Align or attend? Toward more efficient and accurate spoken word discovery using speech-to-image retrieval

— ECE 590 SIP presentation

Liming Wang

Nov.16th, 2020
1 Problem Formulation

2 Methods

3 Experiment
Multimodal Word Discovery (MWD)

Figure: From babyblue.com
Multimodal Word Discovery (MWD)

“Which describes which?”

- Given:
  - Spoken caption: \( x = x_1, \ldots, x_T \)
  - Image regions: \( y = y_1, \ldots, y_L \)
- Find: which spoken frames describes which visual region

Maximum likelihood estimation (MLE)

\[
\max_{\theta} p(x, y | \theta) = \max_{\theta} \sum_{A} p(x, y, A | \theta)
\]

\[
A^* = \arg\max_{A} p(A | x, y, \theta),
\]

where \( A_{ti} = 1 \) if word \( t \) and region \( i \) are aligned.
Two-step approach: MWD via speech-to-image retrieval

**Step 1**
Sentence-level matching (speech-to-image retrieval Harwath and Glass (2015)):

\[
\max_{\theta} p(M|x^{(1:N)}, y^{(1:N)}, \theta) = \prod_{n,m: M_{nm}=1} \frac{p(x^{(n)}, y^{(m)}|\theta)}{\sum_{n'} p(x^{(n)}, y^{(n')}|\theta)},
\]

where \( M_{nm} = 1 \) if caption \( n \) and image \( m \) are matched.

**Step 2**
Word-level matching (MWD):

\[
A^* = \arg\max_A p(A|x, y, \theta).
\]
DAVEnet: State-of-the-art MWD system

Origin
First proposed by Harwath et al. (2018)

Assumptions

1. Dominant (soft) alignment assumption:

   \[ p^{DAVEnet}(x, y|\theta) := \exp \left( \sum_{t,i} A_{ti} s(x_t, y_i) \right) \]

2. Common space assumption:

   \[ s(x_t, y_i) = \frac{\phi_a(x_t)^\top \phi_v(y_i)}{\|\phi_a(x_t)\| \|\phi_v(y_i)\|}, \]

   where \( \phi_a(\cdot), \phi_v(\cdot) \) are learned by two DNNs.
Does DAVEnet always learn good word-level representation?

**Analysis: MLE of DAVEnet**

\[
\max_{\phi_a, \phi_v} s(x, y) = \max \left( \text{Tr} \left( \Phi_a A \Phi_v^\top \right) \right),
\]

where maximum is achieved if, for \( \text{svd}(A) = U, \Sigma, V \):

\[
\Phi_a U = \Phi_v V
\]

**Good sentence embedding \( \neq \) good word embedding**

\[
A \text{ independent of } \phi_v, \phi_a \implies \phi_v^*, \phi_a^* \text{ independent of } x_t \text{ and } y_i \implies \text{bad word-level representation}
\]
1 Problem Formulation

2 Methods

3 Experiment
Fix the common space: Attention mechanisms

Intuition
- Need to make $A$ variable of $\phi_a, \phi_v$
- May still fail to learn good word embedding with variable $A$ (e.g., constant $\phi_v(y_i)$)

Cosine attention

$$
\alpha_{ti} = \frac{\exp(s(x_t, y_i))}{\sum_t \exp(s(x_t, y_i))} \\
\beta_{ti} = \frac{\exp(s(x_t, y_i))}{\sum_i \exp(s(x_t, y_i))}
$$

Additive attention

$$
\alpha_{ti} = \frac{\exp(W_i[\phi_a, t; \phi_v, i; 1])}{\sum_t \exp(W[\phi_a, t'; \phi_v, i; 1])} \\
\beta_{ti} = \frac{\exp(W_i[\phi_a, t; \phi_v, i; 1])}{\sum_i \exp(W_j[\phi_a, t; \phi_v, i; 1])}
$$

