fig. 6.10) The computed Z formed the input features for the discriminative classifier. For the variation autoencoder, I experimented with a single hidden layer MLP, with hidden dimensions ranging from 100 to 700 and found that 500 yielded the best results. Recent cross-validation experiments with hidden dimensions ranging from 100 to 700 with latent dimension 50 showed hidden dimension 100 and 700 as the best choices (See Fig. 6.2). I learned the weights needed to extract the latent embedding representation by applying all the training data to the latent feature VAE.

6.3.2 Experimental Setup

Baselines RGB-D visual classifiers were compared with two baselines. First, in the 'predefined' category classifier [127], visual classifiers were trained for every concept and feature category, as per previous work. For example, "arch" was trained as "arch-as-color," "arch-as-shape," and "arch-as-object" classifiers. The second baseline is a 'category-free' approach, where logistic regression

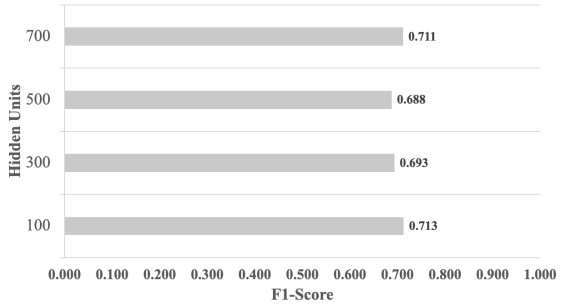

Figure 6.2. Comparison of the F1-score results with hidden dimensions ranging from 100 to 700. Hidden Dimension 100 and 700 with a latent dimension 50 shows better performance compared to other dimensions.

classifiers are trained for every concept with the concatenated feature set. Category-free logistic regression uses the concatenated X features instead of the Z features that UDM uses. Here, "arch" was trained simply as "arch-classifier," accepting as input a concatenated set of all features.

Metrics and Rigor In keeping with prior work, I measured success in grounded concept prediction via the classification performance of the learned concept classifiers. I also used the same label selection process as Pillai and Matuszek [153]. For each learned classifier, I selected three to four positive and four to six negative images from the test set. If the predicted probability for a test image is greater than 0.5, it is considered a positive result. Due to the comparatively small sizes of the datasets, I used four-fold cross-validation, and within each fold, I calculated the average F1-score across 10 trials. I ran the experiments on K20 GPUs, and jobs required no more than 6 GB of memory.

	Predefined category classifer	**Category-free logistic regression**	**Unified discriminative method**		
			Dim 12	**Dim 50**	**Dim 100**
Minimum	0.246	0.233	0.257	**0.456**	0.242
Mean	0.706	0.607	0.659	**0.713**	0.634
Maximum	0.956	0.888	**0.968**	0.963	0.900

Table 6.1. Overall summary of the F1-score distribution comparisons of all concepts. The minimum, mean and the maximum of our method are higher than all baselines, with the UDM with 50 latent dimensions showing better learning especially for difficult categories.

6.3.3 Limited Resource Classifications

Table 6.1 shows the overall summary of the distributional comparison of baselines with discriminative model variants. Here, the RGB-D visual features of 72 objects were used for the analysis. A detailed boxplot can be seen in the later section (See Fig. 6.9). The summary shows the classification performance improvement of the UDM method even for highly noisy, visually varied,

and traditionally underperformed visual classifiers compared to the well-known predefined visual classifier baseline. In a micro-averaged F1 score evaluation, UDM achieved an F1 score of 0.7218 and the category-free approach achieved a score of 0.6699; meanwhile, UDM outperformed the predefined category classifier (0.7192). This demonstrates that UDM performs as strongly as the predefined classifier and eliminates the need to create separate category-specific classifiers.

In addition to this strong performance, UDM improves on the logistic regression-based category-agnostic baseline for concepts that have fewer discriminative training instances. The vocabulary used in our dataset for certain concepts is highly varied, meaning relatively few annotations that use each linguistic token. This is demonstrated in Fig. 6.3, which plots the density estimate of F1 performance for each concept classifier vs. the number of labeled examples provided for the UDM method and category-free logistic regression baseline. As the objective is to achieve good language acquisition with limited annotation, the goal is a high F1-score for learning from a small number of occurrences (upper left quadrant, shaded). The tighter density of UDM (blue triangles) shows that performance is high given limited examples.

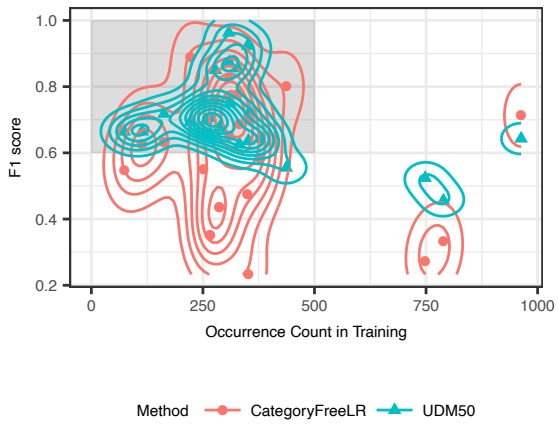

Figure 6.3. The comparison of the F1-score distribution of all concepts of the unified discriminative method vs. category free logistic regression. The goal is a high F1 score with a smaller number of occurrences, so the upper left quadrant (shaded) is the target. F1-score performance of UDM is both high and consistent with limited annotation.

