Image-Text Alignment using Adaptive Cross-attention with Transformer Encoder for Scene Graphs

Juyong Song
jy1004.song@samsung.com
Sunghyun Choi
sh1992.choi@samsung.com

Samsung Research,
Samsung Electronics Co., Ltd.
Seoul, Korea

Abstract

Neural image and text encoders have been proposed to align the abstract image and symbolic text representation. Global-local and local-local information integration between two modalities are essential for an effective alignment. In this paper, we present RELation-aware Adaptive Cross-attention (RELAX) that achieves state-of-the-art performance in cross-modal retrieval tasks by incorporating several novel improvements. First, cross-attention methods integrate global-local information via weighted global feature of a modality (taken as value) for a local feature of the other modality (taken as query). We can make more accurate alignments if we could also consider the global weights of the query modality. To this end, we introduce adaptive embedding to consider the weights. Second, to enhance the usage of scene-graphs that can capture the high-level relation of local features, we introduce transformer encoders for textual scene graphs to align with visual scene graphs. Lastly, we use *NT-XEnt* loss that takes the weighted sum of the samples based on their importance. We show that our approach is effective in extensive experiments that outperform other state-of-the-art models.

1 Introduction

One of the most basic and crucial problems in machine learning is representation learning for aligning symbolic language and abstract vision features. An accurate alignment significantly impacts many visual-language tasks such as image and text retrieval across modalities [6], VQA [30], image captioning [1], image generation from the text [17, 28], and zero-shot learning [19].

The bottom-up and top-down attention mechanism [1] has been proposed to build up local attention features of the image and determine the feature attention weightings using the linguistic global context. The stacked cross-attention network [12] generalizes this approach by interchanging roles of image and text in the cross-modal attention mechanism.

Attention mechanisms, however, are uni-directional in that the query is taken locally from one modality, whereas the key and value are taken globally from the other modality. The values are weighted summed based on the local query and global key attention matrices, which does not consider the global context of the query. We introduce adaptive cross-attention to integrate the global information from both directions.

Another crucial aspect of our method is that it leverages the relationships among objects represented as a scene graph for aligning high-level semantics. Although the scene graph has been conceived as an effective intermediate representation for image retrieval [9, 21], only a few methods that appeared recently [25] were able to show significant performance improvement. Yet, end-to-end methods considering relations that do not explicitly build scene graphs [13, 27] performs better than scene graph methods. We present a method to leverage scene graphs using transformer encoders to consider the context of the triplets.

Triplet ranking margin loss, moreover, averages the negative samples with flat prior, although it performs well with hard negative sampling [6]. We also introduce mutual information lower bound loss for a weighted sum of the negative samples based on their importance. Recent representation learning studies have introduced information-theoretic mutual information lower bound loss, *InfoNCE* [16], sometimes called *N-pair* loss [22] or *NT-XEnt* [3], to relate image augmentations. It has been used in image-text matching [24] and recent studies [17, 19] showed the effectiveness of the *NT-XEnt* for alignment between image and text. We also find that *NT-XEnt* is more effective than the triplet ranking margin loss [6] for our task in the experiments.

In this paper, (i) we propose a novel neural architecture RELation-aware Adaptive Cross-attention (RELAX) with an adaptive cross-attention mechanism. (ii) Efficient transformer encoding for textual scene graphs is suggested and the encoder is shared with the fusion process of the image encoder. (iii) Information-theoretic contrastive loss (*NT-XEnt*) is compared to the triplet ranking loss in the local feature-based method. In summary, we improve global-local and local-local information integration via context-awareness. Context-awareness is also applied to the loss functions through *NT-XEnt*. We show that, through an extensive set of experiments, our proposed model outperforms the state-of-the-art models.

2 Related Works

Using bottom-up attention (BUA), local image features have been widely used to align image and text representations. The local image features resemble those from language models, thus max-attention [10] or stacked cross-attention [12] have been commonly used. Stacked cross-attention takes query from one modality and the key and value from the other modality. The attention weight is calculated by multiplication between the query Q and keys K; then softmax activation produces the probability output. By multiplying the attention weights to the values V, we obtain the representative value corresponding to the query,

$$\hat{V} = \mathrm{softmax}(QK^T)V. \tag{1}$$

To make a score for cross-modal retrieval tasks, the similarity between query and representative value, $S = Q \cdot \hat{V}/\|Q\|\|\hat{V}\|$, is used. Pooling methods aggregate the similarities, and then it results in the final score. Note that each representative value above does not reflect the query's global features in this procedure, so the queries are added in the same weights. To encode the global features of the query, we introduce adaptive embedding [26] for the calculation of the cross-attention. The adaptive embedding is expected to make the weights for the query. We also use sum pooling explored by the previous works [29] to avoid the bias effect caused by the different lengths of the captions.

