arXiv:1604.02125v2 [cs.CV] 5 May 2016

Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes

Gordon Christie[1,*], Ankit Laddha[2,*], Aishwarya Agrawal[1], Stanislaw Antol[1], Yash Goyal[1], Kevin Kochersberger[1], Dhruv Batra[1]

[1]Virginia Tech
[2]Carnegie Mellon University

Abstract. We present an approach to simultaneously perform semantic segmentation and prepositional phrase attachment resolution for captioned images. The motivation for this work comes from the fact that some ambiguities in language simply cannot be resolved without simultaneously reasoning about an associated image. If we consider the sentence "I shot an elephant in my pajamas", looking at the language alone (and not reasoning about common sense), it is unclear if it is the person or the elephant that is wearing the pajamas or both. Our approach involves producing a diverse set of plausible hypotheses for both semantic segmentation and prepositional phrase attachment resolution that are then jointly re-ranked to select the most consistent pair. We show that our semantic segmentation and prepositional phrase attachment resolution modules have complementary strengths, and that joint reasoning produces more accurate results than any module operating in isolation. We also show that multiple hypotheses are crucial to improved multiple-module reasoning. Our vision and language approach significantly outperforms a state-of-the-art NLP system (Stanford Parser [18, 30]) by 17.91% (28.69% relative) in one experiment, and by 12.83% (25.28% relative) in another. We also make small improvements over a state-of-the-art vision system (DeepLab-CRF [15]).

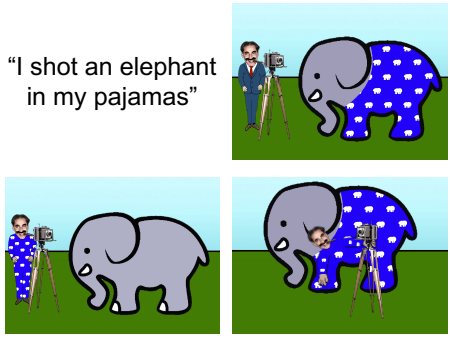
Keywords: Images and language; Reasoning; Grouping; Scene understanding; Semantic image segmentation

1 Introduction

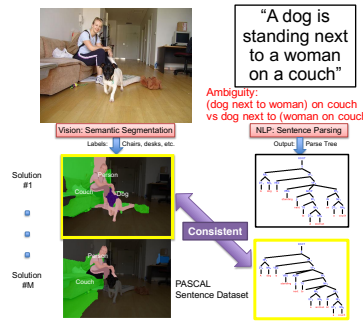
"One morning, I shot an elephant in my pajamas. How he got in my pajamas, I don't know." – Groucho Marx, Animal Crackers (1930).

Perception problems are hard. Whether we are interested in understanding an image or a sentence, our algorithms must operate under tremendous levels of ambiguity. For instance, out of context, a patch from an image may seem like a face, but may simply be an incidental arrangement of tree branches and shadows, causing a face detection/segmentation system to produce nonsensical results, such as hallucinating faces floating

*The first two authors contributed equally.



(a) Prepositional Phrase Attachment Resolution (PPAR) ambiguity.



(b) Joint Segmentation + PPAR.

Fig. 1: (a) An example of an ambiguity in the sentence, which can be resolved by understanding the associated scene. There are at least three possible interpretations of the given sentence. (b) Overview of our approach. We propose a model for simultaneous 2D semantic segmentation and prepositional phrase attachment resolution by reasoning about sentence parses. The language and vision modules each produces M diverse hypotheses, and the goal is to select a pair of consistent hypotheses. In this example the ambiguity to be resolved from the image caption is whether the dog is standing on the couch or next to the couch. Both modules benefit by selecting a pair of compatible hypotheses.

on tree branches and building walls. Similarly, when a human reads the sentence "I eat sushi with tuna", it is clear that the preposition phrase "with tuna" modifies "sushi" and not the act of eating, but this may be ambiguous to a machine. In Natural Language Processing (NLP) this problem – of determining whether a prepositional phrase ("with tuna") modifies a noun phrase ("sushi") or verb phrase ("eating") – is known as Prepositional Phrase Attachment Resolution (PPAR) [56].

Both semantic segmentation and PPAR are well studied problems in the vision [13, 44, 58] and NLP communities [1, 3, 8, 56] respectively. However, prepositions are often fairly visual, and it is clear that some ambiguities cannot be resolved without visual grounding. This suggests a need for solving the two problems simultaneously. Consider the captioned scene shown in Fig. 1b. The caption "A dog is standing next to a woman on a couch" exhibits a PP attachment ambiguity – "(dog next to woman) on couch" vs "dog next to (woman on couch)". Having image segmentations can help resolve this ambiguity, and having the correct PP attachment can help acheive a good segmentation.

There are real-world applications where such joint reasoning is important. Consider providing a robot the instruction "wash the dishes in the sink." The ambiguity here is "dishes in sink" vs. "wash in sink", where the robot would need to pick a sentence parse that does not result in it getting inside of the sink to wash the dishes.

Why is there ambiguity? Ambiguities exist in different domains for different reasons. In vision, ambiguities arise due to limitations in our models (or *Approximation error)* – we do not have models accurate enough to predict the correct segmentation. In sentence parsing, ambiguities arise due to *Bayes error* – there is not enough information available in the input sentence to know which interpretation is correct.

Holistic Reasoning and Challenge. In principle, probabilistic graphical models provide a framework for such holistic reasoning. They allow accumulating evidence from multiple perception modules and converting it into combined beliefs. Unfortunately, idealized probabilistic models that reason about all variables of interest are typically computationally intractable *even for a single module* [14, 57]. In fact, developing specialized inference/learning algorithms for a single module continues to be the major research focus [33, 35]. Jointly reasoning about multiple modules is difficult due to the combinatorial explosion of search space ({exponentially-many segmentations} $\times$ {exponentially-many sentence-parses}).

Proposed Approach and Contributions. In this paper, we address the problem of simultaneous 2D semantic segmentation and PPAR in captioned scenes. To the best of our knowledge this is the first paper to do so.

The inspiration for our approach comes from psycholinguistic evidence suggesting that when people hear ambiguous words ("I saw a bat") they momentarily assess and then rule out their irrelevant meanings (bat = "the sport equipment" or "the flying mammal") [34, 52]. Our approach extracts and leverages a small set of *diverse plausible hypotheses* or guesses from both of the perception modules. Such a hypothesis set has the potential to overcome the search space explosion and increase the chance of extracting at least one accurate solution from the module. Moreover, joint reasoning over each module may simply be restricted to the Cartesian product of the hypothesis sets – {M-segmentations}$\times${M-sentence-parses} – keeping the search space tractable.

Our main thesis is that such a set of plausible hypotheses can serve as a concise interpretable summary of uncertainty in perception modules (What does the semantic segmentation module see in the world? What does the PPAR module describe?) and form the basis for tractable joint reasoning (How do we reconcile what the semantic segmentation module sees in the world with how the PPAR module describes it?). An illustration is shown in Fig. 1b.

Given our two modules with M hypotheses each, how can we integrate beliefs across the segmentation and sentence parse modules to pick the best pair of hypotheses? Our key focus is *consistency* – correct hypotheses from different modules will be correct in a consistent way, but incorrect hypotheses will be incorrect in incompatible ways. Specifically, we develop a MEDIATOR model that scores tuples for consistency and searches over all M^2 tuples to pick the highest scoring one.

We demonstrate our approach on three datasets – PASCAL-50S, ABSTRACT-50S [61], and PASCAL-Context-50S [50]. Our vision+language approach significantly outperforms a state-of-the-art NLP system (Stanford Parser [18, 30]) by 20.66% (36.42% relative) for ABSTRACT-50S, 17.91% (28.69% relative) for PASCAL-50S, and by 12.83% (25.28% relative) for PASCAL-Context-50S. We also make small improvements over a state-of-the-art vision system (DeepLab-CRF [15]).

2 Related Work

Most works at the intersection of vision and NLP tend to be 'pipeline' systems, where vision tasks take 1-best inputs from NLP (*e.g.*, sentence parsings) without trying to im-

prove NLP performance and vice-versa. For instance, [21] uses prepositions to improve finding objects, segmentation and scene classification, but only consider the most-likely parse of the sentence and do not consider resolving ambiguities in text. Analogously, [66] investigates the role of object, attribute, and action classification annotations for generating human-like descriptions. While they achieve impressive results at generating descriptions, they assume perfect vision modules to generate sentences. Our work uses current (still imperfect) vision and NLP modules to reason about images and provided captions, and simultaneously improve both vision and language modules. Similar to our philosophy, an earlier work [5] used images to help disambiguate word senses (*e.g.* piggy banks vs snow banks). In a more recent work, [23] studied the problem of reasoning about an image and a verb, where they attempt to pick the correct sense of the verb that describes the action depicted in the image. In [9], linguistic ambiguities are resolved in sentences coupled with videos that represent different interpretations of the sentences. Perhaps the work closest to us is [36], which leverages information from an RGBD image and its sentential description to improve 3D semantic parsing and resolve ambiguities related to "co-reference resolution" in the sentences (*e.g.*, what "it" refers to). We focus on a different kind of ambiguity – the Prepositional Phrase (PP) attachment resolution. In the classification of parsing ambiguities, co-reference resolution is considered a "discourse ambiguity" [54] (arising out of two different words across sentences for the same object), while PP attachment is considered a "syntactic ambiguity" (arising out of multiple valid sentence structures) and is typically considered much more difficult to resolve [4, 17].

