

000
001
002
003
004
005
006
007
008
009

GeoGlot: Learning to Ground Referential Language in Geometry

Anonymous ICCV submission

Paper ID 2872

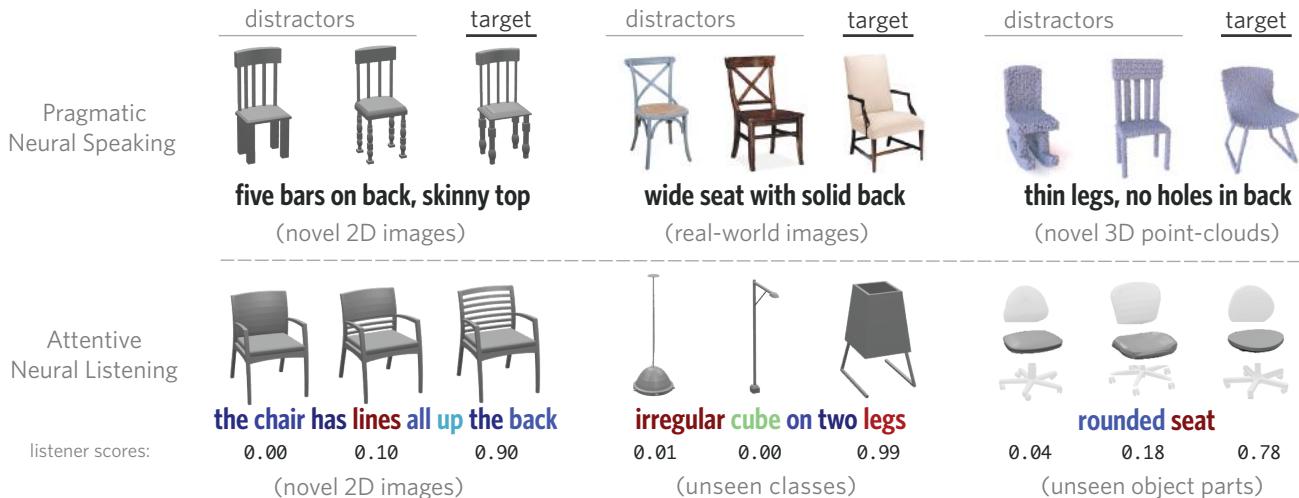


Figure 1: We introduce a corpus of utterances that refer to geometric objects (chairs) and develop neural speakers and listeners with surprising generalization flexibility. Top row: *model-generated* utterances distinguishing a target from two distractors. The speaker is applied (left to right) in unseen: 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.

108

1. Introduction

109

110 People understand visual objects compositionally, in terms of parts and their relations. Language for referring 111 to objects reflects this structure, allowing us to describe 112 fine-grained geometric differences. For instance, we 113 perceive a typical chair as a geometric arrangement of parts 114 like *arms*, *legs*, *seat*, and *back*; and we can compose words 115 to describe its ‘wide seat’, combine these into ‘wide seat 116 with solid back’, and refer to part relations such as ‘a hole 117 between the back and the seat’. We use this rich space of 118 expressions to craft references which are not merely true, 119 but are relevant in the current communication context. For 120 instance choosing referring expressions that *distinguish* one 121 chair from another (Fig. 1). In this paper we investigate the 122 interplay between referential language and intrinsic geometry 123 of complex objects. By gathering fine-grained referring 124 expressions from a reference game experiment, we are 125 able to train language understanding and production models 126 grounded in the geometry of objects. We find that these 127 models use complex part-based language, and transfer 128 surprisingly well to new domains and to real images.

129

130 While a great deal of recent work has explored visually- 131 grounded language understanding, the resulting models 132 have limited capacity to reflect the intrinsic geometry of 133 objects. Indeed, eliciting natural language that refers only 134 to geometry requires carefully controlling the objects, their 135 presentation, and the linguistic task. To address these 136 challenges, we leverage clean 3D representations of objects 137 (CAD models), which allow for flexible and controlled 138 presentation (i.e. textureless, uniform-color objects viewed 139 in a fixed pose). We further make use of the 3D form to 140 construct a reference game task in which the target object is 141 *geometrically* similar to the distractors. The result of this 142 effort is a new multimodal dataset, termed *CiC (Chairs in* 143 *Context)*, comprised of 4,511 unique chairs from ShapeNet 144 and 78,789 referential utterances. In CiC chairs are orga- 145 nized into 4,054 sets of size 3 (called communication 146 contexts) and each utterance is intended to distinguish a chair in 147 context. The visual differences among the grouped objects 148 require a deep understanding of fine-grained object geo- 149 metry; the language that people use to do so is correspondingly 150 complex, exhibiting rich compositionality.

