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

Zejun Li,[1] Zhongyu Wei,[1,4]* Zhihao Fan,[1] Haijun Shan,[2] Xuanjing Huang[3]
[1]School of Data Science, Fudan University, China
[2]Zhejiang Lab, China
[3]School of Computer Science, Fudan University, China
[4]Research Institute of Intelligent and Complex Systems, Fudan University, China
{20210980139, zywei, 18210980005}@fudan.edu.cn, workingshan@163.com, xjhuang@fudan.edu.cn

Abstract

In this paper, we focus on the problem of unsupervised image-sentence matching. Existing research explores to utilize document-level structural information to sample positive and negative instances for model training. Although the approach achieves positive results, it introduces a sampling bias and fails to distinguish instances with high semantic similarity. To alleviate the bias, we propose a new sampling strategy to select additional intra-document image-sentence pairs as positive or negative samples. Furthermore, to recognize the complex pattern in intra-document samples, we propose a Transformer based model to capture fine-grained features and implicitly construct a graph for each document, where concepts in a document are introduced to bridge the representation learning of images and sentences in the context of a document. Experimental results show the effectiveness of our approach to alleviate the bias and learn well-aligned multimodal representations.

Introduction

Image-text matching is one of the fundamental problems in the field of vision and language (Nam, Ha, and Kim 2017; Huang et al. 2018), and the main target is learning to align the semantic spaces of two modalities (Figure 1(a)). Previous works on image-text matching is mainly supervised (Lee et al. 2018; Wang et al. 2019; Zheng et al. 2020), requiring large amounts of annotated image-sentence pairs (Figure 1(b)). Considering labeled pairs of images and sentences are expensive to obtain, progress in developing unsupervised methods is therefore exciting and promising. Although some attempts are made to align image regions and segments of sentences (Karpathy, Joulin, and Fei-Fei 2014; Karpathy and Fei-Fei 2015; Rohrbach et al. 2016; Datta et al. 2019), they still rely on matched image-sentence pairs for distant supervision.

The main challenge in unsupervised image-sentence matching is the lack of such information to distinguish positive and negative samples for model training. (Hessel, Lee, and Mimno 2019) explores to utilize the document-level structural information. In specific, image-sentence pairs in a document (intra-document) are selected as positive samples, and negative samples are drawn from cross-document image-sentence pairs (Figure 1(c)). During training, the model is learning to maximize the distance between the sampled positive and negative pairs. Although this unsupervised sampling strategy is shown to be effective for learning aligned representations for images and sentences to some extent, the sampling bias between the training environment and real environment can not be ignored.

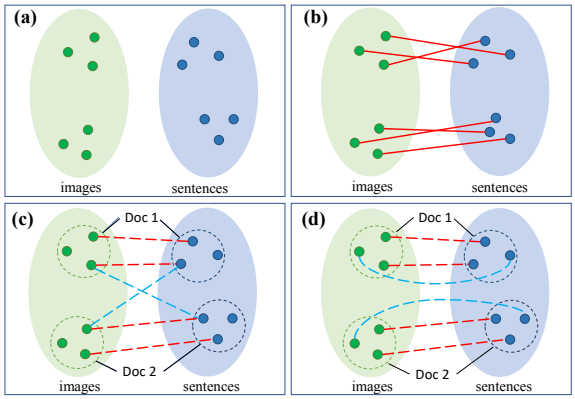
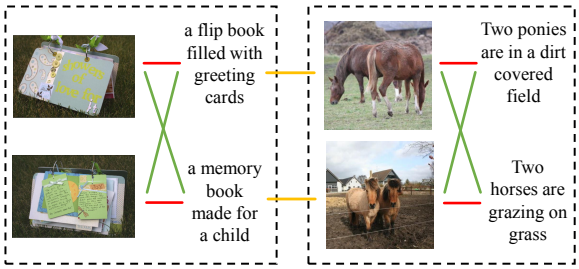


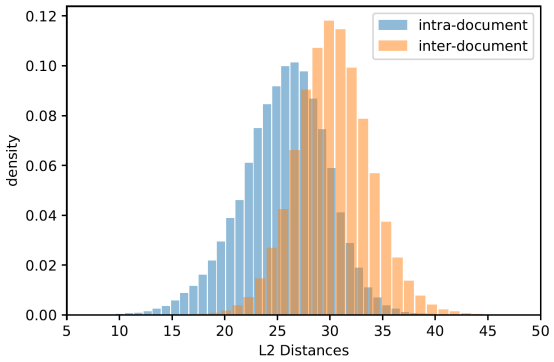
Figure 1: Illustration of different settings for image-sentence matching: (a) multimodal semantic spaces, (b) supervised alignment, (c) unsupervised cross-document objective in (Hessel, Lee, and Mimno 2019), (d) our intra-document objective. Red links denote matched positive pairs, blue links denote negative pairs; solid links represent annotated labels, dashed lines represent pseudo labels detected by unsupervised methods; dashed circles denote image-sets and sentence-sets in documents.

