GeoGlot: Learning to Ground Referential Language in Geometry

Anonymous ICCV submission

Paper ID 2872

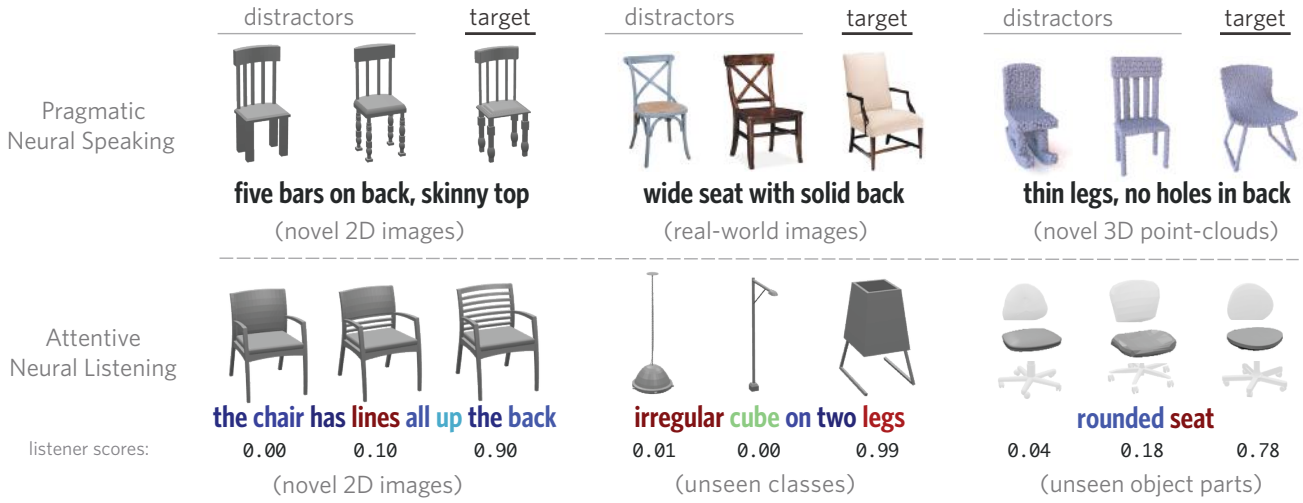


Figure 1: We introduce a corpus of utterances that refer to geometric objects (chairs) and develop neural speakers and listeners with surprising generalization flexibility. Top row: *model-generated* utterances distinguishing a target from two distractors. The speaker is applied (left to right) in unseen: 2D synthetic images, real-world images, and 3D point-clouds. Bottom row: *human-produced* referential utterances color-coded according to the attention placed by our neural listener. Scores indicate model interpretation of which object the utterance refers to. The listener is applied (left to right) in unseen: synthetic images, object categories (here, lamps), object parts.

Abstract

People understand visual objects in terms of parts and their relations. Language for referring to objects can reflect this structure, allowing us to indicate fine-grained geometric differences. Previous studies of grounded language for referring to objects largely miss this structure because reference in those studies is possible using simpler properties like color or relative spacial location. In this work we focus on grounding referential language in the intrinsic geometry of common objects. We first build a large scale, carefully controlled dataset of human utterances that each refer to a 2D rendering of a 3D CAD model within a set of geometrically-similar alternatives. Using this dataset, we develop neural language understanding and production models that vary in their grounding (pure 3D forms via point-clouds vs. rendered 2D images), the degree of pragmatic reasoning captured (e.g. speakers that reason about a listener or not), and the neural architecture (e.g. with or without attention). We find models that perform well with both synthetic and human partners, and with held out utterances and objects. We also find that these models have surprisingly strong generalization capacity to novel object classes (e.g. transfer from training on chairs to test on lamps), as well as to real images drawn from furniture catalogs. Lesion studies suggest that the neural listeners depend heavily on part-related words and associate these words correctly with visual parts of objects (without any explicit training on object parts), and that transfer to novel classes is most successful when known part-words are available. This work both illustrates a practical approach to language grounding and provides a case study in the relationship between geometric and linguistic structure.

1. Introduction

People understand visual objects compositionally, in terms of parts and their relations. Language for referring to objects reflects this structure, allowing us to describe fine-grained geometric differences. For instance, we perceive a typical chair as a geometric arrangement of parts like *arms, legs, seat,* and *back*; and we can compose words to describe its 'wide seat', combine these into 'wide seat with solid back', and refer to part relations such as 'a hole *between* the back and the seat'. We use this rich space of expressions to craft references which are not merely true, but are relevant in the current communication context. For instance choosing referring expressions that *distinguish* one chair from another (Fig. 1). In this paper we investigate the interplay between referential language and intrinsic geometry of complex objects. By gathering fine-grained referring expressions from a reference game experiment, we are able to train language understanding and production models grounded in the geometry of objects. We find that these models use complex part-based language, and transfer surprisingly well to new domains and to real images.

