PRINCIPLES OF GROUNDED LANGUAGE LEARNING

Natalie Parde
parde@uic.edu

CS 594: Language and Vision
Spring 2019

What is grounded language learning?

The process of learning representations for words based on non-linguistic experience.

Orange is the colour between yellow and red on the spectrum of visible light. Human eyes perceive orange when observing light with a dominant wavelength between roughly 585 and 620 nanometres.

- https://en.wikipedia.org/wiki/Orange_(colour)

orange

Origins in Cognitive Science

- If we can understand how *humans* understand language, we can hopefully figure out how to replicate it in computers!
- Sapir-Whorf vs. Chomsky
  - *Is language deterministic?*
- Chomsky: Language is a critical aspect of cognition
- Leading theory driving automatic grounded language learning:
  - *Language is a formal symbol system*

Language of Thought

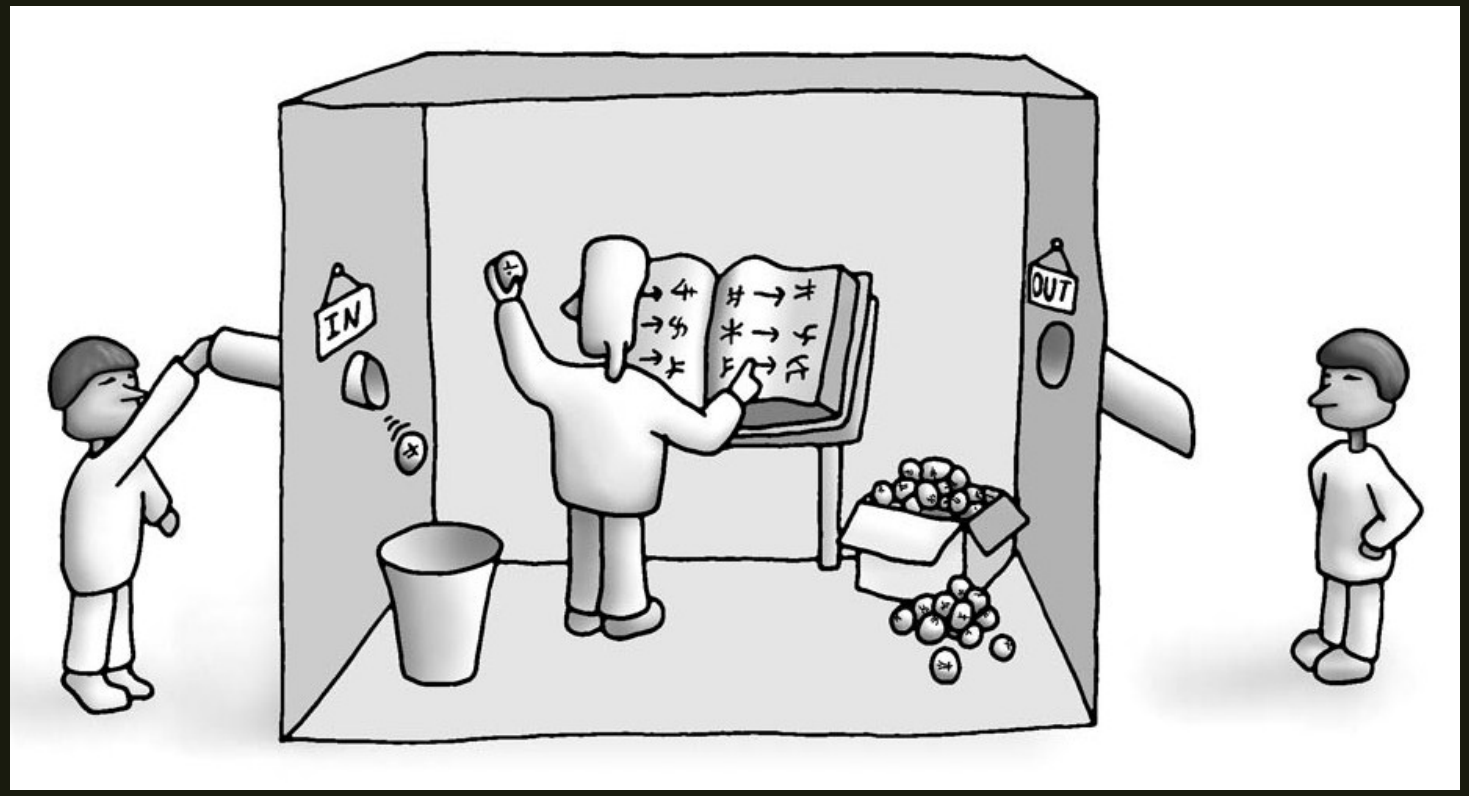
- The Language of Thought Hypothesis:[1]
  - *Thought occurs via structured relationships between symbols in a mental language*
  - *Mental language ≠ specific spoken language*
    - Mental language ⟶ common language across all humans
    - Symbols in specific spoken languages are mapped to symbols in the mental language