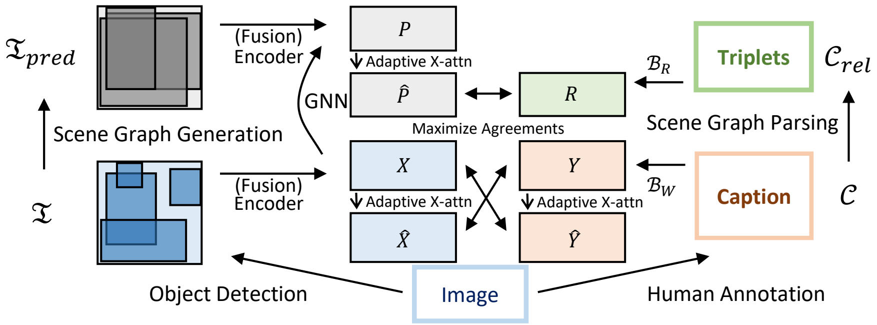

Figure 1: Graphical explanation of RELAX. $\mathfrak{I}$ and $\mathfrak{I}_{pred}$ are the image and label pairs from object detection and scene graph generation, respectively. Image encoders translate the visually grounded images to X and P. $\mathfrak{C}$ is human annotated captions corresponding to the image, whereas $\mathfrak{C}_{rel}$ is the triplets of the textual scene graphs. Caption (Y) and caption relations (R) are encoded by transformer encoders $\mathcal{B}_W$ and $\mathcal{B}_R$. $(\hat{\cdot})$ denotes the embedding after adaptive cross-attention. We maximize agreement between the adaptive embedding and the original query to calculate the similarity scores. Only $\hat{P}$-R is used for the relation alignment because textual scene graphs are not extracted in some cases. Thus, we use a global features of Y as a context vectors of R. See details in the main draft.

In cross-modal image retrieval, using visual [9] and textual scene graphs [21] has been proposed from its early stage. Recent study [25] successfully integrates the scene graph and the neural network using the scene graph matching (SGM) with the graph neural networks (GNN). In SGM, local image object features and words are aligned, while image predicate features and text relations from textual scene graphs are aligned separately. Our model is similar to SGM but with improved embedding using adaptive cross-attention and text embedding from transformer language model.

We use the bidirectional encoder representation from transformer (BERT) [4] as the text encoder. In this model, tokenized sentences start with the [*CLS*] token and end with the [*SEP*] token. For the text relations, each triplet is encoded as a sentence separated by [*SEP*], and we use the embedding of [*SEP*] token as the triplet representation. Recent studies [14, 25] fused the label embedding to the image embedding. In our work, the caption/image object label and caption relations/image predicate label share the encoders.

Mutual information lower bound losses [2, 7, 16] have been recently introduced as self-supervised learning losses. For example, the *NT-XEnt* loss was originally suggested for self-supervised learning that maximizes the agreement of embedding between two different augmentations (viewpoint) of the same single image. In this work, we treat the image and text as the different expressions of the same situation (Fig 1). Because the name 'contrastive' can be adopted to both margin loss and mutual information lower bound losses, we denote the latter as *NT-XEnt* for the rest of the paper. *NT-XEnt* has been applied to recent studies with large-scale datasets [8, 19] to align images and texts. However, these studies regard simple multiplications between global features from image and text, not using attention mechanisms to calculate the similarity matrix. It has also been suggested for local feature alignment in robust retrieval [24]. In this paper, we show that *NT-XEnt* can be applied to the cross-attention model, and yields better performance compared to the margin loss.

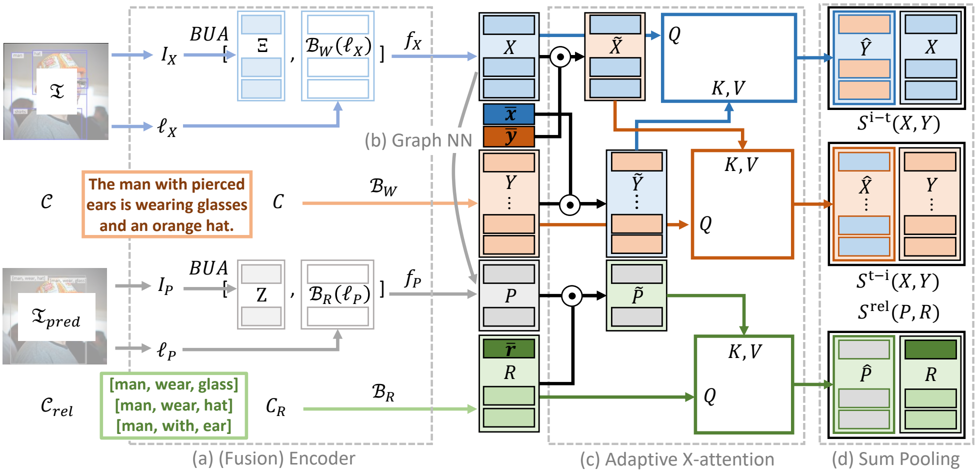

Figure 2: Input data consists of detected image objects (blue), image predicates(gray), caption (orange) and caption relation (green). BERT models and tokenizers for word ($\mathcal{B}_W$, $\mathcal{T}_W$) and relations ($\mathcal{B}_R$, $\mathcal{T}_R$) are presented in fusion module that yields embedding of caption (Y) and caption relations (R). The labels of the image objects (ℓ_X) and predicates (ℓ_P) are integrated to BUA features (Ξ and Z) through the models. $\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{r}}$ are the global embedding of image, caption and relation, and $\tilde{X}, \tilde{Y}, \tilde{P}$ are globally adaptive embedding to counter modalities, $\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}}, \bar{\boldsymbol{r}}$, respectively. After the attention process, dot products yield $\hat{X}, \hat{Y}, \hat{P}$. Black boxes on the rightmost represent the sum of the cosine similarities between the two representations.