151

152 We use this data to train a variety of modern neural lan- 153 guage understanding and production models. These mod- 154 els vary in their grounding (pure 3D forms via point clouds 155 vs. rendered images), the degree of pragmatic reasoning 156 captured, and the neural architecture (e.g. with or with- 157 out self-attention within the utterance encoder). We eval- 158 uate these models on the original language game task with 159 both synthetic and human partners, and with held out ut- 160 terances and objects, finding strong performance. Since 161 language conveys geometric abstractions, such as parts, that 162 are shared between categories, we hypothesized that our 163

164 models could transfer to new classes (e.g. training on chairs 165 while testing on lamps). We find that these models have 166 surprisingly strong generalization capacity to novel object 167 classes, as well as to real images drawn from furniture cat- 168 alogs.

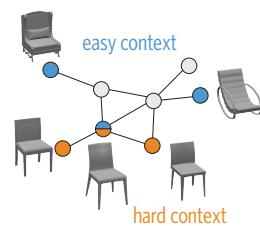
169 Finally, we explore how our models are succeeding on 170 these tasks. We demonstrate that the neural listener learns 171 to prioritize the same geometric information in objects 172 (i.e. properties of chair *parts*) that humans do in solving the 173 communication task, despite never being provided with an 174 explicit decomposition of these objects into parts. We also 175 show that a neural speaker that is *pragmatic*—planning utter- 176 ances in order to convey the right target object to an imagined 177 listener—produces significantly more useful utterances 178 than a *literal* (listener-unaware) speaker, as measured by *hu-* 179 *man* performance in identifying the correct object.

180 Overall, our results demonstrate how to elicit data for 181 linguistic reference to complex visual objects and how to 182 use these data to train neural models that capture structure 183 latent in the language and its grounding.

2. Dataset and task

184 Our dataset consists of triplets of objects coupled with 185 referential utterances that aim to distinguish one object (the 186 “target”) from the remaining two (the “distractors”). To 187 obtain such utterances, we paired participants from 188 Amazon’s Mechanical Turk (AMT) to play an online, reference 189 game [11]. On each round of the game, the two players 190 were shown a triplet of objects. The designated target ob- 191 ject was privately highlighted for one player (the “speaker”) 192 who was asked to send a message through a chat box such 193 that their partner (the “listener”) could successfully select it 194 from the context. To ensure speakers used geometric infor- 195 mation *only*, we scrambled the positions of the objects for 196 each participant and used textureless, uniform-color render- 197 ings of pre-aligned 3D objects taken from the same view- 198 point. To ensure communicative interaction was natural, 199 no constraints were placed on the chat box: referring ex- 200 pressions from the speaker were occasionally followed by 201 clarification questions from the listener or other discourse.

202 A crucial decision in build- 203 ing our dataset concerned 204 the construction of useful 205 contexts that would reliably 206 elicit diverse, *fine-grained* 207 contrastive language. The set 208 of objects must be familiar 209 so we can tap existing visual 210 and linguistic representations. 211 The objects should also be 212 complex and variable to 213 provide diverse geometries and language. To accomplish 214 these goals we utilize the collection of about 7,000 chairs 215



216 from ShapeNet [3]. This class is geometrically complex,
217 highly diverse, and abundant in the real world.
218

