Learning Language from Vision

Candace Ross, Cheahuychou Mao, Yevgeni Berzak, Boris Katz, and **Andrei Barbu**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{ccross,cmao18,berzak,boris,abarbu}@ mit.edu

Abstract

We develop a semantic parser that is trained in a grounded setting using videos captioned with sentences. This setting is both data-efficient, requiring little annotation, and similar to the experience of children where they observe their environment and listen to speakers. The semantic parser recovers the meaning of English sentences despite not having access to any annotated sentences. It does so despite the ambiguity inherent in vision where a sentence may refer to any combination of objects, object properties, relations or actions taken by any agent in a video. For this task, we collected a new dataset for grounded language acquisition. Learning a grounded semantic parser can ultimately lead to a better understanding of child language acquisition. In addition, we demonstrate a new application for grounded lanugage acquisition, zero-shot natural language inference, by incorporating the grounded parser.

1 Introduction

Children learn language from observations that are very different in nature from what parsers are trained on today. Most of the time, children do not receive direct feedback such as annotated sentences or answers to direct questions. Instead, children observe and interact with their environment to learn the meaning and structure of their language. This weak and indirect supervision where most of the information is obtained through passive observation poses a difficult disambiguation problem for learners: how do you know what the speaker is referring to in the environment, i.e., what the speaker means? Speakers can refer to actions, objects, the properties of actions and objects, relations between those actions and objects, as well as other features in the environment and generally do so by combining multiple features into complex sentences. Moreover, speakers need not refer to the most visually salient parts of a scene. In this paper, we induce a semantic parser by simultaneously resolving visual ambiguities and grounding the semantics of language using a corpus of sentences paired with videos without other annotations.

Semantic parsing converts natural language sentences into representations that encode their meanings. These representations – a lambda-calculus expression in our case – can be used for a variety of tasks such as querying databases, understanding references in images and videos, and answering questions. To train the parser presented here we collected a video dataset of variety of actions and objects and asked annotators on Mechanical Turk to produce sentences that are true of these videos. We balanced the co-occurrences of objects and events such that these statistics (e.g., the action *drop* always occurring with a given object, for instance) are not informative. The parser is presented with pairs of short clips and sentences and hypothesizes potential meanings as lambda-calculus expressions. Each hypothesized expression serves as input for a modular vision system that constructs a specific detector to determine the likelihood of the parse being true of the video. The likelihood of the expression with respect to the video supervises the parser. An example of a video/sentence pair is shown in Figure 1. For evaluation, we test each sentence with its annotated ground-truth semantic parse; this information is not available at training time.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

The woman walks by the table with a yellow cup.
$\lambda xyz.\text{woman}\, x, \text{walk}\, x, \text{near}\, x\, y, \text{table}\, y,\ \text{hold}\, x\, z, \text{yellow}\, z, \text{cup}\, z$

Figure 1: We develop a semantic parser trained on video-sentence pairs, *without parses*. At inference time a sentence, *without a video*, is presented and a logical form is produced.

This work makes several contributions. We 1) construct a semantic parser that learns language in a setting closer to that of children; 2) jointly resolve linguistic and visual ambiguities at training time in an adaptable way; 3) demonstrate how such an approach can be used to augment data by bootstrapping with a small number of labeled examples; and 4) release a dataset systematically constructed using Mechanical Turk for visual and language tasks.[1]

2 Task

Given a dataset of captioned videos, D, we train the parameters and lexicon, θ and Λ, of a semantic parser. At training time, we perform gradient descent over the parameters θ and employ *GENLEX* (Zettlemoyer and Collins, 2005) to augment the lexicon Λ. The objective function of the semantic parser is written in terms of a visual-linguistic compatibility between a hypothesized parse p and video v. This compatibility computes the likelihood of the parse being true of the video, $P(v|p)$. At test time, we take as input a sentence without an associated video and produce a semantic parse. We could in principle also take as input the video and produce a targeted parse for that visual scenario. This is a problem similar to that considered by Berzak et al. (2015), but we do not do so here.

We learn a CCG-based (Combinatory Categorical Grammar; Steedman (1996)) semantic parser capable of being trained in this setting. To do so, we adapt the objective function, training procedure, and feature set to this new scenario. The visual-linguistic compatibility function is similar to the Sentence Tracker developed in Siddharth et al. (2014) and Yu et al. (2015). Given a parse, the Sentence Tracker produces a targeted detector that determines if the parse is true of a video, which provides a weak supervision signal for the parser.