A number of recent works have studied problems at the intersection of vision and language, such as Visual Question Answering [2, 24, 48], Visual Madlibs [67], and image captioning [20, 46, 62]. Our work falls in this domain with one key distinction – we joinly reason and produce both vision and NLP outputs.

Our work also has similarities with works on "spatial relation learning" [39, 46], *i.e.* learning a visual representation for noun-preposition-noun triplets ("car on road"). While our approach can certainly utilize such spatial relation classifiers if available, the focus of our work is different. Our goal is to improve semantic segmentation and PPAR by jointly reranking segmentation-parsing solution pairs. Our approach implicitly learns spatial relationships for prepositions ("on", "above", etc) but these are simply emergent latent representations that help our reranker pick out the most consistent pair of solutions. Another work [45] proposes to answer questions about images with a set of hand-defined predicates (*e.g.* 'closeAbove' means vertical distance < threshold) to define spatial relations between objects, while in our approach, these relationships are learned automatically.

There is a rich history in vision and AI in combining perception modules. In the 1970s, several researchers [6, 27, 51, 60, 64] proposed combining a low-level image segmentation module with a high-level image interpretation module.

Recent works have looked at both joint [10, 11, 16, 21, 29, 36, 42, 43, 53, 59, 63, 65] and sequential/cascaded [12, 25, 26, 28, 38, 41] reasoning over different perception modules (*e.g.*, segmentation, recognition, depth estimation, and scene classification). Both directions have their own shortcomings. Joint modeling involves constructing a

single (usually restrictive) probabilistic model that reasons about all variables in all modules at the expense of performance-limiting independence assumptions and limited interactions to ensure tractable inference. Sequential/cascaded models abandon the probabilistic joint-prediction framework altogether and simply feed the output of one module into another, but this results in propagation of errors and general mismanagement of uncertainty. Our proposed approach leverages the best of both worlds, treating each module as a sophisticated black box with a single minimal assumption that it has the ability to produce diverse plausible solutions. It is therefore able to handle simultaneous reasoning about semantic segmentation and PPAR, where no joint model exists in literature. A second layer on top is then free to perform joint reasoning between these modules since the search space is limited to the Cartesian product of these hypothesis sets. Perhaps the closest to our goal is [28], where the single "most probable" outputs of different modules are repeatedly fed as features to other modules in a cascaded manner. Our approach is orthogonal – we have a shallow 2-layer cascade, but each module produces a set of plausible solutions, not just a single one. Combining the two ideas to produce a deep cascade where each step produces multiple hypotheses is an interesting direction for future work.

Comparison to Boosting [22] and Mixture of Experts (MoE) [32]. AdaBoost involves sequentially training weak learners that are all solving the same problem, and finally providing the weighted average of their predictions. Mixture of Experts involves independently training multiple experts/models that all solve the same problem, with a final prediction picked via a 'gating function'. In our approach, the problem of simultaneous segmentation of an image and parsing of an associated caption is broken into two modules, which work independently to produce plausible solutions and then communicate to pick the best tuple of solutions. Neither of these ideas (decomposition, consistency-check) are present in AdaBoost. At a high-level our approach may be thought of as a gating function, but the idea of decomposition into modules is novel and crucial as we move towards more sophisticated problems. We do not have weak learners, but sophisticated systems that are difficult to integrate except via diverse solutions.

3 Approach

In order to emphasize the generality of our approach, and to show that our approach is compatible with a wide class of implementations of semantic segmentation and PPAR modules , we present our approach with the modules abstracted as "black boxes" that satisfy a few general requirements and minimal assumptions. In Section 4, we describe each of the modules in detail, making concrete their respective features, and other details.

3.1 What is a Module?

The goal of a module is to take input variables $\mathbf{x} \in \mathcal{X}$ (images or sentences), and predict output variables $\mathbf{y} \in \mathcal{Y}$ (semantic segmentation or sentence parsing). The two requirements on a module are that it needs to be able to produce *scores* $S(\mathbf{y}|\mathbf{x})$ for potential solutions and a list of *plausible hypotheses* $\mathbf{Y} = \{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^M\}$.

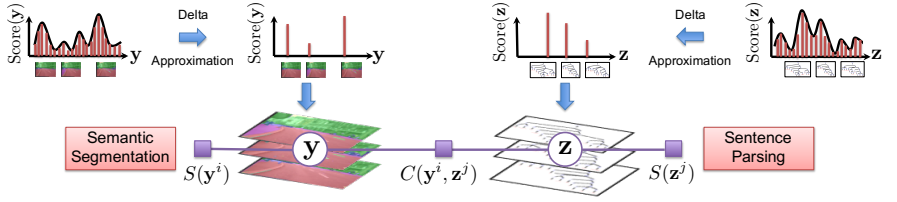


Fig. 2: Illustrative inter-module factor graph. Each node takes exponentially-many or infinitely-many states and we use a 'delta approximation' to limit support.

Multiple Hypotheses. In order to be useful, the set $\mathbf{Y}$ of hypotheses must provide an accurate summary of the score landscape. Thus, the hypotheses should be plausible (*i.e.*, high-scoring) and mutually non-redundant (*i.e.*, diverse). Our approach (described next) is applicable to any choice of diverse hypothesis generators. In our experiments, we use the k-best algorithm of [30] for the sentence parsing module and the DivMBest algorithm [7] for the semantic segmentation module. Once we instantiate the modules in the Experiment section, we describe the diverse solution generation in more detail.

3.2 Joint Reasoning Across Multiple Modules

We now show how to integrate information from multiple modules for holistic reasoning. Recall that our key focus is *consistency* – correct hypotheses from different modules will be correct in a consistent way, but incorrect hypotheses will be incorrect in incompatible ways. Thus, our goal is to search for a pair (semantic segmentation, sentence parsing) that is mutually consistent.

To be concrete, let us consider $n = 2$ modules[1] that each produce a list of M hypotheses $\mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^M\}, \mathbf{Z} = \{\mathbf{z}^1, \ldots, \mathbf{z}^M\}$.

MEDIATOR Model. We develop a "mediator" model that identifies high-scoring hypotheses across modules in agreement with each other. Concretely, we can express the MEDIATOR model as a factor graph where each node corresponds to a module (semantic segmentation and PPAR). Working with such a factor graph is typically completely intractable because each node $\mathbf{y}, \mathbf{z}$ has exponentially-many states (image segmentations, sentence parsing). As illustrated in Fig. 2, in this factor-graph view, the hypothesis sets $\mathbf{Y}, \mathbf{Z}$ can be considered "delta-approximations" for reducing the size of the output spaces.

Unary factors $S(\cdot)$ capture the score/likelihood of each hypothesis provided by the corresponding module for the image/sentence at hand. Pairwise factors $C(\cdot, \cdot)$ represent consistency factors. Importantly, since we have restricted each module variables to just M states, we are free to capture *arbitrary domain-specific high-order relationships* for consistency, without any optimization concerns. In fact, as we describe in our experiments, these consistency factors may be designed to exploit domain knowledge in fairly sophisticated ways.

[1] Our approach is general enough to be applied to an arbitrary n number of modules. However, to keep the exposition concrete, and to match our implementation, we describe the 2-module setting.

Consistency Inference. We perform exhaustive inference over all possible tuples.

$$\underset{i,j\in\{1,\ldots,M\}}{\operatorname{argmax}} \; \mathcal{M}(\mathbf{y}^i, \mathbf{z}^j) = S(\mathbf{y}^i) + S(\mathbf{z}^j) + C(\mathbf{y}^i, \mathbf{z}^j). \tag{1}$$

If we have n modules with M hypotheses each, the number of possible tuples is M^n. In our experiments, we allow each module to take a different value for M, and typically use around 10 solutions for each module. We found that even with such a small set, at least one of the solutions in the set tends to be *highly accurate*, meaning that the hypothesis sets have relatively high recall. Since we only have 2 modules with at most 10 solutions each there are at most a mere 100 pairs to reason about, which is easily enumerable. This shows the power of using a small set of diverse hypotheses. For a large M we can exploit a number of standard ideas from the graphical models literature (*e.g.* dual decomposition or belief propagation). In fact, this is one reason we show the factor in Fig. 2; there is a natural decomposition of the problem into modules.

