DEEP SEMANTIC-VISUAL EMBEDDING WITH LOCALIZATION

Wednesday 27th June, 2018

Martin Engilberge, Louis Chevallier, Patrick Pérez, Matthieu Cord
Tasks

**Visual Grounding of phrases:**
Localyze any textual query into a given image.

**Cross-modal retrieval:**

Query: A cat on a sofa
2D Semantic visual space example:

- Distance in the space has a semantic interpretation.
- Retrieval is done by finding nearest neighbors.
Approach

- Learning image and text joint embedding space.
- Visual grounding relying on the spatial-textual information modeling.
- Cross-modal retrieval leveraging the semantic space and the visual and textual alignment.

**Diagram:**
- **Image:** CNN → adaptation + pooling → projection
- **Text:** tokenisation + embedding → RNN encoding → joint space
Semantic Embedding Model

Visual pipeline:

• ResNet-152 pretrained.
• Weldon spatial pooling.
• Affine projection
• normalization.

Textual pipeline:

• Pretrained word embedding.
• Simple Recurrent Unit (SRU).
• Normalization.

\[ \theta_0, \theta_1, \theta_2 \] and \( \phi \) are the trained parameters

\( \langle X, V \rangle \) cosine sim.
Semantic Embedding Model

Visual pipeline:
- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- Normalization.

Textual pipeline:
- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.

\[ \theta_0, \theta_1, \theta_2 \] and \( \phi \) are the trained parameters.
Pooling mechanisms

Weldon spatial pooling:

• Instead of global average/max pooling.
• Aggregate the min and max of each map.
• Produce activation map with finer localization.
Semantic Embedding Model

Visual pipeline:

• ResNet-152 pretrained.
• Weldon spatial pooling.
• Affine projection
• normalization.

Textual pipeline:

• Pretrained word embedding.
• Simple Recurrent Unit (SRU).
• Normalization.

θ₀: 2 and φ are the trained parameters.
Simple Recurrent Unit: SRU

Recurrent neural network:

- Fixed sized representation for variable length sequence.
- Able to capture long-term dependency between words.
Deep semantic-visual embedding with localization

Semantic Embedding Model

Visual pipeline:

• ResNet-152 pretrained.
• Weldon spatial pooling.
• Affine projection.
• normalization.

Textual pipeline:

• Pretrained word embedding.
• Simple Recurrent Unit (SRU).
• Normalization.

θ₀, 2 \text{ and } φ \text{ are the trained parameters}
Semantic Embedding Model

Visual pipeline:
- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:
- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.

\[ \theta_0, \theta_1, \theta_2 \]

(a, man, in, ski, gear, skiing, on, snow)

\[ \phi \]

\[ \theta_0, \theta_2 \text{ and } \phi \text{ are the trained parameters} \]
Dataset

• MS-CoCo 2014:
  • 110K training images
  • 5 captions per image
  • 2*5k images for validation and test

Dining room table set for a casual meal, with flowers.
Learning strategy: triplet loss

A variant of the standard margin based loss:

- Triplet \((y, z, z')\)
- Anchor: \(y\) (E.g. image representation)
- Positive: \(z\) (E.g. associated caption representation)
- Negative: \(z'\) (E.g. contrastive caption representation)
- Margin parameter \(\alpha\)

\[
\text{loss}(y, z, z') = \max\{ 0, \alpha - \langle y, z \rangle + \langle y, z' \rangle \}
\]
Learning strategy: triplet loss

\[
\text{loss}(y, z, z') = \max\{0, \alpha + d(y, z) - d(y, z')\}
\]
Learning strategy: triplet loss

Hard negative margin based loss:

Loss for a batch $\mathcal{B} = \{(I_n, S_n)\}_{n \in \mathcal{B}}$ of image sentence pairs:

$$
\mathcal{L}(\Theta; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \left( \max_{m \in C_n \cap \mathcal{B}} \text{loss (}x_n, v_n, v_m) + \max_{m \in D_n \cap \mathcal{B}} \text{loss (}v_n, x_n, x_m) \right)
$$

With:

- $C_n$ (resp. $D_n$) set of indices of caption (resp. image) unrelated to $n$-th element.
Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

$$\mathcal{L}(\Theta; B) = \frac{1}{|B|} \sum_{n \in B} \left( \max_{m \in C_n \cap B} \text{loss} (x_n, v_n, v_m) + \max_{m \in D_n \cap B} \text{loss} (v_n, x_n, x_m) \right)$$
Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

\[
\mathcal{L}(\Theta; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \left( \max_{m \in \mathcal{C}_n \cap \mathcal{B}} \text{loss} (\mathbf{x}_n, \mathbf{v}_n, \mathbf{v}_m) + \max_{m \in \mathcal{D}_n \cap \mathcal{B}} \text{loss} (\mathbf{v}_n, \mathbf{x}_n, \mathbf{x}_m) \right)
\]
From training to testing

Training finished:
- Visual-semantic space constructed.
- Parameters of the model are fixed.
- Time for testing.
## Qualitative evaluation: cross-modal retrieval

<table>
<thead>
<tr>
<th>Query</th>
<th>Closest elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>A plane in a cloudy sky</td>
<td><img src="" alt="Image 1" /> <img src="" alt="Image 2" /> <img src="" alt="Image 3" /> <img src="" alt="Image 4" /> <img src="" alt="Image 5" /></td>
</tr>
<tr>
<td>A dog playing with a frisbee</td>
<td><img src="" alt="Image 6" /> <img src="" alt="Image 7" /> <img src="" alt="Image 8" /> <img src="" alt="Image 9" /> <img src="" alt="Image 10" /></td>
</tr>
<tr>
<td>A herd of sheep standing on top of snow covered field.</td>
<td><img src="" alt="Image 11" /> <img src="" alt="Image 12" /> <img src="" alt="Image 13" /> <img src="" alt="Image 14" /> <img src="" alt="Image 15" /></td>
</tr>
<tr>
<td>1. A herd of sheep standing on top of snow covered field.</td>
<td><img src="" alt="Image 16" /> <img src="" alt="Image 17" /> <img src="" alt="Image 18" /> <img src="" alt="Image 19" /> <img src="" alt="Image 20" /></td>
</tr>
<tr>
<td>2. There are sheep standing in the grass near a fence.</td>
<td><img src="" alt="Image 21" /> <img src="" alt="Image 22" /> <img src="" alt="Image 23" /> <img src="" alt="Image 24" /> <img src="" alt="Image 25" /></td>
</tr>
<tr>
<td>3. some black and white sheep a fence dirt and grass</td>
<td><img src="" alt="Image 26" /> <img src="" alt="Image 27" /> <img src="" alt="Image 28" /> <img src="" alt="Image 29" /> <img src="" alt="Image 30" /></td>
</tr>
</tbody>
</table>
Cross-modal retrieval: Evaluated on MS-CoCo image/caption pairs.

Cross-modal retrieval results

<table>
<thead>
<tr>
<th></th>
<th>Caption retrieval</th>
<th>Image retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Way Net [5]</td>
<td>R@1: 55.80%</td>
<td>R@1: 39.70%</td>
</tr>
<tr>
<td></td>
<td>R@5: 75.20%</td>
<td>R@5: 63.30%</td>
</tr>
<tr>
<td>VSE++ [6]</td>
<td>R@1: 64.60%</td>
<td>R@1: 52%</td>
</tr>
<tr>
<td></td>
<td>R@5: 95.70%</td>
<td>R@5: 92%</td>
</tr>
<tr>
<td>Ours</td>
<td>R@1: 69.80%</td>
<td>R@1: 55.90%</td>
</tr>
<tr>
<td></td>
<td>R@5: 91.90%</td>
<td>R@5: 86.90%</td>
</tr>
<tr>
<td></td>
<td>R@10: 96.60%</td>
<td>R@10: 94%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Performance boost coming from:

- Architecture choice: SRU and Weldon spatial pooling.
- Efficient learning strategy: hard negative loss.

