Supplementary material on 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

1 Experimental Details

The visual scene graphs are generated by unbiased scene graph generator [7] or Neural-Motif [8] in this paper. The generators are trained on Visual Genome dataset [4]. $\text{BERT}_{\text{BASE}}$ pretrained on Toronto Book Corpus [9] and Wikipedia is used for the text encoder and it is fine-tuned to the Flickr30k and MS-COCO, while bottom-up attention (BUA) model for the image feature extractor pretrained on Visual Genome [4] is not fine-tuned. The $\text{BERT}_{\text{BASE}}$ encoder has 768 dimensions at the last layer, while the BUA feature is extracted as 2048 dimension vectors. To compare the image and text features, we use 1024 embedding for all the embedding, i.e. X, Y, P and R. The inverse temperature β of *NT-XEnt* is fixed as 10 without additional comment. The relation similarity weight, λ is set to 1 in all cases.

We use Adam optimizer [3] for all models with the learning rate 2.5×10^{-5}. We train our model for 60 epochs, and learning rate decays by 0.1 for every 25 epochs. The inverse temperature β of *NT-XEnt* is fixed as 10 without additional comment. The relation similarity weight, λ is set to 1 in all cases. The batch size is always set to 128. For the margin losses, the margin is set to 0.2 for all models. L2 normalization for image is performed only for t-i attention models.

2 Caption and caption relation embedding

Detailed explanations for caption and caption relation embedding are drawn in Fig. 1. We also explore the caption relation embedding methods, SUM, LAST, and SEP. SUM add all the values of the last layer of $\mathcal{B}_R$ except for the special tokens. LAST picks the last state of the last layer of $\mathcal{B}_R$, *i.e.* just before the [*SEP*] embedding. SEP picks the embedding of [*SEP*] tokens of the last layer of $\mathcal{B}_R$ as depicted in Figure. 1(a). We find that all the methods similarly work, but the SEP method yields better Rsum results than others a little. We get these results with 30 epochs and learning rate update at every 15 epochs, and other details are the same as the experimental details above. Note that this result is coming from a model,

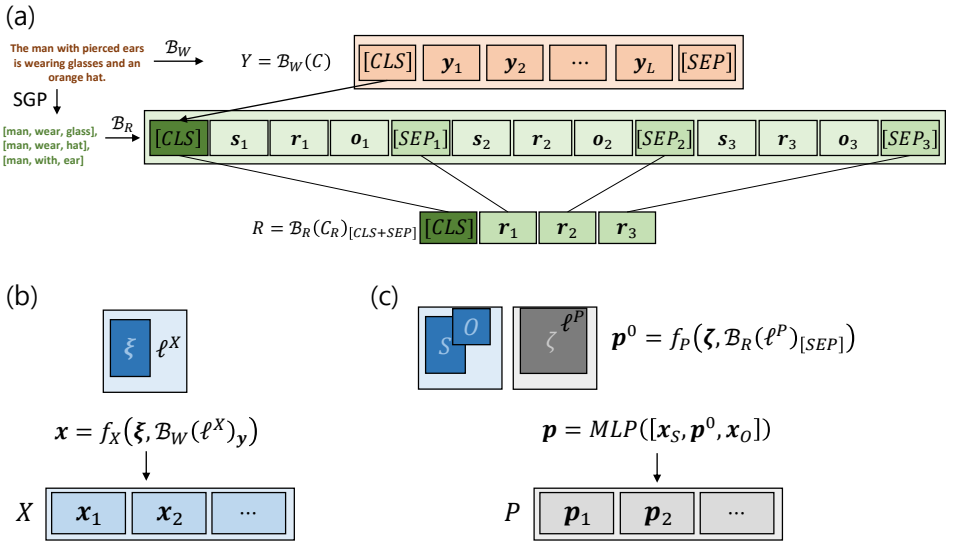

Figure 1: (a) Textual embedding for caption and textual scene graphs. The raw caption and its relations are embedded with two different BERT $\mathcal{B}_W$ and $\mathcal{B}_R$, respectively. For the relation, the [*SEP*] embedding is selected for a relation as the representative one, while [*CLS*] token is replaced by the pooling layer of that of the raw captions to consider the context of all captions. (b) Fusion of the label embedding with vision features of the objects. (c) For the scene graphs, one can fuse the predicate's label with the predicate's visual feature. The subject and object image features are integrated to consider the object-object relation.