Training MEDIATOR. We can express the MEDIATOR score as $\mathcal{M}(\mathbf{y}^i, \mathbf{z}^j) = \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}^i, \mathbf{z}^j)$, as a linear function of *score and consistency features* $\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}^i, \mathbf{z}^j) = [\phi_S(\mathbf{y}^i); \phi_S(\mathbf{z}^j); \phi_C(\mathbf{y}^i, \mathbf{z}^j)]$, where $\phi_S(\cdot)$ are the single-module (semantic segmentation and PPAR module) score features, and $\phi_C(\cdot, \cdot)$ are the inter-module consistency features. We describe these features in detail in the experiments. We learn these consistency weights $\mathbf{w}$ from a dataset annotated with ground-truth for the two modules $\mathbf{y}, \mathbf{z}$.

Let $\{\mathbf{y}^*, \mathbf{z}^*\}$ denote the `oracle` tuple, composed of the most accurate solutions in the hypothesis sets. We learn the MEDIATOR parameters in a discriminative learning fashion by solving the following Structured SVM problem

$$\min_{\mathbf{w},\xi_{ij}} \quad \frac{1}{2}\mathbf{w}^\intercal\mathbf{w} + C\sum_{ij}\xi_{ij} \tag{2a}$$

$$s.t. \quad \underbrace{\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}^*, \mathbf{z}^*)}_{\text{Score of } \texttt{oracle} \text{ tuple}} - \underbrace{\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}^i, \mathbf{z}^j)}_{\text{Score of any other tuple}} \geq \underbrace{1}_{\text{Margin}} - \underbrace{\frac{\xi_{ij}}{\mathcal{L}(\mathbf{y}^i, \mathbf{z}^j)}}_{\text{Slack scaled by loss}} \quad \forall i, j \in \{1, \ldots, M\} \tag{2b}$$

Intuitively, we can see that the constraint (2b) tries to maximize the (soft) margin between the score of the `oracle` tuple and all other tuples in the hypothesis sets. Importantly, the slack (or violation in the margin) is scaled by the loss of the tuple. Thus, if there are other good tuples not too much worse than the `oracle`, the margin for such tuples will not be tightly enforced. On the other hand, the margin between the `oracle` and bad tuples will be very strictly enforced.

Notice that this learning procedure requires us to define the loss function $\mathcal{L}(\mathbf{y}^i, \mathbf{z}^j)$, *i.e.*, the cost of predicting a tuple (semantic segmentation, sentence parsing). We use a weighted average of individual losses:

$$\mathcal{L}(\mathbf{y}^i, \mathbf{z}^j) = \alpha\ell(\mathbf{y}^{gt}, \mathbf{y}^i) + (1-\alpha)\ell(\mathbf{z}^{gt}, \mathbf{z}^j) \tag{3}$$

The standard measure for evaluating semantic segmentation is average Jaccard Index (or Intersection-over-Union) [19], while for evaluating sentence parses w.r.t. their preposi-

tional phrase attachment, we use the fraction of prepositions correctly attached. In our experiments, we report results with such a convex combination of module loss functions (for different values of α). Note that α is not a parameter of our approach; rather it is chosen by the user of the system to reflect their priorities. Our approach can just as easily handle perceptually meaningful but non-decomposable "high-order" coupled loss functions, should they be proposed in future work.

4 Experiments

We now describe the setup of our experiments, provide implementation details of the modules, and describe the consistency features.

Dataset. Access to rich annotated image + captioning datasets is crucial for performing quantitative evaluations. Since this is the first paper to study the problem of joint segmentation and PPAR, we have to curate our own annotations for PPAR on three standard image caption datasets (ABSTRACT-50[61], PASCAL-50S [61], PASCAL Context [50]). To curate the PASCAL PPAR annotations, we first select all sentences that have preposition phrase attachment ambiguities. We then plotted the distribution of prepositions in these sentences. The top 7 prepositions are used, as there is a large drop in the frequencies beyond these. The 7 prepositions are: "on", "with", "next to", "in front of", "by", "near", and "down". We then further sampled sentences to ensure uniform distribution across prepositions. We do a similar filtering for the ABSTRACT-50S dataset (we use top-6 prepositions). Details about the filtering can be found in the appendix. To summarize, the three datasets we use are:

1. **ABSTRACT-50S** [61]: 25,000 sentences (50 per image) with 500 images from abstract scenes made from clipart. Filtering for captions containing the top-6 prepositions resulted in 399 sentences describing 201 unique images. These 6 prepositions are: "with", 'next to", "on top of", "in front of", "behind", and "under".
2. **PASCAL-50S** [61]: 50,000 sentences (50 per image) for the images in the UIUC PASCAL sentence dataset [55]. Filtering for the top-7 prepositions resulted in a total of 30 unique images, and 100 image-caption pairs, where ground-truth PPAR were carefully annotated by two vision + NLP graduate students.
3. **PASCAL-Context-50S** [50]: We use images and captions from PASCAL-50S, but with PASCAL Context annotations (60 categories instead of 21). This makes the vision task more challenging. Filtering this dataset for the top-7 prepositions resulted in a total of 966 unique images and 1822 image-caption pairs. Ground truth annotations for the PPAR module were collected using Amazon Mechanical Turk, where presumably the workers are not trained in linguistics. The differences between the PASCAL-Context-50S and PASCAL-50S datasets illustrate whether expert curation of PPAR is necessary by humans (we find Turkers do a surprisingly good job).

In total, we test our approach on 1197 images and 2321 sentences for the 3 experiments.

Setup. *Single Module:* We first show that visual features help the PPAR by using ABSTRACT-50S dataset, which contains clipart scenes where the extent and position of all the objects in the scene is known. This allows us to consider a scenario with perfect vision system.

Multiple Modules: In this experiment we use imperfect language and vision modules, and show improvements on the PASCAL-50S and PASCAL-Context-50S datasets.

Module 1: Semantic Segmentation (SS) $\mathbf{y}$. We use DeepLab-CRF [15] and DivMBest [7] to produce M diverse segmentations of the images. DeepLab-CRF consists of a fully-convolutional neural network [44] followed by a fully-connected graphical model where mean-field inference can be efficiently performed with high-dimensional convolutions [37]. DivMBest (with Hamming diversity) works by iterating M times, starting with the 1-best solution and adding a constant value, set in cross-validation, to the element in the energies at each pixel indexed by the category picked by the CRF at the previous iteration. More details can be found in [7]. To evaluate we use image-level Jaccard Index (IOU) averaged across categories. We chose semantic segmentation (as opposed to detection boxes) as a module because in some cases segmentations allow better reasoning about prepositions than boxes. For instance, see the example showing that it is easier to identify if a horse is standing vs sitting with segmentations (harder with boxes) in [31]. Of course, semantic segmentation has the disadvantage of not reasoning about instances. We believe instance-level segmentations would be ideal, which our work can be easily extended to in the future.

Module 2: PP Attachment Resolution (PPAR) $\mathbf{z}$. We use a recent version (v3.3.1) of the PCFG Stanford parser module [18, 30] (released in 2014) to produce M parsings of the sentence. In addition to the parse trees, the module can also output *dependencies*, which make syntactical relationships more explicit. Dependencies come in the form *dependency_type(*$word_1$*,* $word_2$*)*, such as the preposition dependency *prep_on(woman-8, couch-11)* (number indicates word position in sentence). To evaluate, we count the percentage of preposition attachments that the parse gets correct. A preposition attachment is deemed correct if a human observer looking at the image deems it valid.

Baselines:

- INDEP. In our experiments, we compare our proposed approach (MEDIATOR) to the highest scoring solution predicted independently from each module. For the semantic segmentation this is the output of DeepLab-CRF [15] and for the PPAR module this is the 1-best output of the Stanford Parser [18, 30]. Since our hypothesis lists are generated by greedy M-Best algorithms, this corresponds to predicting the $(\mathbf{y}^1, \mathbf{z}^1)$ tuple. This comparison establishes the importance of joint reasoning. To the best of our knowledge, there is no existing (or even natural) joint model to compare to.
- DOMAIN ADAPTATION We learn a re-ranker on the parses. Note that domain adaptation is only needed for PPAR since the Stanford parser is trained on Penn Treebank (Wall Street Journal text) and not on text about images (such as image captions). Such domain adaptation is not necessary for segmentation. This is a competitive single-module baseline. Specifically, we use the *same* parse-based features as our approach, and learn a reranker over the the M_z parse trees ($M_z = 10$).

Our approach (MEDIATOR) significantly outperforms both baselines. The improvements over INDEP shows that joint reasoning produces more accurate results than any module (vision or language) operating in isolation.

The improvements over DOMAIN ADAPTATION establish the source of improvements is indeed vision, and not the reranking step. Simply adapting the parse from its original training domain (Wall Street Journal) to our domain (image captions) is not enough.

Ablative Study: Ours-CASCADE: This ablation studies the importance of multiple hypothesis. For each module (say $\mathbf{y}$), we feed the single-best output of the other module $\mathbf{z}^1$ as input. Each module learns its own weight $\mathbf{w}$ using *exactly the same* consistency features and learning algorithm as MEDIATOR and predicts one of the plausible hypotheses $\hat{\mathbf{y}}^{\text{CASCADE}} = \text{argmax}_{\mathbf{y}\in\mathbf{Y}}\, \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}, \mathbf{z}^1)$. This ablation of our system is similar to [28] and helps us in disentangling the benefits of multiple hypothesis and joint reasoning.

