Deep Learning for Category-Free Grounded Language Acquisition

Nisha Pillai **Francis Ferraro** **Cynthia Matuszek**
University of Maryland, Baltimore County, Baltimore, Maryland
npillai1@umbc.edu ferraro@umbc.edu cmat@umbc.edu

Abstract

We propose a learning system in which language is grounded in visual percepts without pre-defined category constraints. We present a unified generative method to acquire a shared semantic/visual embedding that enables a more general language grounding acquisition system. We evaluate the efficacy of this learning by predicting the semantics of ground truth objects and comparing the performance with each of a predefined category classifier and a simple logistic regression classifier. Our preliminary results suggest that this generative approach exhibits promising results in language grounding without pre-specifying visual categories such as color and shape.

1 Introduction

Grounded language acquisition, in which linguistic constructs are paired with visual constructs to learn the perceived world, has been of notable interest in the rapidly growing field of robotics. Joint modeling of language and vision (Matuszek et al., 2012; Pillai and Matuszek, 2018), where 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 are able to learn novel words describing objects, attributes, or actions that are not preexisting in the formal representation language.

While this method learns language groundings from visual features for multiple attribute classes, in most work in this area, classifiers are still trained for specific domains, such as object type or color. However, modeling semantics specific to particular attribute types still constrains language acquisition. Words denote visual classifiers, which are then trained with visual features extracted for a fixed set of semantic categories. The approach is limited to learning a predefined visual category such as color, shape, and object words.

In this work, we present general visual classifiers that learn the language without relying on predefined visual categories. Our method generalizes language acquisition by using novel, generally applicable visual percepts from natural descriptions of real-world objects. Instead of creating classifiers for every high-level category, such as color, shape and object, we use a combination of features in order to create a general classifier for terms that we observe in language used to describe real-world objects. We use 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 (see fig. 1).

Our primary contribution is a proposition to generalize language acquisition by moving away from predefined categories to category-free grounded language acquisition. In order to compare to existing work, we evaluate against attribute-specific color, shape, and object words; however, in contrast to most existing work, the system we present does not rely on these as existing categories. We compare against systems with and without predefined categories.

A high-level view of our approach can be formulated as follows: 1) Join all observed visual features. 2) Use the latent feature discrimination method (Kingma et al., 2014) based on an unsupervised neural variational autoencoder to extract meaningful, representative latent embedding from the cumulative feature set. 3) Learn a general visual classifier using the latent embedding created from the cumulative feature set (See figure 1). These approaches are trained and tested on an RGB-D image dataset (Pillai and Matuszek, 2018) created using a Kinect2 sensor and crowdsourced