Figure 2(a) shows an example. The positive and negative sample pairs for training are much easier to be distinguished (book vs horses), while an image and a sentence in a negative sample from the testing environment can be highly correlated in terms of semantics. We further present the evidence for bias resulted in this sampling strategy in Figure 2(b). For each sentence in the VIST-DII dataset (Hes-

*Corresponding author



(a) Illustration of positive and negative samples for training and evaluation in (Hessel, Lee, and Mimno 2019): links in red/green are negative/positive samples considered during evaluation, while links in yellow are negative samples considered during training.



(b) Distributions of L2 distances between pre-trained CNN features of ground-truth matched and negative images with respect to the same sentence, during inter-document training and intra-document evaluation.

Figure 2: Illustration of the sampling bias.

sel, Lee, and Mimno 2019), we compute the L2 distances between pre-trained CNN features of its ground-truth matched and negative images considered in the training and evaluation process, then visualize the two distributions. The difference between these two distributions is significant[1] and distances of training samples are generally larger. Such a sampling bias makes it hard for trained models to learn good representations for determining the correspondences between similar images and sentences.

To alleviate the issue of sampling bias, we propose to further distinguish positive and negative samples in the same document, which corresponds to Figure 1(d). In practice, we sample negative samples from the pairs with the least semantic similarity in a document. Distinguishing similar images and sentences in a document requires the backbone model's capability to learn cross-modality representations with fine-grained information. Consider the document on the right in Figure 2(a), capturing the object-level details like "dirt" and "grass" is necessary to distinguish these 2 sentences and images. Motivated by the success of recent works on introducing concepts to bridge cross-modal learning (You et al. 2016; Fan et al. 2019), we further explore to model concepts in a document to bridge the semantics of images and sentences. We propose to extract concepts from images and build an intra-document graph composed of images, sentences, and concepts implicitly. A Transformer based model is utilized to model these implicit dependencies and represent images and sentences with context-encoded fine-grained features.

The main contributions of our work are as follows:

- We reveal the sampling bias issue of an unsupervised sampling strategy for image-sentence matching that selects positive and negative samples from intra-document image-sentence and cross-document pairs, respectively.
- To alleviate the sampling bias issue, we propose a strategy to select negative samples and additional positive samples from intra-document pairs and form a new objective for cross-modality representation learning.
- To recognize the complex pattern in intra-document samples, we propose a Transformer based model to capture fine-grained features and integrate concepts into our model to bridge the representation learning of multimodal data in a document.
- We evaluate our method on the task of multi-model link prediction in multi-image, multi-sentence documents. Experiments show the effectiveness of our proposed method to alleviate the bias and learn better multimodal representations in the context of documents for this task.

Unsupervised Sampling Strategy based on Document-Level Structure

We first introduce the setting of document-level structure proposed by Hessel, Lee, and Mimno (2019). We are given a set of documents, each document $d_i = \langle S_i, V_i \rangle$ consists of a set S_i of $|S_i|$ sentences and a set V_i of $|V_i|$ images. On top of this, we define two kinds of image-sentence pairs, namely, intra-document pairs and cross-document pairs. Moreover, we propose three different strategies to sample positive or negative pairs and form three objectives namely, cross-document objective, intra-document objective, and dropout sub-document objective for model training. The overall illustration is shown in Figure 3. $\hat{M}_i$ denotes the similarity matrix of intra-document pairs where each element $\hat{M}_{i,(m,n)}$ is the similarity between $S_{i,m}$ and $V_{i,n}$, $\hat{M}^c_{i,j}$ denotes the similarity matrix of cross-document pairs where each element $\hat{M}^c_{i,j,(m,n)}$ is the similarity between $S_{i,m}$ and $V_{j,n}$.

Cross-Document Objective

The first objective is based on the assumption that co-occurring image-set/sentence-set pairs should be more similar than non-co-occurring image-set/sentence-set pairs.