Parses are represented as lambda-calculus expressions consisting of a set of binders and a conjunction of literal expressions referring to those binders. The domain of the variables are the potential object locations, or object tracks, in the videos. For example, in the parse presented in Figure 1, three potential object track slots are available, represented by the binders x, y, and z. Because of perceptual ambiguities and the large number of possible referents in any one video, we do not explicitly enumerate the space of object tracks. Instead, we rely on a joint-inference process between the parser and the Sentence Tracker. Intuitively, each literal expression of the parse asserts a constraint; for example, if an expression conveys that one object is approaching another, the Sentence Tracker will search the space of object tracks and attempt to satisfy these constraints. In Figure 1, for instance, there is a constraint that for whichever objects are bound to x and z, x must be near y, x must be walking, x must be a person, etc.

3 Model

Given a dataset of captioned videos, we learn the weights and lexicon of a CCG-based semantic parser. The objective function of the semantic parser considers the visual-linguistic compatibility between a hypothesized parse p and video v. This compatibility is determined using a Sentence Tracker, which, given a parse, determines the likelihood of the parse describing the video. This compatability supervises the parser. At test time, we take as input a sentence without an associated video and produce a semantic parse.

[1] Code and data are available at https://github.com/candacelax/grounded-vision-parser

3.1 Semantic Parsing

We adopt a parsing framework similar to Artzi and Zettlemoyer (2013). A CCG-based parser takes a sentence and a lexicon and uses fixed rules to map tokens to constants and semantic/syntactic types. Parsing rules are generic, polymorphic, and language-neutral. The combinatory nature produces multiple hypothesized parse trees. The parser accepts a derivation when the tree reaches a single node. We refer to the single node of the parse tree as the logical form.

We adopt a weighted linear semantic parser to score parses following Zettlemoyer and Collins (2005) and Curran et al. (2007). For each sentence paired with its hypothesized derivation, this approach computes a feature vector ϕ and a parameter vector θ to score valid parses (validation process described in the next section). The highest scoring valid parse is deemed the optimal parse. The loss function is a margin-ranking approach where θ is updated to maximize the margin γ between scores of positive and negative parses.

The lexicon Λ is augmented using the modified *GENLEX* from Artzi and Zettlemoyer (2013), which does not require the ground-truth logical form. At no point is the ground-truth logical form used during training; we rely instead on a visual validation function to compute the margin-violating examples. We describe the visual validation method in the next section.

3.2 Sentence Tracking

To provide visual validation for the parser, we employ a framework similar to that of Yu et al. (2015) called a Sentence Tracker (ST). The ST constructs a parse-specific model by extracting the participants in the scene as well as the relationships and properties of those participants. It builds a graphical model where 1) each participant is localized by an object tracker and 2) each relationship is encoded by temporal models that express the properties of these trackers. The lambda-calculus representation output by the parser is ideal for the vision system. The lambda calculus expression contains a set of binders, whose domain are objects, and a conjunction of constraints that refer to those binders. In essence, this notes which objects should be present in a scene and what static and changing properties and relationships those objects should have with respect to one another.

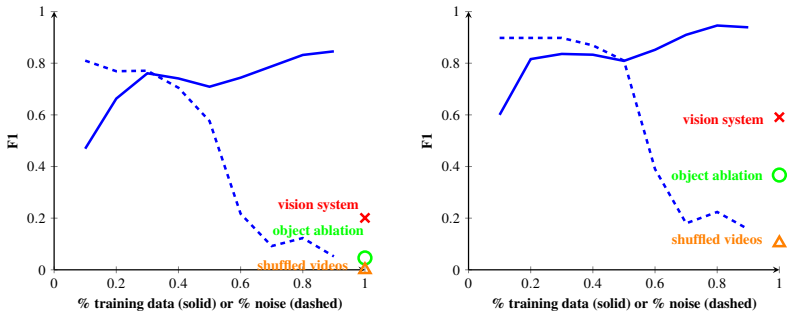
The Viterbi-based ST creates one Hidden Markov model (HMM) for each participant (object detection in the video). Given the mapping between constraints and participants as specified by the parse, it connects the HMMs together forming trackers. Trackers weave these bounding-box object detections into high-scoring object tracks and use constraints to verify if the tracks have the desired properties and relations. We direct the reader to Siddharth et al. (2014) and Yu et al. (2015) for a more thorough mathematical explanation. In short, given a hypothesized parse and a corresponding video, the ST produces a likelihood of the parse describing the video.

