Using Natural Language Processing to Aid Computer Vision

Ray Mooney

Department of Computer Science

University of Texas at Austin

My Initial Motivation for Exploring Grounded Language Learning

- I became frustrated having to manually annotate sentences with formal meaning representations (MRs) to train semantic parsers.
- I hoped to automatically extract MRs from the perceptual context of situated language.
- However, computer vision is generally not capable enough yet to provide the MRs needed for semantic parsing.
- Therefore, I focused on grounded language learning in virtual worlds to circumvent computer vision.

Computer Vision is Hard

- Both language understanding and vision are "AI complete" problems.
- However, in many ways, I think vision is the harder problem.

Text Really Helps Language Learning

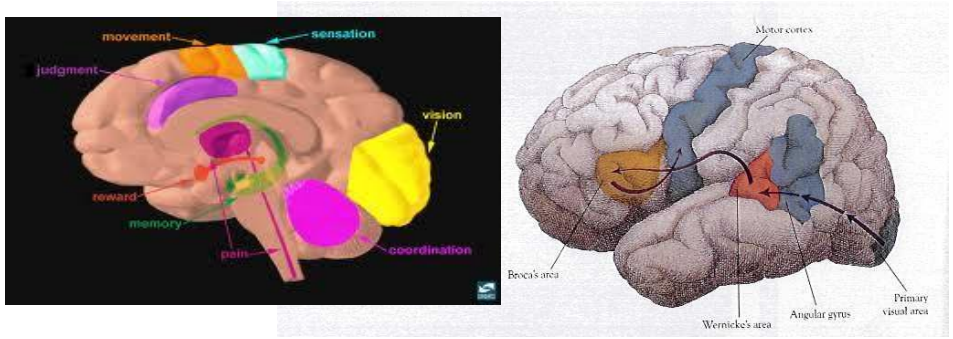
- Both speech and visual perception begin with a complex analog signal (sound vs. light waves).
- However, for language we have orthographic text that intermediates between analog signal and semantic representation.
- Training on large, easily available corpora of text dramatically benefits learning for language.
- No such pervasive, data rich, intermediate representation exists for vision.

Dimensionality
Language vs. Vision

- Language input is fundamentally 1-d, i.e. linear sequences of sounds, phonemes, or words.
- Vision is fundamentally a 2-d (arrays of pixels), 2 ½ - d (depth maps) or 3-d (world model) problem.
- Therefore, the “curse of dimensionality” makes vision a much harder problem.

Biological Hardware
Vision vs. Language

- Humans arguably have a greater amount of neural hardware dedicated to vision than to language.



- Therefore, matching human performance on vision may be computationally more complex.

Biological Evolution
Vision vs. Language

- Vision has a much longer "evolutionary history" than language.
  - First mammals: 200-250 million years ago (MYA)
  - First human language: Homo habilis, 2.3 MYA
- Therefore, a much more complex neural system could have evolved for vision compared to language.

Language Helping Vision

- Now I feel sorry for my poor computer vision colleagues confronting an even harder problem.
- So I'd like to help them, instead of expecting them to help me.

As Jerry Maguire said: Help me, help you!

NL-Acquired Knowledge for Vision

- Many types of linguistic knowledge or knowledge “extracted” from text can potentially aid computer vision.

Textual Retrieval of Images

- Most existing image search is based on retrieving images with relevant nearby text.
- A variety of computer vision projects have used text-based image search to automatically gather (noisy) training sets for object recognition.
- Various ways of dealing with the noise in the resulting data:
  - Multiple-instance learning
  - Cleanup results with crowd-sourcing

Lexical Ontologies

- ImageNet built a large-scale hierarchical database of images by using the WordNet ontology (Deng et al. CVPR-09)
- "Hedging Your Bets" method uses ImageNet hierarchy to classify object as specifically as possible while maintaining high accuracy (Deng et al. CVPR-12)

Subject-Verb-Object (SVO) Correlations

- We developed methods for using probability estimates of SVO combinations to aid activity recognition (Motwani & Mooney, ECAI-12) and sentential NL description (Krishnamoorthy et al., AAAI-13) for YouTube videos.
- SVO probabilities are estimated using a smoothed trigram "language model" trained on several large dependency-parsed corpora.
- Similar statistics can also be used to predict verbs for describing images based on the detected objects (Yang et al., EMNLP 2011).

Object-Activity-Scene Correlations

- Can use co-occurrence statistics mined from text for objects, verbs, and scenes to improve joint recognition of these from images or videos.
- Such statistics have been used to help predict scenes from objects and verbs (Yang et al. EMNLP 2011).

Object-Object Correlations

- Could use object-object co-occurrence statistics mined form text to acquire knowledge that could aid joint recognition of multiple objects in a scene.
- An “elephant” is more likely to be seen in the same image as a “giraffe” than in the same image as a “penguin”

Scripts
Activity-Activity Correlations and Orderings

- Knowledge of stereotypical sequences of actions/events can be mined from text (Chambers & Jurafsky, 2008).
- Such knowledge could be used to improve joint recognition of sequences of activities from video.
  - Opening a bottle is typically followed by drinking or pouring.

Transferring Algorithmic Techniques from Language to Vision

- Text classification suing “bag of words” to image classification using “bag of visual words”
- Linear CRFs for sequence labeling (e.g. POS tagging) for text to 2-d mesh CRFs for pixel classification in images.
- HMMs for speech recognition to HMMs for activity recognition in videos.

Other Ways Language can Help Vision?

Your Idea Here

Conclusions

- Its easier to use language processing to help computer vision than the other way around.
- For a variety of reasons computer vision is harder than NLP.
- Knowledge about or from language can be used to help vision in various ways.

Help me, help you, help me!

Recent Spate of Workshops on Grounded Language

- NSF 2011 Workshop on Language and Vision
- AAAI-2011 Workshop on Language-Action Tools for Cognitive Artificial Agents: Integrating Vision, Action and Language
- NIPS-2011 Workshop on Integrating Language and Vision
- NAACL-2012 Workshop on Semantic Interpretation in an Actionable Context
- AAAI-2012 Workshop on Grounding Language for Physical Systems
- NAACL-2013 Workshop on Vision and Language
- CVPR-2013 Workshop on Language for Vision
- UW-MSR 2013 Summer Institute on Understanding Situated Language