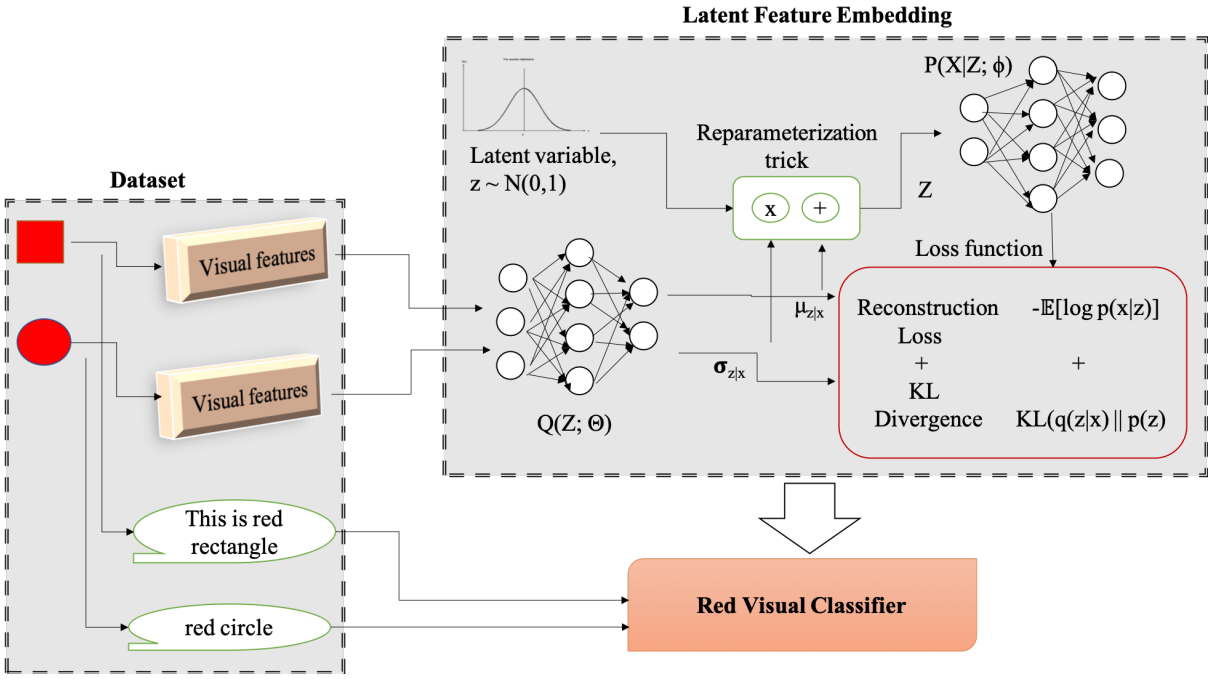

Figure 1: Design diagram of the unified discriminative method. For every object, all kinds of features are extracted and learn it through a latent feature generative model to generate a representative embedding of the feature vector. The extracted embedding is joined with a visual classifier denoted by the "language term."

descriptions of the images obtained from Mechanical Turk. Initial experiments show promising results on generalizing language grounding for terms. Our unsupervised neural variational autoencoder approach provides a matching result compared to the predefined category classifier. This suggests that this work has promise in learning more generalized language groundings.

2 Related Work

Notable research exists in the field of grounded language acquisition (Harnad, 1990) where linguistic constricts are acquired through interacting with the perceptual world. Popular language-vision grounding applications include generating descriptions from images or videos (Vinyals et al., 2015), grounding commands, directions, or action words (Artzi and Zettlemoyer, 2013; Anderson et al., 2018; Misra et al., 2016; Al-Omari et al., 2017; Chai et al., 2018), and visual question answering (VQA) (Antol et al., 2015; Yang et al., 2016; Selvaraju et al., 2017).

Our research is based on Pillai and Matuszek (2018), in which language is learned by jointly connecting with visual characteristics of real world objects. We learn visual attributes (Nyga et al., 2017) from a small dataset which is similar to one-shot learning (Hariharan and Girshick, 2017; Tommasi et al., 2010; Vinyals et al., 2016), which learns from a few samples, and zero-shot learning (Lampert et al., 2014; Elhoseiny et al., 2013) which learns from no samples.

Our experiments are designed to learn color, shape, and object attributes (Berg et al., 2010) from the descriptions of real-world objects without specifying the category of the attributes, whereas Paul et al. (2016); Tellex et al. (2011) propose to ground spatial concepts, Brawer et al. (2018) learn speech joining with context, and Yu et al. (2016, 2017) grounds natural language referring expressions for objects in images. Learning visual attributes such as color and shape is critical in robot object grasping (Rao et al., 2018; Levine et al., 2018) and manipulation tasks. There are studies that ground language by partitioning feature space by context (Thomason et al.), whereas we intend to learn words moving beyond the attribute types without manually separating them.

Deep learning has been successfully applied to image classification (Krizhevsky et al., 2012), object detection (Hatori et al., 2018; Ren et al., 2015), video classification (Karpathy et al., 2014), image to image translation (Isola et al., 2017), linking motion with language (Plappert et al., 2018) and integrating appearance, motion, gaze,

and spatial-temporal context (Balajee Vasudevan et al., 2018). However, this requires a large dataset for processing. Our objective is to gain higher prediction with a small, but complex dataset.

Our architecture predicts visual percepts associated with the language by training the representative latent probability distribution generated from cumulative visual features using a deep generative model. We use a one hidden layer deep generative variational autoencoder model to generate latent embedding from our visual features. Autoencoding has been applied to a number of tasks, including image to image translations (Wang et al., 2018), sign language translation (Cihan Camgoz et al., 2018), 3d shape analysis (Tan et al., 2018), hand pose estimation (Wan et al., 2017), sentence annotations (Ahn et al., 2018), denoising (Mao et al., 2016), and scene understanding (Cadena et al., 2016).