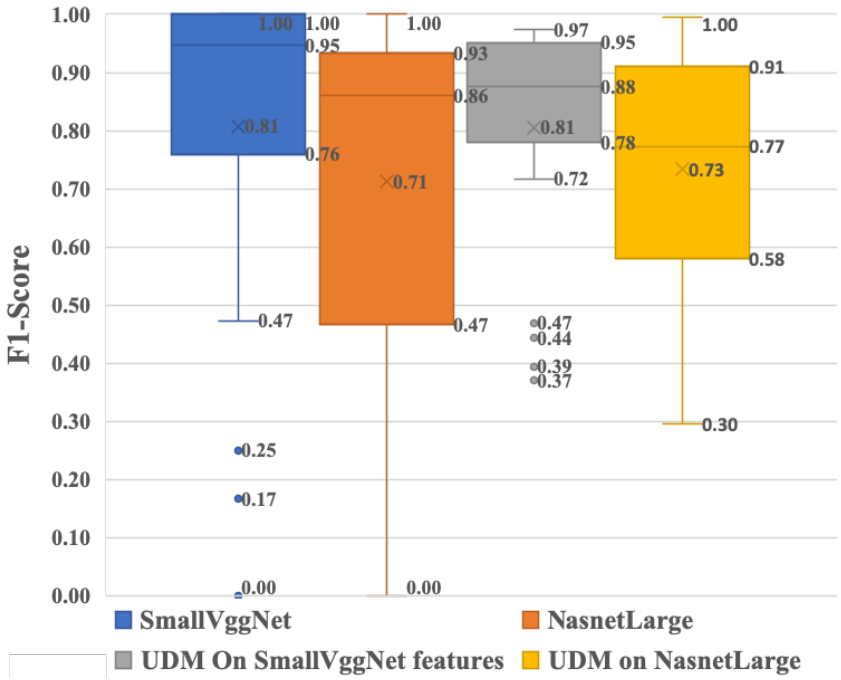

Figure 6.4. The comparison of the F1-score distribution of all concepts of the discriminate method vs. convolutional neural network baselines (leftmost two bars). Though the averaged F1-score of SmallerVGGNet and UDM + SmallerVGGNet is 0.81, SmallerVGGNet scores are as low as 0.0 for some concepts. But the minimum of UDM + SmallerVGGNet is 0.37. F1-score of UDM with NASNetLarge is 0.73 which performs better than NASNetLarge where F1-score is 0.71.

Efficacy over CNN models My analysis with CNN variants shows the quality improvement of UDM when compared to CNN baselines (See Fig. 6.4). Here, I consider the features extracted using SmallVGGNet and NASNetLarge CNN variants to test the efficacy of UDM over CNN features. The results demonstrate that the minimum F1 was improved for UDM (0.37 and 0.30 for SmallVGGNet and NASNetLarge, respectively) compared to the CNN baselines (0.0 for both SmallVGGNet and NASNetLarge). From the results, it is clear that UDM can elevate the base quality in classification over CNN features. I note that SmallerVGGNet performs extremely well on color and object classification (0.75–1.0), whereas it provides low scores for shape classification (0.0–0.47). Shape descriptions are more variable, and so more sparse, compared to color and object description. That results in comparatively poor classification performance from the high-dimensional CNN feature set.

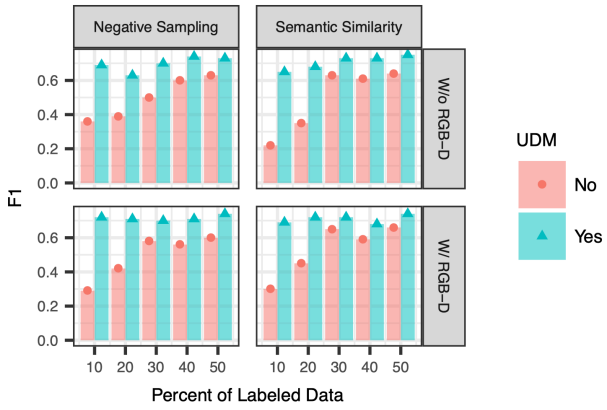

Figure 6.5. Classification performance of UDM in different architecture variations with less training data. This thorough analysis considers two negative sample varieties (semantically dissimilar patterns as negative examples versus all except positive samples as negative examples), and feature input combinations (CNN features with and without RGB-D features).

I also note that NASNetLarge classifiers provide very low color classification scores, especially for concepts such as *red* and *yellow*, for which there exist many variations in the object set. Alternatively, my UDM is able to extract a meaningful representation from NASNetLarge features.

Low Visual and Linguistic Resources The discriminative model (UDM) classified the character-

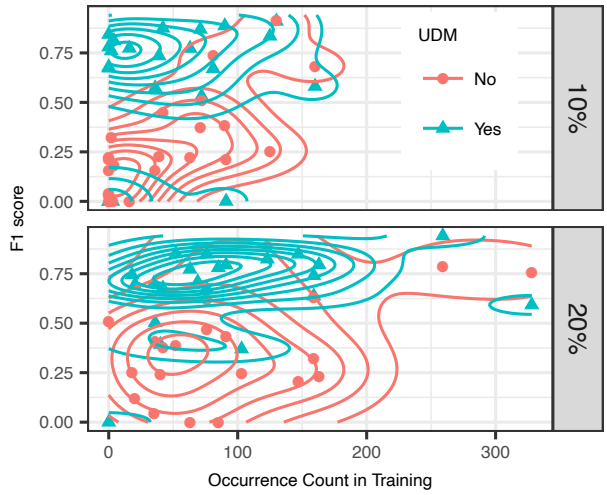

Figure 6.6. F1-score distribution comparison of a CNN variant (SmallVGGNet) vs. UDM, for all concepts with varying annotation frequency (horizontal axis). I operationally defined setting using either 10% or 20% of the labeled data. The performance of the UDM was high and consistent, even with limited annotations.

istics of objects better than the other baselines with less visual and linguistic training data. Fig. 6.5 shows the performance comparison of UDM with baselines on limited training data with different learning parameters. With 10% of the training data, all UDM variants reached an F1-score of ≥ 0.65, whereas baselines were unable to generalize the learning with limited training data. Baselines (with RGB-D) required a minimum of 30% of the training data to learn the groundings for the most important concepts in the dataset. A CNN baseline (without RGB-D) needed 40% of the training data to learn shape concepts, such as "triangular," "rectangular," and "cylinder." Sparsity in the use of particular descriptors results in an F1-score ≤ 0.5 for baseline classification, while low-dimensional representational embeddings yield improved classification for UDM.

Fig. 6.6 shows the high-quality classification performance of UDM as only a small portion of the dataset is made available for discriminative training. The figure shows the comparison of SmallVGGNet architecture vs. UDM. These experiments utilized some percentage of the total training data for learning and verified the efficiency of the model using one fourth of the total data. These cases served to demonstrate how the learning growth of these architectures compared to the UDM approaches.