[1]Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Turing Test ≈ Language Understanding?

- Thought Experiment:
  - *Could you, a human, pass a Turing Test in a language you do not understand?*
  - *If so, how?*

- https://www.google.com/search?q=timer



Searle's Chinese Room

- A person who does not speak Chinese could pass a Turing Test in Chinese, by looking up appropriate outputs for inputs in a Chinese-Chinese dictionary. The person **still would not understand Chinese.**[1]

[1]Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Symbol Grounding Problem

- How do words get their meanings?[1]
  - *Symbols do not represent anything on their own*
  - *Symbols cannot be defined only in terms of other symbols*

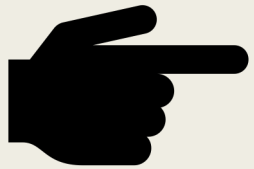
[1]Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Humans understand language by mapping symbols to real-world experiences.

- Human language understanding → far superior to current machine language understanding!
- Grounded language learning attempts to replicate human language understanding using multimodal statistical models, resulting in:
  - *Improved mechanisms for automatic language understanding*
  - *In some cases, new insights to human cognition*

How can machines experience language?

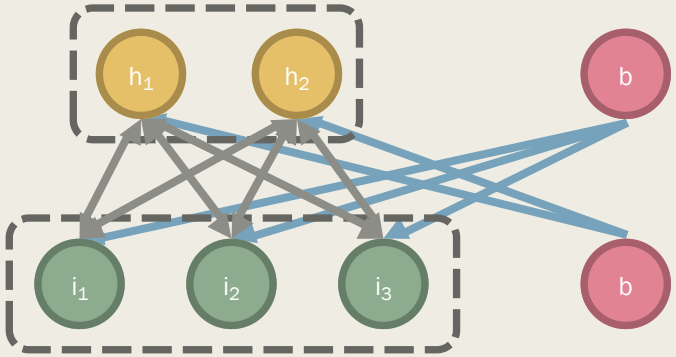
- Sensory input modalities!
  - *Vision*
    - Images
    - Videos
  - *Audio recordings*
  - *Olfactory perception*[1]
  - *Haptic feedback*
- Multimodal: Utilizing more than one input modality



[1]Kiela, D., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Vol. 2, pp. 231-236).

What types of models allow machines to associate language with sensory perceptions?

- In theory, any classification or regression algorithm could work
- Recently, neural networks:
  - *Autoencoders*[1]
  - *Boltzmann Machines*[2]
  - *LSTMs*[3]



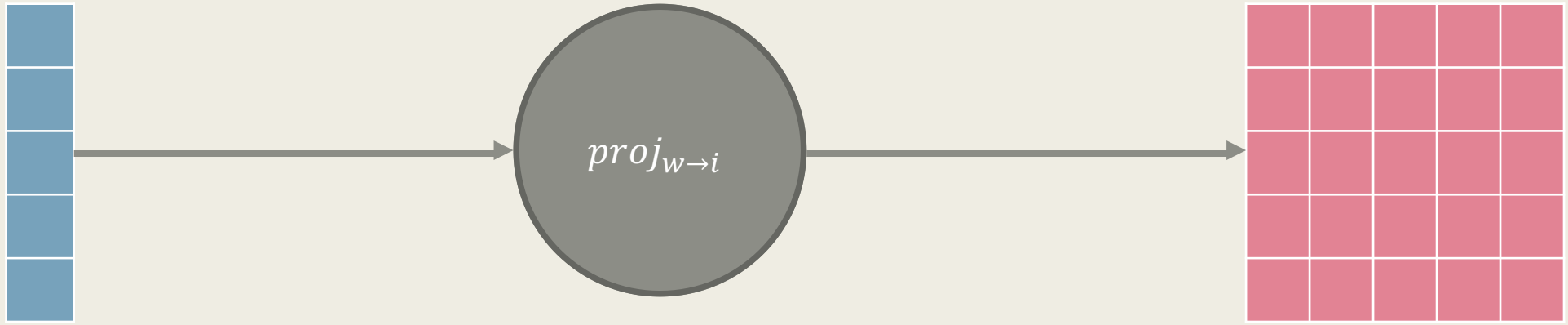
[1]Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 721-732).

