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**[3,1]
[1]Virginia Tech [2]Carnegie Mellon University [3]Georgia Institute of Technology
ankit1991laddha@gmail.com
{gordonac,aish,santol,ygoyal,kbk,dbatra}@vt.edu

Abstract

We present an approach to simultaneously perform semantic segmentation and prepositional phrase attachment resolution for captioned images. Some ambiguities in language cannot be resolved without simultaneously reasoning about an associated image. If we consider the sentence "I shot an elephant in my pajamas", looking at language alone (and not using common sense), it is unclear if it is the person or the elephant wearing the pajamas or both. Our approach produces a diverse set of plausible hypotheses for both semantic segmentation and prepositional phrase attachment resolution that are then jointly reranked 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. Multiple hypotheses are also shown to be crucial to improved multiple-module reasoning. Our vision and language approach significantly outperforms the Stanford Parser (De Marneffe et al., 2006) by 17.91% (28.69% relative) and 12.83% (25.28% relative) in two different experiments. We also make small improvements over DeepLab-CRF (Chen et al., 2015).

1 Introduction

Perception and intelligence problems are hard. Whether we are interested in understanding an image or a sentence, our algorithms must operate under tremendous levels of ambiguity. 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. This problem of determining whether a prepositional phrase ("with tuna") modifies a noun phrase ("sushi") or verb phrase ("eating") is formally known as Prepositional Phrase Attachment Resolution (PPAR) (Ratnaparkhi et al., 1994). Consider the captioned scene shown in Fig-

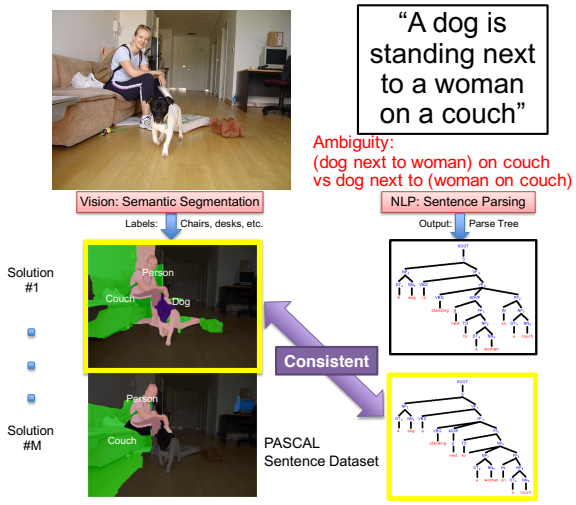


Figure 1: Overview of our approach. We propose a model for simultaneous 2D semantic segmentation and prepositional phrase attachment resolution by reasoning about sentence parses. The language and vision modules each produce M diverse hypotheses, and the goal is to select a pair of consistent hypotheses. In this example the ambiguity to be resolved from the image caption is whether the dog is standing on or next to the couch. Both modules benefit by selecting a pair of compatible hypotheses.

* Denotes equal contribution

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1493–1503,
Austin, Texas, November 1-5, 2016.

ure 1. 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)". It is clear that having access to image segmentations can help resolve this ambiguity, and having access to the correct PP attachment can help image segmentation.

There are two main roadblocks that keep us from writing a single unified model (say a graphical model) to perform both tasks: (1) Inaccurate Models – empirical studies (Meltzer et al., 2005, Szeliski et al., 2008, Kappes et al., 2013) have repeatedly found that models are often inaccurate and miscalibrated – their "most-likely" beliefs are placed on solutions far from the ground-truth. (2) Search Space Explosion – jointly reasoning about multiple modalities 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 object segmentation (also called semantic segmentation) and PPAR in captioned scenes. To the best of our knowledge this is the first paper to do so.

Our main thesis is that a set of diverse plausible hypotheses can serve as a concise interpretable summary of uncertainty in vision and language '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?).

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 pairs for consistency and searches over all M^2 pairs to pick the highest scoring one. We demonstrate our approach on three datasets – ABSTRACT-50S (Vedantam et al., 2014), PASCAL-50S, and PASCAL-Context-50S (Mottaghi et al., 2014). We show that our vision+language approach significantly outperforms the Stanford Parser (De Marneffe et al., 2006) 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 but consistent improvements over DeepLab-CRF (Chen et al., 2015).