3 Model

To improve image-text alignment, we propose a novel structure RELAX: RELation-aware Adaptive Cross-attention with transformer encoders that has three main architectures. (Fig. 2) (a) Shared text encoder with label fusion to combine the external textual knowledge to the image embedding vectors; (b) related objects integration using visual scene graphs via GNN; (c) global feature integration using adaptive embedding. We also effectively integrate the information using sum pooling and *NT-XEnt* loss.

We denote the dataset of detected objects as $\mathfrak{I}$, caption as $\mathfrak{C}$, and their mini-batches as $\mathfrak{I}_b$ and $\mathfrak{C}_b$, respectively. We omit the subscript b for convenience. We then write $\mathfrak{I} = \{(I_X^{(1)}, \ell_X^{(1)}), \cdots, (I_X^{(n)}, \ell_X^{(n)})\}$ for detected objects, and $\mathfrak{C} = \{C^{(1)}, \cdots, C^{(n)}\}$ for caption with the batch-size of n. Note that $I_X^{(k)}$ is the set of cropped images of detected object regions, $\ell_X^{(k)}$ is the corresponding set of the labels, and $C^{(k)}$ is the human-annotated caption for the image.

The visual scene graphs are generated based on detected objects [23, 32]. The visual scene graph comprises a set of triplets. A mini-batch of the visual scene graph is denoted as $\mathfrak{I}_{\text{pred}} = \{(I_P^{(1)}, \ell_P^{(1)}), \cdots, (I_P^{(n)}, \ell_P^{(n)})\}$. Note that $I_P^{(k)}$, called image predicate, is the set of the cropped images of the union regions, i.e. the smallest rectangle that contains the two objects (subject and object), and $\ell_P^{(k)}$ is the corresponding sets of triplet labels. The textual scene graph is composed of several triplets, which can be generated by parsing $\mathfrak{C}$ [21]. We denote a mini-batch of the textual scene graph as $\mathfrak{C}_{\text{rel}} = \{C_R^{(1)}, \cdots, C_R^{(n)}\}$, where each $C_R^{(k)}$ is a set of triplet, called caption relation.

3.1 BERT embedding for textual scene graphs

A RELAX model has two transformer encoders. One is for caption $\mathcal{B}_W$, and the other one is for caption relation $\mathcal{B}_R$. (Fig. 1, 2) The BERT models have an additional perceptron layer to match the dimension to the image encoding later. The caption (C) is encoded as caption embedding (Y) through $\mathcal{B}_W$, $Y = \mathcal{B}_W(C)$. Note that $Y = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_L)$, where $\boldsymbol{y}_j$ is a d-dimensional word embedding vector, and L is the number of tokens of the caption embedding.

For the caption relations (C_R), transformer encoding should be modified because they are triplets. We used $[SEP]$ embedding of BERT right after the triplet as the representative feature of the triplet, or $R = \mathcal{B}_R\left(C_R\right)_{[SEP]}$. Note that $R = \left(\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{L'}\right)$ is caption relation embedding, $\boldsymbol{r}$ is a d-dimensional vector, and L' is the number of caption relations. $[SEP]$ denotes the embedding for $[SEP]$ token. Note that $[SEP]$ is one of the possible choices. See the comparison with other triplet embedding methods and details in supplement material.

3.2 Label Fusion

The images of detected objects I_X is encoded as the intermediate image features Ξ by pre-trained BUA model [□] (Fig. 2a), where $\Xi = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_M)$, $\boldsymbol{\xi}$ is the vector for a local region, and M denotes the number of detected objects. The labels of the objects ℓ_X is encoded by $\mathcal{B}_W$, same as in Sec. 3.1, then the intermediate label features are fused with Ξ by concatenation as follows.

$$\boldsymbol{x} = f_X([\boldsymbol{\xi}, \mathcal{B}_W(\ell_X)_{\boldsymbol{y}}]), \tag{2}$$

where f_X is a perceptron layer, and $\boldsymbol{x}$ is the d-dimensional image embedding vector for the local region. Note that subscript $\boldsymbol{y}$ denotes the label embedding excluding the special tokens, and $[\cdot,\cdot]$ denotes the concatenation of the two features. Then, we get $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M)$, the image embedding set for an image.

Similarly, I_p is encoded as Z by the same BUA model above, where $Z = (\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \cdots, \boldsymbol{\zeta}_{M'})$, each $\boldsymbol{\zeta}$ is the vector for a union region, and M' denotes the number of image predicates of the visual scene graphs. The triplet labels of the relation ℓ_P is encoded by $\mathcal{B}_R$; then the intermediate triplet features are fused with Z by concatenation as follows.

$$\boldsymbol{p}^0 = f_P([\boldsymbol{\zeta}, \mathcal{B}_R(\ell_P)_{[SEP]}]), \tag{3}$$

where f_P is a perceptron layer, and $\boldsymbol{p}^0$ is the d-dimensional vector for the image predicates. $P^0 = (\boldsymbol{p}_1^0, \cdots, \boldsymbol{p}_{M'}^0)$ is the pure image predicate embedding set for an image.