Silberer and Lapata (2014) learns a stacked autoencoder that grounds semantic representation of words by mapping language and vision into an embedding space. Although our objective is similar to theirs, they train stacked autoencoders for every modality by treating them separately and fusing them at the last layer to obtain meaningful representation, whereas we combine all available raw visual features before feeding them into a deep network with no differentiation among the attribute types. Rohrbach et al. (2016) employs a deep network with Long Short-Term Memory network (LSTM) to ground textual phrases in images with no, a few, or all grounding annotations available, but in this work, we intend to ground the semantics of the words without specifying its attribute type, using the annotations from natural descriptions from Amazon Mechanical Turk users.

3 Background on Latent Feature Discriminative Model

Our research uses a deep generative model (Kingma et al., 2014) of variational autoencoder to generate latent embedding for training visual classifiers. A variational autoencoder consists of an encoder, a decoder, and a loss function. The encoder is a neural network which translates input data X into latent (hidden) variables Z. We can view it as $P(Z|X)$. We use a variational distribution $q_\theta(z)$ to approximate $P(Z|X)$. And this $q_\theta(Z)$ is viewed as the encoder. The decoder is also a neural network

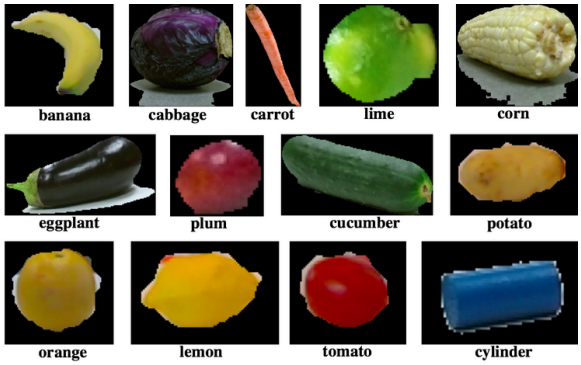

Figure 2: Sample RGB images in the dataset, as taken with a Kinect2 camera and shown to annotators (Pillai and Matuszek, 2018). In this visually varied dataset, shape and object classification are nontrivial.

which attempts to reconstruct X from the latent variables Z. We model it as $P_\phi(X|Z)$ where ϕ are the weight parameters in decoder.

Here our objective is to learn useful and meaningful latent representations (Z), from the input data (inference network/encoder network) to utilize it in our classification. We approximate the posterior probability using a Gaussian function $q_\theta(z|x)$, given by:

$$q_\theta(z|x) = \mathcal{N}(z|\mu_\theta(x), \text{diag}(\sigma^2_\theta(x))$$

Where $\sigma^2(x)$ is a vector of standard deviations, $\mu(x)$ is a vector of means, and $\mu(x)$ and $\sigma^2(x)$ is represented as multilayer perceptrons (MLPs).

The efficiency of the latent representation is enhanced using the loss function, L which is calculated as the sum of reconstruction error (expectation of negative log-likelihood) and the KL divergence of approximation function and prior distribution ($KL(q(z|x)||p(z))$. We can reduce the loss by minimizing the KL divergence. The overall loss function is then:

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

4 Approach

This work is similar to approaches in which language grounding is treated as an association of language tokens (words) with the visual percepts extracted from real world objects. In previous work Pillai and Matuszek (2018), we obtained descriptions of objects, tokenized them, created one visual classifier per category of attribute, and used them to learn real world objects. The 'correct' attribute type was assumed to be the classifier with

the best fit to the training data. Here, instead of building classifiers within specific categories, we create and learn a single visual classifier per term from a single general set of features (e.g., instead of learning separate possible classifiers such as both "cube-as-shape" and "cube-as-object" classifiers, we learn a single "cube" classifier).

This involves three steps: First, concatenate all the visual features as a single vector. Second, generate representative and meaningful embeddings from these cumulative feature vectors using a latent discriminative variational autoencoder. Last, learn one visual classifier per token, associating visual embeddings generated using the learned weights of variational autoencoder (see section 4.2 for details).

4.1 Data Corpus

We use a dataset of images and descriptions that contains color and depth images of real-world objects in 72 categories, divided into 18 classes (figure 2). Objects include food objects such as 'potato,' 'tomato,' and 'corn' and childrens toys in several shapes such as 'cube' and 'triangle'. There are an average of 4.5 images collected for every object. The language dataset includes 6000 descriptions (see figure 3) collected from Amazon Mechanical Turk (85 descriptions per object).

We use a unigram language model in learning visual classifiers. Visual classifiers are associated with unique words extracted from descriptions. These words are produced by tokenizing the descriptions, filtering stopwords, applying stemming and lemmatization processes, and filtering for domain relevance using tf-idf. For example, for a tomato instance, the description could be "This is an image of red tomato," giving "image," "red" and "tomato" as focal tokens.