We construct negative documents by combining non-co-occurring image-set/sentence-set pairs, and the objective for a single positive document can be characterized by hard negative mining with hinge loss:

$$\mathcal{L}_c(S_i, V_i) = \max_{j \neq i} h_\alpha(\text{sim}(S_i, V_i), \text{sim}(S_i, V_j)) + \max_{j \neq i} h_\alpha(\text{sim}(S_i, V_i), \text{sim}(S_j, V_i)) \quad (1)$$

[1] We conduct a two-sample Kolmogorov-Smirnov test where p-value < 0.01.

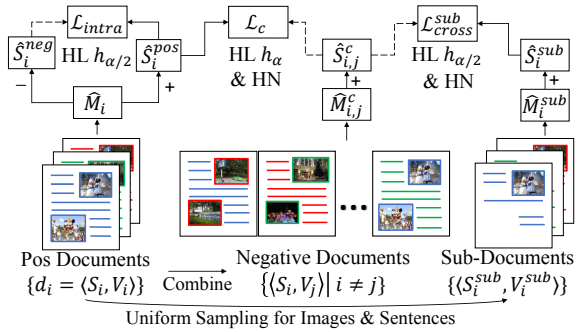


Figure 3: An illustration of the proposed training objectives, lines in documents represent sentences, different colors correspond to different documents. "+" represents the TK function and "-" represents the NegTK function. All $\hat{M}$ in the figure are computed by the same backbone model, and all $\hat{S}$ are the document-level similarities. Dashed lines indicate the negative inputs of hinge loss functions. "HL" and "HN" are short for hinge loss and hard negative, respectively.

where we consider i, j in a mini-batch. $h_\alpha(m, n) = \max(0, n - m + \alpha)$ is the hinge loss function with a margin of α, sim is a similarity function to compute the similarity between an image-set/sentence-set pair by mapping the predicted association matrix $\hat{M}$ between them to a real number, this function will select representative image-sentence pairs according to a specific criterion, and calculate the average similarity of selected pairs as the document-level similarity. We use a function $\mathrm{TK} : \mathbb{R}^{|S|\times|V|} \mapsto \mathbb{R}$ here, where the k most likely sentence-to-image and image-to-sentence edges will be selected based on currently predicted similarity, then compute the average similarity of selected pairs as the document-level similarity. This procedure corresponds to the equation: $\mathrm{sim}(S_i, V_j) = \hat{S}_{i,j}^c = \mathrm{TK}(\hat{M}_{i,j}^c)$.

Intra-Document Objective

The second objective aims to select negative image-sentence pairs from a document, with the assumption that similarities between predicted non-corresponding image-sentence pairs should be lower than predicted matched image-sentence pairs from the same document.

Similar to TK which measures the "positive" similarity, we also introduce a function NegTK to measure the document-level "negative" similarity i.e. how similar are those predicted non-corresponding images and sentences, NegTK will first select the k most unlikely sentence-to-image and image-to-sentence edges based on current predicted similarity, then calculate the average similarity of selected pairs as the document-level "negative" similarity. Then we can characterize this intra-document objective by a hinge loss between the document-level "positive" similarity and "negative" similarity:

$$\begin{aligned} \hat{S}_i^{pos} &= \mathrm{TK}(\hat{M}_i) \\ \hat{S}_i^{neg} &= \mathrm{NegTK}(\hat{M}_i) \\ \mathcal{L}_{intra}(S_i, V_i) &= h_{\frac{\alpha}{2}}(\hat{S}_i^{pos}, \hat{S}_i^{neg}) \end{aligned} \tag{2}$$

where $h(\cdot)$ has the same definition as in Equation 1 but with a smaller margin $\frac{\alpha}{2}$. NegTK can be efficiently implemented with an equivalent definition: $\mathrm{NegTK}(\hat{M}_i) = -\mathrm{TK}(-\hat{M}_i)$.

In essence, adding this complementary objective is equivalent to those intra-document image-sentence pairs with low predicted similarity as negative samples. Generally, it is nearly impossible for all image-sentence pairs to have a semantic association, we believe this strategy will have a high probability to choose those image-sentence pairs without an edge between them in ground-truth.

Dropout Sub-Document Objective

When using TK, only the 2k most probable image-sentence pairs will be regarded as positive samples. Apart from those 2k selected edges, there may exist (weaker) semantic associations between other image-sentence pairs according to the composition of a document.

To utilize that information, we introduce another complementary cross-document objective, under the assumption that even if some sentences and pictures in a document are removed, the dropout sub-document composed of remaining sentences and images will also have a higher document-level similarity than those of negative documents, with a smaller gap.

