Logistics

Project proposals — Monday, March 15th

Assignment 5 … on GANs and Graph Neural Networks later ;)
Semantic Embeddings

Image Embedding

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) : \mathbb{R}^D \rightarrow \mathbb{R}^d \]

Label Embedding

\[ \Psi_L(word_i) = u_i : \{1, \ldots, L\} \rightarrow \mathbb{R}^d \]
**word2vec**: Unsupervised Word Embedding

**Distributional Semantics Hypothesis**: words that are used and occur in the same context tend to have similar meaning.

\[
\Psi_L(word_i) = u_i : \{1, ..., L\} \rightarrow \mathbb{R}^d
\]
word2vec: Unsupervised Word Embedding

**Distributional Semantics Hypothesis:** words that are used and occur in the same context tend to have similar meaning

\[ \Psi_L(\text{word}_i) = u_i : \{1, ..., L\} \rightarrow \mathbb{R}^d \]

*Example:* Horse breeds are loosely divided into three categories

**Skip-gram Model:** unsupervised semantic representation for words

[ Mikolov, Sutskever, Chen, Corrado, Dean, NIPS’13 ]
DeViSE: A Deep Visual-Semantic Embedding Model

\[
\text{loss}(\text{image}, \text{label}) = \sum_{j \neq \text{label}} \max[0, \text{margin} - \tilde{t}_{\text{label}} M \tilde{v}(\text{image}) + \tilde{t}_j M \tilde{v}(\text{image})]
\]
DeViSE: A Deep Visual-Semantic Embedding Model

Supervised Results

<table>
<thead>
<tr>
<th>Model type</th>
<th>dim</th>
<th>Flat hit@k (%)</th>
<th>Hierarchical precision@k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1  2  5 10</td>
<td>2   5   10  20</td>
</tr>
<tr>
<td>Softmax baseline</td>
<td>N/A</td>
<td>55.6 67.4 78.5 85.0</td>
<td>0.452 0.342 0.313 0.319</td>
</tr>
<tr>
<td>DeViSE</td>
<td>500</td>
<td>53.2 65.2 76.7 83.3</td>
<td>0.447 0.352 0.331 0.341</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>54.9 66.9 78.4 85.0</td>
<td>0.454 0.351 0.325 0.331</td>
</tr>
<tr>
<td>Random embeddings</td>
<td>500</td>
<td>52.4 63.9 74.8 80.6</td>
<td>0.428 0.315 0.271 0.248</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>50.5 62.2 74.2 81.5</td>
<td>0.418 0.318 0.290 0.292</td>
</tr>
<tr>
<td>Chance</td>
<td>N/A</td>
<td>0.1 0.2 0.5 1.0</td>
<td>0.007 0.013 0.022 0.042</td>
</tr>
</tbody>
</table>

Zero-shot Results

<table>
<thead>
<tr>
<th>Model</th>
<th>200 labels</th>
<th>1000 labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeViSE</td>
<td>31.8%</td>
<td>9.0%</td>
</tr>
<tr>
<td>Mensink et al. 2012 [12]</td>
<td>35.7%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Rohrbach et al. 2011 [17]</td>
<td>34.8%</td>
<td>-</td>
</tr>
</tbody>
</table>
Semantic Embeddings

Image Embedding

\[ \Psi(I_i) = W \cdot \text{CNN}(I_i; \Theta): \mathbb{R}^D \rightarrow \mathbb{R}^d \]

Label Embedding

\[ \Psi_L(\text{word}_i) = u_i : \{1, \ldots, L\} \rightarrow \mathbb{R}^d \]
**word2vec:** Unsupervised Word Embedding

**Distributional Semantics Hypothesis:** words that are used and occur in the same context tend to have similar meaning

\[
\Psi_L(word_i) = u_i : \{1, \ldots, L\} \rightarrow \mathbb{R}^d
\]

\[L = 310,000\]

e.g., Horse breeds are loosely divided into three categories

**Skip-gram Model:** unsupervised semantic representation for words (trained from 7 billion word linguistic corpus)

[ Fu et al., 2016 ]
**Semi-supervised Vocabulary Informed Learning**

[Fu et al., 2016]

