Constructing Phrase-level Semantic Labels to Form Multi-Grained Supervision for Image-Text Retrieval

Anonymous ACL submission

Abstract

Existing research for image text retrieval mainly relies on sentence-level supervision to distinguish matched and mismatched sentences for a query image. However, semantic mismatch between an image and sentences usually happens in finer grain, i.e., phrase level. In this paper, we explore to introduce additional phrase-level supervision for the better identification of mismatched units in the text. In practice, multi-grained semantic labels are automatically constructed for a query image in both sentence-level and phrase-level. We construct text scene graphs for the matched sentences and extract entities and triples as the phrase-level labels. In order to integrate both supervision of sentence-level and phrase-level, we propose **S**emantic **S**tructure **A**ware **M**ultimodal **T**ransformer (SSAMT) for multi-modal representation learning. Inside the SSAMT, we utilize different kinds of attention mechanisms to enforce interactions of multi-grain semantic units in both sides of vision and language. For the training, we propose multi-scale matching losses from both global and local perspectives, and penalize mismatched phrases. Experimental results on MS-COCO and Flickr30K show the effectiveness of our approach compared to some state-of-the-art models.

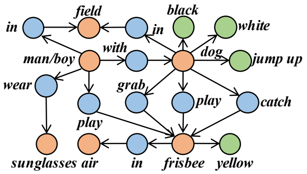
Image		*Matching Score*
Negative Text	*N1: the dog dumps over a man to catch a frisbee.*	*0.50*
	N2: two dogs in a field playing with a frisbee.	*0.40*
	N3: a brown and black dog jumps in the air to catch a frisbee.	*0.53*
	N4: a dog catching a frisbee on a baseball field.	*0.49*
	N5: a dog jumping up to catch a frisbee in a tree.	*0.44*
Positive Text	*P1. a man wearing sunglasses plays frisbee with a dog.*	*0.46*
	P2: a black and white dog jumps up to grab a yellow frisbee.	*0.51*
	P3: the man is playing frisbee with his dog on the field.	*0.48*
	P4: a boy playing catch with a frisbee in a field with a dog.	*0.42*
	P5: a black and white dog is catching a frisbee in the air.	*0.52*
Text Scene Graph		
Phrase-level Semantic Label	*Object-Relation-Object Triplet: (dog play frisbee), (man wear sunglass),…*	
	Attribute-Object Pair: (black white dog)	
	Object-Entity: man, dog, frisbee, sunglasses, …	

Figure 1: An example of a query image, a group of mismatched sentences, a group of matched sentences and their corresponding text scene graph, and augmented labels in phrase-level. Textual segments with underlines stand for mismatching in phrase-level. The matching scores are produced by VSE++ (Faghri et al., 2018).

1 Introduction

Vision and language are two important aspects of human intelligence to understand the world. To bridge vision and language, researchers pay increasing attention to multi-modal tasks. Image-text retrieval (Frome et al., 2013b), one of the fundamental topics, aims to retrieve the matching text (image) for the query image (text). Researchers (Frome et al., 2013b; Kiros et al., 2014; You et al., 2018) extract features from an image-text pair to compute a scalar matching score to measure the similarity. The model is optimized via a triplet loss that makes the representations of the positive image-text pair closer than negative ones. Existing research (Faghri et al., 2018; Lee et al., 2018; Liu et al., 2020; Wei et al., 2020) usually relies on sentence-level supervision for cross-modality representation learning. However, the semantic mismatch usually happens in finer grain, i.e., phrase level.

We show an example in Figure 1, including a query image, some matched sentences and mismatched ones. In terms of matching scores, the model (Faghri et al., 2018) fails to distinguish positive and negative sentences. A closer look at the example shows that mismatched sentences are usually partially irrelevant with phrases of inconsistent semantics (two dogs, baseball filed, etc.). Inspired