219 Perceptually identical objects cannot be distinguished
220 with language at all, while wildly different objects (a
221 chair and a car) can be distinguished with a single word
222 (“chair”). Different contexts must thus contain geometri-
223 cally similar objects to promote the usage of very fine-
224 grained and geometry-related language, in solving the task.
225 To achieve this goal in a scalable manner, we estimated ob-
226 ject similarity between different chairs in the latent space
227 produced by the Point Cloud-AutoEncoder (PC-AE) from
228 [1]. This approach allowed us to leverage the fact that
229 point-clouds extracted from a 3D surface are an intrinsic
230 3D representation of an object, oblique to color or tex-
231 ture. To deal with the inhomogeneity of data in repositories
232 like ShapeNet we used a sampling strategy to construct our
233 triplets. First, we computed the 2-nearest-neighbor graph
234 of all ShapeNet chairs based on their PC-AE embedding
235 distances. On this graph, we selected chairs of highest in-
236 degree which intuitively correspond to the most canonical
237 chairs of the collection, and for each selected chair we gen-
238 erated two kinds of contexts. *Hard* contexts consist of very
239 similar (close in latent space) chairs, while *Easy* contexts
240 consist of less similar (far in latent space) chairs (see inset).¹
241 Human performance on the reference game was high in
242 general, but listeners made significantly more errors in the
243 hard triplets (accuracy 94.2% vs. 97.2%, $z = 13.54, p <$
244 0.001). Significantly longer utterances were used on aver-
245 age to describe targets in hard triplets (approximately 8.4
246 words vs. 6.1, $t = -35, p < 0.001$). A wide spectrum
247 of descriptions was elicited, ranging from the more holis-
248 tic/categorical common for easy triplets (e.g. “the rocking
249 chair”) to more complex, geometric language common for
250 hard triplets (e.g. “thinner legs but without armrests”). In-
251 terestingly, 78% of utterances used at least one part-related
252 word: “back”, “legs”, “seat,” “arms”, or closely related syn-
253 onyms e.g. “armrests”.

254 3. Neural listeners 255

256 Constructing neural listeners that reason about geometric
257 properties is an important contribution of our work. Below
258 we conduct a detailed comparison between three distinct ar-
259 chitectures, highlight the effect of different regularization
260 techniques, and investigate the merits of different *represen-*
261 *tations* of 3D objects for the listening task, namely, 2D ren-
262 dered images and 3D surface point clouds. In what follows,
263 we denote the three objects of a communication context as
264 $O = \{o_1, o_2, o_3\}$, the corresponding word-tokenized utter-
265 ance as $U = u_1, u_2, \dots$ and as $t \in O$ the referential target.

266
267 ¹Additional details about materials and experimental methods can be
268 found in Supplemental Material. The code and data for all experiments in
269 this paper will be publicly available upon acceptance.

270 Our proposed listener is inspired by [21]. It takes as input
271 a (latent code) vector that captures geometric informa-
272 tion for each of the objects in O , and a (latent code) vector
273 for each token of U , and outputs an object–utterance com-
274 patibility score $\mathcal{L}(o_i, U) \in [0, 1]$ for each input object. At
275 its core lies a multi-modal LSTM [12] that receives as input
276 (“is grounded” with) the vector of a single object, processes
277 the word-sequence U , and is read out by a final MLP to
278 yield a single number (the compatibility score). This is re-
279 peated for each o_i , *sharing* all network parameters across
280 the objects. The resulting three scores are soft-max nor-
281 malized and compared to the ground-truth indicator vector of
282 the target under the cross-entropy loss.²

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

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

315 **Incorporating context information** Our proposed
316 baseline listener architecture (*Baseline*, just described) first
317 scores each object *separately* then applies softmax normal-
318 ization to yield a score distribution over the three objects.
319 We also consider two more complex architectures that ex-
320 plicitly encode information about the *entire* context while

321
322 ²Architecture details, hyper-parameter search strategy, and optimal
323 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 generaliza-

³In 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.

tion 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 (~14%) of our test set that use explicitly contrastive language: superlatives (“skinniest”) and comparatives (“skinier”). 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.