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) : \mathbb{R}^D \to \mathbb{R}^d \]

**Label Embedding**

\[ \Psi_L(word_i) = u_i : \{1, \ldots, L\} \to \mathbb{R}^d \]

\[ L = 310,000 \]
Semi-supervised **Vocabulary Informed** Learning  

**Image Embedding**
\[
\Psi(I_i) = \mathbf{W} \cdot \text{CNN}(I_i; \Theta): \mathbb{R}^D \rightarrow \mathbb{R}^d
\]

**Label Embedding**
\[
\Psi_L(\text{word}_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d
\]

\[L = 310,000\]

**Similarity in Embedding Space**
\[
D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2
\]
Semi-supervised **Vocabulary Informed Learning** [Fu et al., 2016]

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta): \mathbb{R}^D \rightarrow \mathbb{R}^d \]

**Label Embedding**

\[ \Psi(L(word_i)) = u_i : \{1, \ldots, L\} \rightarrow \mathbb{R}^d \]

\[ L = 310,000 \]

**Similarity in Embedding Space**

\[ D(u, u') = ||u - u'||_2^2 \]

**Objective Function:**

\[ \min_W \sum_i^N \mathcal{L}_C(W, V, I_i, y_i) + \mathcal{L}_R(W, V, I_i, y_i) + \mu ||V||_F^2 \]
Semi-supervised **Vocabulary Informed** Learning [Fu et al., 2016]

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta): \mathbb{R}^D \rightarrow \mathbb{R}^d \]

**Label Embedding**

\[ \Psi_L(\text{word}_i) = u_i : \{1, \ldots, L\} \rightarrow \mathbb{R}^d \]

\[ L = 310,000 \]

**Similarity in Embedding Space**

\[ D(u, u') = ||u - u'||_2^2 \]

**Objective Function:**

\[ \min_W \sum_i^N \mathcal{L}_C(W, V, I_i, y_i) + \mathcal{L}_R(W, V, I_i, y_i) + \mu ||V||_F^2 \]

\[ \mathcal{L}_C(W, U, x_i, y_i) = \sum [1 + D(Wx_i, u_{y_i}) - D(Wx_i, u_c)] \]
Semi-supervised **Vocabulary Informed** Learning

**Image Embedding**
\[
\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \Theta) : \mathbb{R}^D \rightarrow \mathbb{R}^d
\]

**Label Embedding**
\[
\Psi_L(word_i) = \mathbf{u}_i : \{1, \ldots, L\} \rightarrow \mathbb{R}^d
\]

\( L = 310,000 \)

**Similarity in Embedding Space**
\[
D(u, u') = ||u - u'||^2_2
\]

**Objective Function:**
\[
\min_{\mathbf{W}} \sum_i^N \mathcal{L}_C(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mathcal{L}_R(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mu ||\mathbf{V}||^2_F
\]

\[
\mathcal{L}_C(\mathbf{W}, \mathbf{U}, x_i, y_i) = \sum [1 + D(\mathbf{W}x_i, \mathbf{u}_{y_i}) - D(\mathbf{W}x_i, \mathbf{u}_c)]
\]

[Fu et al., 2016]
Vocabulary Informed Recognition

[ Fu et al., 2016 ]
Vocabulary Informed Recognition

$\mathbf{v}_1$ $\rightarrow$ $\mathbf{v}_2$ $\rightarrow$ $\mathbf{v}_3$

$f(\text{Image})$

unicycle

tricycle

[ Fu et al., 2016 ]
Vocabulary Informed Recognition

\[ f(\text{Image}) \]

- Unicycle
- Tricycle

[ Fu et al., 2016 ]
Vocabulary Informed Recognition

[ Fu et al., 2016 ]
Vocabulary Informed Recognition

[ Fu et al., 2016 ]
Experiments: Datasets

Animals with Attributes

Otter

Polar Bear

... 