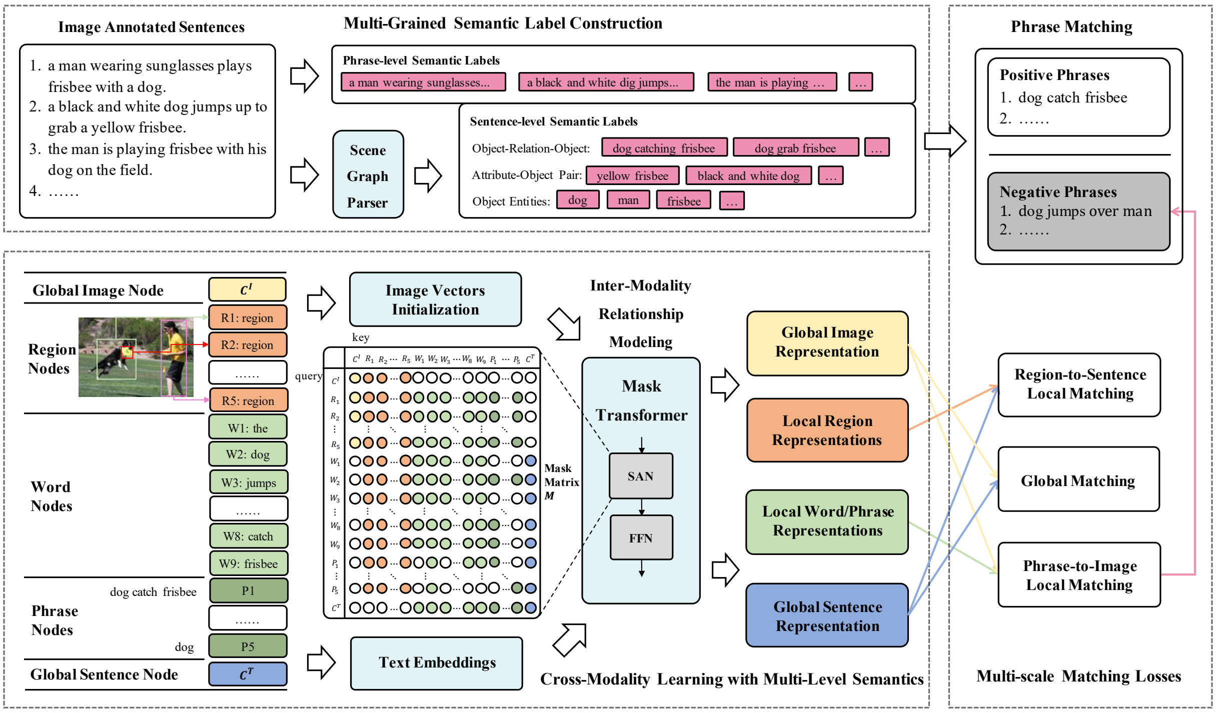


Image Annotated Sentences
1. a man wearing sunglasses plays frisbee with a dog.
2. a black and white dog jumps up to grab a yellow frisbee.
3. the man is playing frisbee with his dog on the field.
4. ……
Multi-Grained Semantic Label Construction
Phrase-level Semantic Labels
a man wearing sunglasses...
a black and white dig jumps...
the man is playing ...
...
Scene Graph Parser
Sentence-level Semantic Labels
Object-Relation-Object:
dog catching frisbee
dog grab frisbee
Attribute-Object Pair:
yellow frisbee
black and white dog
Object Entities:
dog
man
frisbee
Phrase Matching
Positive Phrases
1. dog catch frisbee
2. ……
Negative Phrases
1. dog jumps over man
2. ……
Global Image Node
Region Nodes
Word Nodes
Phrase Nodes
Global Sentence Node
R1: region
R2: region
R5: region
W1: the
W2: dog
W3: jumps
W8: catch
W9: frisbee
dog catch frisbee
P1
dog
P5
Image Vectors Initialization
Text Embeddings
key
query
Mask Matrix M
Inter-Modality Relationship Modeling
Mask Transformer
SAN
FFN
Global Image Representation
Local Region Representations
Local Word/Phrase Representations
Global Sentence Representation
Cross-Modality Learning with Multi-Level Semantics
Region-to-Sentence Local Matching
Global Matching
Phrase-to-Image Local Matching
Multi-scale Matching Losses

Figure 2: The overall framework of our proposed model Semantic Structure Aware Multimodal Transformer (SSAMT). The blank circle in the mask matrix M means that the query node does not attend to the corresponding node.

by this observation, we explore to provide fine-grained supervision in phrase-level for better cross-modality representation learning. In practice, we construct multi-grained semantic labels for a query image of two levels, namely, sentence-level and phrase-level. In sentence-level, we use the whole sentence as the label. In phrase level, we construct the text scene graph of the sentence and extract entities and triples of multiple forms from the graph as labels. Based on these multi-grained semantic labels, we assume the matching model is able to identify fine-grained mismatched semantic units at the same time of distinguishing negative sentences.

In order to utilize the supervision of both sentence-level and phrase-level for cross-modality representation learning, we propose the Semantic Structure Aware Multimodal Transformer (SSAMT) to model multi-grained semantics in vision and language. In language side, we concatenate the sentence and its phrases as input, while both image and its regions are used in vision side. Mask transformer is used to model semantic units of different granularity for both modalities, besides, novel attention mechanisms are presented for interactions of intra-modality and inter-modality. The model learns representations for both modalities of vision (image and regions) and language (sentence and phrases) in multiple scales (global and local). For optimization, we utilize the global matching and local matching for the similarity measurement of image-text pairs, where global matching computes the matching score of the global representations of the image and text, and local matching measures the similarities from fine-grained perspectives including region-to-text and phrase-to-image. In addition, for the phrases extracted from mismatched sentences, we propose phrase-matching to teach the model to increase scores between the matched image-phrase pairs and decrease those mismatched ones. Experiment results on MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) show the effectiveness of our model compared to some state-of-the-art approaches. Further analysis reveals that SSAMT is able to provide better interpretability by locating mismatched phrases of negative sentences.

2 Semantic Structure Aware Multimodal Transformer (SSAMT)

The overall framework of Semantic Structure Aware Multimodal Transformer is shown in Figure 2. It includes three major components, namely, multi-grained semantic labels construction, cross-modality representation learning with multi-grained semantics and multi-scale matching losses. The multi-grained semantic labels construction is to automatically collect semantic labels from annotated sentences of the query image, the cross-modality representation learning with multi-grained semantics is to capture the semantics of different granularity in both modalities, and multi-scale losses are utilized to measure the similarity between the pair of image and sentence. We take the image I_i and text T_j as an example to compute their matching score.