432
433
434
435
436
437

	Architecture	Overall	Subpopulations			486
			Hard	Easy	Sup-Comp	
	<i>Combined-Interpretation</i>	$75.9 \pm 0.5\%$	$67.4 \pm 1.0\%$	$83.8 \pm 0.6\%$	$74.4 \pm 1.5\%$	487
	<i>Early-Context</i>	$79.4 \pm 0.8\%$	$\mathbf{70.1} \pm 1.3\%$	$88.1 \pm 0.6\%$	$75.6 \pm 2.2\%$	488
	<i>Baseline</i>	$\mathbf{79.6} \pm 0.8\%$	$69.9 \pm 1.3\%$	$\mathbf{88.8} \pm 0.4\%$	$\mathbf{76.3} \pm 1.3\%$	489

438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

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	$\mathbf{83.7} \pm 0.3\%$	$\mathbf{79.6} \pm 0.8\%$

452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

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

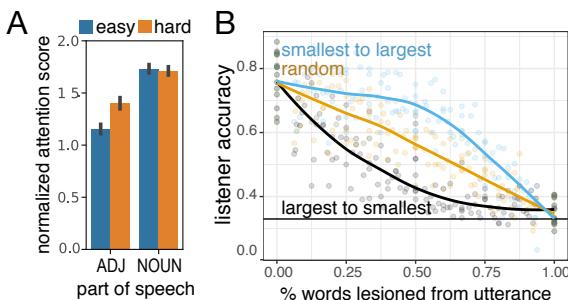
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

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

492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

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	$67.0\% \pm 0.2$	$37.4\% \pm 0.1$

511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

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:

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

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

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

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.

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,

648		distractors	target	distractors	target	distractors	target	702
649								703
650								704
651								705
652								706
653								707
654		listener scores	0.29 0.20 0.51	0.00 0.14 0.86	0.19 0.24 0.57			708
655	pragmatic speaker	it has rollers on the feet		square back, straight legs		thin-est seat		709
656								710
657		listener scores	0.55 0.16 0.29	0.05 0.85 0.10	0.19 0.32 0.49			711
658	literal speaker	the one with the circle on the bottom		the one with the thick-est legs		the chair with the thin-est legs		712
659								713
660								714
661								715
662								716
663								717
664								718
665		distractors	target	distractors	target	distractors	target	719
666								720
667								721
668								722
669								723
670								724
671	gap between the back and the seat			two legs connected		circular arm rests		725
672								726
673		distractors	target	distractors	target	distractors	target	727
674								728
675								729
676								730
677								731
678	very narrow and tall rectangular table with four tapered legs			the bed has a fancy metal headboard and two pillows		this lamp is wire mesh		732
679	listener scores:	0.06 0.07 0.87		0.06 0.02 0.92		0.01 0.02 0.97		733
680								734

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.

681		distractors	target	distractors	target	distractors	target	735
682								736
683								737
684								738
685								739
686								740
687								741
688	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.							742
689								743
690								744
691								745
692								746
693								747
694								748
695								749
696								750
697								751
698								752
699								753
700								754
701								755

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.

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 performance (e.g. “recliner”) do not support transfer to new categories. Indeed, we observe that for transfer

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

761
 762 Table 5: Transfer-learning of neural listeners in novel object
 763 *classes*. The sub-populations denote *entire*: all collected utterances,
 764 *known*: utterances containing *only* chair-training-vocabulary words,
 765 *with-part*: subset of *known*, with utterances containing at least one part-related word,
 766 *without-part* subset of *known* and complement of *with-part*. For reference
 767 the test-chair statistics are shown (first row) but not included in the reported average (last row). The numbers are
 768 *averages* of five listeners trained on different chairs splits.
 769 More details are provided in the Supplemental Material.
 770
 771

Class	Population			
	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

784 **Describing real images** Transfer from synthetic data to
 785 real data is often difficult for modern machine learning
 786 models, that are attuned to subtle statistics of the data. We
 787 explored the ability of our models to transfer to real chair
 788 images (rather than the training images which were ren-
 789 dered without color or texture from CAD models) by cu-
 790 rating a modest-sized (300) collection of chair images from
 791 online furniture catalogs. These images were taken from
 792 a *similar* view-point to that of the training renderings and
 793 have rich color and texture content. We applied the (image-
 794 only) *pragmatic* speaker to these images, after subtracting
 795 the average ImageNet RGB values (i.e. before passing the
 796 images to VGG). Examples of the speaker’s productions are
 797 shown in Figure 4. For each chair, we randomly selected
 798 two distractors and asked 2 AMT participants to guess the
 799 target given the utterance produced by our speaker. Hu-
 800 man listeners correctly guessed the target chair 70.1% of
 801 the time. Our speaker appears to transfer successfully to
 802 real images, which contain color, texture, pose variation,
 803 and likely other differences from our training data.