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 improve NLP performance and vice-versa. For instance, Fidler et al. (2013) use prepositions to improve object segmentation and scene classification, but only consider the most-likely parse of the sentence and do not resolve ambiguities in text. Analogously, Yatskar et al. (2014) investigate 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 by Barnard and Johnson (2005) used images to help disambiguate word senses (*e.g.* piggy banks vs snow banks). In a more recent work, Gella et al. (2016) 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. Berzak et al. (2015) resolve linguistic ambiguities in sentences coupled with videos that represent different interpretations of the sentences. Perhaps the work closest to us is Kong et al. (2014), who leverage information from an RGBD image and its sentential description to improve 3D semantic parsing and resolve ambiguities related to coreference 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, coreference resolution is considered a discourse ambiguity (Poesio and Artstein, 2005) (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 (Bach, 2016, Davis, 2016).

A number of recent works have studied problems at the intersection of vision and language, such as Visual Question Answering (Antol et al., 2015, Geman et al., 2014, Malinowski et al., 2015), Visual Madlibs (Yu et al., 2015), and image captioning (Vinyals et al., 2015, Fang et al., 2015). Our work falls in this domain with a key difference that we produce *both* vision and NLP outputs.

Our work also has similarities with works on 'spatial relation learning' (Malinowski and Fritz, 2014, Lan et al., 2012), *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") but these are simply emergent latent representations that help our reranker pick out the most consistent pair of solutions.

Our work utilizes a line of work (Batra et al., 2012, Batra, 2012, Prasad et al., 2014) on producing diverse plausible solutions from probabilistic models, which has been successfully applied to a number of problem domains (Guzman-Rivera et al., 2013, Yadollahpour et al., 2013, Gimpel et al., 2013, Premachandran et al., 2014, Sun et al., 2015, Ahmed et al., 2015).

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) and $\mathbf{z} \in \mathcal{Z}$ (prepositional attachment expressed in sentence parse). 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* $\mathbf{Y} = \{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^M\}$.

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 Huang and Chiang (2005) for the sentence parsing module and the DivMBest algorithm (Batra et al., 2012) for the semantic segmentation module. Once we instantiate the modules in Section 4, we describe the diverse solution generation in more detail.

3.2 Joint Reasoning Across Multiple Modules

We now show how to intergrate information from both segmentation and PPAR modules. 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.

Let $\mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^M\}$ denote the M semantic segmentation hypotheses and $\mathbf{Z} = \{\mathbf{z}^1, \ldots, \mathbf{z}^M\}$ denote the M PPAR hypotheses.

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 Figure 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. Impor-

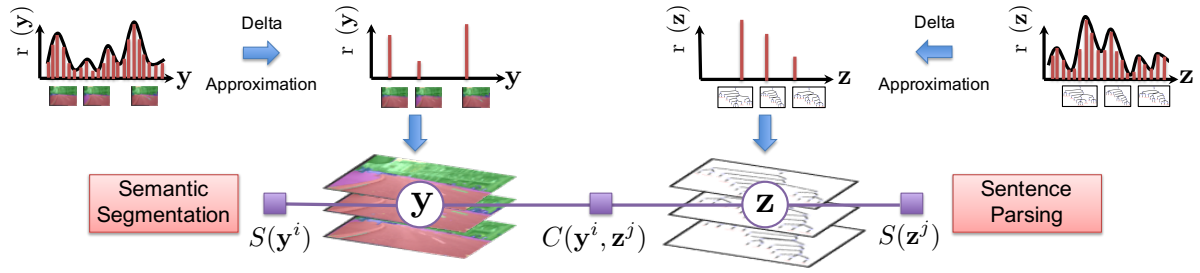


Figure 2: Illustrative inter-module factor graph. Each node takes exponentially-many or infinitely-many states and we use a 'delta approximation' to limit support.

tantly, 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.

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

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

Notice that the search space with M hypotheses each is M^2. In our experiments, we allow each module to take a different value for M, and typically use around 10 solutions for each module, leading to a mere 100 pairs, which is easily enumerable. 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. 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 Figure 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` pair, 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:

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

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

Intuitively, we can see that the constraint (2b) tries to maximize the (soft) margin between the score of the `oracle` pair and all other pairs 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 pairs 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.

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) (Everingham et al., 2010), while for evaluating sentence parses w.r.t. their prepositional 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).