A binary classifier is then trained for each of these terms, In which words used in the descriptions form (positive) binary labels. Tf-idf is used to extract meaningful tokens from this set. Empirically, this helps identify domain-meaningful words for which to learn classifiers. 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.

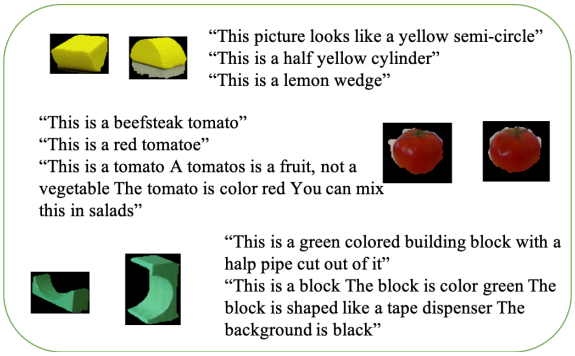

Figure 3: Object samples and language descriptions collected from Amazon Mechanical Turk annotators. The typographical errors are part of the noisy descriptions obtained from Mechanical Turk.

In order to train classifiers, two categories of visual features are used: COLOR (averaged RGB values extracted from the color image of the object) and SHAPE (kernel descriptors) (Bo et al., 2011), extracted from color and depth images respectively. Kernel descriptors model size, 3D shape, and depth edge from the RGB-D depth channel, and are efficient in shape and object classification.

4.2 Unified Discriminative Learning Method

Our objective is to associate the observed natural language W with the set of real-world objects, O. To learn this grounded association, we create a generalized visual feature embedding out of the features extracted from the object instances and use it to learn a general classifier. Components of the unified discriminative model are explained as follows. Language refers to the Amazon Mechanical Turk description used for the object in the figure. The visual groundings section explains the concatenation of visual features and extraction of latent embedding using latent discriminative model. Finally, the category-free visual classifier describes the association between language and visual characteristics. These components are shown in Figure 4.

Language. We use a unigram language model that defines $P(W|S)$, which tokenizes and filters the sentences S into words W. We learn the meaningful and relevant words, $w \in W$, using the tf-idf statistical measure described above. Tf-idf is a well-known metric to measure how relevant a word is in a particular document (in this case written description).

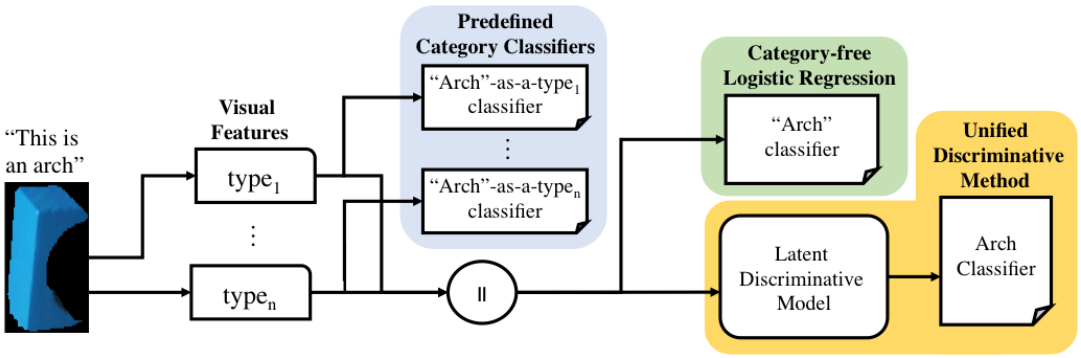

Figure 4: Diagram of approaches. The predefined category classifier—a current approach in the literature—learns one visual classifier for every visual category, accepting input from the respective visual characteristics (blue box). Category-free logistic regression is a baseline that learns a single classifier for each word, accepting cumulative visual features as input (green box). On the far right (yellow box), the Unified Discriminative Method—described in this paper—generates visual classifiers for each word, accepting latent visual embedding generated from cumulative visual features as input.

Category-free Visual Classifiers. Given a learned embedding Z for word w, we learn a binary classifier $P_c(y = 1|Z)$ for the positive items and $P_c(y = 0|Z)$ for the negative items (see section 5). Z is defined as the visual groundings generated from the cumulative features. In our approach, instead of creating a "red-as-color" classifier by training on color features, we create a unified general classifier for the word "red" by associating a generalized probability distribution made from the visual features extracted from the perceived objects.