For a single positive document, we firstly construct a dropout sub-document by randomly removing a certain percentage $(1 - p_{sub})$ of sentences and pictures:

$$\begin{aligned} S_i^{sub} &= \mathrm{Uniform}(S_i, \mathrm{n} = \lfloor p_{sub} \times |S_i| \rfloor) \\ V_i^{sub} &= \mathrm{Uniform}(V_i, \mathrm{n} = \lfloor p_{sub} \times |V_i| \rfloor) \end{aligned} \tag{3}$$

where $\mathrm{Uniform}(A, \mathrm{n} = n_A)$ represents a function to uniformly draw n_A samples from A without replacement. Then we can characterize the objective with a form similar to $\mathcal{L}_c$:

$$\begin{aligned} \mathcal{L}_{cross}^{sub}(S_i, V_i) = &\max_{i \neq j} h_{\frac{\alpha}{2}}(\mathrm{sim}(S_i^{sub}, V_i^{sub}), \mathrm{sim}(S_i, V_j)) \\ &+ \max_{i \neq j} h_{\frac{\alpha}{2}}(\mathrm{sim}(S_i^{sub}, V_i^{sub}), \mathrm{sim}(S_j, V_i)) \end{aligned} \tag{4}$$

At document level, adding this complementary objective means constructing more positive documents; at image-sentence pair level, it is equivalent to a new sampling strategy: select additional positive image-sentence pairs from a document without considering some images and sentences. But these positive samples are regarded weaker and supposed to have a smaller gap.

We combine 3 objectives to the total loss, for a single positive document, it will be:

$$\mathcal{L}(S_i, V_i) = \mathcal{L}_c(S_i, V_i) + \mathcal{L}_{intra}(S_i, V_i) + \mathcal{L}_{cross}^{sub}(S_i, V_i) \tag{5}$$

Cross-Modality Alignment Model

To better leverage information both in intra-document positive and negative image-sentence pairs, the backbone model needs to be able to represent images and sentences with fine-grained features and in the context of a document. Our model will extract representations of images and sentences in a d_{multi}-dimensional multimodal text-image space.

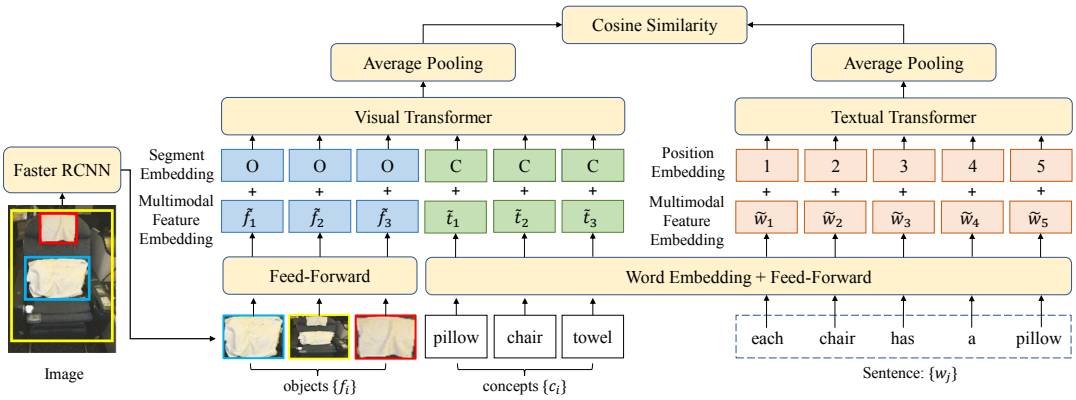


Figure 4: Architecture of our cross-modality alignment model.

Our idea is to introduce another type of nodes "concept" into the original bipartite graph, these nodes represent possible concepts (entities) the document may contain, we utilize these nodes as an intermediary of implicit links between images and sentences, and bridge the representation learning of images and sentences in a document. Correct construction of such kind of graphs for every document is intractable, we thus resort to an implicit implementation with Transformer (Vaswani et al. 2017) – which can be viewed as a densely connected graph model (Xu, Joshi, and Bresson 2019) – and a shared embedding layer between words in sentences and concepts.

Visual Objects and Concepts

To extract possible concepts in a document, we consider visual entities detected from images since concepts in sentences are obscure and may have compound meanings. Following Anderson et al. (2018), we use a pre-trained Faster RCNN (Ren et al. 2015) to extract μ object proposals $\{o_1, ..., o_\mu\}$ for each image, where each object o_i is represented by its 2048-dimensional region-of-interest (ROI) feature f_i. At the same time, predicted labels of these objects are considered as extracted concepts $\{c_1, ..., c_\mu\}$.