Finally, we note that Ours-CASCADE can be viewed as special cases of MEDIATOR. Let MEDIATOR-$(M_{\mathbf{y}}, M_{\mathbf{z}})$ denote our approach run with $M_{\mathbf{y}}$ hypotheses for the first module and $M_{\mathbf{z}}$ for the second. Then INDEP corresponds to MEDIATOR-$(1, 1)$ and CASCADE corresponds to predicting the $\mathbf{y}$ solution from MEDIATOR-$(M_{\mathbf{y}}, 1)$ and the $\mathbf{z}$ solution from MEDIATOR-$(1, M_{\mathbf{z}})$. To get an upper-bound on our approach, we report `oracle`, the accuracy of the most accurate tuple in 10×10 tuples.

In the main paper, our results are presented where MEDIATOR was trained with equally weighted loss ($\alpha = 0.5$), but we provide additional results for varying values of α in the appendix.

MEDIATOR and Consistency Features.

Recall that we have two types of features – (1) score features $\phi_S(y^i)$ and $\phi_S(z^j)$, which try to capture how likely solutions y^i and z^j are respectively, and (2) consistency features $\phi_C(y^i, z^j)$, which capture how consistent the PP attachments in z^j are with the segmentation in y^i. For each ($object_1$, *preposition*, $object_2$) in z^j, we compute 6 features between $object_1$ and $object_2$ segmentations in y^i. Since the humans writing the captions may use multiple synonymous words (*e.g.* dog, puppy) for the same visual entity, we use Word2Vec [49] similarities to map the nouns in the sentences to the corresponding dataset categories.

- **Semantic Segmentation Score Features ($\phi_S(y^i)$) (2-dim)**: We use the ranks and the solution scores from DeepLab-CRF [15].
- **PPAR Score Features ($\phi_S(z^i)$) (9-dim)**: We use the ranks and the log probability of parses from [18], and 7 binary indicators for PASCAL-50S (6 for ABSTRACT-50S) denoting which prepositions are present in the parse.
- **Inter-Module Consistency Features (56-dim)**: For each of the 7 prepositions, 8 features are calculated:
  - One feature is the Euclidean distance between the center of the segmentation masks of the two objects in segmentation that are connected by the preposition. These two objects in the segmentation correspond to the categories with which the soft similarity of the two objects in the sentence is highest among all PASCAL categories.
  - Four features capture the $\max\{0, \text{(normalized-directional-distance)}\}$, where directional-distance measures above/below/left/right displacements between the two objects in segmentation, and normalization involves dividing by height/width.
  - One feature is the ratio of the size of $object_1$ and $object_2$ in segmentation.

- Two features capture the Word2Vec similarity between the two objects in PPAR (say 'puppy' and 'kitty') with their most similar PASCAL category (say 'dog' and 'cat'), where these features are 0 if the categories are not present in segmentation.

A visual illustration for some of these features for PASCAL can be seen in Fig. 3. In the case where an object parsed from z^j is not present in the segmentation y^i, the distance features are set to 0. The ratio of areas features (area of smaller object / area of larger object) are also set to 0 assuming that the smaller object is missing. In the case where an object has two or more connected components in the segmentation, the distances are computed w.r.t. the centroid of the segmentation and the area is computed as the number of pixels in the union of the instance segmentation masks. We also calculate 20 features for PASCAL-50S and 59 features for PASCAL-Context-50S that capture that consistency between y^i and z^j, in terms of presence/absence of PASCAL categories. For each noun in PPAR we compute its Word2Vec[49] similarity with all PASCAL categories. For each of the PASCAL categories, the feature is sum of similarities (with the PASCAL category) over all nouns if the category is present in segmentation, the feature is -1 times sum of similarities over all nouns otherwise. This feature set was not used for ABSTRACT-50S, since these features were intended to help improve the accuracy of the semantic segmentation module. For ABSTRACT-50S, we only use the 5 distance features, resulting in a 30-dim feature vector.

4.1 Single-Module Results

We performed a 10-fold cross-validation on the ABSTRACT-50S dataset to pick M (=10), as shown in Fig. 4a, and the weight on the hinge-loss for MEDIATOR (C). The results are presented in Table 1. Our approach significantly outpeforms 1-best outputs of the Stanford Parser [18, 30] by 20.66% (36.42% relative). This shows a need for diverse hypotheses and reasoning about visual features when picking a sentence parse. `oracle` denotes the best achievable performance using these 10 hypotheses.

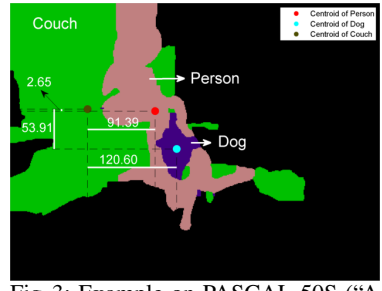


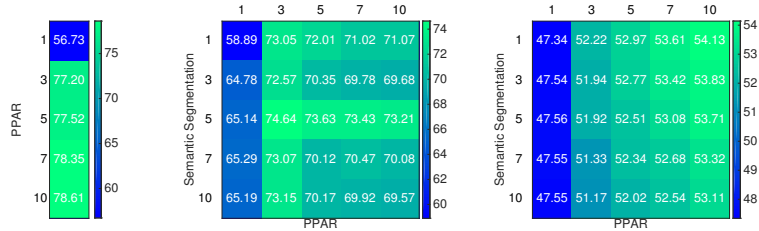
Fig. 3: Example on PASCAL-50S ("A dog is standing next to a woman on a couch."). The ambiguity in this sentence "(dog next to woman) on couch" vs "dog next to (woman on couch)". We calculate the horizontal and vertical distances between the segmentation centers of "person" and "couch" and between the segmentation centers of "dog" and "couch". We see that the "dog" is much further below the couch (53.91) than the woman (2.65). So, if the MEDIATOR model learned that "on" means the first object is above the second object, we would expect it to choose the "person on couch" preposition parsing.

4.2 Multiple-Module Results

We performed 10-fold cross-val for our results of PASCAL-50S and PASCAL-Context-50S, with 8 `train` folds, 1 `val` fold, and 1 `test` fold, where the `val` fold was used to pick $M_{\mathbf{y}}$, $M_{\mathbf{z}}$, and C. Fig. 4 shows the average combined accuracy for the `val` set, which is maximal $M_{\mathbf{y}} = 5, M_{\mathbf{z}} = 3$ for PASCAL-50S (Fig. 4b), and $M_{\mathbf{y}} = 1, M_{\mathbf{z}} = 10$ for PASCAL-Context-50S (Fig. 4c), which are used at test time.

Module	Stanford Parser [18, 30]	Domain Adaptation	Ours	`oracle`
PPAR	56.73	57.23	**77.39**	97.53

Table 1: Results on our subset of the ABSTRACT-50S dataset.



(a) (b) (c)

Fig. 4: (a) Validation accuracies for different values of M on ABSTRACT-50S, (b) for different values of $M_{\mathbf{y}}, M_{\mathbf{z}}$ on PASCAL-50S, (c) for different values of $M_{\mathbf{y}}, M_{\mathbf{z}}$ on PASCAL-Context-50S.

	PASCAL-50S			PASCAL-Context-50S		
	Instance-Level Jaccard Index	PPAR Acc.	Average	Instance-Level Jaccard Index	PPAR Acc.	Average
DeepLab-CRF [15]	66.83	-	-	**43.94**	-	-
Stanford Parser [18, 30]	-	62.42	-	-	50.75	-
Average	-	-	64.63	-	-	47.345
Domain Adaptation	-	72.08	-	-	58.32	-
Ours CASCADE	67.56	75.00	71.28	**43.94**	**63.58**	**53.76**
Ours MEDIATOR	**67.58**	**80.33**	**73.96**	**43.94**	**63.58**	**53.76**
`oracle`	69.96	96.50	83.23	49.21	75.75	62.48

Table 2: Results on our subset of the PASCAL-50S and PASCAL-Context-50S datasets.