Visual Groundings. As described above, we are attempting to form a unified probability distribution, $P(z|o)$ of latent variables (z) out of all feature variables and use it as the general embedding needed to learn the visual classifier. We define the cumulative feature input X as $< f_1, f_2...f_n >$ where f_i is a type of visual feature extracted from the object, o. Here our color features are of dimension 3, and shape features are of dimension 700, reflecting the comparative complexity of shape and the simple colors of the objects in the datset. In this research, our challenge is to find an efficient representation of our feature space $P(Z|X)$ for our grounded learning tasks.

We employ latent feature discriminative variational autoencoder (Kingma et al., 2014) to construct a representative, meaningful low-dimensional embedding, accepting the cumulative feature vector X as input. Employing an encoder function represented by a neural network (see section 3), we learn the encoder weights of the unified discriminative model (UDM) by applying all the training data as input (for more detail on the architecture, see section 5).

The above-mentioned components explain the main parts of our grounded language model. In our framework, for every classifier c, we extract a latent embedding which is the vision grounding Z using the positive and negative object instance features. Vectors of the mean $\mu(x)$ and the standard deviation $\sigma^2(x)$ that are extracted from the generator network define the latent embedding, Z. We used logistic regression on positive and negative groundings to train the binary classifier language learning model for every token.

5 Experimental Results

We use four-fold cross-validation for our experiments. We have image data $X = \{X_1, X_2....X_n\}$, where X_i is a cumulative vector constructed from features of all categories. In this framework, it is the concatenation of the shape and color features.

We selected positive object instances for every meaningful token selected using tf-idf. We consider an object instance a 'positive' example if the object is described by that token in any description. If a token is observed for the first time, we create a new visual classifier named after that token; when new objects are observed with descriptions that include this token, we update visual classifiers with the new positive instance.

In order to find negative training examples,we utilize semantic similarity measures over the descriptions of (Pillai et al., 2018; Pillai and Matuszek, 2018). For this purpose, we treated a con-

catenation of all the descriptions associated with one object as documents, and converted these 'descriptive documents' into vector space using the Distributed Memory Model of Paragraph Vectors (PV-DM) (Mikolov et al., 2013a,b). As semantically similar documents will have similar representation in vector space, we used cosine similarity statistical metric to find the most distant paragraph vectors, and selected the respective object instances as negative examples for our token.

The latent feature discriminative method is a variational autoencoder (see section 3) with a single hidden layer neural network. We experimented with several hidden layer units ranging from 100 to 600, and 500 performed the best. Rectified linear unit (ReLU) non-linear function is a good approximator and is applied as an activation function between the layers. The output layer of the encoder provides the latent embedding for our classification. We experimented with latent embedding lengths ranging from 12 to 100. We learned the weights needed for extracting latent embedding representation by applying all training data to the latent feature variational autoencoder.

Language acquisition success is measured as a funtion of the prediction performance of the learned visual classifiers. Our unified discriminative method is compared with two other approaches. First, in the 'predefined' category classifier, visual classifiers are trained for every token and feature category, as per previous work: For example, "arch" is trained as "arch-as-color," "arch-as-shape," and "arch-as-object" classifiers.

In 'category-free logistic regression,' logistic regression classifiers are trained for every token with the concatenated feature set. Here, "arch" is trained as "arch-classifier," accepting a concatenated set of all features as its input (see figure 4 for the high-level design diagram).

	Predefined category classifier	Category-free logistic regression	Unified Discriminative Method		
			Dim 12	Dim 50	Dim 100
Minimum	0.2460	0.2330	0.2570	0.4560	0.2420
Mean	0.7062	0.6078	0.6590	0.7137	0.6342
Median	0.7240	0.6510	0.6510	0.6990	0.6430
Maximum	0.9560	0.8880	0.9680	0.9630	0.9000

Table 1: Overall summary of the F1-score distribution comparisons plotted in Figure 5. The minimum, mean and the maximum of our method (latent dimension 50) are higher than all other approaches. Our method shows promising results in learning the language beyond category constraints.