Extracting Sentence Representations

Similar to Tan and Bansal (2019), a sentence is first split into words $\{w_1, ..., w_\lambda\}$, then a word w_j and its index i (w_i's absolute position in the sentence) are sent to a 300D word embedding layer and a position embedding layer respectively:

$$\begin{aligned}\tilde{w}_i &= \text{WordEmbed}(w_i)\\ \tilde{u}_i &= \text{PosEmbed}(i)\\ \tilde{h}_i &= \text{LayerNorm}(\tilde{w}_i + \tilde{u}_i)\end{aligned} \tag{6}$$

where the word embedding layer is initialized with GoogleNews-pretrained word2vec embedding (Mikolov et al. 2013), the position embedding layer is randomly initialized from a uniform distribution between -0.02 and 0.02.

We further project the embeddings with a feed-forward layer and encode them by a single-modality N_T-layer Transformer:

$$\begin{aligned}h_i^0 &= W_T\tilde{h}_i + b_T\\ \{h_1^{l+1}, ..., h_\lambda^{l+1}\} &= \text{Transformer}_T^l(\{h_1^l, ..., h_\lambda^l\})\end{aligned} \tag{7}$$

where $W_T \in \mathbb{R}^{d_{\text{multi}} \times 300}$ and $b_T \in \mathbb{R}^{d_{\text{multi}}}$ are parameters of the feed-forward layer.

Finally, we extract the representation of a sentence through an average pooling layer on top of $\{h_1^{N_T}, ..., h_\lambda^{N_T}\}$.

Extracting Image Representations

A cross-modality Transformer is used to model the dependency between extracted visual objects and visual concepts in an image, modeling the links between concepts and images.

In this cross-modality Transformer, each object o_j is represented by its ROI feature f_j and a segment embedding indicating this token is a visual object:

$$\begin{aligned}\tilde{f}_j &= \text{LayerNorm}(W_V f_j + b_V)\\ \tilde{s}_j &= \text{LayerNorm}(\text{SegEmbed}(o_j))\\ \tilde{v}_i &= (\tilde{f}_i + \tilde{s}_i)/2\end{aligned} \tag{8}$$

Each concept c_k is sent to a word embedding layer and then projected through a feed-forward layer to learn a textual concept embedding, and a segment embedding is added to it to indicate this token is a visual concept:

$$\begin{aligned}\tilde{w}_k &= \text{WordEmbed}(c_k)\\ \tilde{t}_k &= \text{LayerNorm}(W_T\tilde{w}_k + b_T)\\ \tilde{s}_k &= \text{LayerNorm}(\text{SegEmbed}(c_k))\\ \tilde{c}_k &= (\tilde{t}_k + \tilde{s}_k)/2\end{aligned} \tag{9}$$

If a concept consists of several words, an average pooling is applied to get the textual embedding. The word embedding layer shares weight with the layer in extracting sentence representations, as in Equation 6. This characterizes the links between concepts and sentences when concepts are directly mentioned in those sentences.

Then all objects and concepts in each image are sent into a N_C-layer cross-modality Transformer:

$$\begin{aligned}\{e_1^0, ..., e_{2\mu}^0\} &= \{\tilde{v_1}, ..., \tilde{v}_\mu, \tilde{c}_1, ..., \tilde{c}_\mu\}\\ \{e_1^{l+1}, ..., e_{2\mu}^{l+1}\} &= \text{Transformer}_T^l(\{e_1^l, ..., e_{2\mu}^l\})\end{aligned} \tag{10}$$

	train/val/test	n_i/m_i	# imgs	density
MSCOCO	25K/2K/2K	10/10	83K	5%
Story-DII	22K/3K/3K	5/5	47K	20%
Story-SIS	37K/5K/5K	5/5	76K	20%

Table 1: Dataset Statistics: density refers to edge density in the ground-truth graph for a document, i.e. the number of ground-truth edges divided by the number of all possible edges ($n_i * m_i$)

Similarly, an average pooling layer on all objects and concepts outputted by the last Transformer layer $\{e_1^{N_C}, ..., e_{2\mu}^{N_C}\}$ is used to extract the representation of an image.

Experiments

We evaluate our unsupervised training strategy on the task of multi-model link prediction in multi-image, multi-sentence documents proposed in (Hessel, Lee, and Mimno 2019).

Multi-image Multi-sentence Linking

For a document $d_i = \langle S_i, V_i \rangle$ consists of a set S_i of $|S_i|$ sentences and a set V_i of $|V_i|$ images, we aim to predict the label for each pair of image and sentence within the document. We generate features for all images and sentences based on trained cross-modality representation learning model and compute the similarity matrix $\hat{M}_i$ where the $(i, j)^{th}$ element is the cosine similarity between the i^{th} sentence representation and j^{th} image representation.