8. Related work

804 **Image labeling and captioning** Our work builds on
 805 recent progress in the development of vision models that
 806 involve some amount of language data, including object cate-

gorization [28, 42] and image captioning [13, 33, 37]. Un-
 810 like object categorization, which pre-specifies a fixed set of
 811 class labels to which all images must project, our systems
 812 use open-ended, referential language. Similarly to other re-
 813 cent works in image captioning [20, 22, 40, 31, 19, 18, 39],
 814 instead of captioning a single image (or entity therein), in
 815 isolation, our systems learn how to communicate across di-
 816 verse communications contexts.

817 **Reference games** In our work we use reference games
 818 [14] in order to operationalize the demand to be relevant
 819 in context. The basic arrangement of such games can be
 820 traced back to the language games explored by Wittgenstein
 821 [35] and Lewis [17]. For decades, such games have been a
 822 valuable tool in cognitive science to quantitatively measure
 823 inferences about language use and the behavioral conse-
 824 quences of those inferences [26, 15, 4, 30]. Recently, these
 825 approaches have also been adopted as a benchmark for dis-
 826 criminative or context-aware NLP [23, 2, 29, 32, 21, 5, 16].

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

9. Conclusion

841 We explored in this paper models of natural language
 842 grounded in the intrinsic geometry of objects. Geometry
 843 is complex and the language we have for referring to
 844 geometry is correspondingly abstract and compositional. This
 845 makes geometry an ideal domain for exploring grounded
 846 language learning, while making language an especially in-
 847 triguig source of evidence for models of geometry. We in-
 848 troduced the Chairs-in-Context corpus of highly descriptive
 849 referring expressions for objects in context. Using this data
 850 we explored a variety of neural listener and speaker models,
 851 finding that the best variants exhibited strong performance.
 852 These models draw on both 2D and 3D object represen-
 853 tations and appear to reflect human-like part decomposi-
 854 tion, though they were never explicitly trained with object parts.
 855 Finally, we found that the learned models are surprisingly
 856 robust, transferring to real images and to new classes of ob-
 857 jects. Future work will be required to understand the trans-
 858 fer abilities of these models and how this depends on the
 859 compositional structure they have learned.

864

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. Paetzl, 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,

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

- 972 S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, 1026
973 G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, 1027
974 A. Rudnick, O. Vinyals, G. S. Corrado, M. Hughes, and 1028
975 J. Dean. Google’s neural machine translation system: Bridg- 1029
976 ing the gap between human and machine translation. *CoRR*, 1030
977 abs/1609.08144, 2016. 6 1031
978 [37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, 1032
979 R. Zemel, and Y. Bengio. Show, attend and tell: Neural 1033
980 image caption generation with visual attention. *CoRR*, 1034
981 abs/1502.03044, 2016. 8 1035
982 [38] L. Yi, H. Su, X. Guo, and L. J. Guibas. Syncspeccnn: Syn- 1036
983 chronized spectral CNN for 3d shape segmentation. *CoRR*, 1037
984 abs/1612.00606, 2016. 5 1038
985 [39] L. Yu, Z. Lin, X. Shen, Y. Jimei, X. Lu, M. Bansal, and L. T. 1039
986 Berg. Mattnet: Modular attention network for referring ex- 1040
987 pression comprehension. *CVPR*, 2018. 8 1041
988 [40] L. Yu, P. Poirson, S. Yang, C. A. Berg, and L. T. Berg. Mod- 1042
989 eling context in referring expressions. *ECCV*, 2016. 8 1043
990 [41] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint speaker- 1044
991 listener-reinforcer model for referring expressions. *CoRR*, 1045
992 abs/1612.09542, 2017. 8 1046
993 [42] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part- 1047
994 based r-cnns for fine-grained category detection. In *Eu- 1048
995 ropean conference on computer vision*, pages 834–849. Springer, 2014. 8 1049
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025