3.3 Graph neural networks for visual scene graphs

We use GNN (Fig. 2b) to integrate the local object information via the visual scene graphs. Once the image predicate embedding (P^0) is extracted, it is mixed with object features (X).

$$\boldsymbol{p} = f_{\mathrm{GNN}}([\boldsymbol{x}_S, \boldsymbol{p}^0, \boldsymbol{x}_O]), \tag{4}$$

where f_{GNN} is a perceptron layer with non-linear activation, such as tanh, and $[\cdot,\cdot,\cdot]$ denotes the concatenation of the features. Note that the index of the subject is S, and the index of the object is O. It can be seen as the graph convolution of bipartite networks of objects and predicates. See detailed graphical explanations in supplement material.

3.4 Adaptive Embedding for Cross-Attention

Cross-attention image-text alignment [12] models have widely been used for retrieval tasks. It can be fulfilled as two kinds of attentions, called text-to-image(t-i) and image-to-text(i-t) attention. T-i attention deals with the word embedding as the query of the attention and the image object embedding as the key and value. In contrast, i-t attention deals with the image object embedding as the query and the word embedding as the key and the value.

In Fig. 2c, to encode the weights of the queries, adaptive embedding [26] integrates the global feature ($\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}$) of the query modality to the local feature of the other (key and value) modality. For instance, $\bar{\boldsymbol{x}}$, the mean feature of the image, can be integrated into the $\boldsymbol{y}$ (Y), and results in $\tilde{\boldsymbol{y}}$ ($\tilde{Y}$) in i-t attention. The global features are linearly projected through two functions, g and b, such as

$$\tilde{\boldsymbol{x}}_i = g_Y(\bar{\boldsymbol{y}}) \odot \boldsymbol{x}_i + b_Y(\bar{\boldsymbol{y}}), \quad \tilde{\boldsymbol{y}}_j = g_X(\bar{\boldsymbol{x}}) \odot \boldsymbol{y}_j + b_X(\bar{\boldsymbol{x}}), \tag{5}$$

where $\odot$ denotes component wise multiplication. By substitution of the key and value of the original cross-attention with the globally adaptive embedding, we can rewrite the attended embedding as following:

$$\hat{X} = \mathrm{softmax}(Y\tilde{X}^T)\tilde{X}, \quad \hat{Y} = \mathrm{softmax}(X\tilde{Y}^T)\tilde{Y}. \tag{6}$$

Note that $\tilde{X} = (\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_M)$ and $\tilde{Y} = (\tilde{\boldsymbol{y}}_1, \cdots, \tilde{\boldsymbol{y}}_L)$. For the i-t model, if the special tokens (especially, $[CLS]$ token) are used as the key and value, it makes an unfair comparison with the other word embedding, so that we erase the special tokens for the attention. Moreover, the similarity score between the two different modalities can be defined as the sum of the cosine-similarities of all the regions(i-t) and words(t-i).

$$S^{\text{i-t}}(X,Y) = \sum_{i=1}^{M} \frac{\boldsymbol{x}_i \cdot \hat{\boldsymbol{y}}_i}{\|\boldsymbol{x}_i\|\|\hat{\boldsymbol{y}}_i\|}, \quad S^{\text{t-i}}(X,Y) = \sum_{j=1}^{L} \frac{\hat{\boldsymbol{x}}_j \cdot \boldsymbol{y}_j}{\|\hat{\boldsymbol{x}}_j\|\|\boldsymbol{y}_j\|}. \tag{7}$$

Note that $\hat{X} = (\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_L)$ and $\hat{Y} = (\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_M)$. We find that the aggregation by sum pooling (Fig. 2d) of the similarity scores is more effective than the average aggregation described in Table 2 (model 1 and 2). For the relation similarities, we only use t-i attention, because occasionally text relations are not extracted. We set $\bar{\boldsymbol{r}}$ to $[CLS]$ embedding of Y as a context vector. Then,

$$S^{\text{rel}}(P,R) = \sum_{j=1}^{L'} \frac{\hat{\boldsymbol{p}}_j \cdot \boldsymbol{r}_j}{\|\hat{\boldsymbol{p}}_j\|\|\boldsymbol{r}_j\|}, \tag{8}$$

where L' is the length of the caption relations. Note that $\hat{P} = \mathrm{softmax}(R\tilde{P}^T)\tilde{P}$ and $\tilde{\boldsymbol{p}}_i = g_R(\bar{\boldsymbol{r}}) \odot \boldsymbol{p}_i + b_R(\bar{\boldsymbol{r}})$, where $\tilde{P} = (\tilde{\boldsymbol{p}}_1, \cdots, \tilde{\boldsymbol{p}}_{M'})$ and $\hat{P} = (\hat{\boldsymbol{p}}_1, \cdots, \hat{\boldsymbol{p}}_{L'})$.