2.1 Multi-Grained Semantic Labels Construction

Each image in vision and language datasets has multiple annotated sentences, for example there are five in MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015). These sentences describe multi-grained semantics of the image including various objects, relations and scenes, and we propose to utilize them to automatically configure corresponding semantic labels.

In practice, we adopt the scene graph parser of SPICE (Anderson et al., 2016) following SGAE[1] (Yang et al., 2019b) to dig object-relation-object triplets, object-attribute pairs and object entities from upper descriptive sentences. For example in Figure 2, the retrieved phrases include "dog catching frisbee" and "yellow frisbee". Moreover, tokens of each sentence are also collected as supplementary of the upper phrases. There phrases and tokens are regarded as semantic labels, L_i, of the image I_i. Semantic labels provide an opportunity to determine the partially irrelevant components in the sentence T_j and we will introduce the details in §2.4.

2.2 Model Input

Text Embeddings For the sentence T_j, we obtain its tokens and phrases as described in §2.1. To initialize each token, we map it to a dense vector using a standard embedding layer following Devlin et al. (2019) which is composed of token embeddings, position embeddings and segment embeddings. For each phrase, we use phrase segment embeddings of three categories, namely, object, attribute and relation as initialization, and think of it as a phrase node that connects to these tokens included in it. We concatenate them with tokens for simultaneous encoding. We do not add context-based embeddings for phrase initialization and introduce the mask mechanism in §2.3 to encode its context as compensation. At the same time, we set up a global sentence node with dense vector C^T as initialization to capture the sentence-level semantics. In summary, our text embeddings have three parts, word embeddings E^W_j, phrase embeddings E^P_j and a global sentence embedding C^T.

Image Vectors Initialization For the image I_i, we employ a pre-trained object detector to extract region features, where each $o_{i,k} \in \mathbb{R}^{d_o}$ is the mean-pooled convolutional feature for the k-th region of I_i and d_o is the hidden size of the detector. We fix the pre-trained model during training. To fit the hidden size of our encoder, we add a fully-connected layer to project each region feature into the same size and get initial image vectors E^I_j. Following the setting of the global sentence node, we also set up a global image node with C^I as initialization to capture the overall semantics of the image.

2.3 Cross-Modality Modeling with Multi-Grained Semantics

To enforce the interaction of multi-grained semantics from both two modalities, we employ inter-modality and intra-modality relationships modeling at the same time, and present the mechanism of mask attention inside the transformer cell to learn the multi-grained semantics with the inherent structure.

Inter-Modality Relationship Modeling Inter-modality relationship model aims to set up the interactions across the two modalities. We use the encoder of transformer (Vaswani et al., 2017) as backbone. In the following equation, we concatenate text embeddings and image vectors in §2.2 as model input.

$$H^0 = [C^T, E^P_j, E^W_j; C^I, E^I_i] \tag{193}$$

In the original setting of transformer, there is no different granularity or structure and each element attends to others without constraints. In our case, we have phrase node to capture the semantics of words in the phrase, and modality-dependent global nodes to model the overall semantics for image and text, respectively. These nodes are utilized for multi-grained semantic modeling and heavily depends on the structure. To keep their meaning, we argue to abandon original attention and employ

[1] https://github.com/yangxuntu/SGAE

mask attention. In implementation, a masking matrix $M \in \mathbb{R}^{|H^0| \times |H^0|}$ is initialized with all 0, and we reset values in specific positions with $-\infty$ to meet these three requirements as following.

(1) In the vision side, each region node is not visible to the global sentence node.

(2) In the language side, each phrase and token node is not visible to the global image node.

(3) Each phrase node is not visible to any other words that not included in the phrase itself.

We add M to the following attention function and utilize it to replace the original one in transformer. We call the new one mask transformer.

$$\text{attention}(Q, K, V, M) = \text{softmax}\left(M + \frac{Q^T K}{\sqrt{d_k}}\right)V$$

After inter-modality relationship modeling, we get a sequence of outputs shown in the following equation.

$$H = \left[h_j^T, H_j^T; h_i^I, H_i^I\right]$$

where h_i^I and h_j^T are global representations for image and text corresponding to global nodes C^I and C^T of image and sentence. H_i^I are representations for regions and H_j^T are for phrases and words. They are local representations for the image and sentence, respectively.

Intra-Modality Relationship Modeling Intra-modality relationship model (Wei et al., 2020; Yang et al., 2019a) is employed to separately encode image and text as a supplementary to the inter-modality relationship modeling, where inputs of image and text are $\left[C^I, E_i^I\right]$ and $\left[C^T, E_j^P, E_j^W\right]$, respectively. We take the outputs of C^I and C^T as intra-modality global representations of the image and text, denoted as a_i^I and a_j^T.