Auxiliary: 40 Animal Classes (annotated)
Target: 10 Animal Classes (NO annotation)

[ Lampert, Nickisch, Harmeling CVPR'09 ]

ImageNet

Auxiliary: 1,000 General Classes (annotated)
Target: 360 General Classes (NO annotation)

[ Deng et al., CVPR'09 ]
## Experiments: Settings

<table>
<thead>
<tr>
<th>AwA/ImageNet</th>
<th>No. Testing Classes</th>
<th>No. Testing Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auxiliary</td>
<td>Target</td>
</tr>
<tr>
<td>SUPERVISED</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ZERO-SHOT</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>OPEN-SET</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The tasks are only separated in **evaluation**; we train **one unified model** for all the settings.

[ Fu et al., 2016 ]
Experiments: Settings

Training

Otter
Polar Bear
Donkey
Poney
Panther

310,000

Testing

Supervised

[ Fu et al., 2016 ]
 convenional object detectors for all these would require millions of example images for each class to be learned. Currently labeled training data, typically hundreds or thousands of images, are required for achieving good classification accuracy. However, in order to achieve good performance, these systems require a lot of manual labor to label training data. Support vector machines, for example, use a combination of powerful low-level features, like SIFT or HOG, and high-level concepts. The system is able to detect objects from a list of high-level attributes. The attributes are extracted from an image, and they allow the transfer of knowledge between object categories: after learning the visual appearance of attributes from any classes with training examples, they are also capable of recognizing these attributes in classes that have not been previously seen.

In this paper, we tackle the problem by introducing a system that is able to detect completely unseen classes when provided with a high-level description. The system is based on a human-specified high-level description of the target objects instead of training images. The description consists of arbitrary semantic attributes, like shape, color, or even geographic information. Because such properties transcend the specific learning task at hand, they can be transferred to previously unseen object classes.

We study the problem of object classification when training images of the target classes are available. This setup has hardly been studied in computer vision research, but it is the rule rather than the exception, because the world contains tens of thousands of different object classes and for only a very few of them image datasets unrelated to the current task. Afterwards, new classes can be detected based on their attribute representation, without the need for a new training phase. In order to evaluate our method and to facilitate research in this area, we have assembled a new large-scale dataset.

The training and test classes are disjoint, i.e., no training examples of them would be as tedious as for all 30,000 animal classes. But, instead of creating a separate training classifier for each of them, it is indeed possible to build a system that does not require any training examples of the target classes. By using an attribute layer it is possible to build a system that performs object detection on their attribute representation, without the need for a new training phase. In order to evaluate our method and to facilitate research in this area, we have assembled a new large-scale dataset.

It has been estimated that humans distinguish between over 30,000 animal classes. But, instead of creating a separate training classifier for each of them, it is indeed possible to build a system that does not require any training examples of the target classes. By using an attribute layer it is indeed possible to build a system that performs object detection on their attribute representation, without the need for a new training phase. In order to evaluate our method and to facilitate research in this area, we have assembled a new large-scale dataset.

Figure 1. A description by attributes of a polar bear and a panther.

**Experiments: Settings**

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Otter</td>
<td>Supervised</td>
</tr>
<tr>
<td>Polar Bear</td>
<td>Zero-shot</td>
</tr>
<tr>
<td>Donkey</td>
<td></td>
</tr>
<tr>
<td>Poney</td>
<td></td>
</tr>
<tr>
<td>Panther</td>
<td></td>
</tr>
</tbody>
</table>

310,000 training images of the target classes.
It has been estimated that humans distinguish between over 30,000 animal classes. But, instead of creating a separate training classifier for each of them would be as tedious as for all the world contains tens of thousands of different object classes and for only a very few of them image datasets unrelated to the current task. Afterwards, new classes can be detected based on their attribute representation, without the need for a new pre-learned, transcendent models. Because such properties serve as an intermediate layer in a classifier cascade and abstraction well from examples, they are also capable of reaching for years to come. Therefore, numerous techniques have been developed, some of which we will discuss in Section 3.