3.5 Loss function

Triplet ranking margin loss. To learn a discriminative representation, triplet margin loss has been suggested [20]. Note that we use online sampling; thus, $\mathcal{X}_N$ and $\mathcal{Y}_N$ are the negative samples in a mini-batch of the image and text embedding, respectively. In [6], the hardest negative sampling in a mini-batch can improve the triplet loss performances.

$$\mathcal{L}_{\text{SH}}(X,Y) = \sum_{X' \in \mathcal{X}_N} \left[-S(X,Y) + S\left(X',Y\right) - m\right]_+ + \sum_{Y' \in \mathcal{Y}_N} \left[-S(X,Y) + S\left(X,Y'\right) - m\right]_+, \tag{9}$$

$$\mathcal{L}_{\text{MH}}(X,Y) = \left[-S(X,Y) + S\left(X'',Y\right) - m\right]_+ + \left[-S(X,Y) + S\left(X,Y''\right) - m\right]_+, \tag{10}$$

where m is the margin value and $[x]_+ = \max(x,0)$. Note that $\mathcal{X}_N$ and $\mathcal{Y}_N$ are the negative samples, X'' and Y'' are the hardest negative samples in a mini-batch of the image and text embedding, respectively. Following [26], we train with interpolation of two margin losses, $\mathcal{L}_{\mathrm{H}} = K\mathcal{L}_{\mathrm{MH}} + (1-K)\mathcal{L}_{\mathrm{SH}}$. Note that K starts from zero and converge to one.

Normalized Temperature XEnt. Mutual information has been a good measure to capture the correlation between two distributions. InfoNCE [16] applied them to self-supervised representation learning. SimCLR [3] introduces temperature as a hyper parameter and insists that the mutual information loss is just the normalized temperature cross-entropy (*NT-XEnt*),

$$\mathcal{L}_{\mathrm{NT}}(X,Y) = -\log\left(\frac{\exp\beta S(X,Y)}{\sum_{X'\in\mathcal{X}}\exp\beta S(X',Y)}\right) - \log\left(\frac{\exp\beta S(X,Y)}{\sum_{Y'\in\mathcal{Y}}\exp\beta S(X,Y')}\right), \tag{11}$$

where β is the inverse temperature, $\mathcal{X}$ is a mini-batch for image embedding, and $\mathcal{Y}$ is a mini-batch for caption embedding. Note that increasing β affects the system more cleanly be separated to make a similar effect of hard negative sampling. *NT-XEnt* computes the probability of the data compared to the other data. In contrast, margin loss assumes that all the data has the same probability (See supplement material).

Considering the image predicate and text relation similarity scores, the total loss becomes $\mathcal{L}^{\mathrm{total}} = \mathcal{L}(X,Y) + \lambda\mathcal{L}(P,R)$, where λ is the weight of relation similarity score.

4 Experiments

4.1 Datasets and Experimental Settings

We evaluate our image-text alignment model on the two publicly available cross-modal retrieval datasets, Flickr30k [31] and MS-COCO [15] following the split of [6, 10, 12]. For both of the datasets, each image has five corresponding human-annotated captions. Flickr30k has 29k training images, 1k validation images, and 1k test images. MS-COCO has 113k training images, 5k validation images, and 5k test images. The final results are reported by averaging over 5 folds for 1k test images or testing on the full 5k test images.

The performance is measured by the standard recall at K (R@K). Following the previous studies, R@1, R@5, and R@10 are measured for image/text retrieval tasks. Note that Rsum is defined as the sum of all measured recall values. RELAX model has all modules described above: fusion, relation, adaptive embedding, and sum pooling. REL-X is RELAX without adaptive embedding. The training details are in supplement material.

4.2 Results on MS-COCO and Flickr30k

In Table 1, performance comparisons on MS-COCO 1k fold5, 5k test set, and Flickr30k 1k test set are presented. Our model outperforms state-of-the-art models for both text retrieval and image retrieval on MS-COCO and Flickr30k. Note that DSRAN [27], CAMERA [18], and ours use BERT embedding while the others use Bi-GRU embedding for the text.

Our model improves 2.5% on image retrieval (R@1) and 2.4% on text retrieval (R@1) for the single models in MS-COCO 1k test set compared to other state-of-the-art models. In practice, our single models already have two image/text encoders (i.e. image encoders for X and P, text encoders for Y and Q). Note that the best single model outperforms other ensemble models in some metrics (IR@1, IR@10).

