Cross-media Structured Common Space for Multimedia Event Extraction

Manling Li*, Alireza Zareian*, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, Shih-Fu Chang

ACL 2020
Introduction

- 33% of images in 100 randomly selected VOA articles contain visual objects that serve as event arguments and are not mentioned in the text.

![Image of event extraction with labels: Event: [Movement.Transport], Visual Arguments: Vehicle, truck, Textual Arguments: Agent, United States, Artifact, soldiers.]

Figure 1: An example of Multimedia Event Extraction. An event mention and some event arguments (Agent and Person) are extracted from text, while the vehicle arguments can only be extracted from the image.
Introduction

- Event extraction is independently studied in Computer Vision (CV) and Natural Language Processing (NLP), significantly different in terms of task definition, data domain, methodology, and terminology.
- ACE dataset:
Introduction

- Event extraction is independently studied in Computer Vision (CV) and Natural Language Processing (NLP), significantly different in terms of task definition, data domain, methodology, and terminology.
- ImSitu dataset:

<table>
<thead>
<tr>
<th>Situations</th>
<th>ImSitu Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>feeding</td>
<td></td>
</tr>
<tr>
<td>agent</td>
<td>verbs</td>
</tr>
<tr>
<td>food</td>
<td>images</td>
</tr>
<tr>
<td>source</td>
<td>situations per image</td>
</tr>
<tr>
<td>eater</td>
<td>total annotations</td>
</tr>
<tr>
<td>place</td>
<td>unique entity types (&gt;3)</td>
</tr>
<tr>
<td>man</td>
<td>unique roles (role types)</td>
</tr>
<tr>
<td>fish</td>
<td>images per verb (range)</td>
</tr>
<tr>
<td>hand</td>
<td>unique situations (&gt;3)</td>
</tr>
<tr>
<td>dolphin</td>
<td></td>
</tr>
<tr>
<td>pool</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

- Event extraction is independently studied in Computer Vision (CV) and Natural Language Processing (NLP), significantly different in terms of task definition, data domain, methodology, and terminology.

=> They propose a new task: **MultiMedia Event Extraction** ($M^2E^2$):

  + An evaluation dataset: 245 fully annotated news articles.
  + A new method for the task that learns a structured multimedia embedding space: Weakly Aligned Structured Embedding (WASE).


**M²E² dataset**

- 245 documents selected from 108,693 multimedia VOA articles.
- Contains 8 ACE types (i.e., 24% of all ACE types), mapped to 98 imSitu types (i.e., 20% of all imSitu types), encompassing 52% ACE events.
- Annotators: 8 with an Inter-Annotator Agreement score of 81.2%.
- This dataset is for **Evaluation Only**.

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Argument Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movement.Transport</td>
<td>Agent (46/64), Artifact (179/103), Vehicle (24/51), Destination (120/0), Origin (66/0)</td>
</tr>
<tr>
<td>Conflict.Attack</td>
<td>Attacker (192/12), Target (207/9), Instrument (37/15), Place (121/0)</td>
</tr>
<tr>
<td>Conflict.Demonstrate</td>
<td>Entity (102/184), Police (3/26), Instrument (0/118), Place (86/25)</td>
</tr>
<tr>
<td>Justice.ArrestJail</td>
<td>Agent (64/119), Person (147/99), Instrument (0/11), Place (43/0)</td>
</tr>
<tr>
<td>Contact.PhoneWrite</td>
<td>Entity (33/46), Instrument (0/43), Place (8/0)</td>
</tr>
<tr>
<td>Contact.Meet</td>
<td>Participant (119/321), Place (68/0)</td>
</tr>
<tr>
<td>Life.Die</td>
<td>Agent (39/0), Instrument (4/2), Victim (165/155), Place (54/0)</td>
</tr>
<tr>
<td>Transaction. TransferMoney</td>
<td>Giver (19/3), Recipient (19/5), Money (0/8)</td>
</tr>
</tbody>
</table>

Table 1: Event types and argument roles in M²E², with expanded ones in bold. Numbers in parentheses represent the counts of textual and visual events/arguments.