Experiment Datasets

We evaluate our proposed method on MSCOCO (Lin et al. 2014) and VIST (Huang et al. 2016). Hessel, Lee, and Mimno (2019) collect images and sentences to compose documents from those crowdlabeled dataset, then construct 3 different datasets for intra-document link prediction: MSCOCO, DII, and SIS. In MSCOCO, each document consists of 5 randomly sampled image-caption pairs, 5 distractor images, and 5 distractor sentences. In DII and SIS, each document consists of 5 images from the same album and 5 sentences of the corresponding description-in-isolation (DII) or story-in-sequence (SIS) story. Statistics of these datasets are given in Table 1.

Implementation Details

For images, we use a Faster-RCNN pre-trained on Visual Genome (Krishna et al. 2017) provided by (Anderson et al. 2018). In DII and SIS, we extract 36 objects and concepts for each image, while the number is adaptive in MSCOCO. We set $d_{multi} = 1024$, $N_T = 3$ and $N_C = 3$. Each layer in the single-modality Transformer and cross-modality Transformer has 8 heads. Mask in Transformers is used to deal with sequences with variable lengths. For training objective, in function TK, we set $k = \min(n_i, m_i)$. The sub-document proportion p_{sub} is set as 0.6 in SIS and DII, 0.8 in MSCOCO. Margin α in hinge loss is set to 0.2. We train our model using Adam optimizer (Kingma and Ba 2014). With a warm-up phase, we linearly increase the learning rate from 1e-7 to the configured max learning rate after several steps. In MSCOCO, max learning rate and warm-up steps are set as 1e-5 and 3000 respectively, while they are set as 5e-5 and 4000 in SIS and DII. After the warm-up phase, we decrease the learning rate by a factor of 5 each time the total loss over the validation set plateaus for more than 3 epochs. The mini-batch size is 11.

Overall Performance

We compare our model with the model proposed in (Hessel, Lee, and Mimno 2019), which are the only existing unsupervised model for this task, 2 baseline models are also proposed by Hessel, Lee, and Mimno (2019) and listed here for a comparison:

- **Object Detection** Each image is represented by the average of the word2vec embeddings of its K most probable labels predicted by pretrained DenseNet169 (Huang et al. 2017), while each sentence is represented by the average of the word2vec embeddings of its words.
- **NoStruct** randomly samples image-caption pairs from a document and treat the similarity between them as the document-level similarity.
- **MulLink** (Hessel, Lee, and Mimno 2019) uses a GRU-CNN based backbone model to encode images and sentences, and it is trained only using the loss in Equation 1.

Table 2 shows the comparative results. Both MulLink and our model show a superior performance than baseline models, reflecting their capability to measure cross-modality similarity more efficiently and sample effective image-sentences pairs in a structural document under the unsupervised setting. In general, Our approach outperforms MulLink in all 3 datasets, which means our proposed sampling strategy and fine-grained backbone model help to improve the performance jointly.

In MSCOCO, there is nearly no bias between intra-document and cross-document negative image-sentence pairs, due to the composition of documents, so MulLink has already achieved nearly perfect performance on the AUC metric. Our model still shows superior performance on $P@1$ and $P@5$, which means MulLink gets troubled in distinguishing some negative intra-document pairs, and our backbone model has a stronger ability to measure cross-modality in finer granularity. In DII, images and sentences in a document are more similar, i.e. a larger differences between cross-document and intra-document image-sentence pairs. In such a setting, distinguishing positive and negative intra-document samples requires fine-grained cross-modality similarity measurement, our approach therefore has obvious improvements on all metrics. In SIS, the main challenge is how to understand each sentence in a story, sentences are tightly related to each other and may even use pronouns to refer to words in other sentences. Our model still outperforms MulLink.