We study the problem of object classification when appearance of attributes from any classes with training examples, of over 30,000 animal classes. But, instead of creating a separate training classifier for each of them would be as tedious as for all the world contains tens of thousands of different object classes and for only a very few of them image datasets unrelated to the current task. Afterwards, new classes can be detected based on their attribute representation, without the need for a new pre-learned, transcendent models. Because such properties serve as an intermediate layer in a classifier cascade and abstraction well from examples, they are also capable of reaching for years to come. Therefore, numerous techniques have been developed, some of which we will discuss in Section 3.

In this paper, we tackle the problem by introducing an attribute layer it is indeed possible to build a training phase. In order to evaluate our method and to facilitate research in this area, we have assembled a new large-scale dataset, which consists of arbitrary semantic attributes, like shape, color, and texture.

### Experiments: Settings

**Training**
- Otter
- Polar Bear
- Donkey
- Poney
- Panther

**Testing**

**Open-set**

- Polar Bear
- Donkey
- Panther
- Other animals

[ Fu et al., 2016 ]
### Experiments: Settings

<table>
<thead>
<tr>
<th>AwA/ImageNet</th>
<th>No. Testing Classes</th>
<th>No. Testing Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auxiliary</td>
<td>Target</td>
</tr>
<tr>
<td>SUPERVISED</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ZERO-SHOT</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>OPEN-SET</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The tasks are only separated in evaluation; we train **one unified model** for all the settings.
<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS-Voc: full instances</td>
<td>CNNOverFeat</td>
<td>78.3</td>
</tr>
<tr>
<td>Akata &lt;i&gt;et al.&lt;/i&gt; CVPR 2015</td>
<td>CNNGoogLeNet</td>
<td>73.9</td>
</tr>
<tr>
<td>TMV-BLP (Fu &lt;i&gt;et al.&lt;/i&gt; ECCV 2014)</td>
<td>CNNOverFeat</td>
<td>69.9</td>
</tr>
<tr>
<td>AMP (SR+SE) (Fu &lt;i&gt;et al.&lt;/i&gt; CVPR 2015)</td>
<td>CNNOverFeat</td>
<td>66.0</td>
</tr>
<tr>
<td>DAP (Lampert &lt;i&gt;et al.&lt;/i&gt; TPAMI 2013)</td>
<td>CNNVGG19</td>
<td>57.5</td>
</tr>
<tr>
<td>PST (Rohrbach &lt;i&gt;et al.&lt;/i&gt; NIPS 2013)</td>
<td>CNNOverFeat</td>
<td>53.2</td>
</tr>
<tr>
<td>DS (Rohrbach &lt;i&gt;et al.&lt;/i&gt; CVPR 2010)</td>
<td>CNNOverFeat</td>
<td>52.7</td>
</tr>
<tr>
<td>IAP (Lampert &lt;i&gt;et al.&lt;/i&gt; TPAMI 2013)</td>
<td>CNNOverFeat</td>
<td>44.5</td>
</tr>
<tr>
<td>HEX (Deng &lt;i&gt;et al.&lt;/i&gt; ECCV 2014)</td>
<td>CNNDECAF</td>
<td>44.2</td>
</tr>
</tbody>
</table>