[2]Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222-2230).

[3]Rajagopalan, S. S., Morency, L. P., Baltrusaitis, T., & Goecke, R. (2016, October). Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision* (pp. 338-353).

Models can learn mappings from one modality to the other, or learn representations based on multiple modalities in the same space.

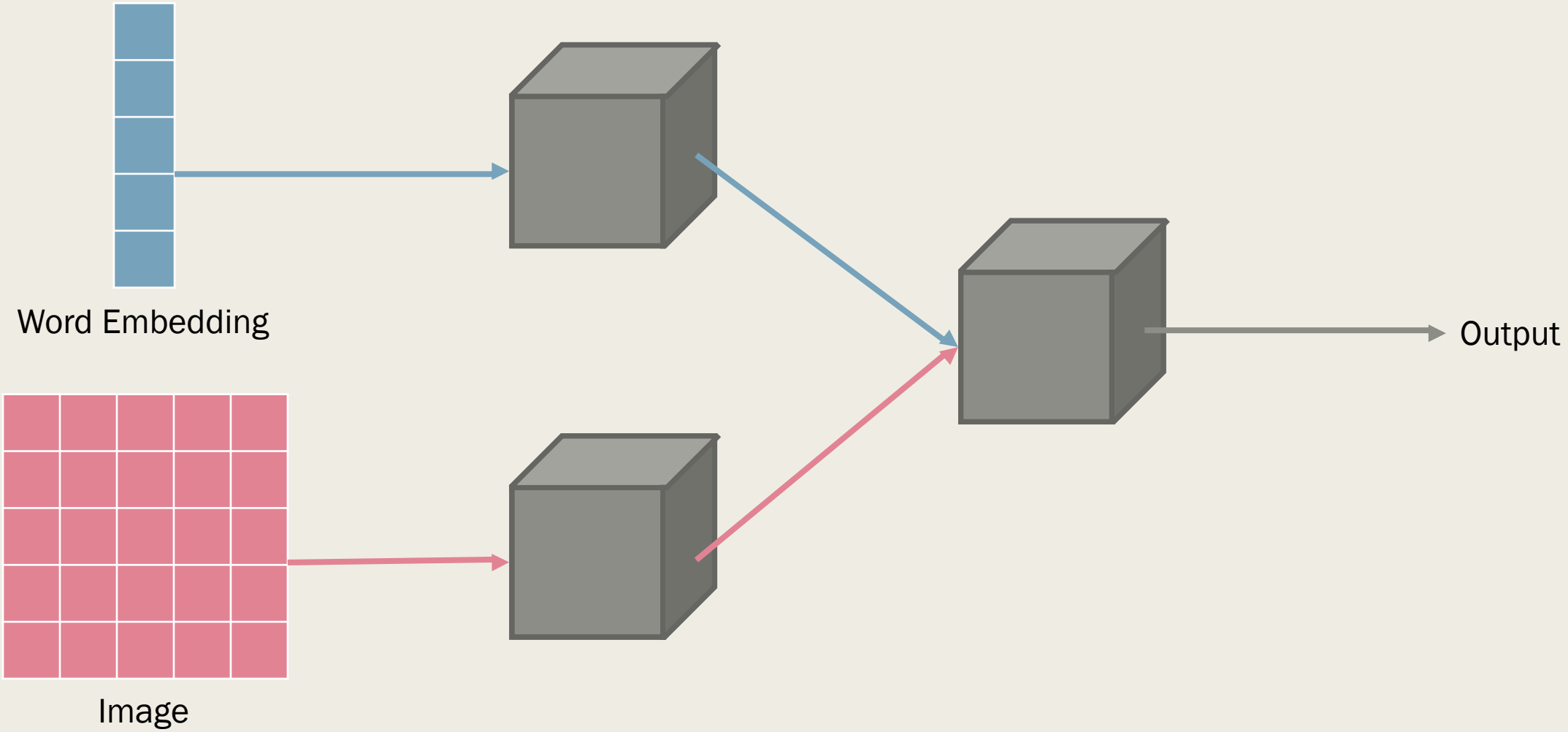
Cross-Modal Mapping



Word Embedding

Image

Joint Representation



Canonical Correlation Analysis

- Popular approach for mapping text and image features to the same space
- Finds pairs of linear projections that maximize the correlation between the text and image features