2.4 Multi-scale Matching Losses

Supposing we have a positive image-text pair (I_i, T_i) with a negative image I_k and a negative sentence T_j. We use triplet loss TriL_α to train our model. In $\text{TriL}_\alpha(u, V, W)$ as following, α is a scalar to control the distance between the cosine score of u and positive samples V and that of negative samples W. The loss is to push $v \in V$ closer to u and push $w \in W$ away from u. Based on multi-grained semantic labels, We measure the similarity of these image-text pairs using three kinds of matching scores, including global, local, and phrase matching.

$$\begin{aligned}&\text{TriL}_\alpha(u, V, W)\\ &= max\left(\alpha - \sum_{v \in V} \frac{cos(u, v)}{|V|} + \sum_{w \in W} \frac{cos(u, w)}{|W|},\ 0\right)\end{aligned}$$

Global Matching Intra-modality and inter-modality relationship modeling both produce representations for global representations of image and sentence, then we have $cos(a_i^I, a_i^T)$ and $cos(h_i^I, h_i^T)$ to measure global similarity of the positive image-text pair (I_i, T_i), and so is for the negative pair (I_i, T_j). The corresponding loss is shown below.

$$\begin{aligned}\mathcal{L}_{\mathit{0}}^G &= \text{TriL}_{\alpha_0}(a_i^I, a_i^T, a_j^T) + \text{TriL}_{\alpha_0}(a_i^T, a_i^I, a_k^I)\\ \mathcal{L}_{\mathit{1}}^G &= \text{TriL}_{\alpha_1}(h_i^I, h_i^T, h_j^T) + \text{TriL}_{\alpha_1}(h_i^T, h_i^I, h_k^I)\end{aligned}$$

Local Matching We utilize local matching which is based on inter-modality relationship modeling to enhance the fine-grained cross-modal matching. It has two parts. (1) Region-to-Sentence: The matching between each region and the sentence. (2) Phrase-to-Image: The similarity of each phrase (token) and the image. We employ $\mathcal{L}_{\mathit{2}}^L$ to make local matching scores of the positive image-text pair larger than the negative one.

$$\mathcal{L}_{\mathit{2}}^L = \text{TriL}_{\alpha_2}(h_i^I, H_i^T, H_j^T) + \text{TriL}_{\alpha_2}(h_i^T, H_i^I, H_k^I)$$

Phrase Matching On the basis of phrase-to-image matching, we employ semantic labels L_i to determine mismatched phrases in negative sentences, and decrease the scores of the mismatched ones and increase the scores of matched ones. In detail, for each phrase or token in T_j, we determine it as positive if it appears in semantic labels L_i or included in some label of L_i, otherwise, it is negative. Through the method, we split the token and phrase of T_j into positive ones H_{j+}^T and negative ones H_{j-}^T. We repeat the same process on the negative pair (I_k, T_i) and get H_{i+}^T and H_{i-}^T. Consider that positive parts are keys to separate mismatched image text pair, we propose $\mathcal{L}_{\mathit{3}}^P$ to further push away negative parts against positive ones in the negative sentence. It also can be interpreted as the penalty on mismatched parts, which is to guide the matching model to make decisions more grounding on them.

$$\mathcal{L}_{\mathit{3}}^P = \text{TriL}_{\alpha_3}(h_i^I, H_{j+}^T, H_{j-}^T) + \text{TriL}_{\alpha_3}(h_k^I, H_{i+}^T, H_{i-}^T)$$

	MS-COCO 1K							Flickr30K						
Model	Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
	R@1	R@5	R@10	R@1	R@5	R@10	RSum	R@1	R@5	R@10	R@1	R@5	R@10	RSum
VSE++	64.7	-	95.9	52.0	-	92.0	-	52.9	79.1	87.2	39.6	69.6	79.5	407.9
CAMP	72.3	94.8	98.3	58.5	87.9	95.0	506.8	68.1	89.7	95.2	51.5	77.1	85.3	466.9
SCAN	72.7	94.8	**98.4**	58.8	88.4	94.8	507.9	67.4	90.3	95.8	48.6	77.7	85.2	465.0
SGM	73.4	93.8	97.8	57.5	87.3	94.3	504.1	71.8	91.7	95.5	53.5	79.6	86.5	478.6
VSRN	74.0	94.3	97.8	60.8	88.4	94.1	509.4	70.4	89.2	93.7	53.0	77.9	85.7	469.9
BFAN	74.9	95.2	-	59.4	88.4	-	-	68.1	91.4	-	50.8	78.4	-	-
MMCA	74.8	95.6	97.7	61.6	**89.8**	95.2	514.7	74.2	**92.8**	96.4	54.8	81.4	87.8	487.4
GSMN	76.1	95.6	98.3	60.4	88.7	95.0	514.1	71.4	92.0	96.1	53.9	79.7	87.1	480.2
SSAMT	**78.2**	**95.6**	98.0	**62.7**	89.6	**95.3**	**519.4**	**75.4**	92.6	**96.4**	**54.8**	**81.5**	**88.0**	**488.7**

Table 1: Comparison results of the cross-modal retrieval on the MS-COCO 1K and Flickr30K in terms of Recall@K(R@K). The comparative models include VSE++ (Faghri et al., 2018), CAMP (Wang et al., 2019b), SCAN (Lee et al., 2018), SGM (Wang et al., 2020), VSRN (Li et al., 2019), BFAN (Liu et al., 2019), MMCA (Wei et al., 2020), GSMN (Liu et al., 2020).