We present our results in Table 2. Our approach significantly outperforms the Stanford Parser [18, 30] by 17.91% (28.69% relative) for PASCAL-50S, and 12.83% (25.28% relative) for PASCAL-Context-50S. We also make small improvements over DeepLab-CRF [15] in the case of PASCAL-50S. To measure statistical significance of our results, we performed paired t-tests between MEDIATOR and INDEP. For both modules (and average), the null hypothesis (that the accuracies of our approach and baseline come from the same distribution) can be successfully rejected at p-value 0.05. For sake of completeness, we also compared MEDIATOR with our ablated system (CASCADE) and found statistically significant differences only in PPAR. These results demonstrate a need for each module to produce a diverse set of plausible hypotheses for our MEDIATOR model to reason about. We show a qualitative example for an input image in Fig. 5. In the case of PASCAL-Context-50S, MEDIATOR performs identical to CASCADE since $M_{\mathbf{y}}$ is chosen as 1 (which is the CASCADE setting) in cross-validation. Recall that MEDIATOR is a larger model class than CASCADE (in fact, CASCADE is a special case of MEDIATOR with $M_y = 1$). It is interesting to see that the large model class does not hurt, and MEDIATOR gracefully reduces to a smaller capacity model (CASCADE) if

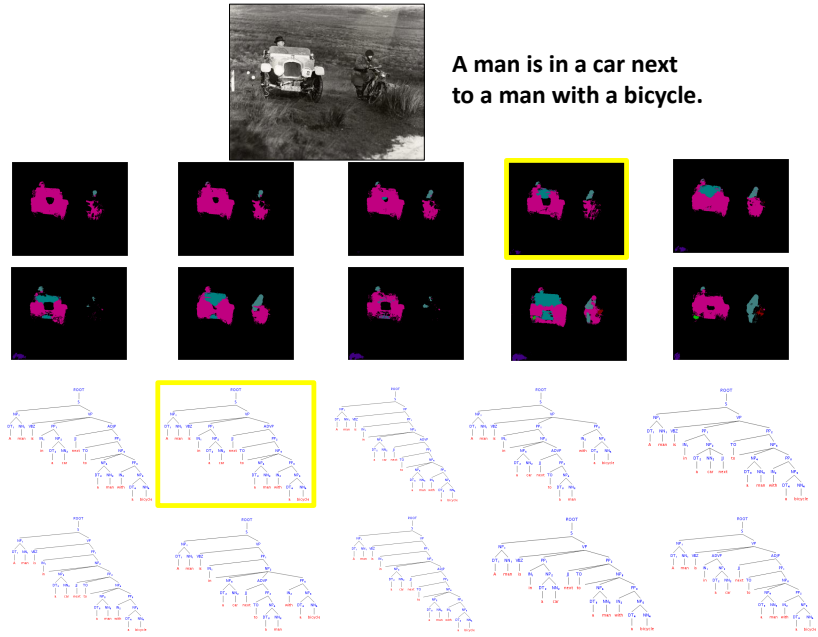


Fig. 5: Qualitative example of our approach for an input image and associated caption. We are able to find a consistent pair of solutions (segmentation, sentence parse), where we pick a segmentation containing the PASCAL categories "person" and "car", and a parse that picks the correct prepositional phrase attachements ("car next to man" and "man with bicycle").

the amount of data is not enough to warrant the extra capacity. We hypothesize that in the presence of more training data, cross-validation may pick a different setting of $M_{\mathbf{y}}$ and $M_{\mathbf{z}}$, resulting in full utilization of the model capacity. Also note that our domain adaptation baseline achieved an accuracy higher than MAP/Stanford-Parser, but significantly lower than our approach for both PASCAL-50S and PASCAL-Context-50S. We also performed this for our single-module experiment and picked M_z (=10) with cross-validation, which resulted in an accuracy of 57.23%. Again, this is higher than MAP/Stanford-Parser (56.73%), but significantly lower than our approach (77.39%). Clearly, this shows that just domain adaptation is not sufficient.

The table also shows that the oracle performance is fairly high (96.50% for PPAR at 10 parses), suggesting that when there is ambiguity and room for improvement, our mediator is able to rerank effectively.

Ablation Study for Features. Table 3 displays the results of an ablation study on PASCAL-50S and PASCAL-CONTEXT-50S to show the importance of the different features. In each row, we retain the module score features and drop a single set of consistency features. We can see that all of the consistency features contribute in the performance of MEDIATOR.

Visualizing Prepositions. Fig. 6 shows a visualization for what our MEDIATOR model has implicitly learned about 3 prepositions ("on", "by", "with"). These visualizations show the score obtained by taking the dot product of distance features (Euclidean and

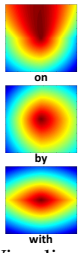


Fig. 6: Visualizations for three different prepositions.

Table 3: Ablation study for different combinations of features on PASCAL-50S and PASCAL-Context-50S. We only show PPAR Acc. for PASCAL-Context-50S because $M_y = 1$.

Feature set	PASCAL-50S		PASCAL-Context-50S
	Instance-Level Jaccard Index	PPAR Acc.	PPAR Acc.
All features	67.58	80.33	63.58
Drop all consistency	66.96	66.67	61.47
Drop Euclidean distance	67.27	77.33	63.77
Drop directional distance	67.12	78.67	63.63
Drop Word2Vec	67.58	78.33	62.72
Drop category presence	67.48	79.25	61.19

directional) between $object_1$ and $object_1$ connected by the preposition with the corresponding learned weights of the model, considering $object_2$ to be at the center of the visualization. Notice that these were learned without explicit training for spatial learning as in spatial relation learning (SRL) works [40, 47]. These were simply recovered as an intermediate step towards reranking SS + PPAR hypotheses. Also note that SRL cannot handle multiple segmentation hypotheses, which our work shows are important (Table 2 CASCADE). In addition our approach is more general.

5 Discussions and Conclusion

We presented an approach to the simultaneous reasoning about prepositional phrase attachment resolution of captions and semantic segmentation in images that integrates beliefs across the modules to pick the best pair of a diverse set of hypotheses. Our full model (MEDIATOR) significantly improves the accuracy of PPAR over the Stanford Parser by 17.91% for PASCAL-50S and by 12.83% for PASCAL-Context-50S, and achieves a small improvement on Semantic Segmentation over DeepLab-CRF for PASCAL-50S. These results demonstrate a need for information exchange between the modules, as well as a need for a diverse set of hypotheses to concisely capture the uncertainties of each module. The large gains in PPAR validate our intuition that vision is very helpful for dealing with ambiguity in language. Furthermore, we see from the oracle accuracies that even larger gains are possible.

While we have demonstrated our approach on a task involving simultaneous reasoning about language and vision, our approach is general and can be used for other applications. Overall, we hope our approach will be useful in a number of settings.

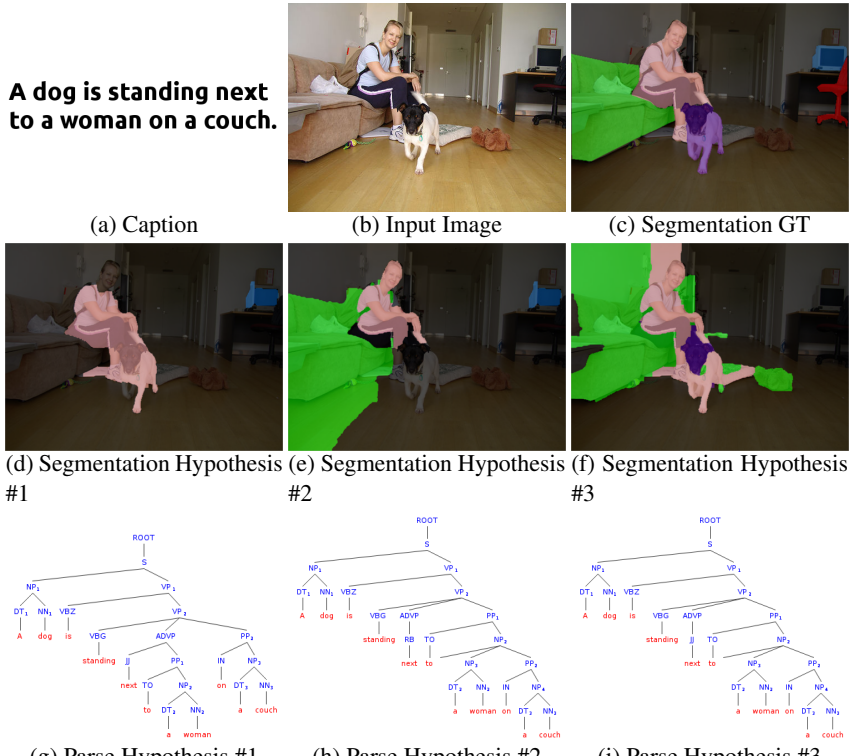
Acknowledgements. We thank Larry Zitnick, Mohit Bansal, Kevin Gimpel, and Devi Parikh for helpful discussions, suggestions, and feedback that were included in this work. A majority of this work was done while AL was an intern at Virginia Tech. This work was partially supported by a National Science Foundation CAREER award, an Army Research Office YIP Award, an Office of Naval Research grant, and GPU donations by NVIDIA, all awarded to DB. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government or any sponsor.

Appendix Overview

In this appendix we provide the following: compact

Appendix I: Additional motivation for our MEDIATOR model.
Appendix II: Background on ABSTRACT-50S.
Appendix III: Details of the dataset curation process for the ABSTRACT-50S, PASCAL-50S, and PASCAL-Context-50S datasets.
Appendix IV: Qualitative examples from our approach.
Appendix V: Results where we study the effect of varying the weighting of each module in our approach.