$$(a'; b') = \underset{a,b}{\operatorname{argmax}} \operatorname{corr}(a^T X, b^T Y)$$

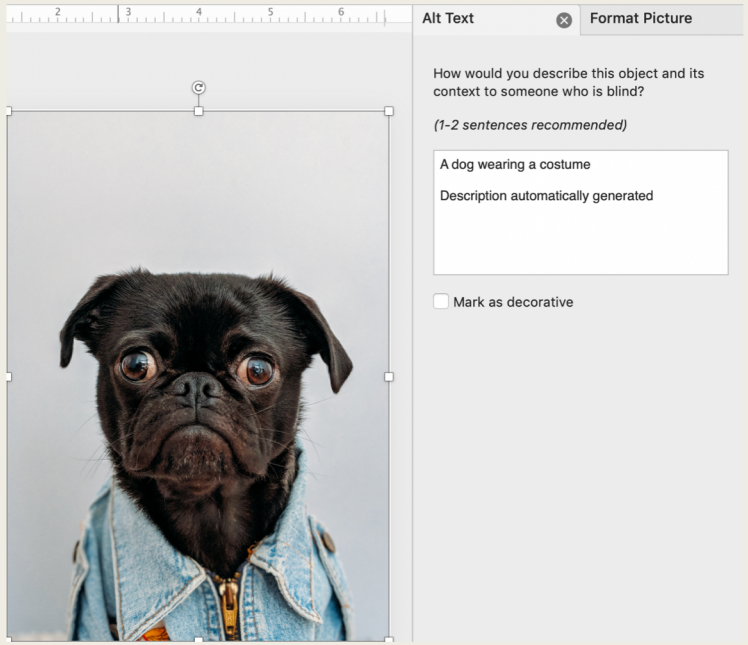
Common Application for Cross-Modal Mapping

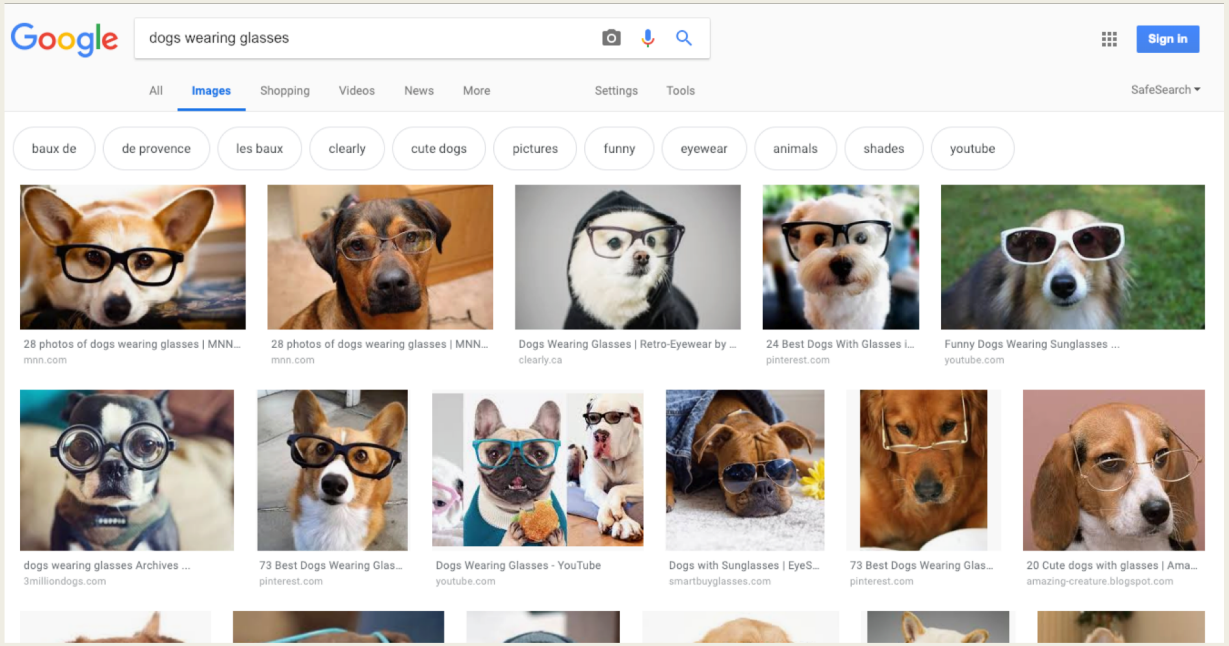
- Zero-shot learning
  - *Building a grounded representation for a word before it's been encountered*
    - "We found a cute, hairy wampimuk sleeping behind the tree."[1]
    - "He put on his sunglasses, rolled down the windows, and sped off into the sunset in his flashy new wampimuk."