While a great deal of recent work has explored visually-grounded language understanding, the resulting models have limited capacity to reflect the intrinsic geometry of objects. Indeed, eliciting natural language that refers only to geometry requires carefully controlling the objects, their presentation, and the linguistic task. To address these challenges, we leverage clean 3D representations of objects (CAD models), which allow for flexible and controlled presentation (i.e. textureless, uniform-color objects viewed in a fixed pose). We further make use of the 3D form to construct a reference game task in which the target object is *geometrically* similar to the distractors. The result of this effort is a new multimodal dataset, termed ***CiC*** (***C**hairs **i**n **C**ontext*), comprised of 4,511 unique chairs from ShapeNet and 78,789 referential utterances. In CiC chairs are organized into 4,054 sets of size 3 (called communication contexts) and each utterance is intended to distinguish a chair in context. The visual differences among the grouped objects require a deep understanding of fine-grained object geometry; the language that people use to do so is correspondingly complex, exhibiting rich compositionality.

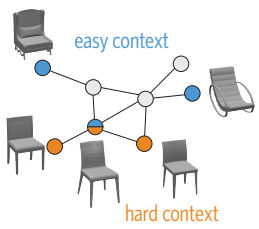
We use this data to train a variety of modern neural language understanding and production models. These models vary in their grounding (pure 3D forms via point clouds vs. rendered images), the degree of pragmatic reasoning captured, and the neural architecture (e.g. with or without self-attention within the utterance encoder). We evaluate these models on the original language game task with both synthetic and human partners, and with held out utterances and objects, finding strong performance. Since language conveys geometric abstractions, such as parts, that are shared between categories, we hypothesized that our models could transfer to new classes (e.g. training on chairs while testing on lamps). We find that these models have surprisingly strong generalization capacity to novel object classes, as well as to real images drawn from furniture catalogs.

Finally, we explore how our models are succeeding on these tasks. We demonstrate that the neural listener learns to prioritize the same geometric information in objects (i.e. properties of chair *parts*) that humans do in solving the communication task, despite never being provided with an explicit decomposition of these objects into parts. We also show that a neural speaker that is *pragmatic*—planing utterances in order to convey the right target object to an imagined listener—produces significantly more useful utterances than a *literal* (listener-unaware) speaker, as measured by *human* performance in identifying the correct object.

Overall, our results demonstrate how to elicit data for linguistic reference to complex visual objects and how to use these data to train neural models that capture structure latent in the language and its grounding.

2. Dataset and task

Our dataset consists of triplets of objects coupled with referential utterances that aim to distinguish one object (the "target") from the remaining two (the "distractors"). To obtain such utterances, we paired participants from Amazon's Mechanical Turk (AMT) to play an online, reference game [11]. On each round of the game, the two players were shown a triplet of objects. The designated target object was privately highlighted for one player (the "speaker") who was asked to send a message through a chat box such that their partner (the "listener") could successfully select it from the context. To ensure speakers used geometric information *only*, we scrambled the positions of the objects for each participant and used textureless, uniform-color renderings of pre-aligned 3D objects taken from the same viewpoint. To ensure communicative interaction was natural, no constraints were placed on the chat box: referring expressions from the speaker were occasionally followed by clarification questions from the listener or other discourse.



A crucial decision in building our dataset concerned the construction of useful contexts that would reliably elicit diverse, *fine-grained* contrastive language. The set of objects must be familiar so we can tap existing visual and linguistic representations. The objects should also be complex and variable to provide diverse geometries and language. To accomplish these goals we utilize the collection of about 7,000 chairs

from ShapeNet [3]. This class is geometrically complex, highly diverse, and abundant in the real world.

Perceptually identical objects cannot be distinguished with language at all, while wildly different objects (a chair and a car) can be distinguished with a single word ("chair"'). Different contexts must thus contain geometrically similar objects to promote the usage of very fine-grained and geometry-related language, in solving the task. To achieve this goal in a scalable manner, we estimated object similarity between different chairs in the latent space produced by the Point Cloud-AutoEncoder (PC-AE) from [1]. This approach allowed us to leverage the fact that point-clouds extracted from a 3D surface are an intrinsic 3D representation of an object, oblique to color or texture. To deal with the inhomogeneity of data in repositories like ShapeNet we used a sampling strategy to construct our triplets. First, we computed the 2-nearest-neighbor graph of all ShapeNet chairs based on their PC-AE embedding distances. On this graph, we selected chairs of highest in-degree which intuitively correspond to the most canonical chairs of the collection, and for each selected chair we generated two kinds of contexts. *Hard* contexts consist of very similar (close in latent space) chairs, while *Easy* contexts consist of less similar (far in latent space) chairs (see inset).[1] Human performance on the reference game was high in general, but listeners made significantly more errors in the hard triplets (accuracy 94.2% vs. $97.2\%, z = 13.54, p < 0.001$). Significantly longer utterances were used on average to describe targets in hard triplets (approximately 8.4 words vs. 6.1, $t = -35, p < 0.001$). A wide spectrum of descriptions was elicited, ranging from the more holistic/categorical common for easy triplets (e.g. "the rocking chair") to more complex, geometric language common for hard triplets (e.g. "thinner legs but without armrests"). Interestingly, 78% of utterances used at least one part-related word: "back", "legs", "seat," "arms", or closely related synonyms e.g. "armrests".

