Semantic Grounding in Visual and World Knowledge

Richard Socher

Chris Manning and Andrew Ng

Joint work with
Danqi Chen, Milind Ganjoo and Quoc Le
Grounding Semantics

- Most semantic word spaces in NLP are learned only from language itself, e.g. distributional word vectors from (neural) language models

50-d word vectors from neural language models
Collobert & Weston, 2008, Turian et al, 2010
Grounding Semantics

- Most semantic word spaces in NLP are learned only from language itself, e.g. distributional word vectors from (neural) language models
Grounding Meaning in World Knowledge

- Some semantic word information is not captured in text
- Explicit relationships between entities, e.g. location of geographic entities, family relations between people, parts of objects (cat, has, tail), ...
- Ability to do logical reasoning

1. Ground Words in Knowledgebases
Grounding Meaning in Visual Knowledge

- Visual information about words cannot be captured in text
- What does green mean?
- How visually similar are object categories?
- What does a typical bird look like?
- How can actions be visualized?

2. Ground words with images
3. Learn joint image - sentence space with Recursive Deep Learning!
Grounding in World Knowledge

REASONING WITH NEURAL TENSOR NETWORKS FOR KNOWLEDGE BASE COMPLETION
Common Sense Reasoning Inside Knowledge Bases

• Question: Can Neural Networks learn to capture logical inference, set inclusions, part-of and hypernym relationships?
Neural Networks for Reasoning over Relationships

• Represent facts in knowledgebase as triplets:
  \[ T = (e_1, R, e_2) \]

• Learning neural network architecture that scores each correct triplet in knowledge base highly

• Neural Tensor Network for scoring triplets

\[
g(e_1, R, e_2) = u^T_R f \left( e^T_1 W^{[1:k]}_R e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R \right)\]
Neural Tensor Network for Scoring Triplets

\[ g(e_1, R, e_2) = u_R^T f \left( e_1^T W_{R}^{[1:k]} e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R \right) \]

Bilinear form works well but some relationships have to relate pairs from different subspaces, e.g. has-part.
Neural Tensor Network for Scoring Triplets

\[ g(e_1, R, e_2) = u^T_R f \left( e_1^T W_R^{[1:k]} e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R \right) \]

Multiple tensor slices allow model to capture different subspace relationships
Neural Tensor Network for Scoring Triplets

\[ g(e_1, R, e_2) = u^T_R f \left( e_1^T W^{[1:k]}_R e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R \right) \]

Apply nonlinearity and sum up for score.
Neural Tensor Network

- Include standard neural network for generality (not always necessary)
- New type of neural network layer
- Related to Ranzato et al. 2010: 3-way RBMs and Deng et al. 2013: tensor networks

\[
U^T \, f( e_1^T \, W^{[1:2]} \, e_2 + V \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} + b )
\]
Neural Networks for Reasoning over Relationships

• Training uses contrastive estimation function (Smith and Eisner 2005) and similar to word vector learning (Collobert and Weston 2008)

• Create corrupted triplets by replacing one entity with another random one

• Cost function:

\[
\sum_{i=1}^{N} \sum_{c=1}^{C} \max \left( 0, 1 - g \left( T^{(i)} \right) + g \left( T_{c}^{(i)} \right) \right) + \lambda \| \Omega \|_2^2,
\]
• Use gradients and mini-batched L-BFGS (AdaGrad improves baseline models but not NTN)
• Tensor gradients:

\[
\frac{\partial g(e_1, R, e_2)}{\partial W[j]} = u_j f'(z_j)e_1 e_2^T,
\]

where \( z_j = e_1^T W[j] e_2 + V_j. \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_j \)
Data and Accuracy of Identifying Correct Relationships

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#R.</th>
<th># Ent.</th>
<th># Train</th>
<th># Dev</th>
<th># Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordnet</td>
<td>11</td>
<td>38,696</td>
<td>112,581</td>
<td>2,609</td>
<td>10,544</td>
</tr>
<tr>
<td>Freebase</td>
<td>13</td>
<td>75,043</td>
<td>316,232</td>
<td>5,908</td>
<td>23,733</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>WordNet</th>
<th>Freebase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance Model</td>
<td>68.3</td>
<td>61.0</td>
</tr>
<tr>
<td>Hadamard Model</td>
<td>80.0</td>
<td>68.8</td>
</tr>
<tr>
<td>Single Layer Model</td>
<td>76.0</td>
<td>85.3</td>
</tr>
<tr>
<td>Bilinear Model</td>
<td>84.1</td>
<td>87.7</td>
</tr>
<tr>
<td>Neural Tensor Network</td>
<td>86.2</td>
<td>90.0</td>
</tr>
</tbody>
</table>