methods	SUM	LAST	SEP
RELAX i-t	495.8	493.7	497.1
RELAX t-i	511.8	509.1	512.4
ensemble	524.6	520.9	525.2

Table 1: Recall sum (Rsum) of different relation embedding methods.

while the results of the main draft is average value over several models and their results are fluctuating. Thus, the Rsum values can be higher or lower than the main draft.

3 Effectiveness of Transformer Embedding

To see the effectiveness of the transformer embedding, we examine the other transformer encoders, RoBERTa [5] and DistilBERT [6]. In Table. 2, RoBERTa performs better than original BERT, while DistilBERT performs worse.

methods	BERT	RoBERTa	DistilBERT
RELAX i-t	497.1	504.1	494.8
RELAX t-i	513.8	516.1	509.5
ensemble	524.7	528.0	521.6

Table 2: Recall sum (Rsum) for various transformer encoders.

4 Gradients of contrastive losses

To show that the *NT-XEnt* is an EM algorithm, we derived the gradients of the loss of the contrastive losses, i.e. *NT-XEnt* and triplet margin loss. Defining X as the image embedding and Y as the text embedding, we can define the probability of X and Y given the anchor of the other modality.

$$P(Y|X) = \frac{\exp \beta S(X,Y)}{\sum_{Y'} \exp \beta S(X,Y')}, \tag{1}$$

$$P(X|Y) = \frac{\exp \beta S(X,Y)}{\sum_{X'} \exp \beta S(X',Y)}, \tag{2}$$

where β is an inverse temperature same as in the main draft. They are not real generative probabilities, but they can be seen as the approximation of the probability of the generative model based on mini-batch samples. Then, the loss of contrastive learning can be defined as the negative log probability.

$$\mathcal{L}_{NT}(Y|X) = -\log P(Y|X), \tag{3}$$

$$\mathcal{L}_{NT}(X|Y) = -\log P(X|Y). \tag{4}$$

Note that we used the final loss as the sum of them, $\mathcal{L}_{NT}(X,Y) = \mathcal{L}_{NT}(Y|X) + \mathcal{L}_{NT}(X|Y)$. The gradient of the total loss can be calculated as following,

$$-\frac{1}{\beta}\frac{\partial \mathcal{L}_{NT}(X,Y)}{\partial \theta} = \frac{\partial S(X,Y)}{\partial \theta} - \left\langle \frac{\partial S}{\partial \theta} \right\rangle_{P(Y|X)} + \frac{\partial S(X,Y)}{\partial \theta} - \left\langle \frac{\partial S}{\partial \theta} \right\rangle_{P(X|Y)}, \tag{5}$$

where $\langle \cdot \rangle_P$ denotes the average over the given probability P. The negative terms can be interpreted as the expectation, while the update procedure as the maximization [1].

For the triplet ranking margin loss, the update equation is similar. Still, it estimates the probability of the negative samples as a flat probability distribution for the samples in the margin range.

$$\mathcal{L}_{SH}(Y|X) = \sum_{Y'} \left[-S(X,Y) + S(X,Y') + m \right]_+, \tag{6}$$

$$\mathcal{L}_{SH}(X|Y) = \sum_{X'} \left[-S(X,Y) + S(X',Y) + m \right]_+. \tag{7}$$