Ablation Study

To better understand the influence of each module in our approach, we perform the ablation study on the DII dataset,

	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**

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

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

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

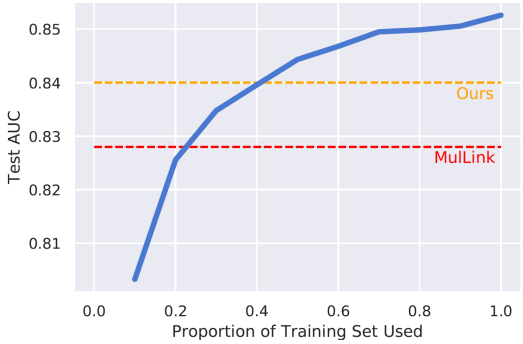


Figure 5: Performance of supervised strategy using different proportions of training data, dashed lines denote performances of unsupervised strategies.

and results are shown in Table 3. In those variations without Transformer, we use a softmax pooling to aggregate all objects (and concepts) features to represent an image, where weights for softmax are computed by a linear layer, w/o concepts means only a sequence of object features are sent to backbone models. Several findings stand out:

- Generally, it is showed that each objective contributes to the performance. Cross-document objective ("C") is the main part since it directly leverages document-level co-occurrence information, other 2 are supplementary objectives to sample more examples with respect to reasonable assumptions, therefore the performance is not satisfying when only using 2 supplementary objectives (see row 8). Intra-document objective ("I") helps to alleviate the bias, dropout sub-document objective ("D") aims to introduce randomness and discover weak cross-modal association, both of them utilize more information and enhance the performance, and the combination of 3 objectives helps the model reach the best performance.
- Without Transformer, just aggregating the concept features into the image representation does not improve performance (see row 2, 3), showing that the implicit graph between concepts and objects modeled by Transformer is necessary to extract better image representations.
- Incorporating concepts into Transformer significantly improves performance on precision (see row 1, 2). Illustrating that modeling of the dependency between objects and concepts is effective. An intuitive case is that our model will easily detect the cross-modal association if sentences involve classes of objects that appear in images.

Further Analysis

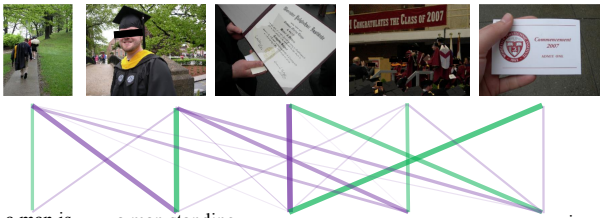
Bias Alleviation

Our proposed sampling strategy aims to alleviate the bias between cross-document training and intra-document evaluation, we conduct error analysis to show the effectiveness of our approach more intuitively. As the "spread" hypothesis in (Hessel, Lee, and Mimno 2019), documents with lower diversity among images/sentences are harder to disambiguate at test time. This hypothesis corresponds to our idea, lower intra-document diversity is equivalent to larger bias between intra-document and cross-document image-sentence pairs, since cross-document image-sentence pairs are always totally uncorrelated.

So we follow the error analysis setting for the "spread" hypothesis, we use DenseNet169 features for images and mean word2vec for sentences, then compute the mean squared distance to their centroid to quantify the spread of a document. An OLS regression of image spread + text spread on test AUC scores is fitted and its R-Square statistic shows how much of the variance in AUC can be explained by the intra-document spread. For DII and SIS, our approach reduces the R-Square from 42% to 26% and 23% to 12% respectively. This experiment does not involve MSCOCO since AUC scores are all large.

These results illustrate that our approach weakens the influence of intra-document diversity (bias between training and evaluating). Accompanied by the superior overall performance, it is strong proof of our approach's effectiveness to alleviate the bias, under the unsupervised setting.

(a) **SIS**; 76 AUC; Concepts: *chair*, *pillow*, *computer*, table, light



(b) **DII**; 87 AUC; Concepts: *umbrella*, *man*, *hat*, hand, *paper*

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 $\hat{M}_i$ (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 the average AUC in corresponding datasets.

Method	AUC	p@1/p@5
1 Transfer from MSCOCO	78.6	66.5/49.5
2 Unsupervised	85.5	77.2/60.1

Table 4: Performance of different methods on DII without explicit labels.

Comparison with Supervised Strategy

To show the efficiency of our unsupervised sampling strategy, we compare the performance with a supervised model, CNN-RNN is used as the backbone model. We vary the proportion of samples used to train supervised models and present the results in Figure 5. It reveals that the difference between fully trained supervised and unsupervised strategies is not that large. And it needs more than 40% of samples for the supervised strategy to generate better performance than our unsupervised approach (20% to beat *MulLink*).

In addition, we compare our model with a supervised model in the setting of transfer learning. We train the Transformer-based model on MSCOCO with ground-truth image-sentence pairs and test it on DII. Results can be seen in Table 4, without ground-truth labels in the target domain, our unsupervised method shows a better performance.

Case Study

To show the effectiveness of using more information provided by additional intra-document samples and appropriate model architecture, we present two illustrative examples in Figure 6, the form of illustration is the same as in (Hessel, Lee, and Mimno 2019). It shows that our models are able to discover fine-grained association by detecting and utilizing objects and corresponding concepts.