[ Fu et al., 2016 ]
<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS-Voc: full instances</td>
<td>CNNOverFeat</td>
<td>78.3</td>
</tr>
<tr>
<td>800 instances (20 inst*40 class);</td>
<td>CNNOverFeat</td>
<td>74.4</td>
</tr>
<tr>
<td>Akata et al. CVPR 2015</td>
<td>CNNGoogleNet</td>
<td>73.9</td>
</tr>
<tr>
<td>TMV-Blp (Fu et al. ECCV 2014)</td>
<td>CNNOverFeat</td>
<td>69.9</td>
</tr>
<tr>
<td>AMP (SR+SE) (Fu et al. CVPR 2015)</td>
<td>CNNOverFeat</td>
<td>66.0</td>
</tr>
<tr>
<td>DAP (Lampert et al. TPAMI 2013)</td>
<td>CNNVGG19</td>
<td>57.5</td>
</tr>
<tr>
<td>PST (Rohrbach et al. NIPS 2013)</td>
<td>CNNOverFeat</td>
<td>53.2</td>
</tr>
<tr>
<td>DS (Rohrbach et al. CVPR 2010)</td>
<td>CNNOverFeat</td>
<td>52.7</td>
</tr>
<tr>
<td>IAP (Lampert et al. TPAMI 2013)</td>
<td>CNNOverFeat</td>
<td>44.5</td>
</tr>
<tr>
<td>HEX (Deng et al. ECCV 2014)</td>
<td>CNNDECAF</td>
<td>44.2</td>
</tr>
</tbody>
</table>

3.3% of training data

Zero-shot Results with AWA

[ Fu et al., 2016 ]
Zero-shot Results

Results with AWA

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS-Voc: full instances</td>
<td>CNNOverFeat</td>
<td>78.3</td>
</tr>
<tr>
<td>800 instances (20 inst*40 class)</td>
<td>CNNOverFeat</td>
<td>74.4</td>
</tr>
<tr>
<td>200 instances (5 inst*40 class)</td>
<td>CNNOverFeat</td>
<td>68.9</td>
</tr>
<tr>
<td>Akata et al. CVPR 2015</td>
<td>CNNGoogLeNet</td>
<td>73.9</td>
</tr>
<tr>
<td>TMV-BLP (Fu et al. ECCV 2014)</td>
<td>CNNOverFeat</td>
<td>69.9</td>
</tr>
<tr>
<td>AMP (SR+SE) (Fu et al. CVPR 2015)</td>
<td>CNNOverFeat</td>
<td>66.0</td>
</tr>
<tr>
<td>DAP (Lampert et al. TPAMI 2013)</td>
<td>CNNVGG19</td>
<td>57.5</td>
</tr>
<tr>
<td>PST (Rohrbach et al. NIPS 2013)</td>
<td>CNNOverFeat</td>
<td>53.2</td>
</tr>
<tr>
<td>DS (Rohrbach et al. CVPR 2010)</td>
<td>CNNOverFeat</td>
<td>52.7</td>
</tr>
<tr>
<td>IAP (Lampert et al. TPAMI 2013)</td>
<td>CNNOverFeat</td>
<td>44.5</td>
</tr>
<tr>
<td>HEX (Deng et al. ECCV 2014)</td>
<td>CNNDECAF</td>
<td>44.2</td>
</tr>
</tbody>
</table>

0.82% of training data

[ Fu et al., 2016 ]
Weakly-supervised **Visual Grounding** of Phrases

Given **image-sentence pairs** learn how to **localize** arbitrary language phrase or sentence in new images

[ Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV'14 ]

[ Xiao et al., 2017 ]
Weakly-supervised **Visual Grounding** of Phrases

Given *image-sentence pairs* learn how to **localize** arbitrary language phrase or sentence in new images.

[ Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV’14 ]

[ Xiao et al., 2017 ]
Weakly-supervised **Visual Grounding** of Phrases

Given **image-sentence pairs** learn how to **localize** arbitrary language phrase or sentence in new images

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

[ Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV’14 ]

[ Xiao et al., 2017 ]
Weakly-supervised Visual Grounding of Phrases

Given **image-sentence pairs** learn how to **localize** arbitrary language phrase or sentence in new images

[ Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV'14 ]

[ Xiao et al., 2017 ]
Weakly-supervised Visual Grounding of Phrases [Xiao et al., 2017]

\[ \Psi_L(\text{phrase}_i) = u_i \]
Weakly-supervised Visual Grounding of Phrases \cite{Xiao2017}

\[ \Psi_L(phrase_i) = u_i \]

Language Encoder

\[ \begin{align*}
\text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} \\
a & \quad \text{man} & \quad \text{that} & \quad \text{is} & \quad \text{cutting} & \quad \text{sandwich}
\end{align*} \]
Weakly-supervised **Visual Grounding** of Phrases

\[
\Psi_L(\text{phrase}_i) = \mathbf{u}_i
\]
Weakly-supervised **Visual Grounding** of Phrases

\[
\Psi_L(\text{phrase}_i) = u_i
\]

- **Label Embedding**
- **Language Encoder**

[ Xiao et al., 2017 ]
Weakly-supervised **Visual Grounding** of Phrases [Xiao et al., 2017]

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) \]

**Label Embedding**

\[ \Psi_L(\text{phrase}_i) = u_i \]

**Language Encoder**

- **a**
- **man**
- **that**
- **is**
- **cutting**
- **sandwich**

**Feature Extractor**

**Purple Bus**

**A Table**

**A Man**
Weakly-supervised Visual Grounding of Phrases [Xiao et al., 2017]

**Image Embedding**

$$\Psi(I_i) = W \cdot CNN(I_i; \Theta)$$

**Label Embedding**

$$\Psi_L(\text{phrase}_i) = u_i$$

**Language Encoder**

- a
- man
- that
- is
- cutting
- sandwich

**Feature Extractor**

16x16

**Latent Attention**

- purple bus
- a table
- a man

**Weakly-supervised Visual Grounding of Phrases**
Weakly-supervised **Visual Grounding** of Phrases  

[ Xiao et al., 2017 ]

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) \]

**Label Embedding**

\[ \Psi_L(\text{phrase}_i) = u_i \]

**Language Encoder**

\[ \text{LSTM} \rightarrow \text{LSTM} \rightarrow \text{LSTM} \rightarrow \text{LSTM} \rightarrow \text{LSTM} \rightarrow \text{LSTM} \]

\[ a \rightarrow \text{man} \rightarrow \text{that} \rightarrow \text{is} \rightarrow \text{cutting} \rightarrow \text{sandwich} \]

**Feature Extractor**

**Latent Attention**  

16x16

```
ψ_L(phrase_i) = u_i
```

```
θ

ϕ(I_i) = W \cdot CNN(I_i; Θ)
```

```
[dog: 0.95]  [frisbee: 0.83]  [outdoor: 0.82]  [grass: 0.81]  [leap: 0.45]
```

**Attributes**

[bananas: 1]  [market: 0.99]  [table: 0.51]  [people: 0.43]
Weakly-supervised **Visual Grounding** of Phrases

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) \]

**Label Embedding**

\[ \Psi_L(phrase_i) = u_i \]

**Similarity in Embedding Space**

\[ D(u, u') = ||u - u'||^2 \]

**Objective Function:**

Combination of previous discriminative similarity and **linguistic regularization**

[ Xiao et al., 2017 ]
Weakly-supervised **Visual Grounding** of Phrases  

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) \]

\[ \Psi_L(\text{phrase}_i) = u_i \]

**Similarity in Embedding Space**

\[ D(u, u') = \|u - u'\|_2^2 \]

**Objective Function:**

Combination of previous discriminative similarity and **linguistic regularization**
Weakly-supervised **Visual Grounding** of Phrases

For **noun phrases**:
- **siblings** should have **disjoint** masks

**Image Embedding**

\[
\Psi(I_i) = W \cdot \text{CNN}(I_i; \Theta)
\]

**Label Embedding**

\[
\Psi_L(\text{phrase}_i) = u_i
\]

**Similarity in Embedding Space**

\[
D(u, u') = ||u - u'||^2_2
\]

