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

Zejun Li¹, Zhongyu Wei¹, Zhihao Fan¹, Haijun Shan², Xuanjing Huang³

¹School of Data Science, Fudan University, China
²Zhejiang Lab, China
³School of Computer Science, Fudan University, China
Motivation

• Cross-modality Alignment:
  • Alignment of 2 semantic spaces (a)
  • Image-sentence matching

• Contrastive Learning:
  • Positive: Matched pairs
  • Negative: Hard negative

• Information is needed to find positive/negative pairs:
  • Supervised: labeled pairs (b)
  • Unsupervised
Unsupervised

• Document-level structural information: co-occurrence of images and sentences.

• In (Hessel, Lee, and Mimno 2019):
  • Use document-level information
  • Positive intra-document pairs
  • Negative cross-document pairs

• Effective but introduce a sampling bias.
Sampling Bias

• Cross-document training:
  • The positive and negative sample pairs are easy to distinguish
  • book vs horses

• Intra-document evaluation:
  • The positive and negative sample pairs are hard to distinguish
  • Book A vs Book B;

Semantic distances between pos/neg images: Training vs Evaluation
Contribution

• An unsupervised strategy:
  • Aiming to alleviate the sampling bias
  • More intra-document pos/neg pairs

• A Transformer based model:
  • Fine-grained features
  • Implicit graph
  • Concepts introduced
Sampling Strategy

- 3 different document-level training objectives
- 3 strategies to sample pseudo positive/negative samples (image-sentence pairs).
Cross-Document Objective

- **Assumption:**
  co-occurring image-set and sentence-set are more semantically similar than non-co-occurring ones

- **Positive:**
  intra-document pairs with the highest similarities in original documents

- **Negative:**
  cross-document pairs with the highest similarities in negative documents
Intra-Document Objective

• Assumption:
  Similarity of predicted unmatched pairs should be lower than predicted matched image-sentence pairs from the same document

• Positive:
  intra-document pairs with the highest similarities in original documents

• Negative:
  intra-document pairs with the lowest similarities in negative documents
Dropout Sub-Document Objective

• Assumption:
  Images and sentences co-occurring in a “sub-document” should be more similar than non-co-occurring ones

• Positive:
  intra-document pairs with the highest similarities in random sub-documents.

• Negative:
  cross-document pairs with the highest similarities in negative documents.
Cross-modality Alignment Model

- **Visual Transformer:**
  Multi-modal embedding (Faster RCNN/Concept Embedding) + Segment embedding

- **Textual Transformer:**
  Multi-modal embedding (Word Embedding) + Position embedding
Graph Constructed

- Implicit graph: tokens + regions + concepts
  - Visual Transformer: regions---concepts
  - Shared Embedding layer: concepts---tokens (hard matching)
Experiment – Tasks & Datasets

- Multimodal link prediction in multi-sentence multi-image documents formulated in (Hessel, Lee, and Mimno 2019):
  - Metrics: AUC and p@1/5
  - MSCOCO, VIST-DII, VIST-SIS

- Evaluation settings:
  - Unsupervised training with our proposed objective
  - Predict intra-document similarities by the trained model.
Methods for Comparison

• **NoStruct:**
  • GRU-CNN
  • randomly samples image-caption pairs from a document and treat the similarity between them as the document-level similarity.

• **Object Detection:**
  • Image: average word2vec embeddings of its top-K ImageNet labels
  • Sentence: average word2vec embeddings of its words
  • no training

• **MulLink (Hessel, Lee, and Mimno 2019):**
  • Backbone: GRU-CNN
  • trained only with the cross-document objective $\ell_C$
  • with the sampling bias
### Overall Performance

<table>
<thead>
<tr>
<th></th>
<th>MSCOCO</th>
<th></th>
<th>Story-DII</th>
<th></th>
<th>Story-SIS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC p@1/p@5</td>
<td>AUC p@1/p@5</td>
<td>AUC p@1/p@5</td>
<td>AUC p@1/p@5</td>
<td>AUC p@1/p@5</td>
<td>AUC p@1/p@5</td>
</tr>
<tr>
<td>Obj Detect</td>
<td>89.5/67.7/45.9</td>
<td>65.3/50.2/35.2</td>
<td>58.4/40.8/28.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoStruct</td>
<td>87.4/50.6/34.3</td>
<td>77.0/60.8/46.3</td>
<td>64.5/42.8/33.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MulLink</td>
<td>99.0/95.0/81.1</td>
<td>82.9/72.0/55.8</td>
<td>68.8/51.8/38.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td><strong>99.3/97.6/86.0</strong></td>
<td><strong>85.5/77.2/60.1</strong></td>
<td><strong>70.2/53.1/39.8</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Overall performance of different models. Numbers in *bold* denote the best performance in each column.

