国际人工智能会议

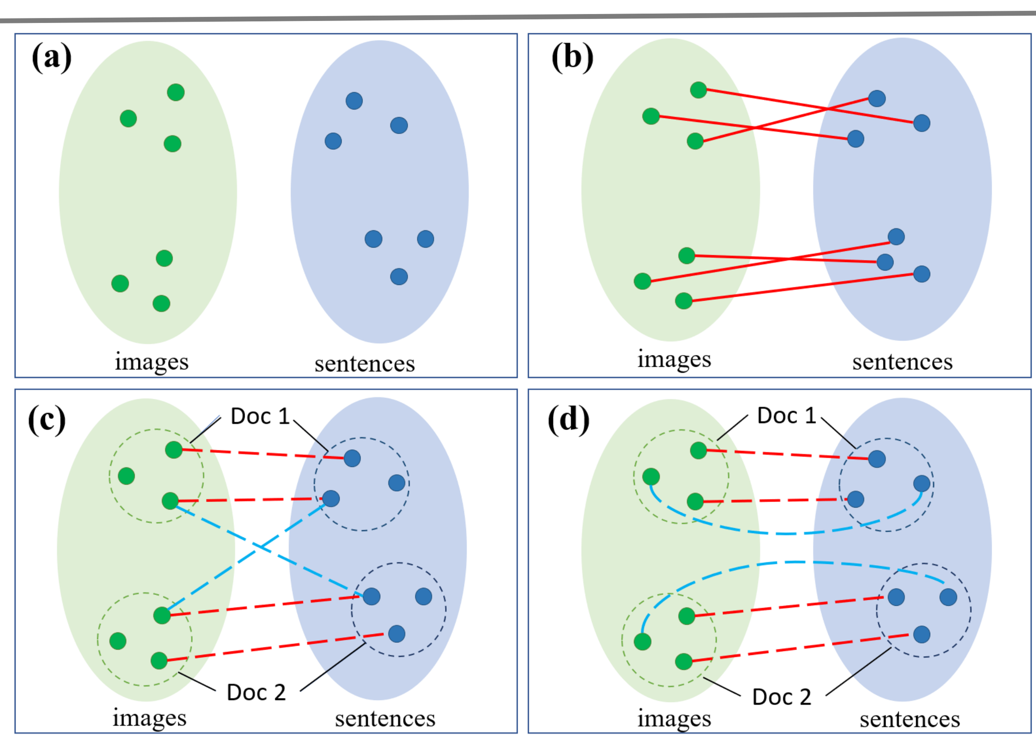
AAAI 2021 论文北京预讲会

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

Zejun Li[1], Zhongyu Wei[1], Zhihao Fan[1], Haijun Shan[2], Xuanjing Huang[1]