- Sutskever et al, Modelling relational data using Bayesian clustered tensor factorization 2009.
- Bordes et al. Learning structured embeddings of knowledge bases 2011
- Jenatton et al. A latent factor model for highly multi-relational data 2012
Entity Representations

- Use word vectors instead of entity vectors
- Initialize word vectors for further small improvement
- Possibly more on out-of-vocabulary
Accuracy Per Relationship

WordNet

- domain topic: 70%
- similar to: 80%
- synset domain topic: 90%
- domain region: 75%
- subordinate instance of: 95%
- has part: 85%
- part of: 90%
- member holonym: 85%
- member meronym: 90%
- type of: 80%
- has instance: 75%

FreeBase

- ethnicity (211): 90%
- religion (107): 95%
- cause of death (170): 80%
- institution (727): 75%
- profession (455): 90%
- nationality (188): 95%
- gender (2): 100%
FreeBase Example

- Francesco Guicciardini: historian, place of birth Florence, nationality Italy
- Francesco Patrizi: male, nationality Italy
- Matteo Rosselli: location Florence

Relationships:
- Francesco Guicciardini and Francesco Patrizi share the relationship of nationality.
- Francesco Guicciardini and Matteo Rosselli share the relationship of location.
- Francesco Guicciardini and Francesco Patrizi share the relationship of gender.
Grounding Semantics in the Visual World

ZERO-SHOT LEARNING THROUGH CROSS-MODAL TRANSFER
What is Zero-Shot Learning?

- Class: Mantis shrimp
  - A marine crustacean that can be up to 30 cm long, has bright neon colors and powerful claws. Their eyes are mounted on mobile stalks and are considered to be the most complex eyes in the animal kingdom.
Zero-Shot Learning

- Which one is the Mantis shrimp?
Zero-Shot Learning

• Which one is the Mantis shrimp?
  1. Eliminate known categories
Zero-Shot Learning

- Which one is the Mantis shrimp?
  2. Classify between the unknown categories
Deep Zero Shot Learning: Step 1

- Build multimodal meaning representations:
  - Start with word vectors learned with an unsupervised model (Bengio et al. 2003, Collobert and Weston 2008, Huang et al. 2012)
Deep Zero Shot Learning: Step 1

• Build multimodal meaning representations by learning a mapping of images to the word vector corresponding to their class (for seen classes)

$$J(\Theta) = \sum_{y \in Y_s} \sum_{x^{(i)} \in X_y} \left\| w_y - \theta^{(2)} f \left( \theta^{(1)} x^{(i)} \right) \right\|^2$$
Deep Zero Shot Learning: Step 2

- For a new image, determine whether it is of a seen or unseen class (via outlier detection methods)
Deep Zero Shot Learning: Step 3

• If image is of a known category, use classifier
• Otherwise, classify based on Euclidean distance to word vectors of zero shot classes
Bayesian Model

- Combine previous steps in single joint seen-unseen model:

\[
p(y|x, X_s, F_s, W, \theta) = \sum_{V \in \{s, u\}} P(y|V, x, X_s, F_s, W, \theta) P(V|x, X_s, F_s, W, \theta)
\]
Results: CIFAR 10
Results

![Bar graph showing zero-shot accuracy for different class pairs](image)

- **cat-dog**: 0.50
- **plane-auto**: 0.60
- **auto-deer**: 0.75
- **deer-ship**: 0.85
- **cat-truck**: 0.90

*Pair of zero-shot classes used*
Results: CIFAR 10

(a) Gaussian model

Accuracy vs. Fraction of points classified as unseen

0.7425
Word Prototype Visualization
COMPOSITIONAL SENTENCE EMBEDDINGS FOR DESCRIBING AND FINDING IMAGES
Grounding Semantic Spaces

• Semantic spaces can be enriched with reasoning capabilities and visual information
• Deep learning techniques are suitable for:
  – Logical reasoning in knowledge bases
  – Zero-Shot Learning using joint word – object spaces
  – Learning joint sentence – scene image mapping