3.3 Joint Model

At training time, we learn using both the semantic parser and the ST. At test time, only the parser is used. Two parameters are learned, a set of weights θ and the lexicon Λ. Λ is used to structure inference. The joint model must learn these parameters despite three sources of noise. First, the ST may simply fail to produce the correct likelihood because machine vision is far from perfect. Overcoming this fallible nature of the ST requires large beam widths to avoid falling into local minima due to these errors.

Second, while parsing a sentence, there are an infinite number of hypotheses true of a video yet not true of the sentence. This makes our environment far less constrained and much more ambiguious than most semantic parsing approaches. This results in two key problems: first, that the lexicon could store many words with "empty" semantics (e.g., chair $\vdash$ *object*, table $\vdash$ *object*), and second, that the lexicon could have excessive polysemy by storing many forms for a given word. We introduce two features to the parser that bias it against empty semantics and against excessive polysemy. Models of communication such as the Rational Speech Acts model (Frank and Goodman, 2012) predict that speakers will avoid inserting meaningless words. One feature counts the number of predicates mapped onto semantic forms which are empty that occur in each parse. The other feature counts the number of new semantic forms created during parsing. As parser performance improves during training, these features begin to bias it against adding empty semantics and new semantic forms.

Third, models in computer vision are computationally expensive. Nonetheless, we require many evaluations of parse-video pairs are required to train a parser. To overcome this, we construct a



(a) **Exact match** (b) **Near miss**

Figure 2: Results from training the grounded semantic parser. In solid blue, *direct supervision* as a function of the amount of training data. In dashed blue, *noisy supervision* uses the whole training set but accepts and rejects parses at random for a given fraction of the time. The red cross is the full vision system while the green o is the object detector ablation. The orange triangle represents *shuffled videos* and shows chance performance. While direct supervision outperforms vision-only supervision, the grounded parser closes the gap and operates like noisy direct supervision with roughly 55% noise.

provably-correct cache that keeps track of failing subexpressions. This is possible because of a feature of this particular vision-language scoring function: the score decreases monotonically with the number of constraints. With these improvements, the modified semantic parser employing vision-language-based validation learns to map sentences into semantic parses despite facing a challenging setting with few examples and much ambiguity.

3.4 Evaluation

For our dataset, we recorded videos of various actors, objects and actions in varying relationships. We used Mechanical Turk to generate captions for the videos. In total, the dataset contains 1200 captions from 401 videos, which selected out of a larger body of sentences collected and pruned as described above. Our training, validation and test sets are 720, 120 and 360 examples respectively; sentences and videos are disjoint across sets. We summarize our experiments and ablation studies in Figure 2. Importantly, note that our model (fully grounded vision parser) far outperforms the baseline and corresponds to direct supervision with around 55% noise.

4 Prior work

Learning to understand language in a multimodal environment is a well-developed task. For example, visual question answering (VQA) datasets have led to a number of systems capable of answering complex questions about scenes (Antol et al., 2015). The goal of our work is not to produce answers for any one set of questions; it is instead to learn to predict the structure of the sentences and their meaning. This is a more general and difficult problem, in particular because at test time we do not receive any visual input, only the sentence. The resulting approach is reusable, generic and more similar to the kind of general-purpose linguistic knowledge that humans have. commands.

Siddharth et al. (2014) and Yu et al. (2015) acquire the meaning of predicates in scenes, but assume a fully-trained semantic parser. Matuszek et al. (2012) similarly present a model to learn the meanings and referents of words yet use restricted attributes and static scenes. setting. Wang et al. (2016) create a language game to learn a parser but do not incorporate visual ambiguity or fallible perception.

Berant et al. (2013) describe semantic parsing with execution by annotating answers to database queries. This learning mechanism provides the same results as the one described here: a parser produces the meanings of sentences at inference time without requiring the database, or in our case a video. Databases have far less ambiguity than videos; there is not a temporal aspect to their contents and there is not a notion of unreliable perception. Berant and Liang (2014) learn to parse sentences from paraphrases; one might consider the work here as concerned with visual and not just linguistic paraphrases. Artzi and Zettlemoyer (2013) consider a setting where a validation function involves the dynamic actions of a simulated robot while sentences describe its actions.