4 Experiments

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

Datasets. Access to rich annotated image + caption datasets is crucial for performing quantitative evaluations. Since this is the first paper to study the problem of joint segmentation and PPAR, no standard datasets for this task exist so we had to curate our own annotations for PPAR on three image caption datasets – ABSTRACT-50S (Vedantam et al., 2014), PASCAL-50S (Vedantam et al., 2014) (expands the UIUC PASCAL sentence dataset (Rashtchian et al., 2010) from 5 captions per image to 50), and PASCAL-Context-50S (Mottaghi et al., 2014) (which uses the PASCAL Context image annotations and the same sentences as PASCAL-50S). Our annotations are publicly available on the authors' webpages. To curate the PASCAL-Context-50S 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 perform a similar filtering for PASCAL-50S and ABSTRACT-50S (using the top-6 prepositions for ABSTRACT-50S). Details are in the supplement. We consider a preposition ambiguous if there are at least two parsings where one of the two objects in the preposition dependency is the same across the two parsings while the other object is different (*e.g.* (dog on couch) and (woman on couch)). To summarize the statistics of all three datasets:

1. **ABSTRACT-50S** (Vedantam et al., 2014): 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". Overall, there are 502 total prepositions, 406 ambiguous prepositions, 80.88% ambiguity rate and 60 sentences with multiple ambiguous prepositions.
2. **PASCAL-50S** (Vedantam et al., 2014): 50,000 sentences (50 per image) for the images in the UIUC PASCAL sentence dataset (Rashtchian et al., 2010). 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. Overall, there are 213 total prepositions, 147 ambiguous prepositions, 69.01% ambiguity rate and 35 sentences with multiple ambiguous prepositions.
3. **PASCAL-Context-50S** (Mottaghi et al., 2014): We use images and captions from PASCAL-50S, but with PASCAL Context segmentation 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 1,822 image-caption pairs. Ground truth annotations for the PPAR were collected using Amazon Mechanical Turk. Workers were shown an image and a prepositional attachment (extracted from the corresponding parsing of the caption) as a phrase ("woman on couch"), and asked if it was correct. A screenshot of our interface is available in the supplement. Overall, there are 2,540 total prepositions, 2,147 ambiguous prepositions, 84.53% ambiguity rate and 283 sentences with multiple ambiguous prepositions.

Setup. *Single Module:* We first show that visual features help PPAR by using the 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 a 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 (Chen et al., 2015) and DivMBest (Batra et al., 2012) to produce M diverse segmentations of the images. To evaluate we use image-level class-averaged Jaccard Index.

Module 2: PP Attachment Resolution (PPAR)

z. We use a recent version (v3.3.1; released 2014) of the PCFG Stanford parser module (De Marneffe et al., 2006, Huang and Chiang, 2005) 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)* (the number indicates the word position in sentence). To evaluate, we count the percentage of preposition attachments that the parse gets correct.

Baselines:

- INDEP. In our experiments, we compare our proposed approach (MEDIATOR) to the highest scoring solution predicted independently from each module. For semantic segmentation this is the output of DeepLab-CRF (Chen et al., 2015) and for the PPAR module this is the 1-best output of the Stanford Parser (De Marneffe et al., 2006, Huang and Chiang, 2005). 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 reranker 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 semantic 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 M_z parse trees ($M_z = 10$).

Our approach (MEDIATOR) significantly outperforms both baselines. The improvements over INDEP show 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 (Heitz et al., 2008) 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 supplement.

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 (Mikolov et al., 2013) 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 ranks and solution scores from DeepLab-CRF (Chen et al., 2015).
- **PPAR Score Features ($\phi_S(z^i)$) (9-dim)**: We use ranks and the log probability of parses from (De Marneffe et al., 2006), and 7 binary indicators for PASCAL (6 for ABSTRACT-50S) denoting which prepositions are present in the parse.

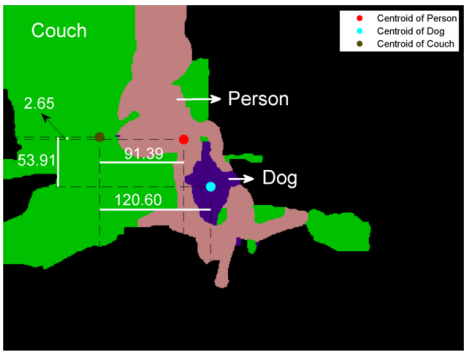


Figure 3: Example on PASCAL-50S ("A dog is standing next to a woman on a couch."). The ambiguity in this sentence "(dog next to woman) on couch" vs "dog next to (woman on couch)". We calculate the horizontal and vertical distances between the segmentation centers of "person" and "couch" and between the 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.