Concept-wise Comparison Figure 6.7 shows the performance comparison of two baselines (See section 6.3.2) and the variants of the unified discriminative method for every meaningful concept for the RGB-D dataset. The predefined category-specific baseline grounds color specific language "terms" exceptionally well compared to other approaches. On average, color classifiers had an F1-score of 0.792 for the predefined category classifier, 0.578 for category-free logistic regression, and 0.611 for UDM with a latent dimension of 50. However, the UDM method with a latent dimension of 50 was able to perform better than the category-free logistic regression when classifier input

	Classifier	Predefined Category Classifier	Category free Logistic Regression	Unified Discriminative Method		
				Latent Dimension 12	*Latent Dimension 50*	*Latent Dimension 100*
Color	blue	0.955	0.803	0.33	0.555	0.356
	green	0.956	0.334	0.436	0.456	0.408
	orange	0.724	0.585	0.496	0.642	0.526
	purple	0.694	0.76	0.853	0.85	0.841
	red	0.807	0.713	0.692	0.643	0.693
	yellow	0.616	0.273	0.257	0.524	0.242
Shape	cube	0.324	0.352	0.642	0.706	0.653
	cylinder	0.522	0.436	0.622	0.676	0.589
	rectangular	0.661	0.633	0.517	0.718	0.483
	triangle	0.716	0.627	0.749	0.665	0.609
	triangular	0.803	0.547	0.651	0.664	0.526
Object	apple	0.79	0.659	0.651	0.699	0.559
	banana	0.246	0.233	0.442	0.637	0.495
	cabbage	0.74	0.651	0.968	0.873	0.813
	carrot	0.843	0.701	0.577	0.725	0.605
	corn	0.639	0.475	0.88	0.923	0.768
	cucumber	0.722	0.687	0.613	0.626	0.643
	eggplant	0.824	0.826	0.914	0.963	0.837
	lemon	0.82	0.778	0.754	0.855	0.825
	lime	0.91	0.888	0.748	0.694	0.855
	potato	0.604	0.55	0.647	0.761	0.683
	tomato	0.742	0.766	0.809	0.749	0.677
	cube	0.324	0.352	0.642	0.706	0.653
	cylinder	0.522	0.436	0.622	0.676	0.589
	triangle	0.716	0.627	0.749	0.665	0.609

Figure 6.7. Averaged macro F1-score comparison of the unified discriminative method against other approaches for every concept with RGB-D features. I segmented the classifiers by category here for ease of analysis as my UDM models do not consider category types. UDM with a latent dimension of 50 can provide promising performance in grounded language acquisition for all categories. Color-specific visual classifiers performed better than the category-free logistic regression baseline. Object and shape classifiers performed well with my method (UDM) with latent dimension 50 compared to other approaches.

was accepted as a vector of raw features. My method with a latent dimension of 50 outperformed both baselines for shape classification, with an average F1-score of 0.69, whereas the category-free logistic regression scored 0.52, and the category-specific approach scored 0.61. With respect to object classification, which is comparatively more complex than shape classification, the method with a latent dimension of 50 performed better than the predefined category classifier and the category-free logistic regression. F1 scores for all methods were as follows: predefined category

Visual Classifiers denoted by "term"	Ground Truth: yellow	purple	triangle	carrot	lemon
"yellow"	0.6619	0.1503	0.4723	0.1467	0.8322
"purple"	0.0070	0.6724	0.1424	0.0474	0.0008
"triangular"	0.3698	0.1911	0.6372	0.4252	0.1363
"carrot"	0.1556	0.0303	0.3211	0.6891	0.0002
"lemon"	0.5149	0.0200	0.2119	0.0097	0.9935

Figure 6.8. Prediction probabilities of selected visual classifiers (x-axis) against ground truth objects (y-axis) selected from a held-out test set with RGB-D features. This confusion matrix exhibits the prediction confidence of the unified discriminative method (UDM) run against real-world objects. Color, shape, and object variations added complexity to performance.

classifier, 0.674; category-free logistic regression, 0.616; and UDM with a latent dimension of 50, 0.754. When the minimum F1-score for UDM with dimension 50 was 0.626, the baseline predefined category classifier's peak F1 was as low as 0.246 and the category free logistic regression F1 was 0.233. These results verify the quality enhancement in visual classification for limited data settings.

Language Prediction Probabilities Figure 6.8 shows the association between the visual classifiers and the ground truth after learning the language and vision components through the unified discriminative method. Color classifiers showed strong performance: for example, the "yellow" classifier was able to predict "yellow" ground truth successfully, as well as associate with "lemon." In the dataset, the variation of "yellow" objects included a diverse set of objects ranging bananas to corn, whereas "purple" objects were limited to eggplant, plum, and cabbage.

Compared to color classifiers, object classifiers were able to predict object instances with great prediction strength. The "lemon" classifier showed a positive association with yellow objects, for example showing strong predictive ability on a lemon. The shape features of a carrot are complex compared to a lemon, so it is unsurprising that the predictive power of the learned "carrot" classifier was not strong compared to that of a "lemon" classifier. From different angles, pictures of carrots

show very different shapes, whereas lemons look almost the same when viewed from all angles. In the case of carrot, some positional angles made it look like a triangle. However, from an elevated view, the angle of the carrot's position made it look like the side of a triangle in my pictures. The complexity of the features substantially affected classification accuracy.

Comparison of Macro F1-score Distributions for All concepts Fig. 6.9 shows the distributional comparison of the other approaches with discriminative model variants. Table 6.1 shows the overall summary of these distribution comparisons, while the boxplot visualizes the median (middle line in the box), two hinges, two whiskers, and all outliers. The lower and upper hinges outline the 25^{th} and 75^{th} percentiles of the data distribution. All scores were higher than the baselines, with a minimum F1-score of 0.4560 for the method with a dimension of 50.

Overall Micro Averaged F1-score Comparisons Fig. 6.10 indicates the micro-averaged F1-score compared with all of the other approaches. My proposed method scored 0.72, which is higher than

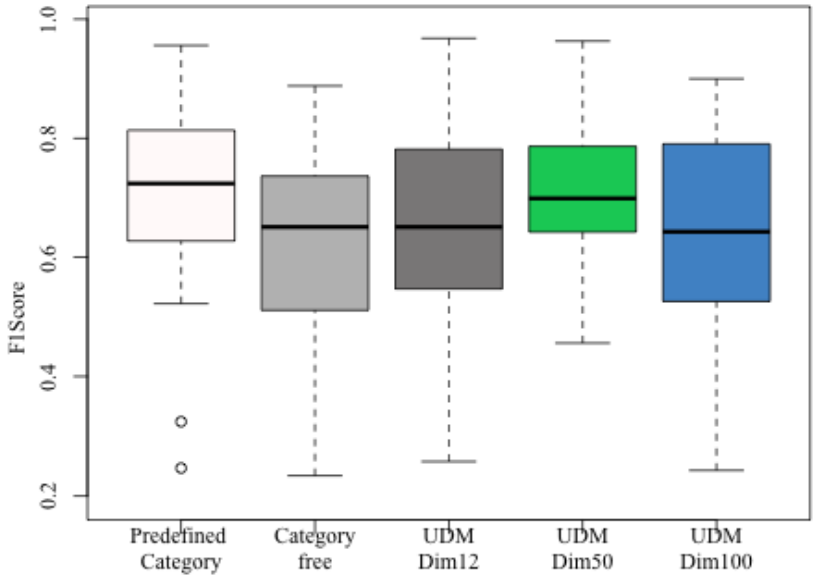

Figure 6.9. The comparison of the F1-score distribution of all concepts of the unified discriminative method vs. baselines (leftmost two bars). Minimum, mean, and maximum F1-score performance of UDM using 50 latent dimensions is both high and consistent compared to both baselines and other latent dimension variants.