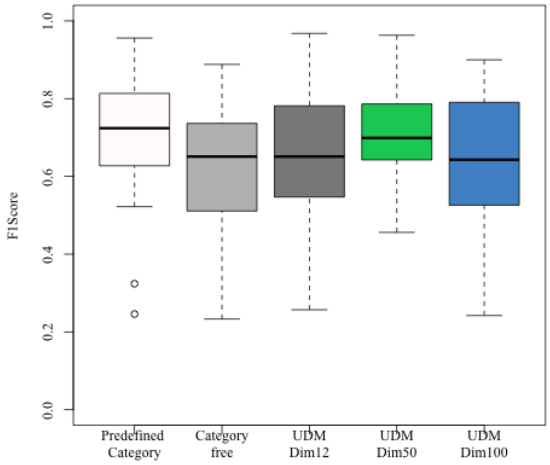

Figure 5: The comparison of the F1-score distribution of all tokens of the unified discriminative method vs. baselines (leftmost two bars). Minimum, mean, and maximum F1-score performance of UDM using 50 latent dimensions is both high and consistent compared to both baselines and other latent dimension variants.

In addition to the comparison with two baselines, we conducted experiments with varying lengths of latent dimensions to ensure the best quality prediction. We experimented unified discriminative method with latent dimensions 12, 50, and 100 to analyze the performance variation in grounded language prediction.

For all analyses, we used 72 objects and 6000 descriptions. Four-fold cross-validation is applied to the all methods. We used the image and language dataset for evaluating all the methods and trained visual classifiers on "color," "shape," and "object" words. For every learned classifier, we selected 3–4 positive and 4–6 negative images from the test set; this number varied with semantic distance. If the predicted probability for a test image is above 0.5 it is considered a positive result. We calculated averaged F1-score conducting 10 experiments for every word.

Language Prediction Probabilities. Table 2 shows the association between visual classifiers and the ground truth after learning the language and vision components through our unified discriminative method. Color classifiers show promising results. The "yellow" classifier is able to predict "yellow" ground truths successfully, as well as "lemon." In our dataset, the variation of

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

Table 2: Prediction probabilities of selected visual classifiers (x-axis) against ground truth objects (y-axis) selected from a held-out test set. This confusion matrix exhibits the prediction confidence of the unified discriminative method (UDM) run against real-world objects. Color, shape, and object variations add complexities to the performance.

"yellow" objects ranged from bananas to corn, while "purple" objects were limited to eggplant, plum, and cabbage (a wider color range, since eggplants are frequently nearly black).

Compared to color classifiers, object classifiers are able to predict object instances with great prediction strength. The "lemon" classifier shows the positive association with 'yellow objects, and 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 is not strong compared to a "lemon" classifier. From different angles, pictures of carrots show very different shapes, while lemons are almost the same from all angles. The complexity of the features affects the classification accuracy substantially.

Performance comparison for specific words. Table 3 shows the performance comparison of two baselines and the variants of the unified discriminative method for every meaningful token selected using tf-idf. As mentioned above, there were 4 trials for every method, and every token is tested 10 times in each trial with 3-6 positive and negative test instances. For every token, F1-score is calculated in every test and is averaged to calculate the overall measure treating every token test result equally.

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, 0.51 for the unified discriminative method with latent dimension 12, 0.611 for UDM with latent dimension 50, and 0.511 for UDM with latent dimension 100. However, our method with latent dimension 50 is able to perform better than the category-free logistic regression where classifier input is accepted as a vector of raw features.

	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

Table 3: Averaged macro F1-score comparison of our unified discriminative method against other approaches for every token. We segment the classifiers by category here for ease of analysis: our UDM models do not consider category types. UDM with latent dimension 50 is able to provide promising performance in grounded language acquisition for all categories. Color-specific visual classifiers perform better compared to the category free logistic regression baseline. Object and shape classifiers perform well with our method (UDM) with latent dimension 50 compared to other approaches.

Our method with latent dimension 50 outperforms both baselines for shape classification, with an average F1-score of 0.69, where category-free logistic regression scored 0.52, UDM with latent dimension 12 scored 0.64, UDM with latent dimension 100 scored 0.57, and category-specific approach scored 0.61.

In the case of object classification, which is comparatively complex, our method with latent dimensions 12 and 50 perform better compared to predefined category classifier and category-free logistic regression. Scores of the methods are as follows: Predefined category classifier: 0.6744, category-free logistic regression: 0.6163, UDM

with latent dimension 12: 0.7154, UDM with latent dimension with 50: 0.7537, and with latent dimension 100: 0.6865. When the minimum F1-score for UDM with dimension 50 is 0.626, the baseline predefined category classifier scores as low as 0.246 and category free logistic regression scores 0.233.

