Unsupervised Sampling Approach for Image-Sentence Matching Using Document-Level Structural Information

Zejun Li¹, Zhongyu Wei², Zhihao Fan¹, Haijun Shan², Xuanjing Huang²
1 Fudan University, China, 2 Zhejiang Lab, China

Introduction

Learning to align semantic spaces of vision and text (a) mainly follows contrastive learning, requiring information to find matched positive pairs (red links) and negative pairs (blue link). Most works are supervised (b) with labeled pairs (solid links), while some unsupervised methods (c) explore to utilize document-level information to sample pseudo pairs (dashed links). Relatively similar intra-document pairs are considered positive and cross-document pairs are negative samples, introducing a sampling bias since cross-document pairs are relatively semantically dissimilar and easy negative samples. We propose strategies to efficiently sample more positive/negative intra-document pairs, and a Transformer-based model to capture fine-grained features, where “concepts” are introduced to bridge the cross-modal representation learning in the context of a document.

Unsupervised Sampling Strategy based on Document-Level Structure

We introduce 3 training objectives, correspond to 3 strategies to sample positive and negative image-sentence pairs:
- Cross-document Objective “C”:
  - Positive: the most similar intra-document pairs
  - Negative: the most similar cross-document pairs
- Intra-document Objective “I”:
  - Positive: the most similar intra-document pairs
  - Negative: the most dissimilar intra-document pairs
- Dropout Sub-Document Objective “D”:
  - Randomly mask some imgs/sents \( \rightarrow \) sub-document
  - Positive: the most similar pairs intra sub-documents
  - Negative: the most similar cross-document pairs
- Combined objectives \( \rightarrow \) aggregated sample pairs

Cross-Modality Alignment Model

A transformer based model is proposed to learn well-aligned cross-modality representations, we enable it to capture fine-grained features and bridge representation learning of images and sentences:
- Visual objects are extracted by Faster RCNN, their corresponding labels are considered “concepts”.
- Concepts and tokens share the same embedding layer to encode conceptually semantic information.
- A densely connected graph between concepts and objects is constructed by Transformer.
- Mean pooling is used to extract overall image/sentence representations.

Experiment & Results

Overall Performance

<table>
<thead>
<tr>
<th></th>
<th>MSCOCO AUC</th>
<th>Story-DB AUC</th>
<th>Story-SIS AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>99.3</td>
<td>97.6</td>
<td>80.5</td>
</tr>
<tr>
<td>NoStruct</td>
<td>85.5</td>
<td>67.7/45.9</td>
<td>65.3</td>
</tr>
<tr>
<td>MulLink</td>
<td>90.0</td>
<td>95.0/80.1</td>
<td>82.9</td>
</tr>
<tr>
<td>Obj Det</td>
<td>87.4</td>
<td>50.6/34.3</td>
<td>77.0</td>
</tr>
<tr>
<td>5 GRU+CNN</td>
<td>85.1</td>
<td>75.58/0.1</td>
<td>82.9</td>
</tr>
</tbody>
</table>

Overall Performance: Obj Detect and NoStruct are baselines. MulLink is the only existing unsupervised model.


Further Analysis

- Ablation Study on DII: C, I, and D correspond to 3 objectives, different combinations used during training. T is short for Transformer

  - Ablation study shows the effectiveness of modules of our alignment model and 3 parts of training objectives (sampled image-sentence pairs).
  - Compared with supervised methods, we are able to utilize more information under the unsupervised setting.

Case Study

- Green/purple links are matched/unmatched pairs in ground truth, line widths are proportional to predicted similarities.