3. Neural listeners

Constructing neural listeners that reason about geometric properties is an important contribution of our work. Below we conduct a detailed comparison between three distinct architectures, highlight the effect of different regularization techniques, and investigate the merits of different *representations* of 3D objects for the listening task, namely, 2D rendered images and 3D surface point clouds. In what follows, we denote the three objects of a communication context as $O = \{o_1, o_2, o_3\}$, the corresponding word-tokenized utterance as $U = u_1, u_2, \ldots$ and as $t \in O$ the referential target.

[1]Additional details about materials and experimental methods can be found in Supplemental Material. The code and data for all experiments in this paper will be publicly available upon acceptance.

Our proposed listener is inspired by [21]. It takes as input a (latent code) vector that captures geometric information for each of the objects in O, and a (latent code) vector for each token of U, and outputs an object–utterance compatibility score $\mathcal{L}(o_i, U) \in [0, 1]$ for each input object. At its core lies a multi-modal LSTM [12] that receives as input ("is grounded" with) the vector of a single object, processes the word-sequence U, and is read out by a final MLP to yield a single number (the compatibility score). This is repeated for each o_i, *sharing* all network parameters across the objects. The resulting three scores are soft-max normalized and compared to the ground-truth indicator vector of the target under the cross-entropy loss.[2]

Object encoders We experimented with three object representations to capture the underlying geometries: (a) the bottleneck representation of a pretrained Point Cloud-AutoEncoder (PC-AE), (b) the embedding provided by a convolutional network operating on single-view images of non-textured 3D objects, or (c) a combination of (a) and (b). Specifically, for (a) we use the PC-AE architecture of [1] trained with single-class point clouds extracted from the surfaces of 3D CAD models, while for (b) we use the activations of the penultimate layer of a VGG-16 [28], pre-trained on ImageNet [6], and fine-tuned on an 8-way classification task with images of objects from ShapeNet. For each representation we project the corresponding code vector to the input space of the LSTM using a fully connected (FC) layer with L_2-norm weight regularization. The addition of these projection-like layers improves the training and convergence of our system.

While there are many ways to simultaneously incorporate the two modalities in the LSTM, we found that the best performance resulted when we ground the LSTM with the image code, concatenate the LSTM's final output (after processing U) with the point cloud code, and finally feed this result into a shallow MLP to produce the compatibility score. We note that grounding the LSTM with point clouds and using images towards the end of the pipeline, resulted in a significant performance drop ($\sim 4.8\%$ on average). Also, proper regularization was critical: adding dropout at the input layer of the LSTM and L_2 weight regularization and dropout at and before the FC projecting layers improved performance $\sim 10\%$. The token codes of each sentence where initialized with the GloVe embedding [25] and fine-tuned for the listening task.

Incorporating context information Our proposed baseline listener architecture (*Baseline*, just described) first scores each object *separately* then applies softmax normalization to yield a score distribution over the three objects. We also consider two more complex architectures that explicitly encode information about the *entire* context while

[2]Architecture details, hyper-parameter search strategy, and optimal hyper-parameters for all experiments are described in the Supplemental.

scoring an object. The first alternative (*Early-Context*), is identical to the proposed architecture, except for the codes used to ground the LSTM. Specifically, if v_i is the image code vector of the i-th object ($o_i \in O$) resulting from VGG, instead of directly using v_i as the grounding vector of o_i, a shallow convolutional network is introduced. This network, of which the output *is* the grounding code, receives the signal $f(v_j, v_k)||g(v_j, v_k)||v_i$, where f, g are the symmetric (and norm-normalized), max-pool and mean-pool functions, $||$ denotes feature-wise concatenation and v_j, v_k the contrastive objects. Here, we use symmetric functions to induce that object-order is irrelevant for our task. The second alternative architecture (*Combined-Interpretation*) first feeds the image vectors for *all* three objects sequentially to the LSTM as inputs and then proceeds to process the tokens of U *once*, to yield the three scores. Similarly to the *Baseline* architecture, point clouds are incorporated in both alternatives via a separate MLP after the LSTM.

Word attention We hypothesized that a listener forced to prioritize a few words in each utterance would learn to prioritize words that express properties that distinguish the target from the distractors (and, thus, perform better). To test this hypothesis, we augment the listener models with a standard *bilinear attention mechanism* [27]. Specifically, to estimate the "importance" of each text-token u_i we compare the output of the LSTM at u_i (denoted as r_i) with the hidden state after the entire utterance has been processed (denoted as h). The relative importance of each word is $a_i \triangleq r_i^T \times W_{\text{att}} \times h$, where W_{att} is a trainable diagonal matrix. The final output of the LSTM uses this attention to combine all latent states: $\sum_{i=1}^{|U|} r_i \odot \hat{a}_i$, where $\hat{a}_i = \frac{\exp(a_i)}{\sum_j^{|U|} \exp(a_j)}$ and $\odot$ is the point-wise product.

