Visually Grounded
Neural Syntax Acquisition

Haoyue Shi*  Jiayuan Mao*  Kevin Gimpel  Karen Livescu

July 29th, 2019
@ACL
When we were children...

A cat is on the lawn.
When we were children...

A cat is on the lawn.

A cat sleeps outside.
When we were children...

A cat is on the lawn.

A cat, as a whole, means something concrete.

A cat sleeps outside.
When we were children...

A cat is on the lawn.
A cat is staring at you.
A cat plays with a ball.

A cat, as a whole, means something concrete.

A cat sleeps outside.
A cat is on the ground.
There is a cat sleeping on the ground.
When we were children...

A cat is on the lawn.
A cat is staring at you.
A cat plays with a ball.

A cat, as a whole, means something concrete.

A cat sleeps outside.
A cat is on the ground.
There is a cat sleeping on the ground.

A cat, as a whole, functions as a single unit in sentences.
When we were children...

A cat is on the lawn.
A cat is staring at you.
A cat plays with a ball.

A cat was chasing a mouse.
A dog was chasing a cat.
A cat was chased by a dog.

... 

A cat sleeps outside.
A cat is on the ground.
There is a cat sleeping on the ground.

A cat, as a whole, functions as a single unit in sentences.
Problem Definition

• Given a large set of parallel image-text data (e.g., MSCOCO), can we generate linguistically plausible structure for the text?

Figure credit: Ding et al. (2018)
Problem Definition

- Given a large set of parallel image-text data (e.g., MSCOCO), can we generate linguistically plausible structure for the text?

* A cat is on the lawn
Problem Definition

- Given a large set of parallel image-text data (e.g., MSCOCO), can we generate linguistically plausible structure for the text?

A cat is on the lawn
Problem Definition

• Given a large set of parallel image-text data (e.g., MSCOCO), can we generate linguistically plausible structure for the text?

A cat is on the lawn
Visually Grounded Neural Syntax Learner

- *Concrete* spans are more likely to be constituents.

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

- $c_1$: a cat
- $c_2$: the lawn
- $c_3$: on the lawn

**Joint Embedding Space**

**Text Encoder**

- $c_1$, $c_2$, $c_3$

**Image Encoder:**

- ResNet 101 (He et al., 2015)

**Estimated Concreteness as Scores**
Visually Grounded Neural Syntax Learner

- *Concrete* spans are more likely to be constituents.

**Caption**: “A cat is on the lawn”

**Constituency Parse Tree**

$c_1 : a$ cat
$c_2 : the$ lawn
$c_3 : on$ the lawn
$...$
Greedy Bottom-Up Parser

a cat is on the lawn
Greedy Bottom-Up Parser

Compute score

$$FFN\left(\begin{bmatrix} v_a \\ v_{cat} \end{bmatrix}\right) = 4.5$$

a cat is on the lawn
Greedy Bottom-Up Parser

Compute score

$$FFN \left( \begin{bmatrix} v_{cat} \\ v_{is} \end{bmatrix} \right) = 0.5$$

a cat is on the lawn
Greedy Bottom-Up Parser

Compute score

\[ FFN \left( \begin{bmatrix} v_{is} \\ v_{on} \end{bmatrix} \right) = 1 \]
Greedy Bottom-Up Parser

Compute score

$$FFN \left( \begin{bmatrix} v_{on} \\ v_{the} \end{bmatrix} \right) = 1$$

a cat is on the lawn
Greedy Bottom-Up Parser

Compute score

\[ FFN \left( \begin{bmatrix} v_{\text{the}} \\ v_{\text{lawn}} \end{bmatrix} \right) = 3 \]
Greedy Bottom-Up Parser

0.45 0.05 0.1 0.1 0.3

a cat is on the lawn

Normalized to a probability distribution
Greedy Bottom-Up Parser

Sample a pair to combine (training)
Greedily combine (inference)

a cat is on the lawn
Greedy Bottom-Up Parser

Textual representation:

Normalized sum of children:

\[ \mathbf{v}_{(a\ cat)} = \frac{\mathbf{v}_a + \mathbf{v}_{cat}}{\|\mathbf{v}_a + \mathbf{v}_{cat}\|_2} \]
Greedy Bottom-Up Parser

