Luke E. Richards, André Nguyen, Kasra Darvish, Edward Raff, and Cynthia Matuszek.
"A Manifold Alignment Approach to Grounded Language Learning." In *the 8th Northeast Robotics Colloquium,* 2019.

A Manifold Alignment Approach to Grounded Language Learning

Luke E. Richards* *UMBC* *Booz Allen Hamilton* richards_luke@bah.com
Andre T. Nguyen* *UMBC* *Booz Allen Hamilton* nguyen_andre@bah.com
Kasra Darvish* *UMBC* kasradarvish@umbc.edu
Edward Raff *UMBC* *Booz Allen Hamilton* raff_edward@bah.com
Cynthia Matuszek *UMBC* cmat@umbc.edu

I. Introduction

As robots become advanced and affordable enough to have in our daily lives, the next question is: How do we make using these machines as intuitive as possible? Language offers an approachable and relatively accessible interface without requiring prior training on the part of the user. We have seen the integration of voice-assistant speakers in homes drastically increase in the recent years. Voice, and more specifically language, is proving to be a preferred method for interacting with AI-enabled assistants.

However, understanding how natural language applies to the physical world is still very much an open problem. Combining language and robotics creates unique challenges that much of the current work on grounded language learning has not addressed. Our proposed approach is to jointly learn language and world representations by learning a projection of both the language and sensor data into a joint space, using a process known as manifold alignment. This will enable learning of more complex grounded language in a domain-independent way. Once completed, this work will provide a bridge between the noisy, multimodal perceived world of the robotic agent and unconstrained natural language.

II. Grounded Language

Acquiring grounded language—learning associations between symbols in language and their referents in the physical world—takes many forms. With some exceptions [17], the majority of current work focuses on grounding language to RGB images [7], [14]. Due to availability of large datasets of parallel RGB images and language [7], [11], these tasks typically operate with a large pool of data. In the grounded language work in robotics, more specifically using RGB-D, large annotated datasets are rare.

Work in grounding language to rich RGB-D images is an active research area. This is a complex problem space, and has been demonstrated successfully in domains as varied as soliciting human assistance with tasks [6], interactive learning [16], and understanding complex spatial expressions [9]. Our own previous work [10], [12] has made simplifying assumptions, using a bag-of-words language model and focusing on using domain-specific visual features for training classifier models. The proposed approach relaxes these assumptions.

This material is based in part upon work supported by the National Science Foundation under Grant Nos. 1657469 and 1813223.

* The first three authors contributed equally to this work.

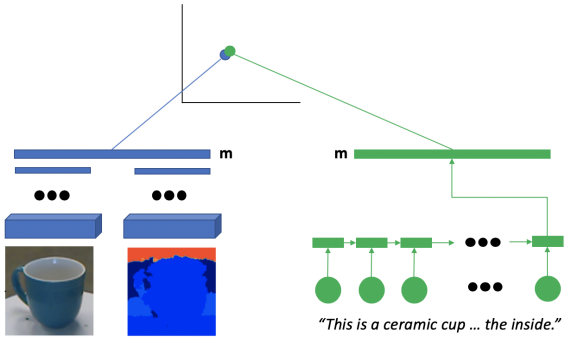

Fig. 1. Our joint learning with vision and language models, using RGB-D images and parallel language descriptions given by people.

III. Approach

Our intention is to treat the language grounding problem as one of manifold alignment—finding a mapping from heterogeneous data sets, such as language and sensor data, to a shared structure in latent space (manifold). This makes the assumption that there is an underlying manifold that data sets share, obtained by leveraging correspondences between paired data elements. We also propose a triplet loss function based on automatically selected negative data.

Manifold Alignment. Jointly learning embeddings for different domains to a shared latent space can yield a consistent representation of concepts across domains. Given n different domains $X_1, ..., X_n$, the manifold alignment task is to find n functions, $f_1, ..., f_n$ such that each function maps each m_i-dimensional space to a shared latent m-dimensional space.

$$f_i : \mathbb{R}^{m_i} \rightarrow \mathbb{R}^m \ , i = 1, ..., n$$

In order to find these mappings, the distance in the embedding space between similar instances within and across domains is minimized. Manifold alignment approaches may be *local* or *global*. In local approaches, such as locally linear embedding (LLE) [13], the goal is to map similar items from different domains closer to each other. Global methods such as [18] preserve not only the local geometry of the data, but

also the global geometry (geodesic distance). There are some hybrid methods that take advantage of both local and global alignment such as local tangent space alignment (LTSA) [19]. Canonical correlation analysis (CCA) [5] is one of the state-of-the-art methods to maximizing the correlation among similar embeddings in the shared manifold, and Deep-CCA [1] applies deep neural networks to CCA.