Appendix I Additional Motivation for MEDIATOR



(a) Caption (b) Input Image (c) Segmentation GT

(d) Segmentation Hypothesis #1 (e) Segmentation Hypothesis #2 (f) Segmentation Hypothesis #3

(g) Parse Hypothesis #1 (h) Parse Hypothesis #2 (i) Parse Hypothesis #3

Fig. 7: In this figure, we illustrate why the MEDIATOR model makes sense for the task of captioned scene understanding. For the caption-image pair (Fig. 7a-Fig. 7b), we see that parse tree #1 (Fig. 7g) shows "standing" (the verb phrase of the noun "dog") connected with "couch" via the "on" preposition, whereas parse trees #2 (Fig. 7h) and #3 (Fig. 7i) show "woman" connected with "couch" via the "on" preposition. This ambiguity can be resolved if we look at an accurate semantic segmentation such as Hypothesis #3 (Fig. 7f) of the associated image (Fig. 7b). Likewise, we might be able to do better at semantic segmentation if we choose a segmentation that is consistent with the sentence, such as Segmentation Hypothesis #3 (Fig. 7f), which contains a person on a couch with a dog next to them, unlike the other two hypotheses (Fig. 7d and Fig. 7e).

An example is shown in Fig. 7, where the ambiguous sentence that describes the image is “A dog is standing next to a woman on a couch”, where the ambiguity is “(dog next to woman) on couch” vs “dog next to (woman on couch)”, which is reflected in parse trees’ uncertainty. Parse tree #1 (Fig. 7g) shows “standing” (the verb phrase of the noun “dog”) connected with “couch” via the “on” preposition, whereas parse trees #2 (Fig. 7h) and #3 (Fig. 7i) show “woman” connected with “couch” via the “on” preposition. This ambiguity can be resolved if we look at an accurate semantic segmentation such as Hypothesis #3 (Fig. 7f) of the associated image (Fig. 7b). Likewise, we might be able to do better at semantic segmentation if we choose a segmentation that is consistent with the sentence, such as Segmentation Hypothesis #3 (Fig. 7f), which contains a person on a couch with a dog next to them, unlike the other two hypotheses (Fig. 7d and Fig. 7e).

Appendix II Background about ABSTRACT-50S

The Abstract Scenes dataset [68] contains synthetic images generated by human subjects via a drag-and-drop clipart interface. The subjects are given access to a (random) subset of 56 clipart objects that can be found in park scenes, as well as two characters, Mike and Jenny, with a variety of poses and expressions. Example scenes can be found in Fig. 8. The motivation is to allow researchers to focus on higher-level semantic understanding without having to deal with noisy information extraction from real images, since the entire contents of the scene are known exactly, while also providing a dense semantic space to study (due to the heavily constrained world). We used the dataset in precisely this way to first test out the PPAR module in isolation to demonstrate that this problem can be helped by a sentence’s corresponding image.

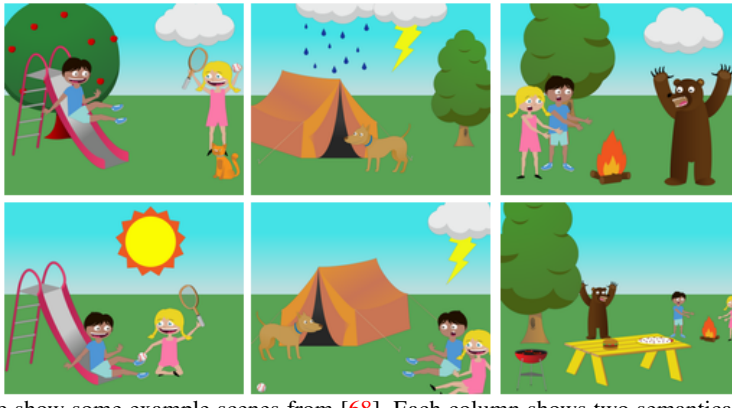


Fig. 8: We show some example scenes from [68]. Each column shows two semantically similar scenes, while the different columns show the diversity of scene types.

Appendix III Dataset Curation and Annotation

The subsets of the PASCAL-50S and ABSTRACT-50S datasets used in the main paper were carefully curated by two vision + NLP graduate students. The subset of the PASCAL-Context-50S dataset used in the main paper was curated by Mechanical Turk workers. The following describes the dataset curation process for each datset.

PASCAL-50S: For PASCAL-50S we first obtained sentences that contain one or more of 7 prepositions (*i.e.*, "with", "next to", "on top of", "in front of", "behind", "by", and "on") that intuitively would typically depend on the relative distance between objects. Then we look for sentences that have preposition phrase attachment ambiguities, *i.e.*, sentences where the parser output has different sets of prepositions for different parsings. (Note that, due to our focus on PP attachment, we do not pay attention to other parts of the sentence parse, so the parses can change while the PP attachments remain the same, as in Fig. 7h and Fig. 7i.) The sentences thus obtained are further filtered to obtain sentences in which the objects that are connected by the preposition belong to one of the 20 PASCAL object categories. Since our Module 1 is semantic segmentation and not instance-level segmentation, we restrict the dataset to sentences involving prepositions connecting two different PASCAL categories. Thus, our final dataset contains 100 sentences describing 30 unique images and contains 16 of the 20 PASCAL categories as described in the paper. We then manually annotated the ground truth PP attachments. Note that such manual labeling by annotators students with expertise in NLP takes a lot of time, but results in annotations that are linguistically high-quality, with any inter-human disagreement resolved by strict adherence to rules of grammar.

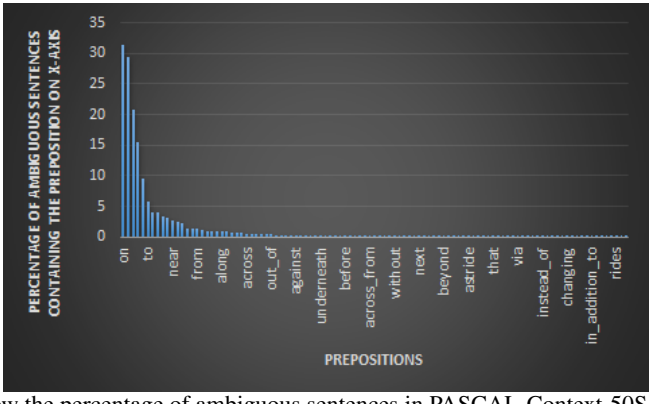


Fig. 9: We show the percentage of ambiguous sentences in PASCAL-Context-50S dataset before filtering for prepositions. We found that there was a drop in the percentage of sentences for prepositions that appear in the sorted list after "down". So, for the PASCAL-Context-50S dataset we only keep sentences that have one or more visual prepositions in the list of prepositions upto "down".

ABSTRACT-50S: We first obtained sentences that contain one or more of 6 prepositions (*i.e.*, "with", "next to", "on top of", "in front of", "behind", "under"). Note, due to the semantic differences between the datasets, not all prepositions found in one were present in the other. Further filtering on sentences was done to ensure that the sentences contain a least one preposition phrase attachment ambiguity that is between the clipart noun categories (*i.e.*, each clipart piece has a name, like "snake", that we search the sentence parsing for). This filtering reduced the original dataset of 25,000 sentences and 500 scenes to our final experiment dataset of 399 sentences and 201 scenes. We then manually annotated the ground truth PP attachments.