- **MSCOCO:**
  - Nearly no bias: MulLink performs well, and the AUC is nearly perfect.

- **Story-DII:**
  - Similar sentences/images in a document → Bias between training and evaluation

- **Story-SIS:**
  - Dependency between sentences of the same document (referring pronouns…).
Ablation Study

• Each objective contributes to the performance---all parts of sampled pos/neg pairs are effective.

• Without Transformer, just aggregating the concept features into the image representation does not improve performance (row 2, 3).

• Incorporating concepts into Transformer significantly improves performance on precision (row 1, 2).

<table>
<thead>
<tr>
<th>backbone</th>
<th>Objectives</th>
<th>AUC</th>
<th>p@1/p@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Ours</td>
<td>C+I+D</td>
<td>85.5</td>
<td>77.2/60.1</td>
</tr>
<tr>
<td>2 w/o Concept</td>
<td>C+I+D</td>
<td>85.3</td>
<td>75.8/59.8</td>
</tr>
<tr>
<td>3 w/o T</td>
<td>C+I+D</td>
<td>85.1</td>
<td>75.0/59.0</td>
</tr>
<tr>
<td>4 w/o T&amp;Concept</td>
<td>C+I+D</td>
<td>85.1</td>
<td>74.6/59.1</td>
</tr>
<tr>
<td>5 GRU+CNN</td>
<td>C+I+D</td>
<td>84.0</td>
<td>72.9/58.0</td>
</tr>
<tr>
<td>6 Ours</td>
<td>C+I</td>
<td>85.2</td>
<td>75.9/59.2</td>
</tr>
<tr>
<td>7 Ours</td>
<td>C+D</td>
<td>85.4</td>
<td>76.2/59.9</td>
</tr>
<tr>
<td>8 Ours</td>
<td>I+D</td>
<td>84.1</td>
<td>73.4/57.8</td>
</tr>
<tr>
<td>9 Ours</td>
<td>C</td>
<td>85.0</td>
<td>75.5/59.4</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on SIS, the “Objectives” column represents different combinations of objectives used during training, where “C”, “T”, and “D” correspond to 3 parts of objectives mentioned, respectively. “T” is short for Transformer, w/o means removing a certain module.
Bias Alleviation

• The "spread" hypothesis in (Hessel, Lee, and Mimno 2019):
  • Lower intra-document diversity = larger bias $\rightarrow$ hard
  • OLS regression of intra-document diversity on test AUC

• Trained with more samples (ours):
  • OLS R-square = influence of bias on the performance
  • DII: 42% $\rightarrow$ 23%
  • SIS: 26% $\rightarrow$ 12%
  • Bias is less influential $\rightarrow$ alleviated
Comparison with Supervised Methods

- Utilize more information in a dataset under unsupervised setting.
- Better performance compared with a transfer model.

![Fig5](https://example.com/fig5.png)

Figure 5: Performance of supervised strategy using different proportions of training data, dashed lines denote performances of unsupervised strategies.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>p@1/p@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Transfer from MSCOCO</td>
<td>78.6</td>
<td>66.5/49.5</td>
</tr>
<tr>
<td>2 Unsupervised</td>
<td>85.5</td>
<td>77.2/60.1</td>
</tr>
</tbody>
</table>

Table 4: Performance of different methods on DII without explicit labels.
Figure 6: Illustrative documents in DII and SIS: Edges in green are true links in ground-truth; edge widths show the magnitude of edges in $\hat{M}_t$ (only positive weights are shown). Main detected concepts are listed and italicized words are directly involved in sentences. Selected documents are representative because their AUC scores match average AUC in corresponding datasets.
THANKS

2020.12.19