[1]Lazaridou, A., Bruni, E., & Baroni, M. (2014). Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1403-1414).

Common Applications for Joint Representations

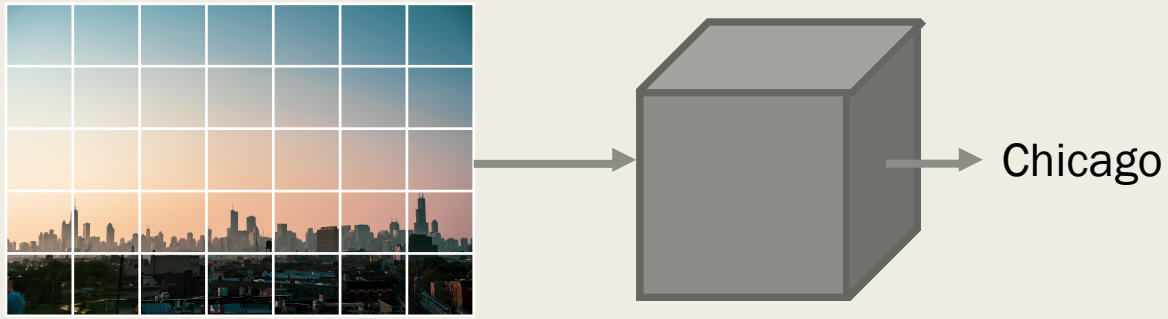
- Image Captioning
- Image Search





How are grounded language models evaluated?

- Model predictions compared with human-provided labels
- Very common:
  - *Accuracy*
- Sometimes:
  - *Precision*
  - *Recall*
  - *F1*



Precision, Recall, and F1

- Precision: Of the values predicted to be X, how many actually *are* X?
- Recall: Of the values that actually *are* X, how many were predicted to be X?
- F1: What is the harmonic mean between precision and recall?

TP: Predicted Positive & Actually Positive	**FP:** Predicted Positive, Not Actually Positive
FN: Predicted Negative, Actually Positive	**TN:** Predicted Negative & Actually Negative

- $P = \frac{TP}{TP+FP}$
- $R = \frac{TP}{TP+FN}$
- $F_1 = 2\left(\frac{P \times R}{P+R}\right)$

Resources

Targeted Workshops:

- Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics: https://splu-robonlp.github.io/
- Visually Grounded Interaction and Language: https://nips2018vigil.github.io/
- Language Grounding for Robotics: https://robo-nlp.github.io/2017_index.html

Relevant Lectures:

- How We Teach Computers to Understand Pictures, by Fei Fei Li: https://youtu.be/40riCqvRoMs
- From Naïve Physics to Connotation: Learning and Reasoning about the World Using Language, by Yeijin Choi: https://youtu.be/V1vRmKnjagw
- Robots that Learn Grounded Language through Interactive Dialog, by Ray Mooney: https://youtu.be/8ZUkF3dNURQ

ImageNet: http://www.image-net.org/

COCO: http://cocodataset.org

Wrapping up....

- Overview
- Cognitive Science Origins
  - *Language of Thought*
  - *Searle's Chinese Room*
  - *Symbol Grounding Problem*
- Multimodality
- Grounded Language Models
  - *Cross-Modal Mapping*
  - *Joint Representation*
- Sample Applications
- Evaluation Metrics
- Resources