Self attention

$$
\alpha_{tt'}^{(m)} = \frac{\exp(\Phi_a^{(m)^T} \Phi_a^{(m)})_{tt'}}{\sum_{t''} \exp(\Phi_a^{(m)^T} \Phi_a^{(m)})_{tt''}}
$$
Fix the common space: Change the space

### DNN-HMM-DNN Model by Wang and Hasegawa-Johnson (2020)

- **Additional hidden variables:**
  - \( z = [z_1, \ldots, z_L] \): image concept of each image region
  - \( \phi = [\phi_1, \ldots, \phi_T] \): acoustic unit label of each speech segment

- **Conditional likelihood:**
  \[
  p(y|x, \theta) = \sum_{z,A,\phi} p(z|y)p(A, \phi, x|z, L)
  \]

- Learn to recognize concepts and phones with two **DNNs** \( \psi_a \) and \( \psi_v \)
- Learn to align concepts and phones with an **HMM**
DNN-HMM-DNN as learning a common probabilistic space

Rewrite the conditional likelihood using matrix operations:

\[
\max_{A} p(y|x, A, \theta) = \max_{A_t \in \Delta_L, \forall t} \text{Tr} \left( \Psi_a^T P \Psi_v A \right),
\]

Guarantee

As long as the latent word/concept classifiers are sufficiently accurate, it can be shown that the SMT is a consistent estimator when learning many-to-one relations between spoken words and image regions.

Figure: \( C_{ZW} := (\Psi_v A \Psi_a)_z \)
1 Problem Formulation

2 Methods

3 Experiment
Dataset

- **Flickr8k** (Hodosh et al. (2010)): Split according to Karpathy et al. (2014), 30000 image-caption pairs for training, 1000 images for evaluation

- **SpeechCOCO** (Havard et al. (2017)): 80 image concept classes, 80000 image-caption pairs for training, 1000 images randomly chosen from the MSCOCO 2014 validation set for evaluation

- **Preprocessing**: Filter the most frequent 2000 word types, not including stop words
Features

- **Speech features:**
  - **Retrieval:** Mel filter-bank features with 25ms window and 10ms skip step
  - **MWD:** last layer of the speech encoder averaged over each word segment, compressed to 300-d vectors with PCA

- **Image features:** 2048-d ResNet50 features for the top 10 ROIs proposed by the Faster-RCNN pretrained on Visual Genome and ImageNet
### Implementation details

#### NMT

- **TDNN-based systems**: 5 convolution layers, 1024-d embedding, default settings of the DAVEnet implementation by (Harwath et al. (2018))
- **BiLSTM-based system**: 3 convolution layers, 1000-d embedding
- **Transformer-based system**: 3 self-attention layers, 1024-d embedding, implementation from ESPnet (Watanabe et al. (2018))
- **Loss function**: masked margin softmax loss (Ilharco et al. (2019))

#### SMT

- Softmax distributions with Gaussian kernels for encoders, 400 latent word types, 80 latent image concepts for SpeechCOCO; 600 latent image concepts for Flickr
## Results: Speech-to-image retrieval

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>S2I @1</th>
<th>@5</th>
<th>@10</th>
<th>I2S @1</th>
<th>@5</th>
<th>@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAVEnet MISA</td>
<td>COCO</td>
<td>12</td>
<td>38</td>
<td>57</td>
<td>12</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>DAVEnet (phones)</td>
<td>COCO</td>
<td>32</td>
<td>66</td>
<td>79</td>
<td>32</td>
<td>66</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>Flickr</td>
<td>17</td>
<td>42</td>
<td>55</td>
<td>18</td>
<td>39</td>
<td>51</td>
</tr>
<tr>
<td>Cosine+DAVEnet</td>
<td>COCO</td>
<td>13</td>
<td>42</td>
<td>60</td>
<td>14</td>
<td>43</td>
<td>61</td>
</tr>
<tr>
<td>Additive+DAVEnet</td>
<td>COCO</td>
<td>9</td>
<td>31</td>
<td>48</td>
<td>10</td>
<td>35</td>
<td>53</td>
</tr>
<tr>
<td>Normalized+DAVEnet</td>
<td>COCO</td>
<td>10</td>
<td>32</td>
<td>48</td>
<td>9</td>
<td>33</td>
<td>48</td>
</tr>
<tr>
<td>LSTM</td>
<td>COCO</td>
<td>10</td>
<td>30</td>
<td>45</td>
<td>11</td>
<td>32</td>
<td>45</td>
</tr>
<tr>
<td>NMT+Transformer</td>
<td>COCO</td>
<td>5</td>
<td>17</td>
<td>26</td>
<td>4</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>SMT+DAVEnet</td>
<td>COCO</td>
<td>3</td>
<td>13</td>
<td>20</td>
<td>0.1</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>SMT (phones)</td>
<td>COCO</td>
<td>7</td>
<td>24</td>
<td>36</td>
<td>4</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Flickr</td>
<td>7</td>
<td>19</td>
<td>29</td>
<td>3</td>
<td>11</td>
<td>19</td>
</tr>
</tbody>
</table>
## Results: MWD