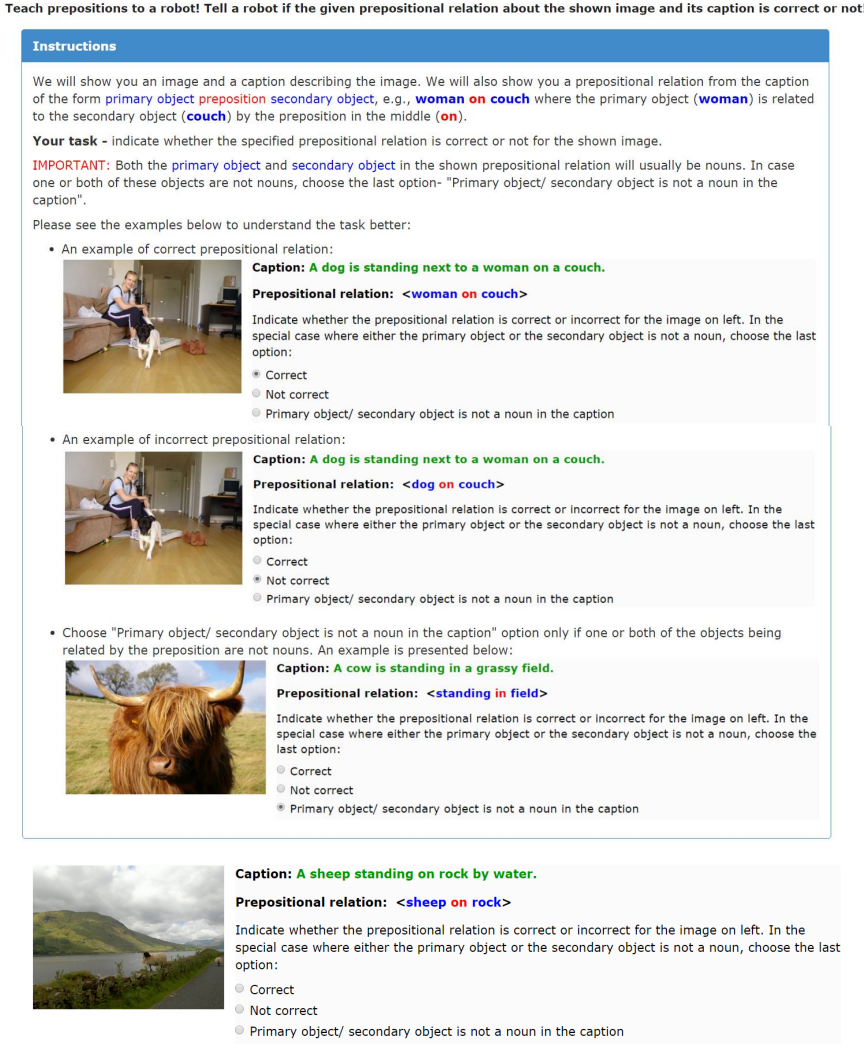


Fig. 10: The AMT interface to collect ground truth annotations for prepositional relations.

PASCAL-Context-50S

For PASCAL-Context-50S, we first selected all sentences that have preposition phrase attachment ambiguities. We then plotted the distribution of prepositions in these sentences (see Fig. 9.). We found that there was a drop in the percentage of sentences for prepositions that appear in the sorted list after “down". Therefore, we only kept sentences that have one or more 2-D visual prepositions in the list of prepositions upto “down". Thus we ended up with the following 7 prepositions: “on”, “with”, “next to”, “in front of”, “by”, “near”, and “down”. We then further sampled sentences to ensure uniform distribution across prepositions. Note that unlike PASCAL-50S, we did not filter sentences based on whether the objects connected by the prepositions belong to one of 60 PASCAL Context categories or not. Instead, we used the Word2Vec [49] similarity between the objects in the sentence and the PASCAL Context categories as one of the features. Thus, our final dataset contains 1822 sentences describing 966 unique images.

The ground truth PP attachments for these 1822 sentences were annotated by Amazon Mechanical Turk workers. For each unique prepositional relation in a sentence, we showed the workers the prepositional relation of the form primary object preposition secondary object and its associated image and sentence and asked them to specify whether the prepositional relation is correct or not correct. We also asked them to choose the third option - “Primary object/ secondary object is not a noun in the caption” in case that happened. The user interface used to collect these annotations is shown in Fig. 10. We collected five answers for each prepositional relation. For evaluation, we used the majority response. We found that 87.11% of human responses agree with the majority response, indicating that even though AMT workers were not explicitly trained in rules of grammar by us, there is relatively high inter-human agreement.

Appendix IV Qualitative Examples

Fig. 11-Fig. 15 show qualitative examples for our experiments. Fig. 11-Fig. 12 show examples for the multiple modules examples (semantic segmentation and PPAR), and Fig. 13-Fig. 15 show examples for the single module experiment. In each figure, the top row shows the image and the associated sentence. For the multiple modules figures, the second and third row show the diverse segmentations of the image, and the bottom two rows show different parsings of the sentence (last two rows for single module examples, as well). In these examples our approach uses 10 diverse solutions for the semantic segmentation module and 10 different solutions for the PPAR module. The highlighted pairs of solutions show the solutions picked by the MEDIATOR model. Examining the results can give you a sense of how the parsings can help the semantic segmentation module pick the best solution and vice-versa.

A young couple sit with their puppy on the couch.

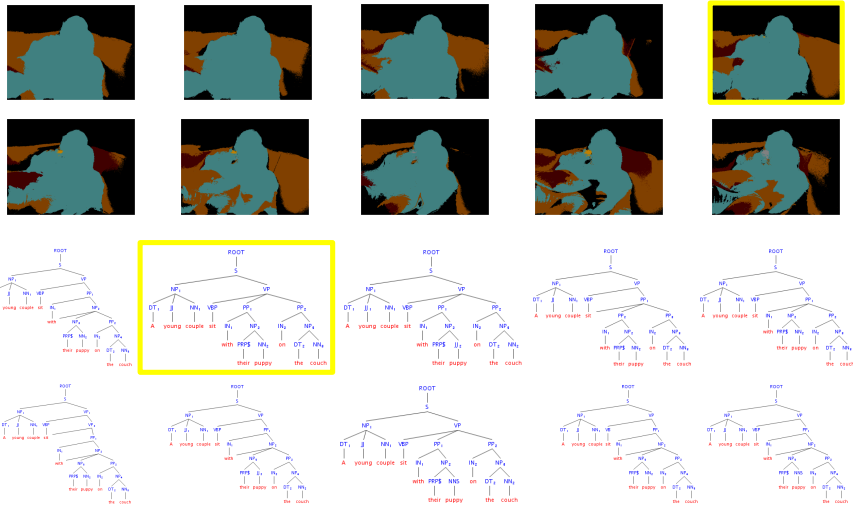
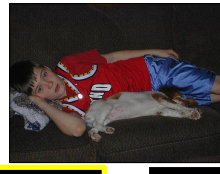


Fig. 11: Example 1 – multiple modules (SS and PPAR)



Boy lying on couch with a sleeping puppy curled up next to him.

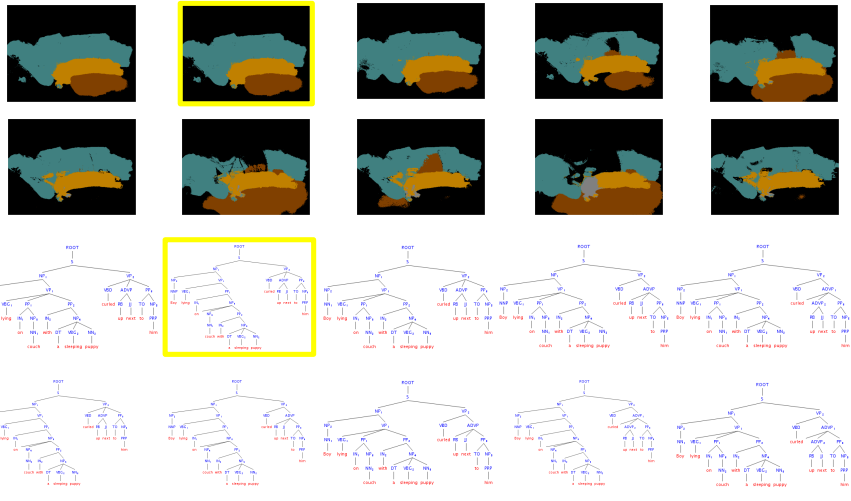


Fig. 12: Example 2 – multiple modules (SS and PPAR)

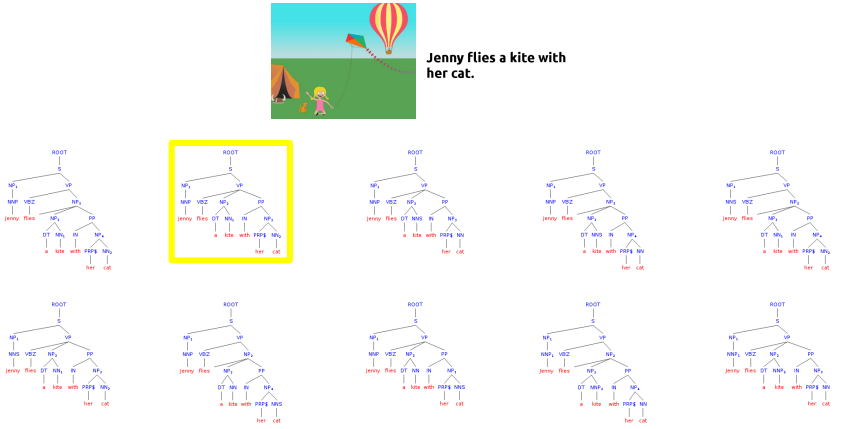


Fig. 13: Example 1 – single module (PPAR)

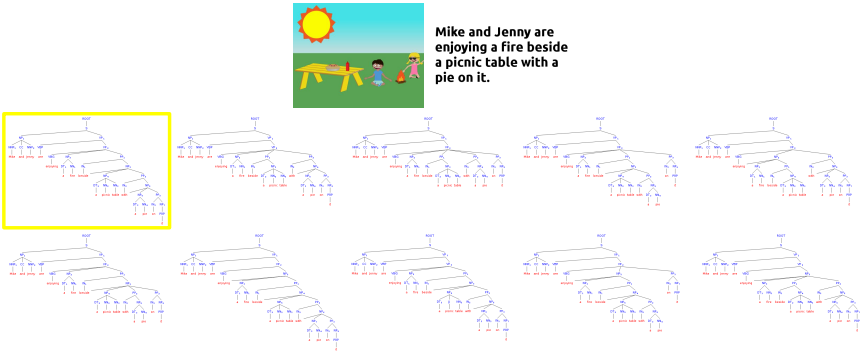


Fig. 14: Example 2 – single module (PPAR)