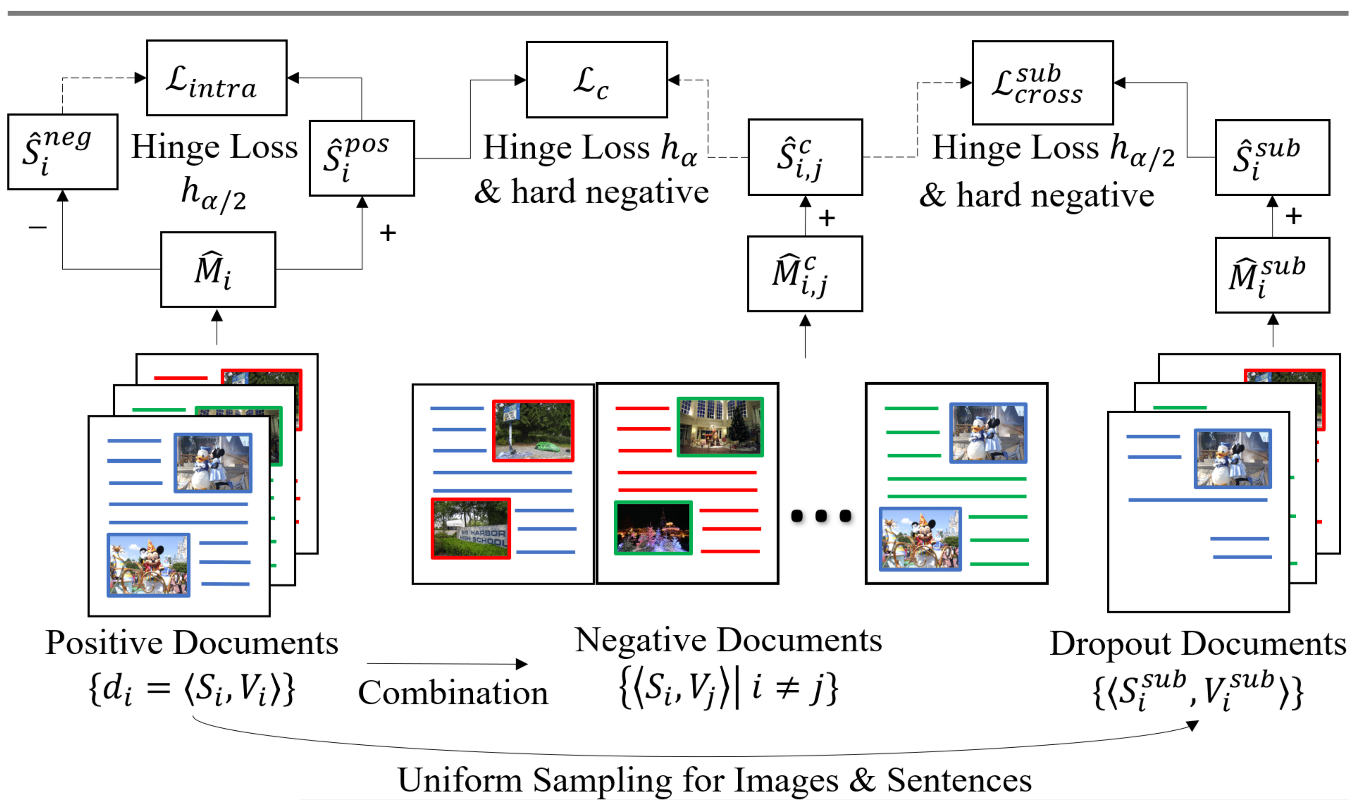
1 Fudan University, China, 2 Zhejiang Lab, China

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.

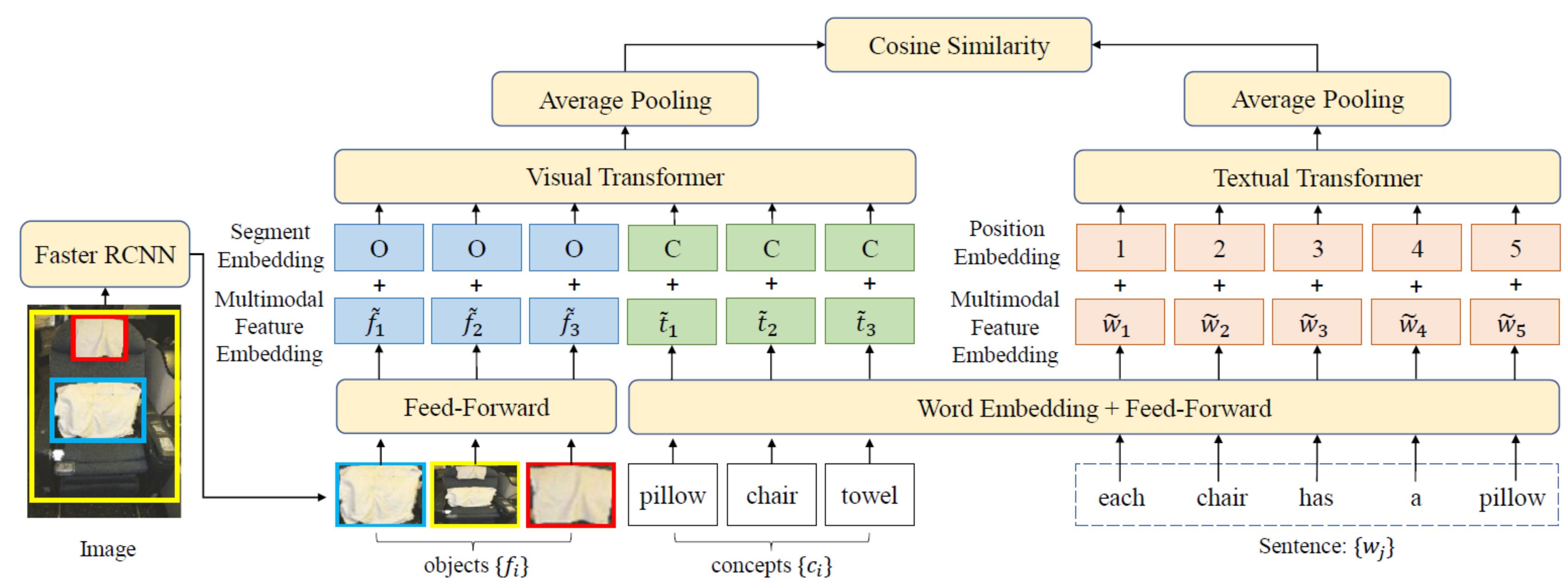
Unsupervised Sampling Strategy based on Document-Level Structure