Ablation study: cross modal retrieval results

<table>
<thead>
<tr>
<th></th>
<th>Caption retrieval</th>
<th>Image retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard Neg + WLD + SRU 4</td>
<td>R@1: 69.80%</td>
<td>R@1: 55.90%</td>
</tr>
<tr>
<td></td>
<td>R@5: 91.90%</td>
<td>R@5: 86.90%</td>
</tr>
<tr>
<td></td>
<td>R@10: 96.60%</td>
<td>R@10: 94%</td>
</tr>
<tr>
<td>Hard Neg + GAP + SRU 4</td>
<td>R@1: 64.50%</td>
<td>R@1: 51.20%</td>
</tr>
<tr>
<td></td>
<td>R@5: 90.20%</td>
<td>R@5: 84.00%</td>
</tr>
<tr>
<td></td>
<td>R@10: 95.50%</td>
<td>R@10: 92.00%</td>
</tr>
<tr>
<td>Hard Neg + WLD + GRU 1</td>
<td>R@1: 63.80%</td>
<td>R@1: 52.20%</td>
</tr>
<tr>
<td></td>
<td>R@5: 90.20%</td>
<td>R@5: 84.90%</td>
</tr>
<tr>
<td></td>
<td>R@10: 96%</td>
<td>R@10: 92.60%</td>
</tr>
<tr>
<td>Classic + WLD + SRU 4</td>
<td>R@1: 49.50%</td>
<td>R@1: 39.60%</td>
</tr>
<tr>
<td></td>
<td>R@5: 81%</td>
<td>R@5: 77.30%</td>
</tr>
<tr>
<td></td>
<td>R@10: 90.10%</td>
<td>R@10: 89.10%</td>
</tr>
</tbody>
</table>
### Evaluation: cross-modal retrieval and limitations

<table>
<thead>
<tr>
<th>Query</th>
<th>Closest elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple wooden spoons are shown on a table top.</td>
<td><img src="image1.png" alt="Image" /> <img src="image2.png" alt="Image" /> <img src="image3.png" alt="Image" /> <img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td>The plane is parked at the gate at the airport terminal.</td>
<td><img src="image5.png" alt="Image" /> <img src="image6.png" alt="Image" /> <img src="image7.png" alt="Image" /> <img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td>1. Two elephants in the eld moving along during the day.</td>
<td><img src="image9.png" alt="Image" /></td>
</tr>
<tr>
<td>2. Two elephants are standing by the trees in the wild.</td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td>3. An elephant and a rhino are grazing in an open wooded area.</td>
<td><img src="image11.png" alt="Image" /></td>
</tr>
<tr>
<td>1. A harbor filled with boats floating on water</td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td>2. A small marina with boats docked there</td>
<td><img src="image13.png" alt="Image" /></td>
</tr>
<tr>
<td>3. A group of boats sitting together with no one around</td>
<td><img src="image14.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Localization

Visual grounding module:

• Weakly supervised, with no additional training.
• Localize a textual query in an image.
• Using the embedding space to select convolutionnal activation maps.

Source image

Text query

two glasses

Visual grounding Heat map
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection.
- Normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.

\( \theta_0, \theta_1 \) and \( \phi \) are the trained parameters.

(a, man, in, ski, gear, skiing, on, snow)

\( \langle X, V \rangle \) cosine sim.
Localization

Generation of heatmap $\mathbf{H}$:

$$G'[i, j, :] = AG[i, j, :], \forall (i, j) \in [1, w] \times [1, h]$$

$$H = \sum_{u \in K(v)} |v[u]| \ast G'[:, :, u]$$

$K(v)$ the set of the indices of its $k$ largest entries.
Qualitative evaluation: localization

Visual grounding examples:

• Generating multiple heat maps with different textual queries.
The pointing game: Localizing phrases corresponding to subregions of the image.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Center&quot; baseline</td>
<td>19.50%</td>
</tr>
<tr>
<td>Linguistic structure [7]</td>
<td>24.40%</td>
</tr>
<tr>
<td>Ours</td>
<td>33.80%</td>
</tr>
</tbody>
</table>
Toward zero-shot localization:

• Emergence of colors understanding:

• Even on artificial images:
Toward zero-shot localization:

- Generalization to unseen elements:
Summary:

- Semantic-visual embedding model.
- Effective on the cross-modal retrieval task.
- Visual grounding of text with no extra supervision.

Thank you!

Paper - Finding beans in burgers: Deep semantic-visual embedding with localization