(a cat) is on the lawn

0.45 0.05 0.1 0.1 0.3

Textual representation:

\[ \text{Normalized sum of children} \]

\[ \mathbf{v}_{(a \ cat)} = \frac{\mathbf{v}_a + \mathbf{v}_{cat}}{\|\mathbf{v}_a + \mathbf{v}_{cat}\|_2} \]
Greedy Bottom-Up Parser

Compute probability

(a cat) is on the lawn

a cat is on the lawn
Greedy Bottom-Up Parser

(a cat) is on (the lawn)

(a cat) is on the lawn

Combine

0.25 0.15 0.15 0.45

0.45 0.05 0.1 0.1 0.3

a cat is on the lawn
Greedy Bottom-Up Parser

((a cat) (is (on (the lawn))))

... 

(a cat) is on (the lawn)

\[0.25 \quad 0.15 \quad 0.15 \quad 0.45\]

(a cat) is on the lawn

\[0.45 \quad 0.05 \quad 0.1 \quad 0.1 \quad 0.3\]

a cat is on the lawn

Finished!
Visually Grounded Neural Syntax Learner

- *Concrete* spans are more likely to be constituents.

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

\[ c_1 : \text{a cat} \]
\[ c_2 : \text{the lawn} \]
\[ c_3 : \text{on the lawn} \]

...
Visually Grounded Neural Syntax Learner

- *Concrete* spans are more likely to be constituents.

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

\[
\begin{align*}
  & a cat \\
  & \text{is} \\
  & \text{on} \\
  & \text{the lawn} \\
\end{align*}
\]

- \(c_1\) : a cat
- \(c_2\) : the lawn
- \(c_3\) : on the lawn
- ...
Visually Grounded Neural Syntax Learner

- *Concrete* spans are more likely to be constituents.

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

- $c_1$: a cat
- $c_2$: the lawn
- $c_3$: on the lawn

**Joint Embedding Space**

**Image Encoder:** ResNet 101 (He et al., 2015)
The Joint Embedding Space

Hinge-based triplet loss between images and captions for visual semantic embeddings (VSE; Kiros et al., 2015):

\[ \mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} \left[ \text{sim}(i', c) - \text{sim}(i, c) + \delta \right]_+ + \left[ \text{sim}(i, c') - \text{sim}(i, c) + \delta \right]_+ \]

\[ \left[ \cdot \right]_+ = \max(\cdot, 0) \]

\[ \text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot) \]
The Joint Embedding Space

Hinge-based triplet loss between images and captions for visual semantic embeddings (VSE; Kiros et al., 2015):

\[
\mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]_+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]_+
\]

\[
[\cdot]_+ = \max(\cdot, 0) \quad \text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot)
\]

A cat is on the lawn.
The Joint Embedding Space

Hinge-based triplet loss between images and captions for visual semantic embeddings (VSE; Kiros et al., 2015):

\[ \mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} \left[ \text{sim}(i', c) - \text{sim}(i, c) + \delta \right]_+ + \left[ \text{sim}(i, c') - \text{sim}(i, c) + \delta \right]_+ \]

A cat is on the lawn.

\[ \mathcal{L}(\cdot) = \max(\cdot, 0) \]

\[ \text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot) \]
The Joint Embedding Space

Hinge-based triplet loss between images and captions for visual semantic embeddings (VSE; Kiros et al., 2015):

\[ \mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]_+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]_+ \]

\[ [\cdot]_+ = \max(\cdot, 0) \]

\[ \text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot) \]
Concreteness Estimation in the Joint Embedding Space

Hinge-based triplet loss between images and captions constituents for visual semantic embeddings:

\[
\mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]^+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]^+
\]

\[\cdot^+ = \max(\cdot, 0)\]

\[\text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot)\]
Concreteness Estimation in the Joint Embedding Space

Hinge-based triplet loss between images and captions constituents for visual semantic embeddings:

\[ \mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]^+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]^+ \]

\[ [\cdot]^+ = \max(\cdot, 0) \]
\[ \text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot) \]
Concreteness Estimation in the Joint Embedding Space