4. Listener experiments

We begin our evaluation of the proposed listeners using two reference games tasks based on different data splits. In the *language generalization* task, we test on target objects that were seen as targets in at least one context during training but ensure that all utterances in the test split are from unseen speakers. In the more challenging *object generalization* task, we restrict the set of objects that appeared as targets in the test set to be *disjoint* from those in training such that all speakers *and* objects in the test split are new. For each of these tasks, we evaluate choices of input modality and word attention, using $[80\%, 10\%, 10\%]$ of the data, for training, validating and testing purposes.

Baseline listener accuracies are shown in Table 2.[3] Overall the model achieves good performance. As expected, all listeners have higher accuracy on the language generalization task (3.2% on average). The attention mechanism on words yields a mild performance boost, as long as images are part of the input. Interestingly, images provide a significantly better input than point-clouds when only one modality is used. This may reflect the higher-frequency content of images (we use point-clouds with only 2048 points), or the fact that VGG was pre-trained while the PC-AE was not. However, we find significant gains in accuracy (4.1% on average) from exploiting the two object representations *simultaneously*, implying a complementarity among them.

Next, we evaluate how the different approaches to incorporating context information described in Section 3 affect listener performance. We focus on the more challenging object generalization task, using models that include attention and both object modalities. We report the findings in Table 1. We find that the *Baseline* and *Early-Context* models perform best overall, outperforming the *Combined-Interpretation* model, which does not share weights across objects. This pattern held for both hard and easy trial types in our dataset. We further explored the small portion ($\sim$14%) of our test set that use explicitly contrastive language: superlatives ("skinniest") and comparatives ("skinnier"). Somewhat surprisingly we find that the *Baseline* architecture remains competitive against the architectures with more explicit context information. The *Baseline* model thus achieves highest performance while being simplest and most flexible (it can be applied to *arbitrary-sized* contexts); we focus on this architecture in the explorations below.

4.1. Exploring learned representations

Linguistic ablations Which aspects of a sentence are most critical for our listener's performance? To inspect the properties of words receiving the most attention, we ran a part-of-speech tagger on our corpus. We found that the highest attention weight is placed on *nouns*, controlling for the length of the utterance. However, adjectives that *modify* nouns received more attention in hard contexts (controlling for the average occurrence in each context), where nouns are often not sufficient to disambiguate (see Fig. 2A). To more systematically evaluate the role of higher-attention tokens in listener performance, we conducted an utterance lesioning experiment. For each utterance in our dataset, we successively replaced words with the <UNK> token according to three schemes: (1) from highest attention to lowest, (2) from lowest attention to highest, and (3) in random order. We then fed these through an equivalent listener trained *without* attention. We found that up to 50% of words can be removed without much performance degradation, but only if these are low attention words (see Fig. 2B). Our word-attentive listener thus appears to rely on context-appropriate content words to successfully disambiguate the referent.

[3] In all results mean accuracies and standard errors across 5 random seeds are reported, to control for the data-split populations and the initialization of the neural-network.

Architecture	Overall	Subpopulations		
		Hard	Easy	Sup-Comp
Combined-Interpretation	$75.9 \pm 0.5\%$	$67.4 \pm 1.0\%$	$83.8 \pm 0.6\%$	$74.4 \pm 1.5\%$
Early-Context	$79.4 \pm 0.8\%$	$\mathbf{70.1} \pm 1.3\%$	$88.1 \pm 0.6\%$	$75.6 \pm 2.2\%$
Baseline	$\mathbf{79.6} \pm 0.8\%$	$69.9 \pm 1.3\%$	$\mathbf{88.8} \pm 0.4\%$	$\mathbf{76.3} \pm 1.3\%$

Table 1: Comparing different ways to include context. The simplest *Baseline* model performs as well as more complex alternatives. Subpopulations are the subset of test data containing: hard trials (geometrically similar distractors), easy trials, superlatives or comparatives.

	Input Modality	Language Task	Object Task
No Attention	Point Cloud	$67.6 \pm 0.3\%$	$66.4 \pm 0.7\%$
	Image	$81.2 \pm 0.5\%$	$77.4 \pm 0.7\%$
	Both	$83.1 \pm 0.4\%$	$78.9 \pm 1.0\%$
With Attention	Point Cloud	$67.4 \pm 0.3\%$	$65.6 \pm 1.4\%$
	Image	$81.7 \pm 0.5\%$	$77.6 \pm 0.8\%$
	Both	$\mathbf{83.7} \pm 0.3\%$	$\mathbf{79.6} \pm 0.8\%$

Table 2: Performance of the *Baseline* listener architecture using different object representations and with/without word level attention, in two generalization tasks.

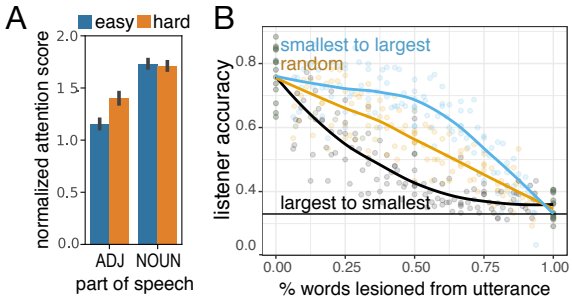


Figure 2: (A) The listener places more attention on adjectives in hard (orange) triplets than easy (blue) ones. (B) Lesioning highest attention words to lowest worsens performance more than lesioning random words or lesioning lowest attention words.