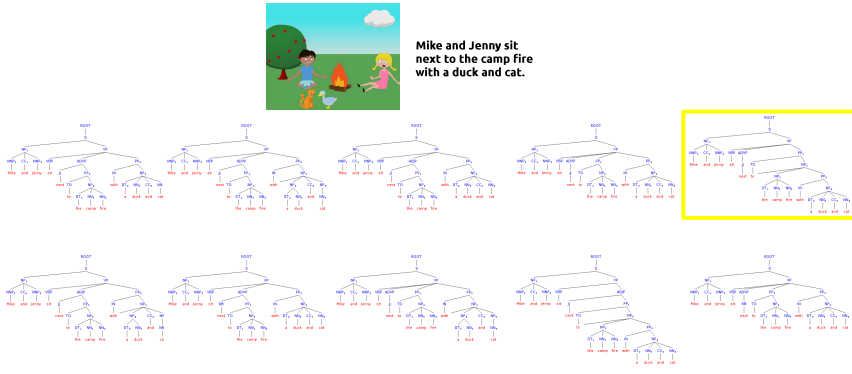


Fig. 15: Example 3 – single module (PPAR)

Appendix V Effect of Different Weighting of Modules

So far we have used the "natural" setting of $\alpha = 0.5$, which gives equal weight to both modules. Note that α is not a parameter of our approach; it is a design choice that the user/experiment-designer makes. To see the effect of weighting the modules differently, we tested our approach for various values of α. Fig. 16 shows how the accuracies of each module vary depending on α for the MEDIATOR model for PASCAL-50S and PASCAL-Context-50S. Recall that α is the coefficient for the semantic segmentation module and 1-α is the coefficient for the PPAR resolution module in the loss function. We see that as expected, putting no or little weight on the PPAR module drastically hurts performance for that module. Our approach is fairly robust to the setting of α, with a peak lying but any weight on it performs fairly similar with the peak lying somewhere between the extremes. The segmentation module has similar behavior, though it is not as sensitive to the choice of α. We believe this is because of small "dynamic range" of this module – the gap between the 1-best and oracle segmentation is smaller and thus the MEDIATOR can always default to the 1-best as a safe choice.

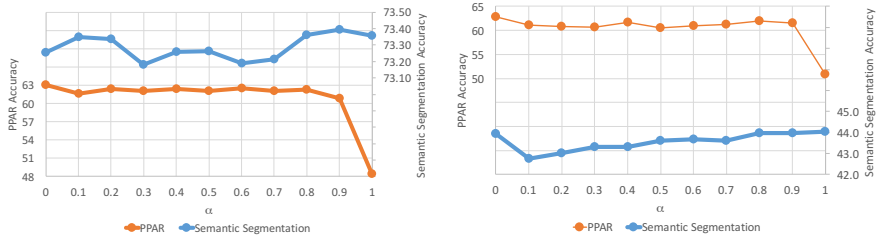


Fig. 16: Accuracies MEDIATOR (both modules) vs α, where α is the coefficient for the semantic segmentation module and 1-α is the coefficient for the PPAR resolution module in the loss function. Our approach is fairly robust to the setting of α, as long as it is not set to either extremes, since that limits the synergy between the modules. (a) shows the results for PASCAL-50S, and (b) shows the results for PASCAL-Context-50S.

Bibliography

[1] E. Agirre, T. Baldwin, and D. Martinez. Improving parsing and pp attachment performance with sense information. In *ACL*, 2008.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. *ICCV*, 2015.

[3] M. Atterer and H. Schütze. Prepositional phrase attachment without oracles. *Computational Linguistics*, 2007.

[4] K. Bach. http://online.sfsu.edu/kbach/ambguity.html. Routledge Encyclopedia of Philosophy entry.

[5] K. Barnard, M. Johnson, and D. Forsyth. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data-Volume 6*, pages 1–5. Association for Computational Linguistics, 2003.

[6] H. G. Barrow and J. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17(75–116), 1981.

[7] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-Best Solutions in Markov Random Fields. In *ECCV*, 2012.

[8] Y. Belinkov, T. Lei, R. Barzilay, and A. Globerson. Exploring compositional architectures and word vector representations for prepositional phrase attachment. *Transactions of the Association for Computational Linguistics (TACL)*, 2014.

[9] Y. Berzak, A. Barbu, D. Harari, B. Katz, and S. Ullman. Do you see what i mean? visual resolution of linguistic ambiguities. *arXiv preprint arXiv:1603.08079*, 2016.

[10] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzàlez. Harmony potentials - fusing global and local scale for semantic image segmentation. *IJCV*, 96(1):83–102, 2012.

[11] D. Bradley. *Learning In Modular Systems*. PhD thesis, Carnegie Mellon University, 2009.

[12] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011.

[13] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, pages 430–443, 2012.

[14] V. Chandrasekaran, N. Srebro, and P. Harsha. Complexity of inference in graphical models. In *24th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[16] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *CVPR*, 2013.

[17] E. Davis. http://cs.nyu.edu/faculty/davise/ai/ambiguity.html. Notes on Ambiguity.

[18] M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010.

[20] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. In *CVPR*, 2015.

[21] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013.

[22] Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[23] S. Gella, M. Lapata, and F. Keller. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv preprint arXiv:1603.09188*, 2016.

[24] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In *PNAS*, 2014.

[25] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009.

[26] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, pages 1030–1037, 2009.

[27] A. R. Hanson and E. M. Riseman. VISIONS: A computer system for interpreting scenes. In *Computer Vision Systems*. Academic Press, 1978.

[28] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008.

[29] D. Hoiem, A. A. Efros, and M. Hebert. Closing the loop in scene interpretation. In *CVPR*, 2008.

[30] L. Huang and D. Chiang. Better k-best parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology (IWPT)*, pages 53–64, 2005.

[31] H. Izadinia, F. Sadeghi, S. K. Divvala, H. Hajishirzi, Y. Choi, and A. Farhadi. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10–18, 2015.

[32] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In *International Joint Conference on Neural Networks*, 1993.

[33] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, T. Kröger, J. Lellmann, et al. A comparative study of modern inference techniques for structured discrete energy minimization problems. *International Journal of Computer Vision*, pages 1–30, 2014.

[34] M. I. Khawalda and E. M. Al-Saidat. Structural ambiguity interpretation: A case study of arab learners of english. *Global Journal of HUMAN SOCIAL SCIENCE*, 2012.

[35] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2011.

[36] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014.

[37] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NIPS*, 2011.

[38] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010.

[39] T. Lan, W. Yang, Y. Wang, and G. Mori. Image retrieval with structured object queries using latent ranking svm. In *Computer Vision–ECCV 2012*, pages 129–142. Springer, 2012.

[40] T. Lan, W. Yang, Y. Wang, and G. Mori. Image retrieval with structured object queries using latent ranking svm. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV'12, 2012.

[41] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[42] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*, 2010.

[43] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

[44] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[45] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems*,

2014.

[46] M. Malinowski and M. Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv preprint arXiv:1411.5190*, 2014.

[47] M. Malinowski and M. Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *CoRR*, abs/1411.5190, 2014.

[48] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[49] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[50] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[51] Y. Ohta, T. Kanade, and T. Sakai. An analysis system for scenes containing objects with substructures. In *Proceedings of the Fourth International Joint Conference on Pattern Recognitions*, pages 752–754, 1978.

[52] C. V. Petten. Words and sentences: Event-related brain potential measures. *Psychophysiology*, 32, 1994.

[53] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition via multiple hypotheses rescoring. *Journal of Machine Learning Research*, 16, 2015.

[54] M. Poesio and R. Artstein. Annotating (anaphoric) ambiguity. In *Corpus Linguistics Conference*, 2005.

[55] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[56] A. Ratnaparkhi, J. Reynar, and S. Roukos. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the workshop on Human Language Technology*, pages 250–255. Association for Computational Linguistics, 1994.

[57] S. E. Shimony. Finding MAPs for belief networks is np-hard. *Artificial Intelligence*, 68(2):399–410, August 1994.

[58] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 2009.

[59] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.

[60] J. Tenenbaum and H. Barrow. Experiments in interpretation-guided segmentation. *Artificial Intelligence*, 8(3):241 – 274, 1977.

[61] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2014.

[62] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015.

[63] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, 2008.

[64] Y. Yakimovsky and J. A. Feldman. A semantics-based decision theory region analyzer. In *IJCAI*, pages 580–588, 1973.

[65] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012.

[66] M. Yatskar, M. Galley, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, 2014.

[67] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank image generation and question answering. *arXiv*, 2015.

[68] C. L. Zitnick and D. Parikh. Bringing Semantics Into Focus Using Visual Abstraction. In *CVPR*, 2013.