Single model	MS-COCO 1k test						MS-COCO 5k test						Flickr30k 1k test					
	Text Retrv			Image Retrv			Text Retrv			Image Retrv			Text Retrv			Image Retrv		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN i-t [12]	68.4	93.9	98.0	54.8	86.1	93.3	46.4	77.4	87.2	34.4	63.7	75.7	67.7	88.9	94.0	44.0	74.2	82.6
SCAN t-i [12]	70.9	94.5	97.8	56.4	87.0	93.9	-	-	-	-	-	-	61.8	87.5	93.7	45.8	74.4	83.0
SGM [25]	73.4	93.8	97.8	57.5	87.3	94.3	50.0	79.3	87.9	35.3	64.9	76.5	71.8	91.7	95.5	53.5	79.6	86.5
ADAPT i-t [26]	74.5	94.2	97.9	62.0	90.4	95.5	-	-	-	-	-	-	70.2	90.8	95.8	55.5	82.7	89.8
ADAPT t-i [26]	75.3	95.1	98.4	63.3	90.0	95.5	-	-	-	-	-	-	73.6	93.7	96.7	57.0	83.6	90.3
CAMERA [18]	75.9	95.5	98.6	62.3	90.1	95.2	53.1	81.3	89.8	39.0	70.5	81.5	76.5	95.1	97.2	58.9	84.7	90.2
DSRAN [27]	78.8	96.1	98.5	62.9	89.9	95.3	56.3	84.2	90.7	40.3	70.9	81.3	78.6	95.6	97.6	57.3	84.8	90.9
REL-X i-t (m)	77.9	95.8	98.6	60.2	88.7	94.8	57.3	83.8	90.6	36.4	67.5	79.1	75.5	93.2	96.5	54.3	81.3	88.2
REL-X i-t (NT)	77.0	95.4	98.5	60.1	88.6	94.7	57.1	83.0	90.5	37.4	67.6	78.5	75.6	92.8	96.6	54.5	81.2	88.0
RELAX i-t (m)	78.6	96.3	**98.8**	62.3	89.7	95.6	58.1	84.4	91.4	40.0	70.1	80.2	77.5	93.4	96.7	56.8	82.9	89.4
RELAX i-t (NT)	78.2	95.7	98.4	62.3	89.7	95.5	58.0	84.0	90.9	40.3	70.2	80.8	77.1	93.4	97.0	57.5	82.8	89.3
REL-X t-i (m)	76.0	95.7	98.5	62.4	90.0	95.7	54.1	82.7	90.6	39.9	69.9	80.9	**81.2**	**95.8**	**98.0**	61.6	85.8	91.6
REL-X t-i (NT)	**81.2**	**96.3**	98.4	**65.8**	**91.1**	**96.1**	**61.8**	**87.2**	**92.8**	**44.2**	**73.3**	**83.1**	**81.2**	95.5	97.7	62.1	85.9	91.6
RELAX t-i (m)	77.4	96.0	98.6	62.0	89.8	95.9	55.7	83.4	90.8	40.0	69.6	80.3	80.1	95.5	97.5	**62.2**	**86.4**	**91.8**
RELAX t-i (NT)	80.8	96.3	98.7	65.3	90.6	95.8	60.4	85.9	92.4	43.8	72.5	82.1	80.8	95.3	97.6	**62.2**	86.1	**91.8**

Ensemble	Text Retrv			Image Retrv			Text Retrv			Image Retrv			Text Retrv			Image Retrv		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [12]	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4	67.9	90.3	95.8	48.6	77.7	85.2
VSRN [13]	76.2	94.8	98.2	62.8	89.7	95.1	53.0	81.1	89.4	40.5	70.6	81.1	71.3	90.6	96.0	54.7	81.8	88.2
ADAPT [26]	76.5	95.6	98.9	62.2	90.5	96.0	-	-	-	-	-	-	76.6	95.4	97.6	60.7	86.6	92.0
SGRAF [5]	79.6	96.2	98.5	63.2	90.7	96.1	57.8	-	91.6	41.9	-	81.3	77.8	94.1	97.4	58.5	83.0	88.8
CAMERA [18]	77.5	96.3	98.8	63.4	90.9	95.8	55.1	82.9	91.2	40.5	71.7	82.5	78.0	95.1	97.9	60.3	85.9	91.7
DSRAN [27]	80.6	96.7	98.7	64.5	90.8	95.8	57.9	85.3	92.0	41.7	72.7	82.8	80.5	95.5	97.9	59.2	86.0	91.9
REL-X (m)	81.2	96.8	98.8	65.8	91.8	96.6	62.2	86.9	93.1	43.4	73.5	83.7	82.8	95.8	98.0	63.3	87.0	92.3
REL-X (NT)	82.3	96.8	98.8	67.2	91.8	96.5	64.7	88.3	93.4	45.5	74.7	84.2	82.7	95.9	98.0	63.2	86.9	92.3
RELAX (m)	81.9	**97.0**	**99.2**	66.5	92.2	**96.8**	63.4	87.9	93.6	44.7	74.0	83.6	84.0	**96.2**	**98.2**	64.9	**88.1**	92.9
RELAX (NT)	83.0	96.7	99.0	**67.5**	92.1	96.7	64.8	87.7	92.9	**46.6**	**75.2**	**84.6**	**84.4**	96.1	98.1	**65.0**	88.0	**93.0**
Best models	**83.1**	**97.0**	99.1	**67.5**	**92.4**	96.7	**65.0**	**88.7**	**93.7**	46.3	**75.2**	**84.6**	-	-	-	-	-	-

Table 1: Cross-modal retrieval results on MS-COCO 1k test set, 5k test set, and Flickr30k 1k test set. The bold numbers denote the best models for each metric. REL-X and RELAX are our models, while RELAX is the same as REL-X except for adaptive embedding. (m) and (NT) denote the models using margin and *NT-XEnt* loss. Ensemble models use the average similarity matrix of t-i and i-t attention models. In the last line, we also examine the ensemble of the best models, RELAX i-t (m) and REL-X t-i (NT) for MS-COCO.