Based on these three types of matching methods and corresponding losses, we get the overall loss with hyperparameters λ_0, λ_1, λ_2 and λ_3 to balance these losses.

$$\mathcal{L}^S = \lambda_0 \mathcal{L}^G_0 + \lambda_1 \mathcal{L}^G_1 + \lambda_2 \mathcal{L}^L_2 + \lambda_3 \mathcal{L}^P_3$$

Previous image-text retrieval models usually take the hardest image (text) from in-batch data as the negative image (text), which requires the matching score computation of all pairwise image-text combinations in batch. This is expensive in inter-modality relationship modeling, thus we sample negative instances through intra-modality matching scores to reduce the computation cost.

Inference During inference, we utilize the following score(I_i, T_j) for ranking.

$$\text{score}(I_i, T_j) = cos(a^I_i, a^T_j) + \mu_1 cos(h^I_i, h^T_j) + \sum_{h^T_{j,k} \in H^T_j} \frac{\mu_2 cos(h^I_i, h^T_{j,k})}{|H^T_i|} + \sum_{h^V_{i,k} \in H^I_i} \frac{\mu_2 cos(h^T_j, h^I_{i,k})}{|H^I_i|}$$

where μ_1 and μ_2 are hyperparameters.

3 Experiment

3.1 Experimental Setup

Datasets We evaluate our proposed model on MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015). Each image of MS-COCO is accompanied with 5 human annotated captions. We split the dataset into training, validation and test sets respectively with $113,287/5,000/5,000$ images following (Karpathy and Fei-Fei, 2015). For MS-COCO 1K, the testing set is further divided into 5 splits and the performance reported are the average over the 5 folds of 1K test images (Faghri et al., 2018). Flickr30K (Plummer et al., 2015) consists of 31000 images collected from the Flickr website. Each image contains 5 descriptive sentences. We take the same splits for training, validation and testing sets as in Karpathy and Fei-Fei (2015), 1000 images for validation and 1000 images for testing, while the rest for training.

Model	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CAMP	50.1	82.1	89.7	39.0	68.9	80.2
SCAN	50.4	82.2	90.0	38.6	69.3	80.4
SGM	50.0	79.3	87.9	35.3	64.9	76.5
MMCA	54.0	82.5	90.7	38.7	69.7	**80.8**
SSAMT	**57.7**	**84.2**	**90.8**	**40.8**	**70.5**	80.5

Table 2: Comparison results of the cross-modal retrieval on the MS-COCO in terms of Recall@K(R@K).

Evaluation Metric The performance of image text retrieval is evaluated by the standard recall at K (R@K), $K = 1, 5, 10$. It is defined as the fraction of queries for which the correct item belongs to the top-K retried items. In image text retrieval, we can take image or text as query to retrieve matched texts or images, corresponding to the settings of image-to-text and text-to-image. We also take RSum which is the sum of R@1+R@5+R@10 in both image-to-text and text-to-image as an overall metric.

3.2 Implementation Details

For the image, we employ Faster-RCNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017) to extract region features following BUTD[2] (Anderson et al., 2018). We prune the vocabulary by dropping words that appear less than five times. Both of our intra-relationship model and inter-relationship model has 2 layers, the hidden dimension is 1024, the head of attention is 16 and the inner dimension of feed-forward network is $2,048$. The number of parameters in our model is 64.7M. More details are in appendix.

3.3 Overall Performance

We compare our model with some classic and state-of-the-art approaches, including VSE++ (Faghri et al., 2018), CAMP (Wang et al., 2019b), SCAN (Lee et al., 2018), SGM (Wang et al., 2020), VSRN (Li et al., 2019), BFAN (Liu et al., 2019), MMCA (Wei et al., 2020), GSMN (Liu et al., 2020). The results on MS-COCO 1K and Flickr30K are presented in Table 1, and that on MS-COCO is shown in Table 2. We can see that our proposed SSAMT outperforms all existing methods, with the best R@1= 78.2% for image-to-text retrieval and R@1= 62.7% for text-to-image retrieval on MS-COCO 1K. For MS-COCO, the proposed approach maintains the superiority with an improvement of more than 3% on the R@1 of image-to-text retrieval. In Flickr30K, our model achieves the best performance with image-to-text R@1 of 75.4%.

3.4 Ablation Study

We perform ablation study on MS-COCO 1K to explore the effectiveness of phrase-level labels and different matching scores. We compare two groups of settings by controlling the usage of phrase-level labels, namely, mask transformer w/o phrases and w/ phrases. Under each group of settings, we take mask transformer (MT) with global matching loss (GM) as baseline. On top of it, we add local matching loss (LM) and phrase matching loss (PM) in sequence to justify their influences. Experiment results are shown in Table 3. We can see that performance increases as components are gradually added. Moreover, these models based on mask transformer w/ phrases perform better than their count-parts without phrases. These facts demonstrate the effectiveness of different components of SSAMT.