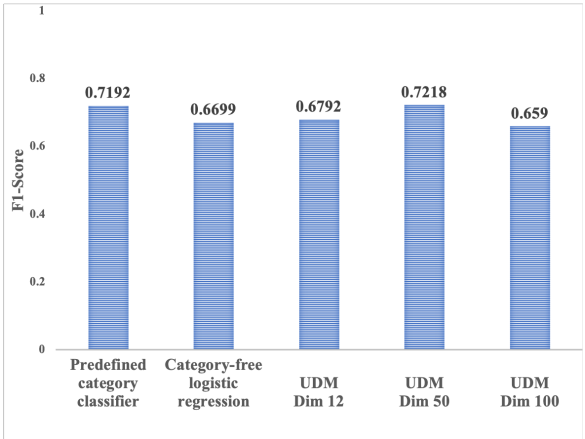

Figure 6: Averaged micro F1-score performance of visual classifiers. The unified discriminative method (UDM) shows improved performance than predefined category classifier where classifiers are learned per category and the category-free logistic regression where the concatenated feature set is learned per word.

Comparison of macro F1-score distributions for all tokens. The plot 5 shows the distributional comparison of other approaches with discriminative model variants. Table 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 are higher than baselines, with a minimum F1-score performance is 0.4560 for our method with dimension 50.

Overall micro averaged F1-score comparisons. The plot 6 indicates the micro-averaged F1-score of performance comparison of our method with all other approaches. Our method scores 0.72 which is high compared to all other performances. This indicates that extracting meaningful embedding from existing features is an efficient method to carry out grounded language prediction.

6 Conclusion

An unconstrained general classifier is essential for learning language in association with the features extracted from objects. In this work, we present an approach for learning a category-free unified language grounding model. Our results indicate that learning meaningful embedding from the cumulative unified feature set is a great approach for learning linguistic constructs beyond constrained domains. We demonstrate that such a unified model, when using carefully chosen parameters, can efficiently ground linguistic concepts with unconstrained natural language using sensor data. This reduces the need to learn classifiers with specific category features.

Analysis with our unified discriminative method, which extracts the relevant representation of feature sets, suggests that the method is effective. The efficient use of such a learning system can potentially reduce the need for selecting important tokens from a large corpora. In the future, we intend to run the approach with a more varied and complex dataset. We also intend to compare our results with deep net classifier with no auto-encoder layers. In addition, we plan to compare our models with pre-trained word and resnet models. We also aim to spot semantic similarity in words using word vectors without visual data.

References

Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. 2018. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5. IEEE.

Muhannad Al-Omari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Proceedings of the 31st National Conference on Artificial Intelligence (AAAI)*, pages 4349–4356.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick,

and Devi Parikh. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics (TACL)*, 1:49–62.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2018. Object referring in videos with language and human gaze. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

T. Berg, A. Berg, and J. Shih. 2010. Automatic attribute discovery and characterization from noisy web data. *Computer Vision–ECCV 2010*.

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

Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati. 2018. Situated human–robot collaboration: predicting intent from grounded natural language. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 827–833. IEEE.

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

Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9.

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

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *The IEEE International Conference on Computer Vision (ICCV)*.

Bharath Hariharan and Ross Girshick. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *The IEEE International Conference on Computer Vision (ICCV)*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436.

Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems*, pages 2802–2810.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations Workshop Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*.

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

Daniel Nyga, Mareike Picklum, and Michael Beetz. 2017. What no robot has seen beforeprobabilistic interpretation of natural-language object descriptions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4278–4285. IEEE.

Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M Howard. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Proceedings of Robotics: Science and Systems (R:SS) 2016*. Robotics: Science and Systems (RSS).

Nisha Pillai, Francis Ferraro, and Cynthia Matuszek. 2018. 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*.

Nisha Pillai and Cynthia Matuszek. 2018. Unsupervised end-to-end data selection for grounded language learning. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA.

Matthias Plappert, Christian Mandery, and Tamim Asfour. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26.

Achyutha Bharath Rao, Krishna Krishnan, and Hongsheng He. 2018. Learning robotic grasping strategy based on natural-language object descriptions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 882–887. IEEE.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 721–732.

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

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4):64–76.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multimodal grounded linguistic semantics by playing“ i spy”.

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. 2010. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3081–3088. IEEE.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. 2017. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 680–689.

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

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290.