**Objective Function:**

Combination of previous discriminative similarity and **linguistic regularization**
Weakly-supervised Visual Grounding of Phrases

For noun phrases:
• siblings should have disjoint masks

Image Embedding

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) \]

Label Embedding

\[ \Psi_L(\text{phrase}_i) = u_i \]

Similarity in Embedding Space

\[ D(u, u') = ||u - u'||_2^2 \]

Objective Function:

Combination of previous discriminative similarity and linguistic regularization
Weakly-supervised **Visual Grounding** of Phrases  

For **noun phrases**:

- siblings should have **disjoint** masks
- parents should be **union of children** masks

**Image Embedding**

\[ \Psi(I_i) = W \cdot CNN(I_i; \Theta) \]

**Label Embedding**

\[ \Psi_L(\text{phrase}_i) = u_i \]

**Similarity in Embedding Space**

\[ D(u, u') = \|u - u'\|^2 \]

**Objective Function:**

Combination of previous discriminative similarity and **linguistic regularization**
For **noun phrases**:

- **siblings** should have **disjoint** masks.
- **parents** should be the **union** of children masks.

**Image Embedding**

$$\Psi(I_i) = W \cdot CNN(I_i; \Theta)$$

**Label Embedding**

$$\Psi_L(\text{phrase}_i) = u_i$$

**Similarity in Embedding Space**

$$D(u, u') = ||u - u'||^2_2$$

**Objective Function:**

Combination of previous discriminative similarity and **linguistic regularization**

[Xiao et al., 2017]
Qualitative Results

Input:

guy in green t-shirt holding skateboard

[Xiao et al., 2017]
Qualitative Results

Input:
guy in green t-shirt holding skateboard

NO linguistic constraints
Qualitative Results

Input:
guy in green t-shirt holding skateboard

NO linguistic constraints

[ Xiao et al., 2017 ]
Qualitative Results

Input: guy in green t-shirt holding skateboard

NO linguistic constraints

Our Model

[Xiao et al., 2017]
Qualitative Results

Input:

a person driving a boat


NO linguistic constraints

Our Model

[Xiao et al., 2017]
Qualitative **Results**

**Input:**

a child wearing black protective helmet

[ Xiao et al., 2017 ]

---

**NO** linguistic constraints

Our Model
## Quantitative Results

Segmentation performance on COCO dataset

[ Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV’14 ]

<table>
<thead>
<tr>
<th></th>
<th>IoU@0.3</th>
<th>IoU@0.4</th>
<th>IoU@0.5</th>
<th>Avg mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-strcutred</td>
<td>0.302</td>
<td>0.199</td>
<td>0.110</td>
<td>0.203</td>
</tr>
<tr>
<td>Parent-Child</td>
<td>0.327</td>
<td>0.213</td>
<td>0.118</td>
<td>0.219</td>
</tr>
<tr>
<td>Sibling</td>
<td>0.316</td>
<td>0.203</td>
<td>0.114</td>
<td>0.211</td>
</tr>
<tr>
<td>Ours</td>
<td>0.347</td>
<td>0.246</td>
<td>0.159</td>
<td>0.251</td>
</tr>
</tbody>
</table>
Order Embeddings

[Vendrov et al., 2016]
Multimodal Representation Types

**Joint** representations:

- Simplest version: modality concatenation (early fusion)
- Can be learned supervised or unsupervised

**Coordinated** representations:

- Similarity-based methods (e.g., cosine distance)
- Structure constraints (e.g., orthogonality, sparseness)
- CCA (unsupervised), joint embeddings (supervised)

*slide from Louis-Philippe Morency*
Final Words …

**Joint** representations
- Project modalities to the same space
- Use when all the modalities are present during test time
- Suitable for multi-model fusion

**Coordinated** representations
- Project modalities to their own coordinated spaces
- Use when only one of the modalities is present during test-time
- Suitable for multimodal translation
- Good for multimodal retrieval

*slide from Louis-Philippe Morency*