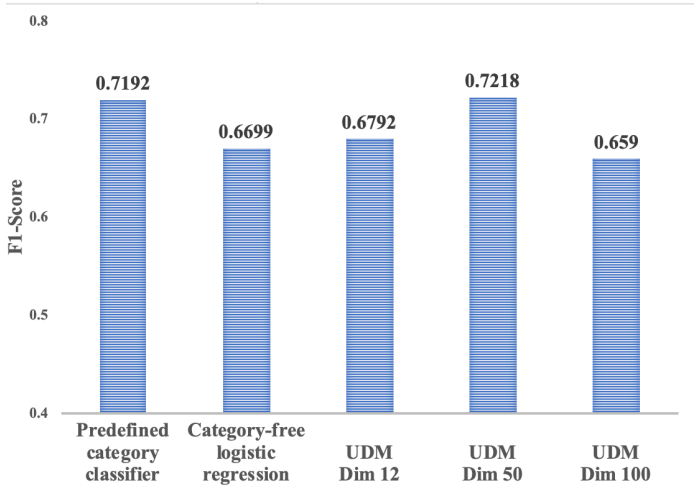

Figure 6.10. Averaged micro F1-score performance of visual classifiers. The unified discriminative method (UDM) shows improved performance than predefined category classifier where classifiers are learned per category and the category-free logistic regression where the concatenated feature set is learned per concept.

the performances of all other methods. This indicates that extracting meaningful embeddings from existing features is an efficient method for conducting grounded language concept learning.

Concept-wise Classification Analysis In this section, I provide a deeper analysis of the results of Fig. 6.7 and additional qualitative observations. In general, my results demonstrate that color concepts are learned well by the predefined category classifier. Low-dimensional training features were used to learn color concepts in the case of a predefined category classifier, and lower variability in these training features led to better classification. Low-dimensional features lack representation compared to complex high-dimensional features when all features are combined into representational embeddings. In the experiments, the UDM combined less varied color features and other highly varied features. This combination led to worse performance in the case of color concepts. In the case of "blue" concepts, for example, less variability in the training features and less noise in the annotations produced higher-quality classifiers for predefined category classifiers that used only color-related training features. However, the addition of highly varied shape features confused the

quality of the UDM classifier. This case is similar to the case of green classification. The dataset contained highly varied visual objects for the green concept, including vegetables and children's toys. Due to the influence of high-dimensional and highly varied shape features and noisy annotations, the green concept with the UDM classifier selected most of the vegetables as "green." Orange concept classification became noisy as the original annotators[1] described most of the red and yellow objects as orange. This produced a highly varied dataset for the orange classifier, which resulted in a weak classification for UDM. The predefined category classifier for orange as a color concept showed good performance, whereas orange as an object concept showed weak performance. Less noise in the purple concept annotation and low variability of the training items produced a strong "purple" classifier for all methods. Moreover, UDM extracted a meaningful latent representation for "purple" concepts. In the case of "red" concept grounding, high variability in the dataset and average noise in the annotation played a significant role in classification. Less noise in the annotations seemed to be a key component in UDM concept grounding. Similarly, the original annotators described a variety of vegetables such as corn, potato, orange, and banana as "yellow" concepts. This results in highly variable visual training data to the classifier, causing difficulties for the UDM in extracting meaningful representations.

Experiments demonstrated that noise in the annotations affected the quality of the performance of "object" type concepts as well. Plums, potatoes, tomatoes, oranges, and limes were often described as apples, as noisy annotations brought highly varied visual sets into the training data. At the same time, a lack of noise in the annotations resulted in a robust classifier for the "banana" concept, despite each image of a banana potentially being quite varied (i.e., based on the angle at which

[1] "Original annotators" refers to the crowdsourced workers who provided the descriptions for the Pillai and Matuszek [153] dataset, which were then converted into the concepts I classified.

the image was taken). Furthermore, less noise in the annotations and less varied visual features made the “cabbage” classifier a strong one. Again, less noise and less visual variation helped make the “corn” concept a robust classifier. It is interesting to see that banana was sometimes selected as “cucumber” as I learned some bananas are green and long like cucumbers. Although the shape features are not exactly the same, the shape of the cucumber is visually similar to that of several other vegetables. It is therefore interesting to observe that the low-dimensional representation of cucumber training features appeared identical for several vegetables of similar color and shape. Similar to the “cabbage” classification, the “eggplant” concept received less noise and less visual variation. In the case of “lemon,” the original annotators were confused between oranges and lemons. Moreover, tomatoes were annotated as lemons, on one occasion. Although the annotation was noisy, the visual features did not vary considerably. This aided the UDM in extracting a useful latent embedding for the “lemon” classifier. At times, lemon objects and lime objects were both annotated as “lime”. The similarity of shape features to many other vegetables led UDM latent representation to choose similar objects as “lime.” Noiseless annotations, but common round shape features, made the “potato” concept an interesting one to learn. UDM could extract a generalizable latent embedding and perform a reasonable classification for this vegetable. Similarly, the visual concept for “tomato” held a common representation, yet the noisy annotation limited the performance of the UDM.

6.3.4 Multilingual Verification

My objective here was to show that this simple discriminative method is generalizable to multilingual visual classification (see Tab. 6.2). I used Spanish and Hindi descriptions [89] collected

Language	Sampling	10%	20%	30%	40%	50%	60%	70%
Spanish	Category-free LR	0.05	0.14	0.23	0.41	0.43	0.49	0.48
	UDM	0.14	0.24	0.32	0.45	0.45	0.48	0.52
Hindi	Category-free LR	0.038	0.160	0.228	0.334	0.437	0.504	0.518
	UDM	0.187	0.290	0.413	0.490	0.516	0.536	0.552

Table 6.2. F1-score performance of UDM in multilingual classification with less training data. UDM provided a consistent improvement compared to the category-free logistic regression baseline with both Spanish and Hindi training data.

from non-trained humans for 72 RGB-D objects in a dataset [153] for this experiment. Because the current top performance on this Hindi and Spanish grounding problem [89] uses logistic regression, I used category-free logistic regression as a baseline here. With both Spanish and Hindi descriptions, the UDM achieved consistent performance improvements compared to the baseline. This validates that using a generic visual classification approach is useful regardless of the language or complexity of the input data.