To calculate the ensemble results, SCAN, ADAPT, and ours use the average similarity matrix of t-i and i-t attention models, while DSRAN and CAMERA use the same architecture. The best ensemble model also improves 3.0% on image retrieval (R@1) and 2.5% on text retrieval (R@1) on MS-COCO 1k test set. For MS-COCO 5k test set, our model improves 4.7% on image retrieval (R@1) and 7.1% on text retrieval (R@1).

For Flickr30k, the RELAX t-i models outperform the other image retrieval models, while the REL-X t-i models are better in text retrieval. The best results improve 3.3% on image retrieval (R@1) and 2.6% on text retrieval (R@1). For the ensemble results, RELAX models result in the best scores for all metrics. Our models improve 4.5% on image retrieval (R@1) and 3.8% on text retrieval (R@1).

For the adaptive embedding, compared to REL-X i-t model, RELAX i-t model improves the image retrieval (R@1) from 60.1 to 62.3 for MS-COCO 1k test set. In Fig. 3, the image global adaptive embedding for the i-t models (RELAX i-t) are more effective than text global adaptation for t-i models. In practice, REL-X t-i performs better than RELAX t-i models in most of the cases. A global text information using adaptive X-attention can annoy the system, while global image information is essential for better performance in (i-t) attention. Applying the loss *NT-XEnt* is also effective, especially in MS-COCO. The standard errors are plotted for the MS-COCO 1k test for the fold5 model (Fig. 3a) and four independent models for the Flickr30k 1k test (Fig. 3c).

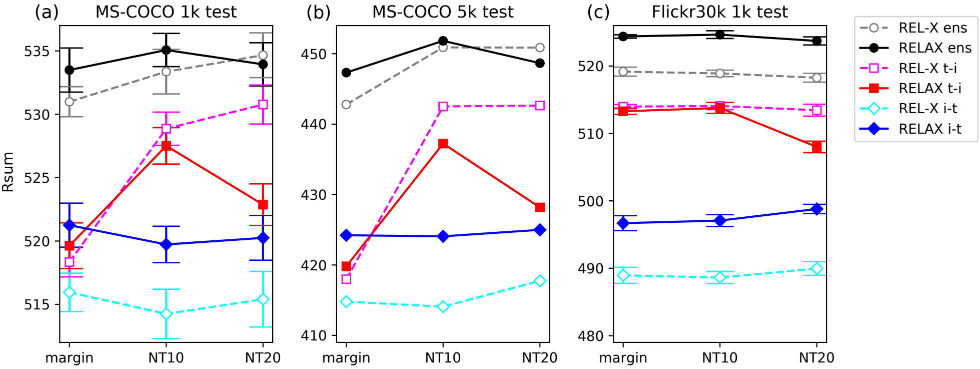

Figure 3: Recall sum (Rsum) on (a) MS-COCO 1k fold5, (b) 5k, and (c) Flickr30k 1k test set depending on the changing of the loss function where NT10 and NT20 are the *NT-XEnt* with the inverse temperature $\beta = 10$ and 20. Solid lines with filled points represent RELAX with adaptive attention models, while dotted lines with empty points represent REL-X models with original cross-attention. Red squares are t-i attention models, while blue lines are i-t attention models. Black circles are the ensemble of t-i and i-t models by aggregation on similarity matrices. We plot standard errors for the same model on fold5 MS-COCO test sets in (a), for four different models on Flickr 1k test set in (c).

4.3 Ablation study

In Table 2, we explore the effect of each module. We find that the sum pooling of similarity scores of cross-attention is an essential factor. Average pooling can induce a bias when the lengths of the captions are varied by ignoring the crucial features for longer sentences. When there exist many captions to compare, average and sum pooling are not identical, and the absolute values become important.

Comparing models 2/3 and 6/8, fusion with a shared model also improves the performances that enable easy alignment. We also examine the effect of adaptive embedding by comparing models 3/4 and 7/8. *NT-XEnt* sometimes improves the performance (models 5 and 8), sometimes it doesn't (models 4 and 7). Note that ensembling the best models for t-i ($\beta = 10$) and i-t ($\beta = 20$) attention among the various temperatures can perform slightly better (0.4% in Rsum) than model 8.

model	Pooling	Fusion	Relation	Adapt	Loss	Text Retrv			Image Retrv			Rsum
						R@1	R@5	R@10	R@1	R@5	R@10	
1	Mean	✗	✗	✗	Margin	70.4	90.6	94.9	47.7	75.1	83.0	461.7
2	Sum	✗	✗	✗	Margin	81.2	95.1	97.5	60.4	85.0	91.0	510.2
3	Sum	✓	✗	✗	Margin	82.9	95.5	97.7	62.4	86.2	91.8	516.4
4	Sum	✓	✓	✗	Margin	82.8	95.8	98.0	63.3	87.0	92.3	519.2
5	Sum	✓	✓	✓	Margin	84.0	96.2	98.2	64.9	88.1	92.9	524.4
6	Sum	✗	✓	✓	NT-XEnt	83.8	95.9	98.0	63.4	87.2	92.4	520.7
7	Sum	✓	✓	✗	NT-XEnt	82.7	95.9	98.0	63.2	86.9	92.3	518.9
8	Sum	✓	✓	✓	NT-XEnt	84.4	96.1	98.1	65.0	88.0	93.0	524.7

Table 2: Ablation study on Flickr30k dataset. We check on the ensemble models using the mean similarity matrix of the t-i and i-t attention models. The values are the average over four different models.