Visual ablations To test the extent to which our listener is relying on the same semantic *parts* of the object as humans, we next conducted a lesion experiment on the visual input. We took the subset of our test set where (1) all chairs had complete part annotations available [38] and (2) the corresponding utterance mentioned a *single* part (17.5% of our test set). We then created lesioned versions of all three objects on each trial by removing pixels of images (and/or points when point-clouds are used), corresponding to parts according to two schemes: *removing* a single part or *keeping* a single part. We did this either for the mentioned one, or another part, chosen at random. We report listener accuracies on these lesioned objects in Table 3. We found that removing random parts hurts the accuracy by 11.5% on average, but removing the mentioned part dropped accuracy more than three times as much, nearly to chance. Conversely, keeping only the mentioned part while lesioning the rest of the image merely drops accuracy by 11.9% while keeping a non-mentioned (random) part alone brings accuracy down close to chance. In other words, on trials when participants depended on information about a part to communicate the object to their partner, we found that visual information about that part was both *necessary and sufficient* for the performance of our listener model.

	Single Part Lesioned	Single Part Present
Mentioned Part	$41.8\% \pm 0.1$	$66.6\% \pm 0.1$
Random Part	$67.0\% \pm 0.2$	$37.4\% \pm 0.1$

Table 3: Evaluating the part-awareness of neural listeners by lesioning object *parts*. Results shown are for image-only listeners, with average accuracy of 78.5% when *intact* objects are used. Similar findings regarding point-cloud-based lesioning are provided in the Supplemental Material.

5. Neural speakers

Architecture We next explore models that learn to generate an utterance that refers to the target and which distinguishes it from the distractors. Similarly to a neural listener the heart of these models is an LSTM which encodes the objects of a communication context, and then decodes an utterance. Specifically, for an *image-based* model, on the first three time steps, the LSTM input is the VGG code of each object. Correspondingly, for a *point-cloud-based* model, the LSTM input is the object codes extracted from a PC-AE. During training and after the objects are encoded, the remaining input to the LSTM is the 'current' utterance token, while the output of the LSTM is compared with the 'next' utterance token, under the cross-entropy loss [34]. The target object is always presented last, eliminating the need to represent the index of the target separately. To find the best model hyper-parameters (e.g. L_2-weights, dropout-rate and # of LSTM neurons) and the optimal stopping epoch, we

sample synthetic utterances from the model during training and use a pretrained *listener* to select the result with the highest listener accuracy. We found this approach to produce models and parameters that yield better quality utterances than evaluating with listening-unaware metrics like BLEU [24].

Variations The above (*literal*) speakers can learn to generate language that discriminates targets from distractors. To test the degree to which distractor objects are used for generation, we experiment with *context-unaware* speakers that are provided the encoding of the target object *only*, and are otherwise identical to the above models. Motivated by the recursive social reasoning characteristic of human pragmatic language use (as formalized in the Rational Speech Act framework [8]), we create *pragmatic* speakers that choose utterances according to their capacity to be discriminative, as judged by a pretrained "internal" listener. In this case, we sample utterances from the (*literal*) speakers, but score (i.e. re-rank) them with:

$$\beta \log(P_L(t|U,O)) + \frac{(1-\beta)}{|U|^\alpha} \log(P_S(U|O,t)), \qquad (1)$$

where P_L is the listener's probability to predict the target (t) and P_S is the likelihood of the *literal* speaker to generate U. The parameter α controls a length-penalty term to discourage short sentences [36], while β controls the relative importance of the speaker's vs. the listener's opinions.

6. Speaker experiments

Qualitatively, our speakers produce good object descriptions, see Fig. 3 for examples, with the pragmatic speakers yielding more discriminating utterances.[4] To quantitatively evaluate the speakers we measure their success in reference games with two different kinds of partners: with an independently-trained listener model and with human listeners. To conduct a *fair* study when we used a neural listener, we split the training data in half. The evaluating listener was trained using one half, while the scoring (or "internal") listener used by the pragmatic speaker was trained on the remaining half. For our human evaluation, we used the *literal* and *pragmatic* variants to generate referring expressions on the test set (we use all training data to train the internal listeners here). We then showed these referring expressions to participants recruited on AMT and asked them to select the object from context that the speaker was referring to. We collected approximately 2.2 responses for each triplet (we use 1200 unique triplets from the *object-generalization* test-split, annotated separately by each speaker model). The synthetic utterances used were the highest scoring ones (Eq. 1) for each model with optimal (per-validation) α and a $\beta = 1.0$. We note that while the *point-based* speakers operate *solely* with point-cloud representations, we display their produced utterances to AMT participants accompanied by CAD rendered images, to keep the human-side presentation identical across experiments.

[4]The anonymized project webpage contains additional qualitative results: https://www.bit.ly/geo_glot

Table 4: Evaluating neural speakers operating with Point Cloud or Image object representations.