6.3.5 Highly Complex, Multi-Colored Resource Verification

In this experiment, an RGB-D dataset with 300 objects was used for testing. As a baseline, the CNN variant NASNetLarge was selected for evaluating language acquisition performance. Even with 10% of the total training data, the discriminative method delivered better results than the CNN approach. My approach scored 0.46 F1-score for the classification of all concepts, whereas NASNetLarge was able to score only 0.39. This shows that my approach is effective in learning a better representative embedding from the visual features and is generalizable to any dataset.

6.4 Discussion

In this work, I have presented a simple, yet strong approach for learning a unified language grounding model that is not constrained to predefined attribute categories. I show that pre-training a straightforward Gaussian variational autoencoder efficiently grounds linguistic concepts found in unconstrained natural language to real sensor data. To compare against previous, more limited work, my evaluation primarily focuses on prediction of color, shape, and object descriptions. I also present experimental results demonstrating successful learning of a broad range of concepts from a well-studied RGB+D dataset. I hope that the improvements in low-resource settings will provide tools and insights for future work.

Analysis with my unified discriminative method, which extracts the relevant representation of feature sets, suggests that the method is effective. Moreover, its efficacy and performance improvements, especially in low-resource settings, are striking. The efficient use of such a learning system can potentially reduce the need to manually select important concepts from large corpora.

Chapter 7

Conclusion

This thesis describes a set of works for a grounded language learning system that utilizes limited resources to effectively associate natural expressions with real-world observations. Incorporating such methodologies benefits language acquisition in estimating the complexities of the system, automatically extracting counter-perspectives of the world, organizing the order of perspectives to reduce the severity of training, and finally comprehending the representations in a generalized structure.

My research thesis states that an intelligent learning system can construct an optimized, representative, unified semantic model of the perceived world from noisy, ambiguous, complex, and limited language channels using carefully selected probabilistic and active machine learning techniques.

7.1 Future Work

A natural way of acquiring knowledge and having a conversation with novice humans in an interactive environment is always challenging for artificially intelligent systems. This research is motivated by the necessity for effortless learning and intercommunication in person-centric situations in human–robot interactions. To develop an end-to-end human–robot interactive language learning framework, more research should be conducted in the following areas:

Conversation based Language Learning: My thesis used the language descriptions collected from common users using the AMT crowdsourcing tool. In a real-world human–robot interaction scenario, the robot needs to incorporate the nuances of dialogue-based information. The incorporation of unsupervised negative sample generation would enhance the quality of the conversation, and therefore of the language learning. Additionally, the human response adds certainty to negative selection.

Semantic based Language Model: My work used the "concept-as-classifier" language model to extract meaningful and important concepts from noisy human descriptions. It used tf*idf to filter through the significance of the concepts. Although effective, the system demands a better context-sensitive language model to better incorporate significant and noisy natural language in real-world communication.

Active Learning Systems with Prior Knowledge: My research has demonstrated that an interactive robot with no prior information can achieve enhanced learning performance with limited data using unsupervised techniques. Semi-supervised models can further improve my learning model from a small number of labeled samples to classify a large number of unlabeled real-world objects. However, the relationship between the labeled and unlabeled distributions could significantly affect learning performance. Another future enhancement should explore active learning techniques that improve language efficiency when the system has prior knowledge. Selecting representative samples from the unlabeled population using VAEs might also effectively bolster the learning model's efficiency and cost.

Improved Human–Robot Interaction: Human–Robot interaction serves an essential role in many robotics fields, including manufacturing, surgery, aviation, military, agriculture, and education. Future human–robot interaction research should pay careful attention to engineering "entertainment" in social conversations. Technologies should enhance the how entertaining the learning process is by collecting inputs from human–robot interaction and applying them to machine learning techniques. Such improved interactions could build reliable and trustworthy relations that help foster common understanding. Designing an interaction model should also address legal, moral, and ethical issues such as privacy, manipulation, and bias that exist in society. The reflection of societal norms and personal preference in conversations and their potentially adverse effects on the human–robot relationship should also be well-studied.

Other Applications: As virtual media is prevalent, learning in simulated environments has become increasingly popular. Simulated environments also require a large number of visuals with varying parameters. Fusing different latent representations using VAEs has the potential to successfully generate a variety of images for gaming, architecture, mapping, and virtual reality applications.

7.2 Synopsis

This dissertation proposed several research methodologies to advance current multimodal grounded learning systems. The proposed system enables a robot to encapsulate the semantics of novel concepts in a generic form from a multimodal representation, and it enhances learning methodologies to achieve effective learning from minimal data.

The primary contribution of this research is to further the development of robot learning,

advancing the ways in which robots learn during human–robot interactions to approach the same way that humans learn from each other. This work will contribute to the future of artificial intelligence, especially in relation to social/human behavior.

Bibliography

[1] Omri Abend, Tom Kwiatkowski, Nathaniel Smith, Sharon Goldwater, and Mark Steedman. Bootstrapping language acquisition. *Cognition*, 2017.

[2] Raja Hafiz Affandi, Emily Fox, Ryan Adams, and Ben Taskar. Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning (ICML)*, 2014.

[3] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] Muhannad Al-Omari, Paul Duckworth, David C Hogg, and Anthony G Cohn. Natural language acquisition and grounding for embodied robotic systems. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[5] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 2014.

[6] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Lora Aroyo and Chris Welty. Truth is a lie: crowd truth and the seven myths of human annotation. *AI Magazine*, 2015.

[8] Zhila Esna Ashari and Hassan Ghasemzadeh. Mindful active learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[9] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 1954.

[10] Daniel J Barber, Thomas M Howard, and Matthew R Walter. A multimodal interface for real-time soldier-robot teaming. In *SPIE Defense+ Security*. International Society for Optics and Photonics, 2016.

[11] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven J. Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell W. Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. Video in sentences out. *Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (AUAI)*, 2012.

[12] Leonor Becerra-Bonache, Henning Christiansen, and M. Dolores Jiménez-López. A gold standard to measure relative linguistic complexity with a grounded language learning model. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, 2018.

[13] Anja Belz, Adrian Muscat, Pierre Anguill, Mouhamadou Sow, Gaétan Vincent, and Yassine Zinessabah. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In *Proceedings of the 11th International Conference on Natural Language Generation*, 2018.