- **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 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 $\max\{0,$ (normalized -directional-distance)$\}$, where directional-distance measures above/below/left/right displacements between the two objects in the segmentation, and normalization involves dividing by height/width.
  - One feature is the ratio of sizes between object$_1$ and object$_2$ in the 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 Figure 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 similarity with all PASCAL categories. For each of the PASCAL categories, the feature is the sum of similarities (with the PASCAL category) over all nouns if the category is present in segmentation, and is -1 times the 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

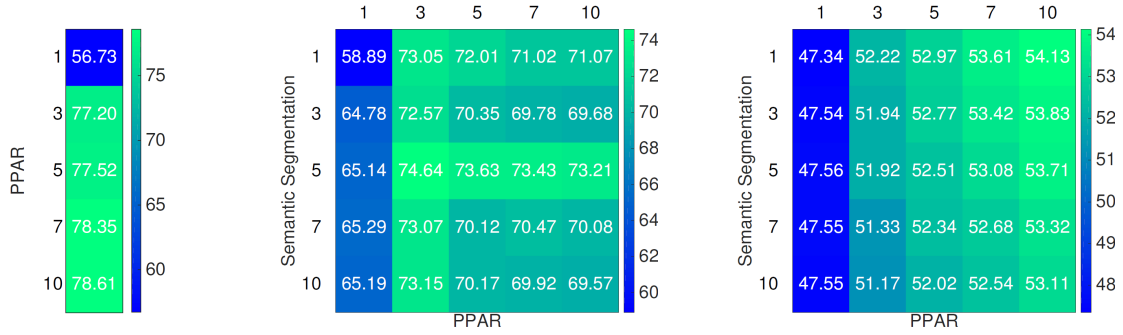
We performed a 10-fold cross-validation on the ABSTRACT-50S dataset to pick M (=10) and the weight on the hinge-loss for MEDIATOR (C). The results are presented in Table 1. Our approach significantly outpeforms 1-best outputs of the Stanford Parser (De Marneffe et al., 2006) by 20.66% (36.42% relative). 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.

Module	Stanford Parser	Domain Adaptation	Ours	`oracle`
PPAR	56.73	57.23	**77.39**	97.53

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

4.2 Multiple-Module Results

We performed 10-fold cross-val for our results of PASCAL-50S and PASCAL-Context-50S, with 8



(a) ABSTRACT-50S (b) PASCAL-50S (c) PASCAL-Context-50S

Figure 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	66.83	-	-	**43.94**	-	-
Stanford Parser	-	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 are able to significantly outperform the Stanford Parser and make small improvements over DeepLab-CRF for PASCAL-50S.

`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. Figure 4 shows the average combined accuracy on `val`, which was found to be maximal at $M_{\mathbf{y}} = 5, M_{\mathbf{z}} = 3$ for PASCAL-50S, and $M_{\mathbf{y}} = 1, M_{\mathbf{z}} = 10$ for PASCAL-Context-50S, which are used at test time.

We present our results in Table 2. Our approach significantly outperforms the Stanford Parser (De Marneffe et al., 2006) 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 (Chen et al., 2015) 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. 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 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,

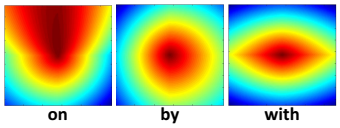


Figure 5: Visualizations for 3 different prepositions (red = high scores, blue = low scores). We can see that our model has implicitly learned spatial arrangements unlike other spatial relation learning (SRL) works.

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

Table 3: Ablation study of different feature combinations. Only PPAR Acc. is shown for PASCAL-Context-50S because $M_y = 1$.

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, domain adaptation alone is not sufficient. We also see that `oracle` performance is fairly high, suggesting that when there is ambiguity and room for improvement, MEDIATOR is able to rerank effectively.

Ablation Study for Features. Table 3 displays 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 all consistency features contribute to the performance of MEDIATOR.

Visualizing Prepositions. Figure 5 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 directional) between $object_1$ and $object_2$ 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 (Malinowski and Fritz, 2014, Lan et al., 2012). 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. Large gains in PPAR validate our intuition that vision is very helpful for dealing with ambiguity in language. Furthermore, we see even larger gains are possible from the oracle accuracies.