Hinge-based triplet loss between images and captions constituents for visual semantic embeddings:

$$\mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]_+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]_+$$

$$[\cdot]_+ = \max(\cdot, 0)$$
$$\text{sim}(\cdot, \cdot) = \cos(\cdot, \cdot)$$
Concreteness Estimation in the Joint Embedding Space

Hinge-based triplet loss between images and captions **constituents** for visual semantic embeddings:

\[
\mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]_+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]_+
\]

Abstractness: local hinge loss between constituents and images.

\[
\text{abstract}(c; i) = \mathcal{L}(i, c)
\]

\[
[\cdot]_+ = \max(\cdot, 0) \\
\text{sim}(\cdot; \cdot) = \cos(\cdot; \cdot)
\]
Concreteness Estimation in the Joint Embedding Space

Hinge-based triplet loss between images and captions *constituents* for visual semantic embeddings:

$$
\mathcal{L}(i, c) = \sum_{(i', c') \neq (i, c)} [\text{sim}(i', c) - \text{sim}(i, c) + \delta]_+ + [\text{sim}(i, c') - \text{sim}(i, c) + \delta]_+
$$

Abstractness: local hinge loss between constituents and images.

$$
\text{abstract}(c; i) = \mathcal{L}(i, c)
$$

Concreteness is defined similarly:

$$
\text{concrete}(c; i) = \sum_{(i', c') \neq (i, c)} [-\text{sim}(i', c) + \text{sim}(i, c) - \delta]_+ + [-\text{sim}(i, c') + \text{sim}(i, c) - \delta]_+
$$

$$
[\cdot]_+ = \max(\cdot, 0)
\text{sim}(\cdot; \cdot) = \cos(\cdot; \cdot)
$$

Concreteness is defined similarly:
Visually Grounded Neural Syntax Learner

• *Concrete* spans are more likely to be constituents.

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

\[
c_1 : \text{a cat} \\
c_2 : \text{the lawn} \\
c_3 : \text{on the lawn} \\
\ldots
\]

**Joint Embedding Space**

**Text Encoder**

**Image Encoder:** *ResNet 101* (He et al., 2015)
**Visually Grounded Neural Syntax Learner**

- *Concrete* spans are more likely to be constituents.

---

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

- $c_1$: a cat
- $c_2$: the lawn
- $c_3$: on the lawn

**Joint Embedding Space**

**Text Encoder**

- $c_1$ -> $c_2$ -> $c_3$

**Image Encoder:**

- ResNet 101 (He et al., 2015)

**Estimated Concreteness as Scores**
Visually Grounded Neural Syntax Learner

- *Concrete* spans are more likely to be constituents.
- REINFORCE (Williams, 1992) as gradient estimator.

**Caption:** “A cat is on the lawn”

**Constituency Parse Tree**

- $c_1$: a cat
- $c_2$: the lawn
- $c_3$: on the lawn

**Joint Embedding Space**

**Text Encoder**

- $c_1$ (a cat)
- $c_2$ (the lawn)
- $c_3$ (on the lawn)

**Image Encoder:**

- ResNet 101 (He et al., 2015)
Where should function words go?

((A cat) on) (the lawn)  (A cat) (on (the lawn))  ✓

Fact #1: *On* is the head of *on the lawn*.

Fact #2: English is strongly head-initial.
Many other Indo-European languages are head-initial as well.

Fact #3: Under the setting of visual grounding, most abstract words are function words (e.g., prepositions, determiners, complementizers).

Empirical Solution (Head-Initial; HI):
Discourage abstract words from combining to the front.

\[
\begin{align*}
c &= [c_{left}; c_{right}] \\
reward(c) &= concrete(c; i) \\
\end{align*}
\]

\[
\begin{align*}
\text{reward}(c) &= \frac{concrete(c; i)}{\lambda \cdot abstract(c_{right}; i) + 1}, \\
\lambda &> 0
\end{align*}
\]
Training and Evaluation

Each model takes 5 runs, with different random seeds

$F_1$: Average agreement with Benepar (Kitaev and Klein, 2018)