We propose using heterogeneous models for the different domains of corresponding data. In order to extract features from sensor data (vision+depth), we take advantage of CNNs (convolutional neural networks); text embeddings (natural language domain) are extracted with RNNs (recurrent neural networks). Triplet loss is then constructed using these heterogeneous features.

Deep Metric Learning and Triplet Loss. Deep metric learning uses deep neural networks to learn a projection of data to an embedding space where intra-class distances are smaller than inter-class distances. Our intention is that the learned metric and embedding capture the semantics of the paired data. Triplet loss functions directly encode the desire that data from a common class be 'closer together' than data from other classes [2], [15]. In particular, triplet loss seeks to minimize the distance between an *anchor point* and a positive point belonging to the same class as the anchor, while maximizing the distance between the anchor point and a negative point belonging to a different class. Given an anchor x_a, positive x_p, and negative x_n, we seek to minimize the following triplet loss (where d is a distance metric and α is a margin enforced between positive and negative data pairs):

$$L(x_a, x_p, x_n) = \max\{d(x_a, x_p) - d(x_a, x_n) + \alpha, 0\}$$

Previous work has used triplet loss for learning metric embeddings (e.g., [4] maps similar data from homogeneous domains closer to each other in a shared lower-dimensional latent space). Our approach, in contrast, is to use data from heterogeneous domains to learn the metric embeddings based on triplet loss learning.

Triplet Selection. Previous work in triplet loss learning has chosen negatives pairs based on a lack of positive labels [3], [4]. This approach is ineffective for natural language, in which things tend not to be labeled exhaustively. Negative examples are chosen through semantic distances between natural language descriptions using the method of [10].

We define a triplet as (x_a, x_p, x_n) where x_a is a RGB-D image, x_p is a description that was given for the RGB-D image, and x_n is a sentence description that is negative as defined by the semantic distance threshold. Our triplet mining algorithm will respect this constraint while potentially taking a batch-hard approach in which we select the furthest x_p and x_n with respect to x_a.

IV. Discussion and Future Work

Applying manifold alignment to learning groundings between language and physical context is a relatively novel approach; the only exceptions we are aware of focus on the cooking domain [3], [14]. Our proposed method is distinct in that we will use deep neural networks and triplet loss to learn the shared latent space, as well as selecting negative examples from paired data using unsupervised document embeddings. Our main goal is to push the research objective to incorporate a true joint learning of the two domains. Experiments will be run on the well-known RGB-D object dataset [8], extended to include natural language descriptions.

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[2] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks." in *Bmvc*, vol. 1, no. 2, 2016, p. 3.

[3] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2018.

[4] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[5] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[6] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus, "Recovering from failure by asking for help," *Autonomous Robots*, vol. 39, no. 3, pp. 347–362, Oct 2015.

[7] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2016.

[8] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1817–1824.

[9] R. Paul, J. Arkin, D. Aksaray, N. Roy, and T. M. Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms," *The International Journal of Robotics Research*, vol. 37, no. 10, pp. 1269–1299, 2018.

[10] N. Pillai and C. Matuszek, "Unsupervised selection of negative examples for grounded language learning," in *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA, 2018.

[11] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[12] L. E. Richards and C. Matuszek, "Learning to understand non-categorical physical language for human-robot interactions," in *Proc. of the R:SS 2019 workshop on AI and Its Alternatives in Assistive and Collaborative Robotics (RSS: AI+ACR)*, Freiburg, Germnany, June 2019.

[13] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, 2000.

[14] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3020–3028.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[16] L. She and J. Y. Chai, "Interactive learning of grounded verb semantics towards human-robot communication," in *ACL*, 2017.

[17] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, "Learning multi-modal grounded linguistic semantics by playing" i spy"." in *IJCAI*, 2016, pp. 3477–3483.

[18] C. Wang and S. Mahadevan, "Manifold alignment preserving global geometry," in *23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[19] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM journal on scientific computing*, vol. 26, no. 1, pp. 313–338, 2004.