[14] Jaafar BenAbdallah, Juan C Caicedo, Fabio A Gonzalez, and Olfa Nasraoui. Multimodal image annotation using non-negative matrix factorization. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2010.

[15] Shane Bergsma and Benjamin Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[16] Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *IEEE Big Data*, 2017.

[17] Michael Bloodgood. Support vector machine active learning algorithms with query-by-committee versus closest-to-hyperplane selection. *IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018.

[18] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Kernel descriptors for visual recognition. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2010.

[19] Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. Object recognition with hierarchical kernel descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011.

[20] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017.

[21] Melissa Bowerman. The'no negative evidence'problem: How do children avoid constructing an overly general grammar? In *Explaining language universals*. J. Hawkins (Ed.), 1988.

[22] Kalesha Bullard, Andrea L Thomaz, and Sonia Chernova. Towards intelligent arbitration of diverse active learning queries. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[23] Kalesha Bullard, Yannick Schroecker, and Sonia Chernova. Active learning within constrained environments through imitation of an expert questioner. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[24] Vannevar Bush et al. As we may think. *The atlantic monthly*, 1945.

[25] Cesar Cadena, Anthony R Dick, and Ian D Reid. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and Systems*, 2016.

[26] Maya Cakmak and Andrea L Thomaz. Designing robot learners that ask good questions. In *Proceedings of the ACM/IEEE international conference on Human-Robot Interaction*, 2012.

[27] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development (TAMD)*, 2010.

[28] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 1986.

[29] Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: Towards interactive task learning with physical agents. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[30] Jean Charbonnier and Christian Wartena. Predicting word concreteness and imagery. *Proceedings of the 13th International Conference on Computational Semantics*, 2019.

[31] Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2013.

[32] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*. Springer, 2018.

[33] Xinlei Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] Xu Chen and Brett Wujek. Autodal: Distributed active learning with automatic hyperparameter selection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[35] Y. Cheng, Y. Shi, Z. Sun, D. Feng, and L. Dong. An interactive scene generation using natural language. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[36] Rishi Chhatwal, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. Empirical evaluations of active learning strategies in legal document review. In *IEEE Big Data*, 2017.

[37] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014.

[38] Grzegorz Chrupala, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Association for Computational Linguistics*, 2017.

[39] Istvan Chung, Oron Propp, Matthew R Walter, and Thomas M Howard. On the performance of hierarchical distributed correspondence graphs for efficient symbol grounding of robot instructions. In *Intelligent Robots and Systems (IROS)*, 2015.

[40] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Vanya Cohen, B. Burchfiel, Thao Nguyen, Nakul Gopalan, Stefanie Tellex, and G. Konidaris. Grounding language attributes to objects using bayesian eigenobjects. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.

[42] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.

[43] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, E. Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Association for Computational Linguistics (ACL)*, 2020.

[44] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[45] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.

[48] Haris Dindo and Daniele Zambuto. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[49] Don Donderi. Visual complexity: A review. *Psychological Bulletin*, 2006.

[50] Felix Duvallet, Matthew R Walter, Thomas Howard, Sachithra Hemachandra, Jean Oh, Seth Teller, Nicholas Roy, and Anthony Stentz. Inferring maps and behaviors from natural language instructions. In *Experimental Robotics*, 2016.

[51] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, 2001.

[52] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Empirical Methods in Natural Language Processing*, 2013.

[53] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *ArXiv*, 2016.

[54] Yansong Feng and Mirella Lapata. How many words is a picture worth? automatic caption generation for news images. In *Association for Computational Linguistics*, 2010.

[55] Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[56] Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Kenneth Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. A survey of current datasets for vision and language research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[57] Alex Forsythe, Gerry Mulhern, and Martin Sawey. Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods*, 2008.

[58] Efstratios Gavves, Thomas Mensink, Tatiana Tommasi, Cees G. M. Snoek, and Tinne Tuytelaars. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[59] Yonatan Geifman and Ran El-Yaniv. Deep active learning with a neural architecture search. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[60] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[61] Jennifer A Gillenwater, Alex Kulesza, Sergei Vassilvitskii, and Zelda E. Mariet. Maximizing induced cardinality under a determinantal point process. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[62] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[63] Víctor Gonzalez-Pacheco, María Malfaz, A Castro-Gonzalez, José Carlos Castillo, F Alonso, and Miguel Angel Salichs. Analyzing the impact of different feature queries in active learning for social robots. *International Journal of Social Robotics*, 2018.

[64] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge extraction. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 2013.

[65] Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. Using natural language for reward shaping in reinforcement learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[66] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[67] Denis Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, and Sotaro Tsukizawa. Deep active learning for biased datasets via fisher kernel self-supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[68] Yuchen Guo, Guiguang Ding, Yue Gao, and Jungong Han. Active learning with cross-class similarity transfer. In *Conference on Artificial Intelligence (AAAI)*, 2017.

[69] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 1990.

[70] Peter M Hastings and Steven L Lytinen. The ups and downs of lexical acquisition. In *AAAI Conference on Artificial Intelligence (AAAI)*, 1994.

[71] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[72] Sachithra Hemachandra and Matthew R Walter. Information-theoretic dialog to improve spatial-semantic representations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[73] Sachithra Hemachandra, Felix Duvallet, Thomas M Howard, Nicholas Roy, Anthony Stentz, and Matthew R Walter. Learning models for following natural language directions in unknown environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[74] Jack Hessel, David Mimno, and Lillian Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[75] Jack Hessel, David Mimno, and Lillian Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2194–2205, 2018.

[76] John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. Learning translations via images with a massively multilingual image dataset. In *Association for Computational Linguistics*, 2018.

[77] Padraig Higgins, Gaoussou Youssouf Kebe, Kasra Darvish, Don Engel, Francis Ferraro, and Cynthia Matuszek. Towards making virtual human-robot interaction a reality. In *3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI)*, 2021.

[78] Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. Multimodal pivots for image caption translation. *ArXiv*, 2016.

[79] Hajar Homayouni, Sudipto Ghosh, Indrakshi Ray, and Michael G Kahn. An interactive data quality test approach for constraint discovery and fault detection. In *IEEE Big Data*, 2019.

[80] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Association for Computational Linguistics (ACL)*, 2019.

[81] Baichuan Huang, Deniz Bayazit, Daniel Ullman, Nakul Gopalan, and Stefanie Tellex. Flight, camera, action! using natural language and mixed reality to control a drone. *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[82] Er-Chen Huang, Hsing-Kuo Pao, and Yuh-Jye Lee. Big active learning. In *IEEE Big Data*, 2017.

[83] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[84] Vihan Jain, Gabriel Magalhães, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Association for Computational Linguistics (ACL)*, 2019.