Std: Standard deviation of $F_1$ scores

Self-$F_1$: Average agreement across the $\binom{5}{2}$ pairs of models

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Language</th>
<th># Image (train/dev/test)</th>
<th># Caption (train/dev/test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCOCO (Lin et al., 2014)</td>
<td>EN</td>
<td>80K/1K/1K</td>
<td>400K/5K/5K</td>
</tr>
<tr>
<td>Multi30K (Elliott et al., 2016)</td>
<td>EN, DE, FR</td>
<td>28K/1K/1K</td>
<td>28K/1K/1K</td>
</tr>
</tbody>
</table>
## Unsupervised/Naturally Supervised Parsing

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. F1</th>
<th>Std</th>
<th>Self-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>27.1</td>
<td>32.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Left</td>
<td>23.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>22.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRPN</td>
<td>52.5</td>
<td>60.3</td>
<td>2.6</td>
</tr>
<tr>
<td>ON-LSTM</td>
<td>69.3</td>
<td></td>
<td>3.3</td>
</tr>
<tr>
<td>VG-NSL</td>
<td>50.4</td>
<td>87.1</td>
<td>0.3</td>
</tr>
<tr>
<td>VG-NSL+HI</td>
<td>53.5</td>
<td>90.2</td>
<td>0.2</td>
</tr>
<tr>
<td>VG-NSL+HI+FastText</td>
<td>54.4</td>
<td>89.8</td>
<td>0.4</td>
</tr>
</tbody>
</table>

PRPN: Shen et al. (2018)
ON-LSTM: Shen et al. (2019)
FastText: Joulin et al. (2016)
Data Efficiency

F1 (agreement with Benepar)

Self-F1

Benepar: Kitaev and Klein (2018)
Performance on Multiple Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>PRPN</th>
<th>ON-LSTM</th>
<th>VG-NSL (ours)</th>
<th>VG-NSL+HI (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>30.8</td>
<td>33.5</td>
<td>38.7</td>
<td>38.7</td>
</tr>
<tr>
<td>French</td>
<td>27.5</td>
<td>34.9</td>
<td>34.3</td>
<td>38.1</td>
</tr>
<tr>
<td>German</td>
<td>31.5</td>
<td>27.7</td>
<td>36.3</td>
<td>38.3</td>
</tr>
</tbody>
</table>

PRPN: Shen et al. (2018)
ON-LSTM: Shen et al. (2019)
Conclusions

- VG-NSL: Simple yet effective model for naturally supervised parsing

Future Work:
- Extension to abstract domains
- Advanced perception modules for relations and quantities
- Advanced parsing modules for non-continuous constituents
Thank you!
Why not a stronger parser?
How is Benepar on MSCOCO?

• Manually labeled 50 random captions in MSCOCO following the PTB principles (Bies et al., 1995)

• $F_1 = 95.65\%$

• More details: appendix D and the project page
Agreement with Linguistic Concreteness

Turney et al. (2011): Semi-supervised concreteness estimation
Brysbaert et al. (2014): Manually labeled concreteness scores

<table>
<thead>
<tr>
<th>Pearson $\rho$</th>
<th>Turney et al. (2011)</th>
<th>Brysbaert et al. (2014)</th>
<th>VG-NSL+HI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turney et al. (2011)</td>
<td>1.00</td>
<td>0.84</td>
<td>0.72</td>
</tr>
<tr>
<td>Brysbaert et al. (2014)</td>
<td>--</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td>VG-NSL+HI (ours)</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Agreement with Linguistic Concreteness
Extension to Abstract Domains

• Dependency based word embeddings (Levy and Goldberg, 2014)
  • Syntactically similar words are near in the embedding space
  • Estimate the embeddings for unknown words based on known ones.

• Unsupervised dependency word embeddings?
Performance on Multiple Languages

Fact: German is less strongly head-initial than English (Baker, 2001).
More Recent Work on Unsupervised Parsing

• Deep Inside-Outside Recursive Autoencoders (Drozdov et al., 2019)
• Unsupervised Recurrent Neural Network Grammars (Kim et al., 2019)
• Compound Probabilistic Context-Free Grammars for Grammar Induction (Kim et al., 2019)
• An Imitation Learning Approach to Unsupervised Parsing (Li et al., 2019)