5 Discussion

We suggest a novel structure RELAX with scene graphs for integrating the relations between objects. The visual grounding information is integrated by the fusion of the labels with shared transformer encoders. To integrate the global context of a query, adaptive embedding is examined to the key and value embedding. By virtue of each module, RELAX results in a noticeable improvement compared to the other state-of-the-art models.

Adaptive value features can encode the importance of the query features, while the previous cross attention method just sums up the queries. The adaptive cross-attention matrix can be simplified as $Qf(\bar{\boldsymbol{q}}^T)K^T$, where Q, K are query and key, $\bar{\boldsymbol{q}}$ is the global feature of the Q. When we attend the relation between Q and $\bar{\boldsymbol{q}}$, the essential queries that lead the global features are emphasized by this adaptation. In our experiments, global text feature adaptation to local image features seems less critical than global image feature adaptation. For BERT embedding already considers the importance of the query features by self-attention, the effect may be doubled by the adaptation, which coincides with our explanation. Thus, self-attention for image features can be considered for future work. Also, adaptive attention encourages the key features to have the global information of the query modality. The global query information in adaptive attention can lead the key features, similar to *entrainment* in biological systems [11].

Though a scene-graph, a neuro-symbolic approach, for both image and text, has been suggested as an essential factor of the complex query retrieval for the relation detection, the performance on the retrieval tasks was worse than the end-to-end methods. Recently, the knowledge of scene graphs can be effectively integrated using neural networks, and the scene graph generator is also improved. Above all, neuro-symbolic approaches rely more on text information than end-to-end models; thus, an efficient text encoder brings more improvement than the other recent approaches using a similar text encoder.

In practice, our model is based on [25]. Still, we change the encoder from bi-GRU to BERT and share the encoder for the captions($\mathcal{B}_W$ for $\mathfrak{C}$)/relations($\mathcal{B}_R$ for $\mathfrak{C}_{rel}$) and the labels of the detected objects (i.e. the labels of $\mathfrak{I}$ and the triplets of $\mathfrak{I}_{pred}$). The fusion can deliver the knowledge of the pre-trained models that helps the alignment. Only the extracted labels from the object detection and scene graph generation are not good enough to align the image and complex queries. Still, it is helpful to align the dense features between two modalities efficiently because the captions and extracted labels are in the same domain, i.e. language.

The mutual information lower bound loss is also examined for better and efficient training. *NT-XEnt* is a kind of log-likelihood maximization when we interpret the softmax of the similarity scores as the probability of the corresponding positive image(text) given a text(an image). Thus, it can be interpreted as an expectation-maximization (EM). It seems to converge faster (See supplementary material) than margin loss that puts the embedding of negative samples far from that of the positive sample. The temperature of *NT-XEnt* should be tuned for better performance because the temperature controls the effect of hard-negative sampling in *NT-XEnt*. We can explore better loss designs such as adaptive temperature for future work.

6 Acknowledgement

This work is supported by SAMSUNG Research, Samsung Electronics Co.,Ltd. We want to express our special thanks to Prof. Kee-Eung Kim for helpful advice and discussion. We also thank Daniel D. Lee, Joohyung Lee, Jungmin Kwon, Byungchul Kang and many colleagues for their supports and feedbacks.

References

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

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In ICML, 2018.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In AAAI, 2021.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In BMVC, 2018.

[7] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In ICLR, 2019.

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918, 2021.

[9] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In CVPR, 2015.

[10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.

[11] Peter Lakatos, Joachim Gross, and Gregor Thut. A new unifying account of the roles of neuronal entrainment. Current Biology, 29(18):R890–R905, 2019.

[12] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In ECCV, 2018.

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

[14] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In ECCV, 2020.

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

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

[17] OpenAI. Dall·e: Creating images from text, Jan 2021. URL https://openai.com/blog/dall-e/.

[18] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. Context-aware multi-view summarization network for image-text matching. In ACM Multimedia, 2020.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. Image, 2:T2.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, 2015.

[21] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In Workshop on Vision and Language (VL15), 2015.

[22] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In NIPS, 2016.

[23] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In CVPR, 2020.

[24] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In ECCV, 2020.

[25] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In WACV, 2020.

[26] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros. Adaptive cross-modal embeddings for image-text alignment. In AAAI, 2020.

[27] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. TCSVT, 2020.

[28] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In CVPR, 2018.

[29] Artem Babenko Yandex and Victor Lempitsky. Aggregating local deep features for image retrieval. In ICCV, 2015.

[30] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In CVPR, 2016.

[31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.

[32] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In CVPR, 2018.