Speaker Architecture	Modality	Neural Listener	Human Listener
Context Unaware	Point Cloud	$59.1 \pm 2.0\%$	-
	Image	$64.0 \pm 1.7\%$	-
Literal	Point Cloud	$71.5 \pm 1.3\%$	66.2
	Image	$76.6 \pm 1.0\%$	68.3
Pragmatic	Point Cloud	$90.3 \pm 1.3\%$	69.4
	Image	**92.2** $\pm 0.5\%$	**78.7**

We found (see Table 4) that our *pragmatic* speakers perform best with both synthetic and human partners. While their success with the synthetic listener model may be unsurprising, given the architectural similarity of the internal listener and the evaluating listener, *human* listeners were 10.4 percentage points better at picking out the target on utterances produced by the *pragmatic* vs. *literal* speaker for the best-performing (*image-based*) variant. We also found an asymmetry between the listening and speaking tasks: while context-unaware listeners achieved high performance, we found that context-unaware speakers fare significantly worse than context-aware ones. Last, we note that both literal and pragmatic speakers produce *succinct* descriptions (average sentence length 4.21 vs. 4.97) but the pragmatic speakers use a much richer vocabulary (14% more unique nouns and 33% more unique adjectives, after controlling for average length discrepancy).

7. Out-of-distribution transfer learning

Language is abstract and compositional. These properties make language use generalizable to new situations (e.g. using concrete language in novel scientific domains) and robust to low-level perceptual variation (e.g. lighting). In our final set of experiments we examine the degree to which our neural listeners and speakers learn representations that are correspondingly *robust*: that capture associations between the visual and the linguistic domains permit generalization out of the training domain.

Understanding out-of-class reference To test the generalization of listeners to novel stimuli, we collected referring expressions in communication contexts comprised of objects in Shapenet drawn from new classes: beds, lamps,

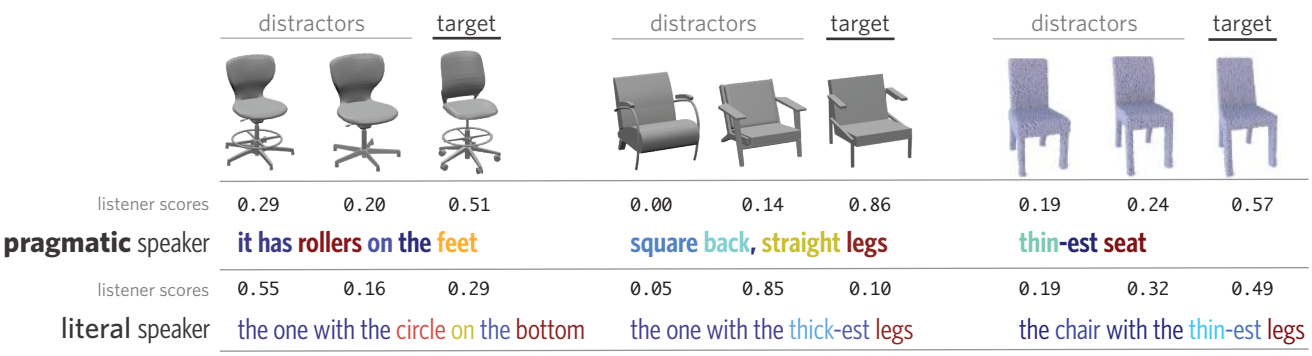


Figure 3: *Pragmatic* vs. *literal* speakers in *unseen* ('hard') contexts. The pragmatic generations successfully discern the target, even in cases where the literal ones fail. The two left-most examples are based on image-based speakers/listeners, the right-most with point-cloud-*based*. The utterances are color coded according to the attention placed by an evaluating neural listener whose classification scores are shown above each corresponding utterance.