<table>
<thead>
<tr>
<th>Source</th>
<th>Event Mention</th>
<th>Argument Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentence</td>
<td>image</td>
<td>textual</td>
</tr>
<tr>
<td>6,167</td>
<td>1,014</td>
<td>1,297</td>
</tr>
</tbody>
</table>

Table 2: M²E² data statistics.
Overall architecture

VOA Image-Caption data

ACE data

ImSitu data

Shared Event & Role Classifiers
Text Event Extraction

- Use the CAMR parser (Wang et al., 2015b,a, 2016) to obtain an AMR graph for each sentence.
- Word representation = Glove + POS + NER + Position
- GCN:

\[
\begin{align*}
    w^{(k+1)}_i &= f\left( \sum_{j \in N(i)} g^{(k)}_{ij} (W_{E(i,j)}w^{(k)}_j + b^{(k)}_{E(i,j)}) \right), \\
    P(y_e|w) &= \frac{\exp (W_e w^c + b_e)}{\sum_{e'} \exp (W_{e'}w^c + b_{e'})}, \\
    P(y_a|t) &= \frac{\exp (W_a [t^c; w^c] + b_a)}{\sum_{a'} \exp (W_{a'}[t^c; w^c] + b_{a'})}.
\end{align*}
\]

- Prediction:
Image Event Extraction

- Produce a situation graph (similar to ARM) for each image:
  > central node is labeled as a verb.
  > neighbor nodes are arguments labeled as (noun, role).
- Propose two ways to construct a situation graphs:
  > Object-based (predefined-type object detection).
  > Attention-based (role-driven object detection).
Object detection: use a Faster R-CNN \((\text{Ren et al., 2015})\) trained on Open Images with 600 object types.

Use VGG-16 CNN \((\text{Simonyan and Zisserman, 2014})\) to obtain image/object representation.

Image representation: \(\mathbf{m}\); object representations: \(\mathbf{O}_i\)

Embedding layer: \(\hat{\mathbf{m}} = \text{MLP}_m(\mathbf{m}), \hat{\mathbf{o}}_i = \text{MLP}_o(\mathbf{o}_i)\)

Verb and noun prediction:

\[
P(v|m) = \frac{\exp(\hat{\mathbf{m}})}{\sum_{\mathbf{m}'} \exp(\hat{\mathbf{m}}')},
\]

\[
P(n|\mathbf{o}_i) = \frac{\exp(\hat{\mathbf{o}}_i \mathbf{n})}{\sum_{\mathbf{n}'} \exp(\hat{\mathbf{o}}_i \mathbf{n}')},
\]

Argument role labeling: \(P(r_i|\mathbf{o}_i) = \sigma(\text{MLP}_r(\hat{\mathbf{o}}_i))\)

\[
\mathcal{L}_v = - \log P(v^*|m),
\]

\[
\mathcal{L}_r = - \log (P(r_i^*|\mathbf{o}_i) + P(n_i^*|\mathbf{o}_i))
\]
Many object types are not covered by the pretrained R-CNN model.

Use VGG-16 CNN to obtain key vectors $k_i$ for 7x7 local regions of the input image.

For each possible role of the event type (i.e., verb), form the query vector: $q_r = W_q[r; m] + b_q$

Attention is then done over the 7x7 regions to obtain object representations:

$$h_i = \frac{\exp(q_r k_i)}{\sum_{j \in 7 \times 7} \exp(q_r k_j)}$$

$$o_r = \sum_i h_i m_i$$

Verb and noun predictions are done similar as in object-based method.
Cross-media joint training

- Form the structured common space by learning to map captions <-> images via VOA image-caption pair data.
- Soft alignment from each words to image objects and vice versa:
  \[
  \alpha_{ij} = \frac{\exp(w_i^C o_j^C)}{\sum_j \exp(w_i^C o_j^C)}, \beta_{ji} = \frac{\exp(w_i^C o_j^C)}{\sum_i \exp(w_i^C o_j^C)}
  \]
- Representations aligned to the common space:
  \[
  w_i' = \sum_j \alpha_{ij} o_j^C, o_j' = \sum_i \beta_{ji} w_i^C
  \]
- Alignment cost:
  \[
  \langle s, m \rangle = \sum_i ||w_i - w_i'||_2^2 + \sum_j ||o_j - o_j'||_2^2
  \]
- Training objective:
  \[
  \mathcal{L}_c = \max(0, 1 + \langle s, m \rangle - \langle s, m^- \rangle)
  \]
Training