	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
	mask transformer w/o phrases					
MT+GM	73.9	93.7	97.5	57.7	87.0	94.1
+LM	75.5	94.3	97.7	59.8	87.9	94.5
+PM	76.3	94.7	98.0	61.2	88.8	94.8
	mask transformer w/ phrases					
MT+GM	74.5	94.1	97.3	58.2	87.7	94.4
+LM	76.8	95.0	97.9	61.4	88.7	94.6
+PM	**78.2**	**95.6**	**98.0**	**62.7**	**89.6**	**95.3**

Table 3: Ablation study for SSAMT on MS-COCO 1K. There are two groups of settings, namely, mask transformer w/o phrases and mask transformer w/ phrases. In each group, we take mask transformer with global matching loss as baseline. Components are added on top of the previous setting one by one from the first row to the bottom one.

4 Further Analysis

In this section, we dive into SSAMT to further analyze the characteristics of semantic labels.

4.1 Influence of Phrase-level Labels on Inference Power of Various Matching Scores

Multi-grained semantic labels are used in computing the multi-scale scores for a given image text pair as $\mathcal{L}_1^G$, $\mathcal{L}_2^L$ and $\mathcal{L}_3^P$. Experiments have shown their effectiveness on improving the overall performance in image-text retrieval. We would like to further explore contributions of different matching scores in the inference process. In addition to the full version of SSAMT, we also train SSAMT without phrase matching loss (denoted as SSAMT w/o PM) as comparison. SSAMT w/o PM has the same architecture as SSAMT and is able to compute phrase matching score during inference. In the experiment of image-to-text (i2t) setting, for each positive image text pair in the test set of MS-COCO (Lin et al., 2014), we sample another sentence to form a negative pair. So is text-to-image (t2i). We compute the mean of classification accuracy across the test set.

Experiment results are presented in Table 4. We use one of three matching scores (global, local and phrase) to make the decision of classification. Two findings standout. (1) The performance of SSAMT w/o PM drops significantly when it computes phrase matching to determine the irrelevant

[2] https://github.com/peteanderson80/bottom-up-attention

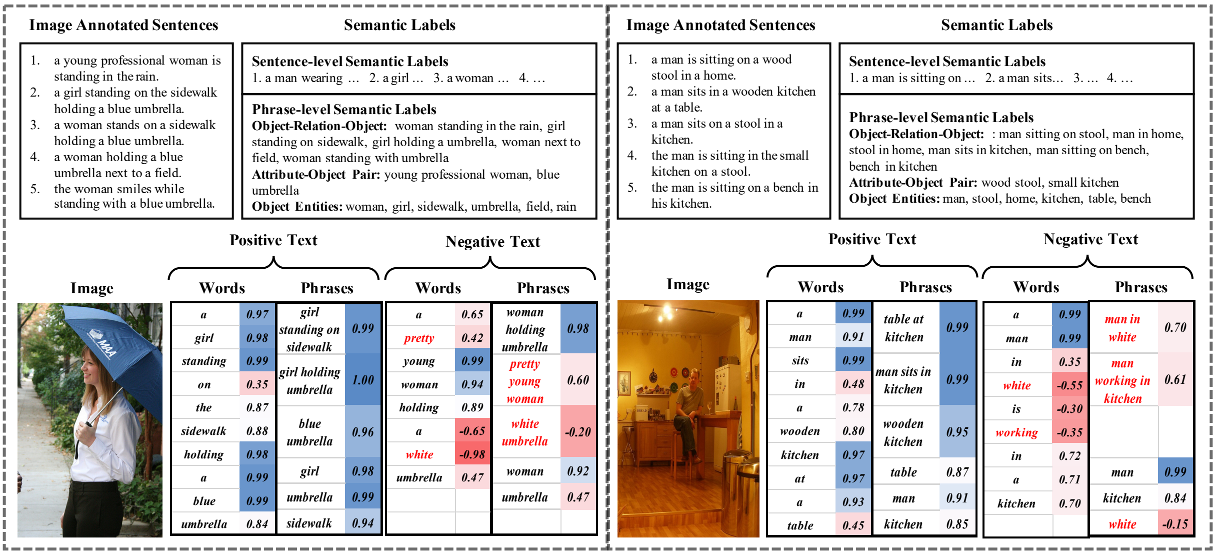


Image Annotated Sentences

1. a young professional woman is standing in the rain.
2. a girl standing on the sidewalk holding a blue umbrella.
3. a woman stands on a sidewalk holding a blue umbrella.
4. a woman holding a blue umbrella next to a field.
5. the woman smiles while standing with a blue umbrella.

Semantic Labels

Sentence-level Semantic Labels
1. a man wearing … 2. a girl … 3. a woman … 4. …

Phrase-level Semantic Labels
Object-Relation-Object: woman standing in the rain, girl standing on sidewalk, girl holding a umbrella, woman next to field, woman standing with umbrella
Attribute-Object Pair: young professional woman, blue umbrella
Object Entities: woman, girl, sidewalk, umbrella, field, rain

Image

Positive Text

Words	
a	*0.97*
girl	*0.98*
standing	*0.99*
on	*0.35*
the	*0.87*
sidewalk	*0.88*
holding	*0.98*
a	*0.99*
blue	*0.99*
umbrella	*0.84*

Phrases	
girl standing on sidewalk	*0.99*
girl holding umbrella	*1.00*
blue umbrella	*0.96*
girl	*0.98*
umbrella	*0.99*
sidewalk	*0.94*

Negative Text

Words	
a	*0.65*
pretty	*0.42*
young	*0.99*
woman	*0.94*
holding	*0.89*
a	*-0.65*
white	*-0.98*
umbrella	*0.47*

Phrases	
woman holding umbrella	*0.98*
pretty young woman	*0.60*
white umbrella	*-0.20*
woman	*0.92*
umbrella	*0.47*

Image Annotated Sentences

1. a man is sitting on a wood stool in a home.
2. a man sits in a wooden kitchen at a table.
3. a man sits on a stool in a kitchen.
4. the man is sitting in the small kitchen on a stool.
5. the man is sitting on a bench in his kitchen.