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 → sub-document
  - Positive: the most similar pairs intra sub-documents
  - Negative: the most similar cross-document pairs
- Combined objectives → 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

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

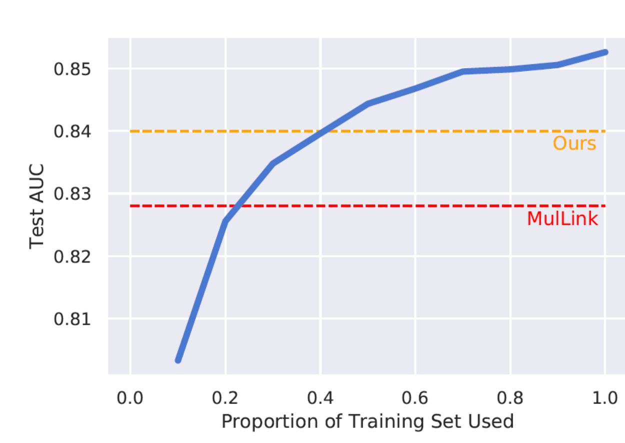
Overall Performance: Obj Detect and NoStruct are baslines, MulLink is the only existing unsupervised model.

- Evaluation on the task of unsupervised multi-image multi-sentence linking among a document: our method shows a superior performance.

Further Analysis

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

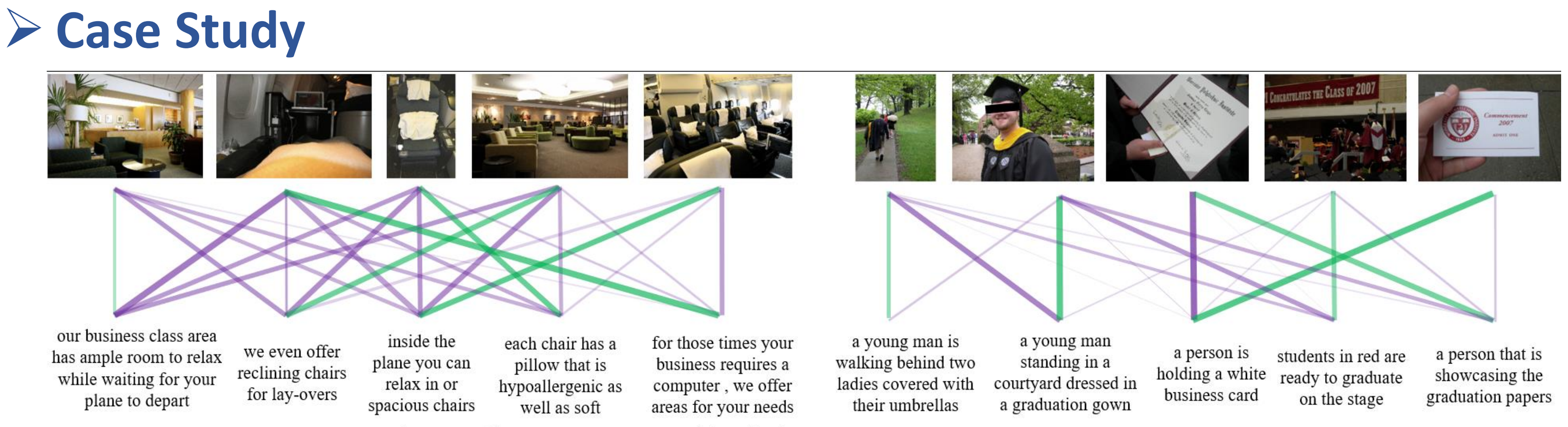
Ablation Study on DII: C, I, and D correspond to 3 objectives, different combinations used during training, T is short for Transformer.



Comparison with supervised methods (blue)

- 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



(a) **SIS**; 76 AUC; Concepts: *chair*, *pillow*, *computer*, table, light (b) **DII**; 87 AUC; Concepts: *umbrella*, *man*, *hat*, hand, *paper*

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

主办方：中国中文信息学会青年工作委员会

承办方：智源社区