- Training objective for verb and noun prediction:
  \[ \mathcal{L}_v = -\log P(v^*|m), \]
  \[ \mathcal{L}_r = -\log (P(r^*_i|o_i) + P(n^*_i|o_i)) \]

- Training objectives for shared classifiers:
  \[ \mathcal{L}_e = -\sum_w \log P(y_e|w) - \sum_m \log P(y_e|m), \]
  \[ \mathcal{L}_a = -\sum_t \log P(y_a|t) - \sum_o \log P(y_a|o), \]

- Training objective for shared common space:
  \[ \mathcal{L}_c = \max(0, 1 + \langle s, m \rangle - \langle s, m^- \rangle) \]

- Overall training objective:
  \[ \mathcal{L} = \mathcal{L}_v + \mathcal{L}_r + \mathcal{L}_e + \mathcal{L}_a + \mathcal{L}_c \]
Inference

- Given a multimedia document with:
  - a set of sentences: \( S = \{s_1, s_2, \ldots\} \)
  - a set of images: \( M = \{m_1, m_2, \ldots,\} \)

- First, compute pair-wise similarities \( \langle s, m \rangle \)
  - select the closest image for each sentence.
  - select the closest sentence for each image.

- Compute the aligned representations for words (and objects similarly):
  \[
  \gamma = \exp(-\langle s, m \rangle) \quad \text{and} \quad w_i'' = (1 - \gamma)w_i + \gamma w_i'
  \]

- Predictions are then done with the aligned representations.
## Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>Text-Only Evaluation</th>
<th></th>
<th>Image-Only Evaluation</th>
<th></th>
<th>Multimedia Evaluation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
<td>Argument Role</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(P)</td>
<td>(R)</td>
<td>(F_1)</td>
<td>(P)</td>
<td>(R)</td>
<td>(F_1)</td>
</tr>
<tr>
<td>Text</td>
<td>JMEE</td>
<td>42.5</td>
<td>58.2</td>
<td>48.7</td>
<td>22.9</td>
<td>28.3</td>
<td>25.3</td>
</tr>
<tr>
<td>Image</td>
<td>GAIL</td>
<td>43.4</td>
<td>53.5</td>
<td>47.9</td>
<td>23.6</td>
<td>29.2</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>WASE(^f)</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
<td>21.4</td>
<td>30.1</td>
<td>24.9</td>
</tr>
<tr>
<td>Multimedia</td>
<td>WASE(^f)_att</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WASE(^f)_obj</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8</td>
<td>39.4</td>
<td>16.6</td>
<td>24.7</td>
<td>19.8</td>
</tr>
<tr>
<td>Flat(^a)tt</td>
<td>34.2</td>
<td>63.2</td>
<td>44.4</td>
<td>20.1</td>
<td>27.1</td>
<td>23.1</td>
<td>27.1</td>
</tr>
<tr>
<td>Flat(^a)obj</td>
<td>38.3</td>
<td>57.9</td>
<td>46.1</td>
<td>21.8</td>
<td>26.6</td>
<td>24.0</td>
<td>26.4</td>
</tr>
<tr>
<td>WASE(^a)tt</td>
<td>37.6</td>
<td>66.8</td>
<td>48.1</td>
<td>27.5</td>
<td>33.2</td>
<td><strong>30.1</strong></td>
<td>32.3</td>
</tr>
<tr>
<td>WASE(^a)obj</td>
<td>42.8</td>
<td>61.9</td>
<td><strong>50.6</strong></td>
<td>23.5</td>
<td>30.3</td>
<td>26.4</td>
<td>43.1</td>
</tr>
</tbody>
</table>

15