[85] Patrick Jenkins, Rishabh Sachdeva, Gaoussou Youssouf Kebe, Padraig Higgins, Kasra Darvish, Edward Raff, Don Engel, John Winder, Francisco Ferraro, and Cynthia Matuszek. Presentation and analysis of a multimodal dataset for grounded language learning. *arXiv preprint*, 2021.

[86] Maria Dolores Jiménez-López and Leonor Becerra-Bonache. Could machine learning shed light on natural language complexity? In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 2016.

[87] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 1988.

[88] Caroline Kery. Esta es una naranja atractiva: Adventures in adapting an english language grounding system to non-english data. Master's thesis, University of Maryland, Baltimore County, 2019.

[89] Caroline Kery, Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. Building language-agnostic grounded language learning systems. In *28th International Conference on Robot and Human Interactive Communication (Ro-Man)*, 2019.

[90] S Mohammad Khansari-Zadeh and Aude Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 2011.

[91] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *International Conference on Neural Information Processing Systems*, 2018.

[92] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[93] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[94] Ross A Knepper, Stefanie Tellex, Adrian Li, Nicholas Roy, and Daniela Rus. Recovering from failure by asking for help. *Autonomous Robots*, 2015.

[95] W. Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *International Conference on Social Robotics*, 2013.

[96] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 2002.

[97] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Textual description of human activities by tracking head and hand motions. In *Pattern Recognition*, 2002.

[98] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *ACM/IEEE International Conference on Human-robot Interaction*, 2010.

[99] Quan Kong, Bin Tong, Martin Klinkigt, Yuki Watanabe, Naoto Akira, and Tomokazu Murakami. Active generative adversarial network for image classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[100] Aryeh Kontorovich, Sivan Sabato, and Ruth Urner. Active nearest-neighbor learning in metric spaces. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[101] Bartosz Krawczyk and Alberto Cano. Adaptive ensemble active learning for drifting data stream mining. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[102] Ranjay Krishna, Yuke Zhu, Oliver Groth, J. M. Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2016.

[103] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 2013.

[104] Alex Kulesza and Ben Taskar. Structured determinantal point processes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2010.

[105] Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *International Conference on Machine Learning (ICML)*, 2011.

[106] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 2012.

[107] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[108] Howard Lasnik. On certain substitutes for negative data. In *Learnability and linguistic theory*. Springer, 1989.

[109] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, 2014.

[110] Yoad Lewenberg, Yoram Bachrach, Ulrich Paquet, and Jeffrey Rosenschein. Knowing what to ask: A bayesian active learning approach to the surveying problem. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[111] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994.

[112] Molly L Lewis and Michael C Frank. The length of words reflects their conceptual complexity. *Cognition*, 2016.

[113] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Computational Natural Language Learning*, 2011.

[114] Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C. Lee Giles. Investigating active learning for concept prerequisite learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[115] Hongru Liang, Haozheng Wang, Jun Wang, Shaodi You, Zhe Sun, Jin-Mao Wei, and Zhenglu Yang. JTAV: Jointly learning social media content representation by fusing textual, acoustic, and visual features. In *Association for Computational Linguistics (ACL)*, 2018.

[116] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020.

[117] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[118] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer plsa for multimodal image retrieval. In *ACM International Conference on Image and Video Retrieval*, 2009.

[119] Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[120] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 2019.

[121] Zhao-Yang Liu and Sheng-Jun Huang. Active sampling for open-set classification without initial annotation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[122] Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. Predicting concreteness and imageability of words within and across languages via word embeddings. In *The Third Workshop on Representation Learning for NLP*, 2018.

[123] Shuming Ma, Dongdong Zhang, and Ming Zhou. A simple and effective unified encoder for document-level machine translation. In *The Association for Computational Linguistics*, 2020.

[124] Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. Computerized measures of visual complexity. *Acta psychologica*, 2015.

[125] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[126] Cynthia Matuszek. Grounded language learning: Where robotics and nlp meet. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[127] Cynthia Matuszek*, Nicholas FitzGerald*, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *29th International Conference on Machine Learning (ICML)*, 2012.

[128] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proc. of the 28th National Conference on Artificial Intelligence (AAAI)*, March 2014.

[129] Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *The Twenty-First International Conference on Machine Learning (ICML)*, 2004.

[130] Prem Melville and Raymond J Mooney. Creating diversity in ensembles using artificial data. *Information Fusion*, 2005.

[131] Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM)*, 2004.

[132] Prem Melville, Stewart M. Yang, Maytal Saar-Tsechansky, and Raymond Mooney. Active learning for probability estimation using jensen-shannon divergence. In *Machine Learning: ECML*, 2005.

[133] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR) Workshop*, 2013.

[134] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Conference on Neural Information Processing Systems*, 2013.

[135] Aliaksei Miniukovich and Antonella De Angeli. Pick me!: Getting noticed on google play. In *CHI Conference on Human Factors in Computing Systems*, 2016.

[136] Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. In *Int'l Journal of Robotics Research (IJRR)*, 2016.

[137] Raymond J Mooney. Learning language from perceptual context: A challenge problem for ai. In *Proceedings of the 2006 AAAI Fellows Symposium*, 2006.

[138] Raymond J. Mooney. Learning to connect language and perception. In Dieter Fox and Carla P. Gomes, editors, *AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[139] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE conference on computer vision and pattern recognition*, 2017.

[140] Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, Chad Marston, and Jean-Francois Puget. Hybridization of active learning and data programming for labeling large industrial datasets. In *IEEE Big Data*, 2018.

[141] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NeurIPS workshop on deep learning and unsupervised feature learning*, 2011.

[142] Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 1976.

[143] Andre T Nguyen, Luke E Richards, Gaoussou Youssouf Kebe, Edward Raff, Kasra Darvish, Frank Ferraro, and Cynthia Matuszek. Practical cross-modal manifold alignment for robotic grounded language learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 2021.

[144] Khanh Nguyen and Hal Daumé. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *ArXiv*, 2019.

[145] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[146] Thao Nguyen, Nakul Gopalan, Roma Patel, Matt Corsaro, Ellie Pavlick, and Stefanie Tellex. Robot object retrieval with contextual natural language queries. In *Robotics: Science and Systems*, 2020.

[147] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2011.

[148] Aishwarya Padmakumar, Peter Stone, and Raymond J. Mooney. Learning a policy for opportunistic active learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[149] B. Pan, H. Dong, W. Chen, and C. Xu. Semiparametric clustering: A robust alternative to parametric clustering. *IEEE transactions on neural networks and learning systems*, 2019.

