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

Gordon Christie1, Ankit Laddha2, Aishwarya Agrawal1, Stanislaw Antol1, Yash Goyal1, Kevin Kochersberger1, Dhruv Batra3,4
1Virginia Tech 2Carnegie Mellon University 3Georgia Institute of Technology 4* denotes equal contribution

Overview

Objective:
• Simultaneously reason about language with prepositional phrase attachment resolution (PPAR), and vision with semantic segmentation (SS)
• PPAR: pick a sentence parse with the correct prepositional attachments (e.g. elephant in pajamas, person in pajamas, or both in pajamas)
• SS: label each image pixel with a category (e.g. elephant)

Approach

Learning & Inference

Module Features
ranks, solution scores, binary presence features

Consistency Features
object distances, preps in segmentation?

Prepositional Phrase Attachment Resolution (PPAR)

DeepLab-CRF (Chen et al. ’15)

Semantic Segmentation

Estimating

Multiple hypotheses

DnMBest (Batra et al. ’12)

Problem

• Need vision to resolve language ambiguities
• Two roadblocks for writing a single unified model:
  (1) Search Space Explosion
  (2) Inaccurate models

Our Idea

• We curate 3 datasets, and keep (image, sentence) pairs where prepositional attachment ambiguities exist
• Only the top 10 parses are used in each dataset
• We further sample to ensure uniform distribution across prepositions

Datasets

Single Module

PASCAL-50S
Top-6 prepositions used: "with", "next to", "on", "over", "in front of", "behind", and "under"
396 sentences, 201 unique images

PASCAL-Context-50S
Same prepositions as PASCAL-50S
2,182 sentences, 966 unique images

Multiple Modules

PASCAL-50S
Ground truth PPAR annotated by two vision+NLP grad students
PASCAL image annotations

PASCAL-Context-50S
Ground truth PPAR annotated on Amazon Mechanical Turk
PASCAL Context image annotations

Results

Approach Comparisons

Single Module Results

Language alone does not suffice

Multiple Module Results

Language alone does not suffice

Visualizations

Further improvements possible

• Our approach implicitly learns the spatial arrangements of the prepositions
• Visualizations are shown for 3 prepositions
• These show where we expect object, to be when connected to object, (center of the visualizations) by one of the prepositions

Approach

• Our approach, each module generates diverse solutions and then we pick the best pair.

Approximation

• Baseline: Each module outputs its 1-best solution. Improvements over this establish importance of joint reasoning.

Visualization

Scores:

- red = High
- blue = Low

DeepLab-CRF

Data available!