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 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 N00014-14-1-0679, and GPU donations by NVIDIA, all awarded to DB. GC was supported by DTRA grant HDTRA1-13-1-0015 provided by KK. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of the U.S. Government or any sponsor.

References

Faruk Ahmed, Dany Tarlow, and Dhruv Batra. 2015. Optimizing Expected Intersection-over-Union with Candidate-Constrained CRFs. In *ICCV*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Kent Bach. 2016. Routledge encyclopedia of philosophy entry. `http://online.sfsu.edu/kbach/ambguity.html`.

Kobus Barnard and Matthew Johnson. 2005. Word Sense Disambiguation with Pictures. *Artificial Intelligence*, 167.

Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. 2012. Diverse M-Best Solutions in Markov Random Fields. In *ECCV*.

Dhruv Batra. 2012. An Efficient Message-Passing Algorithm for the M-Best MAP Problem. In *UAI*.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. Do You See What I Mean? Visual Resolution of Linguistic Ambiguities. In *EMNLP*.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2015. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*.

Ernest Davis. 2016. Notes on ambiguity. `http://cs.nyu.edu/faculty/davise/ai/ambiguity.html`.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC*.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2).

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *CVPR*.

Sanja Fidler, Abhishek Sharma, and Raquel Urtasun. 2013. A Sentence is Worth a Thousand Pixels. In *CVPR*.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised Visual Sense Disambiguation for Verbs using Multimodal Embeddings. In *NAACL HLT*.

Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. 2014. A Visual Turing Test for Computer Vision Systems. In *PNAS*.

K. Gimpel, D. Batra, C. Dyer, and G. Shakhnarovich. 2013. A Systematic Exploration of Diversity in Machine Translation. In *EMNLP*.

Abner Guzman-Rivera, Pushmeet Kohli, and Dhruv Batra. 2013. DivMCuts: Faster Training of Structural SVMs with Diverse M-Best Cutting-Planes. In *AISTATS*.

Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. 2008. Cascaded Classification Models: Combining Models for Holistic Scene Understanding. In *NIPS*.

Liang Huang and David Chiang. 2005. Better k-best Parsing. In *IWPT*, pages 53–64.

Jörg H. Kappes, Bjoern Andres, Fred A. Hamprecht, Christoph Schnörr, Sebastian Nowozin, Dhruv Batra, Sungwoong Kim, Bernhard X. Kausler, Jan Lellmann, Nikos Komodakis, and Carsten Rother. 2013. A Comparative Study of Modern Inference Techniques for Discrete Energy Minimization Problems. In *CVPR*.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? Text-to-Image Coreference. In *CVPR*.

Tian Lan, Weilong Yang, Yang Wang, and Greg Mori. 2012. Image Retrieval with Structured Object Queries Using Latent Ranking SVM. In *ECCV*.

Mateusz Malinowski and Mario Fritz. 2014. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv preprint arXiv:1411.5190*.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *ICCV*.

Talya Meltzer, Chen Yanover, and Yair Weiss. 2005. Globally Optimal Solutions for Energy Minimization in Stereo Vision Using Reweighted Belief Propagation. In *ICCV*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*.

Massimo Poesio and Ron Artstein. 2005. Annotating (Anaphoric) ambiguity. In *Corpus Linguistics Conference*.

Adarsh Prasad, Stefanie Jegelka, and Dhruv Batra. 2014. Submodular meets Structured: Finding Diverse Subsets in Exponentially-Large Structured Item Sets. In *NIPS*.

Vittal Premachandran, Daniel Tarlow, and Dhruv Batra. 2014. Empirical Minimum Bayes Risk Prediction:

How to extract an extra few% performance from vision models with just three more parameters. In *CVPR*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the workshop on Human Language Technology*. ACL.

Qing Sun, Ankit Laddha, and Dhruv Batra. 2015. Active Learning for Structured Probabilistic Models With Histogram Approximation. In *CVPR*.

Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. 2008. A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *PAMI*, 30(6).

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*.

Payman Yadollahpour, Dhruv Batra, and Greg Shakhnarovich. 2013. Discriminative Re-ranking of Diverse Segmentations. In *CVPR*.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See No Evil, Say No Evil: Description Generation from Densely Labeled Images. In *Lexical and Computational Semantics*.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual Madlibs: Fill in the Blank Description Generation and Question Answering. In *ICCV*.