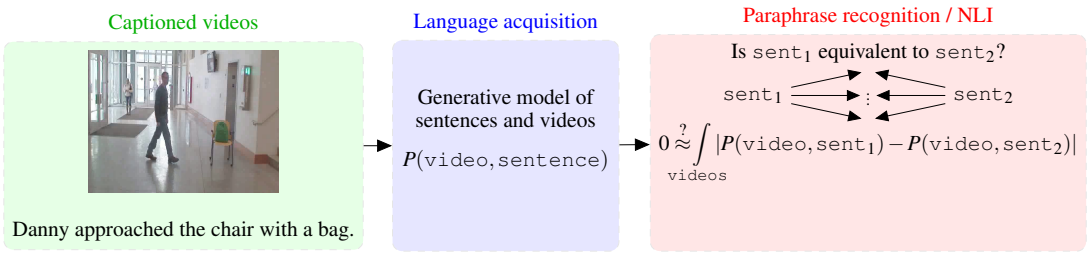


Figure 3: We use a dataset of captioned videos to train a grounded language model. To perform paraphrasing, we use the intuition that if two sentences (derived from different scenes) have identical meaning, they should always be true or false of the same scenes. We sample videos conditioned on the sentences, as a proxy for integrating over all possible videos, and ask if the likelihoods of the two sentences are the same. If they are, the two sentences have identical meaning.

5 Application to Natural Language Inference

Such models of language and vision can be used to not just acquire language but to perform new tasks. Generally, creating models that tackle new tasks requires new task-specific training data. We show that the model used earlier, can perform a new task with its acquired knowledge, even without any training data for that task. In particular, that if we truly understand the semantics of two sentences, we should be able to compare them.

Given two sentences, we use a variant of the model described thus far as a generative model; see Figure 3. With a language model, we parse sentences, and then structure the model of the Sentence Tracker according to the parses of that model. This provides a collection of HMMs that can recognize if a sentence is true of a video, as described earlier. While we have used the model so far as a classifier, we can instead draw samples from this model while conditioning it on a sentence. That allows the model to image new, never-before-seen videos. One sentence is used to sample videos while the other sentence is scored against those videos. If two sentences mean the same thing, they should be true of the same videos, while if one implies the other this should also be apparent in the pattern of likelihoods. In preliminary experiments we achieve around 73.8% on a new corpus we're developing of grounded sentences while BERT achieves 73.2% accuracy. Note that our model is not trained on a single NLI example, while BERT is trained on several tens of thousands of such sentences. This work points the way toward novel applications of grounded reasoning and language learning to account for more facets of human intelligence.

6 Discussion

We present a semantic parser that learns the structure of language using weak supervision from vision. At test time, the model parses sentences without the need for visual input. Learning by passive observation in this way extends the capabilities of semantic parsers and points the way to a more cognitively plausible model of language acquisition. Several limits remain. Evaluating parses as correct or incorrect depending on a match to a human-annotated logical form is an overly strict criterion and is a problem that also plagues fully-supervised syntactic parsing (Berzak et al., 2016). Since two logical forms may express the same meaning, it is not yet clear what an effective evaluation metric is for these grounded scenarios. In addition, learning in such a passive scenario is hard as correlations between events, e.g., every *pick up* event involves a *touch* event, are very difficult to disentangle. Finally, we demonstrate a new application of grounded language acquisition to solving new tasks without requiring task-specific training data.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Artzi, Y. and Zettlemoyer, L. (2013). Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions. *Transactions of the Association for Computational Linguistics*, pages 49–62.

Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic Parsing on Freebase from Question-Answer Pairs. *Empirical Methods for Natural Language Processing*.

Berant, J. and Liang, P. (2014). Semantic parsing via paraphrasing. In *Annual Meeting of the Association for Computational Linguistics*.

Berzak, Y., Barbu, A., Harari, D., and Katz, B. (2015). Do You See What I Mean ? Visual Resolution of Linguistic Ambiguities. *Conference on Empirical Methods on Natural Language Processing*, (September):1477–1487.

Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., and Katz, B. (2016). Anchoring and agreement in syntactic annotations. *Conference on Empirical Methods in Natural Language Processing*.

Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 33–36. Association for Computational Linguistics.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., and Fox, D. (2012). A joint model of language and perception for grounded attribute learning. In *International Conference on Machine Learning*, pages 1671–1678. ACM.

Siddharth, N., Barbu, A., and Mark Siskind, J. (2014). Seeing what you're told: Sentence-guided activity recognition in video. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Steedman, M. (1996). *Surface Structure and Interpretation*. The MIT Press.

Wang, S. I., Liang, P., and Manning, C. D. (2016). Learning language games through interaction. *Meeting of the Association for Computational Linguistics*.

Yu, H., Siddharth, N., Barbu, A., and Siskind, J. M. (2015). A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*.

Zettlemoyer, L. and Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 658–666, Arlington, Virginia. AUAI Press.