<table>
<thead>
<tr>
<th>System</th>
<th>Alignment Recall</th>
<th>Alignment Precision</th>
<th>Alignment F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT+DAVEnet</td>
<td>60</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>SMT+Transformer</td>
<td>21.8</td>
<td>43</td>
<td>29</td>
</tr>
<tr>
<td>SMT (phones)</td>
<td>37.9</td>
<td>19</td>
<td>25.5</td>
</tr>
<tr>
<td>NMT+DAVEnet</td>
<td>54.9</td>
<td>27.8</td>
<td>36.9</td>
</tr>
<tr>
<td>NMT+Transformer</td>
<td><strong>62.7</strong></td>
<td><strong>31.8</strong></td>
<td><strong>42.2</strong></td>
</tr>
</tbody>
</table>

**Table**: Word discovery performance of various systems on MSCOCO; Results are evaluated only with words that describe one of the 80 concepts
Tradeoff between retrieval and word discovery

Figure: Alignment and Retrieval precision-recall curves for various models
Figure: Word discovery results of different systems on the image-caption pair “a woman eating a piece of pastry in a market area.” The texts are not available in the first two figures during training and are shown for ease of understanding.
Discussion

- Averaging vs. peak detection: the right approach for extracting word embedding from DAVEnet?
- Common space clustering vs. probabilistic alignment/clustering
- Discriminative training vs Maximum likelihood training
### Discriminative training of SMT

\[
\max \text{Tr}(\Psi_a^\top P\Psi_v A)
\]
\[
s.t. \quad \text{Tr}(\bar{\Psi}^\top P\Psi_v) = 1,
\]
\[
P_w \in \mathbb{R}^{K^+}, \forall w
\]

where \(\bar{\Psi} := \sum_{n=1}^{N} A\Psi_a^{(n)\top}\).

### Solution

\[
P_w^* = \frac{(\Psi_v^\top A\Psi_a)_{z^*w}}{(\bar{\Psi}^\top \Psi_a)_{z^*z^*}} e_{z^*},
\]

where \(z^* = \arg \max_z \frac{(\Psi_a^\top A\Psi_v)_{z^*w}}{(\Psi_a^\top \Psi_a)_{z^*z^*}} \approx \arg \max_z \frac{\bar{p}(z^*|w)}{\bar{p}(z^*)},\) where \(\bar{p}(\cdot)\) is the empirical distribution.
A speech embedding learned using a TDNN gives the highest speech-to-image retrieval scores, but that embedding learned using a self-attention Transformer model gives higher scores for word discovery.

In both cases, accuracy is boosted by using an NMT-based attention mechanism with self-attention layers, which helps the retrieval model to learn better alignments for visual words.

From our results, we believe a joint retrieval-discovery is important for developing better word discovery systems.
Future Direction

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. *Automatic Speech Recognition and Understanding*.


M. Hodosh, P. Young, and J. Hockenmaier. 2010. Framing image description as a ranking task: data, models and evaluation metrics. In *Journal of Artificial Intelligence Research*.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.