Semantic Labels

Sentence-level Semantic Labels
1. a man is sitting on … 2. a man sits… 3. … 4. …

Phrase-level Semantic Labels
Object-Relation-Object: : man sitting on stool, man in home, stool in home, man sits in kitchen, man sitting on bench, bench in kitchen
Attribute-Object Pair: wood stool, small kitchen
Object Entities: man, stool, home, kitchen, table, bench

Image

Positive Text

Words	
a	*0.99*
man	*0.91*
sits	*0.99*
in	*0.48*
a	*0.78*
wooden	*0.80*
kitchen	*0.97*
at	*0.97*
a	*0.93*
table	*0.45*

Phrases	
table at kitchen	*0.99*
man sits in kitchen	*0.99*
wooden kitchen	*0.95*
table	*0.87*
man	*0.91*
kitchen	*0.85*

Negative Text

Words	
a	*0.99*
man	*0.99*
in	*0.35*
white	*-0.55*
is	*-0.30*
working	*-0.35*
in	*0.72*
a	*0.71*
kitchen	*0.70*

Phrases	
man in white	*0.70*
man working in kitchen	*0.61*
man	*0.99*
kitchen	*0.84*
white	*-0.15*

Figure 3: Two examples of SSAMT. Red words in negative text mean they are negative that does not appear in semantic labels. Score on the right of each phrase or word is the corresponding phrase-to-image local matching score. Color of scores from blue to red denotes that phrases or words are more and more irrelevant to the image.

Matching Type	SSAMT	SSAMT w/o PM
global (i2t)	76.9%	75.2%
global (t2i)	83.9%	82.6%
local (i2t)	77.2%	75.4%
local (t2i)	84.2%	82.6%
phrase (i2t)	97.1%	30.0%
phrase (t2i)	97.2%	28.4%

Table 4: Accuracy of global matching, local matching and phrase matching with respect to SSAMT and SSAMT w/o PM. i2t and t2i mean image-to-text and text-to-image, respectively.

parts of the negative sentence. This indicates that if there is no explicit supervision to guide the model to be grounding on the truly irrelevant parts during training, it hardly unsupervised learns the grounding. (2) SSAMT outperforms SSAMT w/o PM in all six metrics. This demonstrates that, through guiding the model to be more grounding on the truly irrelevant parts, phrase-level labels contribute to all three matching during the inference.

4.2 Influence of Phrase-level Labels on the Improvement of Training Datab Efficiency

In the application of semantic labels, we exploit more supervision signals for each query image in phrase level. We would like to see the influence of these phrase-level labels on the training efficiency. We compare three versions of SSAMT, namely, SSAMT, SSAMT w/o PM and SSAMT w/o PM&LM, with different sets of matching losses respectively. In practice, we randomly pick out 75%, 50% and 25% data from Flickr30K (Plummer et al., 2015), which respectively contain 21, 750, 14, 500 and 7, 250 images and 108, 750, 72, 500 and 36, 250 image-text pairs. We present RSum score of three models on test set of Flickr30K for evaluation.

Model	100%	75%	50%	25%
SSAMT w/o PM&LM	483.3	462.4	436.5	372.5
SSAMT w/o PM	485.1	464.6	437.7	373.4
SSAMT	488.7	471.2	446.4	383.8

Table 5: RSum of different models with different sizes of training data.

Experiment results are shown in Table 5. We find that RSum of all three models decreases as dataset size decreases from 100% to 25%. The performance gain of SSAMT w/o PM over SSAMT w/o PM&LM gets smaller as the dataset size decreases, which ranges from 1.8 to 0.9. However, the gain of SSAMT over SSAMT w/o PM gets larger, which ranges from 3.6 to 10.4. This demonstrates that phrase-level labels can improve the training efficiency, and the improvement is more significant with a small dataset scale.

When the size of training dataset reduces from 29, 000 (100%) to 7, 250 (25%), the size of vocabulary changes from 9, 568 to 8, 608, of which the reduction is smaller than the dataset. It has less impact on phrase matching. This means that the supervision of phrase matching does not decrease that rapidly. Thus, semantic labels are more beneficial for image-text retrieval under the condition of small data amount.

4.3 Case Study

We show two examples in Figure 3. For each image, we show its annotated sentences on top of it. Based on these sentences, we follow the instruction in §2.1 to automatically construct corresponding multi-grained semantic labels, and list them on the right of these image annotated sentences. We show two texts that one is positive (matched) and the other is negative(mismatched) on the right of each image. Each of these texts is separated into words (left) and phrases (right), and they are accompanied with corresponding phrase-to-image local matching scores. We use blue and red to highlight positive and negative scores respectively, the darker the higher (lower in negative). We observe that phrase scores of the positive sentence are large, and those of negative one are relatively small, especially those mismatched parts, such as "white umbrella", "man in white" and "man working in kitchen". These facts verify that SSAMT can produce more faithful results.