[150] Hankui Peng and Nicos G Pavlidis. Subspace clustering with active learning. In *IEEE Big Data*, 2019.

[151] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[152] Nisha Pillai and Cynthia Matuszek. Identifying negative exemplars in grounded language data sets. *Robotics: Science and Systems Workshop on Spatial-Semantic Representations in Robotics*, 2017.

[153] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[154] Nisha Pillai, Karan K Budhraja, and Cynthia Matuszek. Improving grounded language acquisition efficiency using interactive labeling. In *Robotics: Science and Systems workshop on Model Learning for Human-Robot Communication*, 2016.

[155] Nisha Pillai, Francis Ferraro, and Cynthia Matuszek. Optimal semantic distance for negative example selection in grounded language acquisition. *Robotics: Science and Systems Workshop on Models and Representations for Natural Human-Robot Communication*, 2018.

[156] Nisha Pillai, Francis Ferraro, and Cynthia Matuszek. Deep learning for category-free grounded language acquisition. In *NAACL Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, 2019.

[157] Nisha Pillai, Edward Raff, Francis Ferraro, and Cynthia Matuszek. Sampling approach matters: Active learning for robotic language acquisition. *IEEE BigData (special session on machine learning in big data)*, 2020.

[158] Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. Neural variational learning for grounded language acquisition. In *IEEE International Conference on Robot and Human Interactive Communication (Ro-Man)*, 2021.

[159] Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. Measuring perceptual and linguistic complexity in multilingual grounded language data. In *The International FLAIRS Conference Proceedings*, 2021.

[160] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[161] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[162] Mattia Racca and Ville Kyrki. Active robot learning for temporal task models. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018.

[163] Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: an empirical study of common baselines. *Data mining and knowledge discovery*, 2017.

[164] Luke E. Richards and Cynthia Matuszek. Learning to understand non-categorical physical language for human-robot interactions. In *Robotics: Science and Systems 2019 workshop on AI and Its Alternatives in Assistive and Collaborative Robotics (RSS: AI+ACR)*, 2019.

[165] Luke. E. Richards, Kasra. Darvish, and Cynthia. Matuszek. Learning object attributes with category-free grounded language from deep featurization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

[166] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 2016.

[167] Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 2002.

[168] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[169] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 1988.

[170] Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. McGraw - Hill Book Company, 1983.

[171] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[172] Connor Schenck and Dieter Fox. Towards learning to perceive and reason about liquids. In *International Symposium on Experimental Robotics*, 2017.

[173] David Schlangen, Sina Zarrieß, and Casey Kennington. Resolving references to objects in photographs using the words-as-classifiers model. In *Association for Computational Linguistics*, 2016.

[174] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 1980.

[175] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012.

[176] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. *Association for Computational Linguistics*, 2019.

[177] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[178] Weishi Shi and Qi Yu. Integrating bayesian and discriminative sparse kernel machines for multi-class active learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[179] Elaine Schaertl Short, Adam Allevato, and Andrea L Thomaz. Sail: simulation-informed active in-the-wild learning. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019.

[180] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Robotics: Science and Systems*, 2018.

[181] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.

[182] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Visually grounded meaning representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[183] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[184] Danijel Skočaj, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M Kruijff, Marc Hanheide, Nick Hawes, Jeremy L Wyatt, Thomas Keller, Kai Zhou, et al. An integrated system for interactive continuous learning of categorical knowledge. *Journal of Experimental & Theoretical Artificial Intelligence*, 2016.

[185] Richard Socher, Andrej Karpathy, Quoc Le, Christopher Manning, and Andrew Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014.

[186] Lingyun Song, Jun Liu, Buyue Qian, and Yihe Chen. Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[187] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[188] Ying-Peng Tang and Sheng-Jun Huang. Self-paced active learning: Query the right thing at the right time. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[189] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

[190] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Toward information theoretic human-robot dialog. *Robotics*, 2013.

[191] Stefanie Tellex, Pratiksha Thaker, Joshua Joseph, and Nicholas Roy. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 2013.

[192] Stefanie Tellex, Ross A Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. In *Robotics: Science and systems*, 2014.

[193] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 2020.

[194] Jesse Thomason. Continuously improving natural language understanding for robotic systems through semantic parsing, dialog, and multi-modal perception. *Doctoral Dissertation Proposal*, 2016.

[195] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by playing“ i spy”. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2016.

[196] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J Mooney. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning (CoRL)*, 2017.

[197] Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[198] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. *Conference on Robot Learning (CoRL)*, 2019.

[199] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011.

[200] Mycal Tucker, Derya Aksaray, Rohan Paul, Gregory J Stein, and Nicholas Roy. Learning unknown groundings for natural language interaction with mobile robots. In *Robotics Research: The 18th International Symposium ISRR*, 2017.

[201] A. M. Turing. Computing machinery and intelligence. *Mind*, 1950.

[202] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[203] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[204] Hanmo Wang, Runwu Zhou, and Yi-Dong Shen. Bounding uncertainty for active batch selection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[205] Yaxing Wang, Joost van de Weijer, and Luis Herranz. Mix and match networks: Encoder-decoder alignment for zero-pair image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[206] Zhiqiang Wang, Catherine Da Cunha, Mathieu Ritou, and Benoit Furet. Comparison of k-means and gmm methods for contextual clustering in hsm. *Procedia Manufacturing*, 2019.

[207] Baoyuan Wu, Weidong Chen, Peng Sun, Wei Liu, Bernard Ghanem, and Siwei Lyu. Tagging like humans: Diverse and distinct image annotation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[208] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[209] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 2016.

[210] Wenqiang Xu, Haiyang Wang, Fubo Qi, and Cewu Lu. Explicit shape encoding for real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[211] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. The label complexity of active learning from observational data. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[212] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision (ECCV)*, 2016.

[213] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[214] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *ACM International Conference on Multimedia*, 2020.

[215] Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[216] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[217] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Association for Computational Linguistics (ACL)*, 2020.

[218] Chen Yu and Dana H Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 2004.

[219] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[220] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[221] ZDNet. An eternal springtime for ai. *ZDNet*, 2018. URL https://www.zdnet.com/article/andrew-ng-sees-an-eternal-springtime-for-ai/.

[222] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[223] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[224] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[225] Mingyang Zhou, Josh Arnold, and Zhou Yu. Building task-oriented visual dialog systems through alternative optimization between dialog policy and language generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2019.

[226] Justin Zobel and Alistair Moffat. Exploring the similarity space. In *ACM SIGIR Forum*. ACM, 1998.

[227] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.