Note that *SH* is from 'sum hinge' in [2]. The gradient of the margin loss can be written as,

$$\begin{aligned} -\frac{\partial \mathcal{L}_{SH}(X,Y)}{\partial \theta} = n' & \left(\frac{\partial S(X,Y)}{\partial \theta} - \frac{1}{n'} \sum_{Y' \in \mathcal{Y}_m} \frac{\partial S(X,Y')}{\partial \theta} \right) \\ + n'' & \left(\frac{\partial S(X,Y)}{\partial \theta} - \frac{1}{n''} \sum_{X' \in \mathcal{X}_m} \frac{\partial S(X',Y)}{\partial \theta} \right). \end{aligned} \tag{8}$$

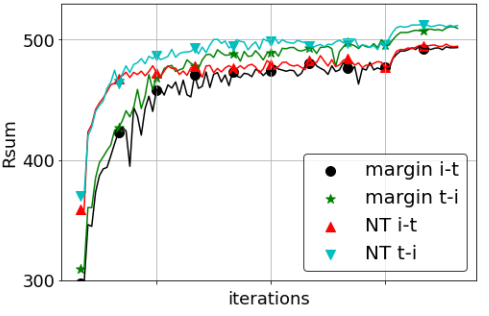

Figure 2: Validation recall sum (Rsum) of first 30 epochs. The models using NT-XEnt can be trained faster than margin losses, while the final results are similar.

Note that $\mathcal{X}_m$ and $\mathcal{Y}_m$ are the negative samples that are in the range of margin, n' and n'' are the length of $\mathcal{Y}_m$ and $\mathcal{X}_m$. The terms of negative samples are the simple average with flat distribution when dividing with their length.

The hardest negative sampling method substitutes the distribution as the delta function of the hardest negative sample.

$$\mathcal{L}_{MH}(Y|X) = \left[-S(X,Y) + S(X,Y'') + m\right]_+, \tag{9}$$

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

Note that X'' and Y'' are the hardest negative samples in each modality and MH is from 'max hinge' in [2]. And its gradient is given as,

$$-\frac{\partial \mathcal{L}_{MH}(X,Y)}{\partial \theta} = n'\left(\frac{\partial S(X,Y)}{\partial \theta} - \frac{\partial S(X,Y'')}{\partial \theta}\right) + n''\left(\frac{\partial S(X,Y)}{\partial \theta} - \frac{\partial S(X'',Y)}{\partial \theta}\right). \tag{11}$$

Note that n' and n'' are one when the hardest negative samples (Y'' and X'') are in the margin range, otherwise zero. Comparing the gradients of margin loss and Eq. 8, *NT-XEnt* is expected to be trained faster (Fig. 2).

5 Qualitative analysis

We visualize top-3 image retrieval results of RELAX (i-t attention) model given text queries on Flickr30k in Fig. 3 and Fig. 4. Both of object and relation attentions (mean value of softmax (XY^T)) are also presented. From these examples, we find that RELAX retrieves the images using both object-sentence pairs and predicate-relation pairs.

References

[1] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

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

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

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

[9] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015.

Query (a) : Five people wearing winter jackets and helmets stand in the snow, with snowmobiles in the background.
Parsed relations : [['people', 'snow', 'stand in'], ['people', 'snowmobile', 'stand with'], ['people', 'jacket', 'wear'], ['people', 'helmet', 'wear'], ['snowmobile', background', 'in']]

Query (b) : Two men sitting on the roof of a house while another one stands on a ladder.
Parsed relations : [['one', 'ladder', 'stand on'], ['man', 'roof', 'sit on'], ['roof', 'house', 'of']]

Query (c) : A baseball player coming into home plate with knee held high and the catcher right in front of him waiting for the ball.
Parsed relations : [['player', 'ball', 'wait for'], ['right', 'ball', 'wait for'], ['player', 'plate', 'come into'], ['player', 'knee', 'come with']]

Query (d) : A woman in a blue helmet and red pants riding a motorcycle.
Parsed relations : [['woman', 'motorcycle', 'ride'], ['woman', 'helmet', 'in'], ['woman', 'pants', 'in']]

Figure 3: Image retrieval results of top-3. The retrieved images are recalled at top-1.

Figure 4: Image retrieval results of top-3. The retrieved images are not recalled at top-1, but the wrong answers are relevant.