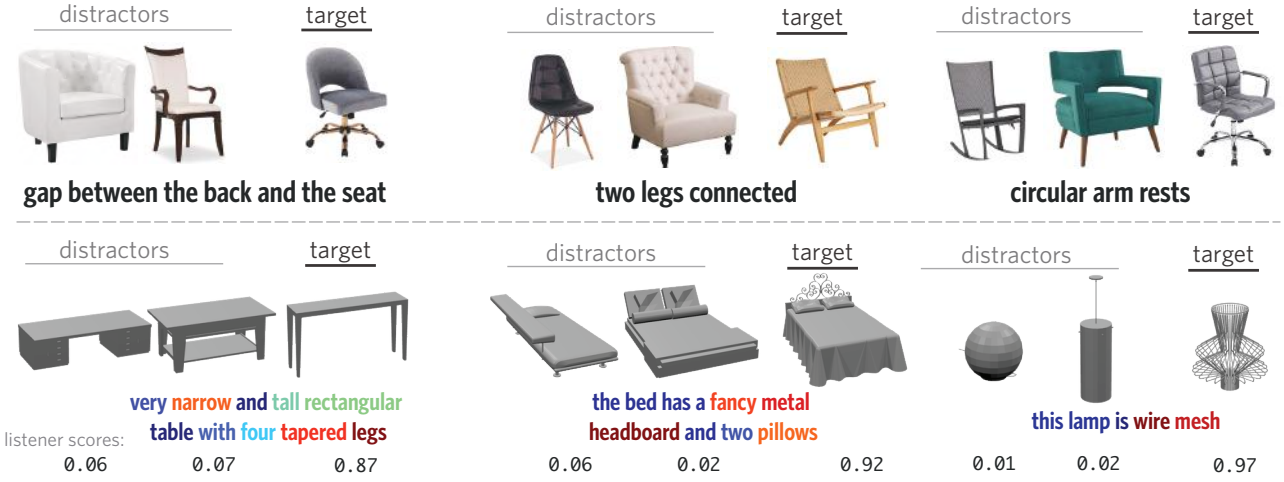


Figure 4: Examples of *out-of-distribution* neural speaking and listening. Top row: model generations for real-world catalogue images. The speaker successfully detects fine grained geometric differences despite large variations in color. Bottom row: responses of a word-attentive listener with CAD renderings of objects from unseen classes and human-produced utterances. The listener detects the (often localized) visual elements that are mentioned by humans, despite the large visual discrepancy of these stimuli from the chairs of the training-domain.

sofas and tables. These classes are distinct from chairs, but share some parts and properties, making transfer possible for a sufficiently compositional model. For each of these classes we created 200 contexts comprised of random triplets of objects; we collected 2 referring expressions for each target in each context (from participants on AMT). Examples of visual stimuli and collected utterances are shown in Fig. 4 (bottom-row). To this data, we applied an (image-only, with/without-attention) listener trained on the CiC (i.e. chairs) data. We avoid using point-clouds since unlike VGG which was finetuned with multiple ShapeNet classes, the PC-AE was pre-trained on a single-classs.

As shown in Table 5, the average accuracy is well above chance in all transfer categories (56% on average). Moreover, constraining the evaluation to utterances that contain *only* words that are in the CiC training vocabulary (75% of all utterances, column: *known*) only slightly improves the results. This is likely because utterances with unknown words still contain enough known vocabulary for the model to determine meaning. We further dissect the *known* population into utterances that contain part-related words (*with-part*) and their complement (*without-part*). For the training domain of chairs without-part utterances yield slightly higher accuracy. However the useful subcategories that support this perfromance (e.g. "recliner") do not support transfer to new categories. Indeed, we observe that for transfer

classes (except sofa) the listener performs better when part-related words are present. Furthermore, the performance gap between the two populations appears to become larger as the perceptual distance between the transfer and training domains increases (compare sofas to lamps).

Table 5: Transfer-learning of neural listeners in novel object *classes*. The sub-populations denote *entire*: all collected utterances, *known*: utterances containing *only* chair-training-vocabulary words, *with-part*: subset of *known*, with utterances containing at least one part-related word, *without-part* subset of *known* and complement of *with-part*. For reference the test-chair statistics are shown (first row) but not included in the reported average (last row). The numbers are *averages* of five listeners trained on different chairs splits. More details are provided in the Supplemental Material.

	Population			
Class	entire	known	with part	without part
chair	77.4	77.8	77.0	**80.5**
bed	56.4	55.8	**63.8**	51.5
lamp	50.1	51.9	**60.3**	47.1
sofa	53.6	55.0	**55.1**	54.7
table	63.7	65.5	**68.3**	62.7
average	56.0	57.1	**61.9**	54.9

Describing real images Transfer from synthetic data to real data is often difficult for modern machine learning models, that are attuned to subtle statistics of the data. We explored the ability of our models to transfer to real chair images (rather than the training images which were rendered without color or texture from CAD models) by curating a modest-sized (300) collection of chair images from online furniture catalogs. These images were taken from a *similar* view-point to that of the training renderings and have rich color and texture content. We applied the (image-only) *pragmatic* speaker to these images, after subtracting the average ImageNet RGB values (i.e. before passing the images to VGG). Examples of the speaker's productions are shown in Figure 4. For each chair, we randomly selected two distractors and asked 2 AMT participants to guess the target given the utterance produced by our speaker. Human listeners correctly guessed the target chair 70.1% of the time. Our speaker appears to transfer successfully to real images, which contain color, texture, pose variation, and likely other differences from our training data.

8. Related work

Image labeling and captioning Our work builds on recent progress in the development of vision models that involve some amount of language data, including object categorization [28, 42] and image captioning [13, 33, 37]. Unlike object categorization, which pre-specifies a fixed set of class labels to which all images must project, our systems use open-ended, referential language. Similarly to other recent works in image captioning [20, 22, 40, 31, 19, 18, 39], instead of captioning a single image (or entity therein), in isolation, our systems learn how to communicate across diverse communications contexts.

Reference games In our work we use reference games [14] in order to operationalize the demand to be relevant in context. The basic arrangement of such games can be traced back to the language games explored by Wittgenstein [35] and Lewis [17]. For decades, such games have been a valuable tool in cognitive science to quantitatively measure inferences about language use and the behavioral consequences of those inferences [26, 15, 4, 30]. Recently, these approaches have also been adopted as a benchmark for discriminative or context-aware NLP [23, 2, 29, 32, 21, 5, 16].