Related Work

Image-sentence matching is one of the fundamental tasks in the field of vision and language (Nam, Ha, and Kim 2017; Huang et al. 2018). A rich line of early studies focus on one-to-one matching (Yan and Mikolajczyk 2015; Klein et al. 2015; Faghri et al. 2017; Gu et al. 2018), usually extract global representations for image and sentence, then measure their similarities in a joint semantic space through. With the success of deep learning, employing CNN and RNN as modality-specific encoders becomes the mainstream. To learn an aligned multimodal semantic space where matched image-sentence pairs have small distances or high similarities, proposed training strategies usually use triplet ranking loss (Yan and Mikolajczyk 2015; Kiros, Salakhutdinov, and Zemel 2014; Klein et al. 2015; Peng and Qi 2019), while hard negative mining is showed to significantly improve the performance in (Faghri et al. 2017).

To capture fine-grained cross-modality association, most existing many-to-many matching methods try to incorporate relationships between image regions and sentence words (Karpathy, Joulin, and Fei-Fei 2014; Karpathy and Fei-Fei 2015; Huang, Wang, and Wang 2017; Lee et al. 2018; Wu et al. 2019). Some works align image segments and portions of a sentence without explicit labels (Karpathy, Joulin, and Fei-Fei 2014; Karpathy and Fei-Fei 2015; Rohrbach et al. 2016; Datta et al. 2019).

Generally, most previous works follow a retrieval paradigm within a large dataset (Lin et al. 2014; Young et al. 2014), where images and sentences are independent. Hessel, Lee, and Mimno (2019) formulate the task of multimodal intra-document links prediction in multi-image multi-sentence documents, some documents are collected from the datasets of visual storytelling, which is another task requiring modeling for intra-document dependency (Huang et al. 2016; Wang et al. 2020).

Conclusion and Future Work

In this work, we focus on the problem of unsupervised image-sentence matching. In order to alleviate the sampling bias introduced by the existing unsupervised training strategy, we propose a new sampling strategy to efficiently sample additional positive and negative intra-document samples. In addition, we propose to use a Transformer based model to learn cross-modality representations for images and sentences. Our approach improves the matching accuracy of an unsupervised multimodal link prediction task across different datasets. In the future, we would like to explore more downstream tasks using our unsupervised matching strategy. Besides, it is interesting to investigate few-shot semantic concept detection in an unsupervised way.

Acknowledgments

This work is partially supported by Ministry of Science and Technology of China (No.2020AAA0106701), Science and Technology Commission of Shanghai Municipality Grant (No.20dz1200600, 17JC1420200). We would also like to thank Ruize Wang and the anonymous reviewers for their constructive feedback.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Datta, S.; Sikka, K.; Roy, A.; Ahuja, K.; Parikh, D.; and Divakaran, A. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, 2601–2610.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* .

Fan, Z.; Wei, Z.; Wang, S.; and Huang, X.-J. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6514–6524.

Gu, J.; Cai, J.; Joty, S. R.; Niu, L.; and Wang, G. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7181–7189.

Hessel, J.; Lee, L.; and Mimno, D. 2019. Unsupervised Discovery of Multimodal Links in Multi-Image, Multi-Sentence Documents. In *EMNLP*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, T.-H. K.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Devlin, J.; Agrawal, A.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual Storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Huang, Y.; Wang, W.; and Wang, L. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2310–2318.

Huang, Y.; Wu, Q.; Song, C.; and Wang, L. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6163–6171.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.

Karpathy, A.; Joulin, A.; and Fei-Fei, L. F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, 1889–1897.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* .

Klein, B.; Lev, G.; Sadeh, G.; and Wolf, L. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4437–4446.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123(1): 32–73.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 201–216.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Nam, H.; Ha, J.-W.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 299–307.

Peng, Y.; and Qi, J. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15(1): 1–24.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 817–834. Springer.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, R.; Wei, Z.; Li, P.; Zhang, Q.; and Huang, X. 2020. Storytelling from an Image Stream Using Scene Graphs. In *AAAI*, 9185–9192.

Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; and Shao, J. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, 5764–5773.

Wu, Y.; Wang, S.; Song, G.; and Huang, Q. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2088–2096.

Xu, P.; Joshi, C. K.; and Bresson, X. 2019. Multi-graph transformer for free-hand sketch recognition. *arXiv preprint arXiv:1912.11258* .

Yan, F.; and Mikolajczyk, K. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3441–3450.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651–4659.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2: 67–78.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020. Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16(2): 1–23.