5 Related Work

Most works in image-text retrieval focus on capturing the cross-modal semantic association by better feature extraction and cross-modality interaction. Nam et al. (2017), Fan et al. (2018), Ji et al. (2019) and Fan et al. (2019) represent the image by semantics gathered from block-based attention, or by region-level. Lee et al. (2018), Wang et al. (2019b), Wang et al. (2019a), Li et al. (2019), Wang et al. (2020) and Wei et al. (2020) detect objects in images by pre-trained Faster R-CNN (Ren et al., 2015) following the bottom-up manner proposed by Anderson et al. (2018). For text processing, Klein et al. (2015) explore to use Fisher Vectors as discriminative representations. Language models like Skip-Gram are employed to extract word representations (Frome et al., 2013a), and text segments are generally encoded by recurrent neural network (RNN) (Kiros et al., 2014; Faghri et al., 2018; Chen and Luo, 2020). To emphasize local structure of visual (text), VSRN (Li et al., 2019) uses the region relationship modeling for vision localness modeling and GRU (Cho et al., 2014) for visual global semantic modeling. SGM (Wang et al., 2020) employs GCN (Defferrard et al., 2016) to model visual and textual scene graph, then computes the similarity between two graphs. Besides, GSMN (Liu et al., 2020) fuses neighborhood associations for graph structure matching. Different from these GCN based methods, our model process token and phrases in parallel to jointly model in localness and globalness, which is simpler and more efficient. MMCA (Wei et al., 2020) utilizes transformer for cross-modality relationship encoding, which is widely applied in vision-language pre-trained models, such as UNITER (Chen et al., 2020), Unicoder-VL (Li et al., 2020) and ERINE-ViL (Yu et al., 2020). They take token and region as transformer input, but do not explicitly introduce phrases. Compare with them, we introduce more fine-grained supervision to find out irrelevant phrases, and build mask transformer with phrase nodes for better local semantic modeling. Moreover, we show that our model can produce explainable outputs.

6 Conclusion

In this paper, to make full use of negative sentences in both phrase-level and sentence-level, we explore to build multi-grained semantic labels, in which phrase-level ones are automatically constructed through extracting phrases of object entities, object-attribute pairs and object-relation-object triplets from images annotated sentences. We concatenate the sentence and its phrases in language side, while image and its regions in vision side, then present mask transformer for jointly cross-modality modeling with multi-grained semantics. We have multi-scale matching losses to capture the image-to-text matching and region-to-sentence/phrase-to-image matching. Based on the phrase-to-image matching, we utilize the semantic labels to determine the non-correspondence between phrases and image, and adjust scores between the image-phrase pairs. Experiment results show the effectiveness of our model on MS-COCO and Flickr30K. Further analysis reveals that semantic labels improve the efficiency of data exploitation and guide the model to discriminate mismatching sentences with more grounding on mismatched parts.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Tianlang Chen and Jiebo Luo. 2020. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29:3844–3852.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, and Xuan-Jing Huang. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6514–6524.

Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013a. Devise: A deep visual-semantic embedding model.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. 2013b. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, volume 26, pages 2121–2129. Curran Associates, Inc.

Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. 2019. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5754–5763.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *ICCV*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1508–1517.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019a. Position focused attention network for image-text matching. In *IJCAI*.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019b. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5764–5773.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019a. Context-aware self-attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019b. Auto-encoding scene graphs for image captioning. In *CVPR*.

Quanzeng You, Zhengyou Zhang, and Jiebo Luo. 2018. End-to-end convolutional semantic embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5735–5744.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.

A Implementation Details

Our training has two phases. For all experiments, the dropout rate is 0.3 and the Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ is taken as the optimizer of our model. The linear-decay learning rate scheduler is employed with 10K update steps, 1K warm-up steps. In MS-COCO, we first train the intra-r model with $\alpha_0 = 0.1$, the maximum instance (#region+#token) per batch is $32,768$, the accumulation step is 1 and the peak learning rate is 2e-4. Then, we fix the intra-modality relationship model to train the inter-modality relationship model with $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, $\alpha_3 = 0.2$ and $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.1$. The maximum instance per batch is $8,192$, the accumulation step is 16 and the peak learning rate is 6e-4. During testing, $\mu_1 = 0.2$ and $\mu_2 = 0.1$. In Flickr30K, we first train the intra-modality relationship model with $\alpha_0 = 0.05$, the maximum instance per batch is $32,768$, the accumulation step is 1, and the peak learning rate is 1e-4. Then, we fix the intra-modality relationship model to train the inter-modality one with $\alpha_1 = 0.05$, $\alpha_2 = 0.05$, $\alpha_3 = 0.1$ and $\lambda_1 = 1.0$, $\lambda_2 = 0.1$ $\lambda_3 = 0.1$. The maximum instance per batch is $8,192$, the accumulation step is 4 and the peak learning rate is 8e-4. During testing, $\mu_1 = 0.3$ and $\mu_2 = 0.1$.