Rational speech acts framework Our models draw on recent formalization of human language use in the Rational Speech Acts (RSA) framework [8]. At the core of RSA is the Gricean proposal [10] that speakers are agents who select utterances that are parsimonious yet informative about the state of the world. RSA formalizes this notion of informativity as the expected reduction in the uncertainty of an (internally simulated) listener, as our pragmatic speaker does. The literal listener in RSA uses semantics that measure compatibility between an utterance and a situation, as our baseline listener does. Previous work has shown that RSA models account for context sensitivity in speakers and listeners [9, 21, 41, 7]. Our results add evidence for the effectiveness of this approach in complex domains.

9. Conclusion

We explored in this paper models of natural language grounded in the intrinsic geometry of objects. Geometry is complex and the language we have for referring to geometry is correspondingly abstract and compositional. This makes geometry an ideal domain for exploring grounded language learning, while making language an especially intriguing source of evidence for models of geometry. We introduced the Chairs-in-Context corpus of highly descriptive referring expressions for objects in context. Using this data we explored a variety of neural listener and speaker models, finding that the best variants exhibited strong performance. These models draw on both 2D and 3D object representations and appear to reflect human-like part decomposition, though they were never explicitly trained with object parts. Finally, we found that the learned models are surprisingly robust, transferring to real images and to new classes of objects. Future work will be required to understand the transfer abilities of these models and how this depends on the compositional structure they have learned.

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. J. Guibas. Learning representations and generative models for 3d point clouds. *Proceedings of the 35th International Conference on Machine Learning*, 2018. 3

[2] J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. *CoRR*, 2016. 8

[3] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 3

[4] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986. 8

[5] R. Cohn-Gordon, N. Goodman, and C. Potts. Pragmatically informative image captioning with character-level reference. *CoRR*, abs/1804.05417, 2018. 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[7] D. Fried, J. Andreas, and D. Klein. Unified pragmatic models for generating and following instructions. *CoRR*, abs/1711.04987, 2017. 8

[8] N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818 – 829, 2016. 6, 8

[9] C. Graf, J. Degen, R. X. D. Hawkins, and N. D. Goodman. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2016. 8

[10] H. P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, pages 43–58. Academic Press, New York, 1975. 8

[11] R. X. D. Hawkins. Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4):966–976, 2015. 2

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[13] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 8

[14] S. Kazemzadeh, V. Ordonez, M. Mark, and B. L. Tamara. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 8

[15] R. M. Krauss and S. Weinheimer. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1964. 8

[16] A. Lazaridou, K. M. Hermann, K. Tuyls, and S. Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *CoRR*, abs/1804.03984, 2018. 8

[17] D. Lewis. *Convention: A philosophical study*. Harvard University Press, 1969. 8

[18] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. *CVPR*, 2018. 8

[19] R. Luo and G. Shakhnarovich. Comprehension-guided referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 8

[20] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and M. Kevin. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283, 2016. 8

[21] W. Monroe, R. X. Hawkins, N. D. Goodman, and C. Potts. Colors in context: A pragmatic neural model for grounded language understanding. *CoRR*, abs/1703.10186, 2017. 3, 8

[22] K. V. Nagaraja, I. V. Morariu, and D. S. Larry. Modeling context between objects for referring expression understanding. *ECCV*, 2016. 8

[23] M. Paetzel, D. N. Racca, and D. DeVault. A multimodal corpus of rapid dialogue games. In *Language Resources and Evaluation Conference (LREC)*, 2014. 8

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 6

[25] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[26] S. Rosenberg and B. D. Cohen. Speakers' and listeners' processes in a word-communication task. *Science*, 1964. 8

[27] S. Shen and H. Lee. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *CoRR*, abs/1604.00077, 2016. 4

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8

[29] J.-C. Su, C. Wu, H. Jiang, and S. Maji. Reasoning about fine-grained attribute phrases using reference games. *CoRR*, abs/1708.08874, 2017. 8

[30] K. van Deemter. *Computational models of referring: a study in cognitive science*. MIT Press, 2016. 8

[31] R. Vedanta, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. *CoRR*, abs/1701.02870, 2017. 8

[32] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. In *Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. 8

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2015. 8

[34] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1989. 5

[35] L. Wittgenstein. *Philosophical investigations*. Macmillan, 1953. 8

[36] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser,

S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. S. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. 6

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2016. 8

[38] L. Yi, H. Su, X. Guo, and L. J. Guibas. Syncspeccnn: Synchronized spectral CNN for 3d shape segmentation. *CoRR*, abs/1612.00606, 2016. 5

[39] L. Yu, Z. Lin, X. Shen, Y. Jimei, X. Lu, M. Bansal, and L. T. Berg. Mattnet: Modular attention network for referring expression comprehension. *CVPR*, 2018. 8

[40] L. Yu, P. Poirson, S. Yang, C. A. Berg, and L. T. Berg. Modeling context in referring expressions. *ECCV*, 2016. 8

[41] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint speaker-listener-reinforcer model for referring expressions. *CoRR*, abs/1612.09542, 2017. 8

[42] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014. 8