IMAGE-TEXT ALIGNMENT AND RETRIEVAL USING LIGHT-WEIGHT TRANSFORMER

Wenrui Li *Xiaopeng Fan*

Harbin Institute of Technology, School of Computer Science & Technology

ABSTRACT

With the increasing demand for multi-media data retrieval in different modalities, cross-modal retrieval algorithms based on deep learning are constantly updated. However, most of them have trouble with large model parameters and insufficient intrinsic nature between different modalities. We proposed a Light-weight Transformer Alignment Network (LTAN), which adopts the current mainstream visual and textual feature extraction methods. With convolutional neural network combined with light-weight transformer architecture and fully connected neural network, LTAN improves the generalization ability of the model while maintaining high performance. In order to extract visual features that lay emphasis on global details, enhancement paths are constructed to fuse precise location signals stored in low-level features with semantic information extracted from high-level to improve the model retrieval accuracy. It obtains the state-of-the-art results on image and sentence retrieval on MS-COCO and Flickr30k datasets. On the MS-COCO 1K test set, our model obtains an improvement of $\mathbf{3.9}\%$ and $\mathbf{2.5}\%$ respectively for the image and sentence retrieval tasks on the Recall@1 metric. The size of our model is $\mathbf{15}\%$ smaller than models using standard transformer as backbone.

***Index Terms*—** Cross-modal retrieval, deep learning, multi-modal matching, computer vision, natural language

1. INTRODUCTION

For cross-modal retrieval tasks, this paper focuses on visual and textual modalities. The image-text matching task has been studied extensively [1, 2, 3, 4, 5]. Recently many works use mainstream architectures for visual and textual processing. On the one hand, Convolutional Neural Networks (CNN) has dominated computer vision modeling for a long time [6]. CNN architecture becomes more powerful by greater scale, more extensive connections and increasing complexity like more sophisticated forms of convolution. On the other hand, transformer is the popular natural language processing architecture which deals with sequence modeling and translation task [7, 8]. Transformer has the self-attention mechanism to modeling data long-range dependencies.

VSE++ incorporates hard negatives in the loss function without any additional cost of mining [9]. MMCA adopts a novel cross-attention module by jointly modeling both inter-modality and intra-modality relationships in a unified deep model [10]. However, both of them need to study early interaction in raw image pixels. TERAN introduces a transformer-based model which processes images and sentences independently, matches them into the same common space and forces a fine-grained word-region alignment. This method reaches the state-of-the-art results in image-sentence matching but is heavy [11]. In light of this, we proposed Light-weight Transformer Alignment Network (LTAN), using enhancement paths and light-weight transformer layers to explore intrinsic nature between different modalities and reduce the model parameters, respectively.

The main contributions of this paper are as follows:

- We introduce the Light-weight Transformer Alignment Network (LTAN) which outperforms state-of-the-art models in terms of Recall@K (with K = $\{1, 5, 10\}$) on MS-COCO and Flickr30k datasets.
- We proposed enhancement paths to explore intrinsic nature in raw image pixels. Light-weight transformer architecture and fully connected neural network are used in LTAN to make it learn more extensively with fewer parameters.
- We perform an ablation study on slight modifications of our LTAN model, including weight sharing in last fully connected layers, the impact of using enhancement paths and different activation functions during training.

The rest of this paper is organized as follows. In Section 2, overall description is given. In Section 3, we show results on datasets and the ablation study. Finally, some conclusions are drawn in Section 4.

2. LIGHT-WEIGHT TRANSFORMER ALIGNMENT NETWORK (LTAN)

The overall architecture of LTAN is shown in Fig. 1, the visual and textual features processed by light-weight trans-

This work was supported in part by the National Key Research and Development Program of China (2021YFF0900500), and the National Science Foundation of China (NSFC) under grants 61972115 and 61872116.

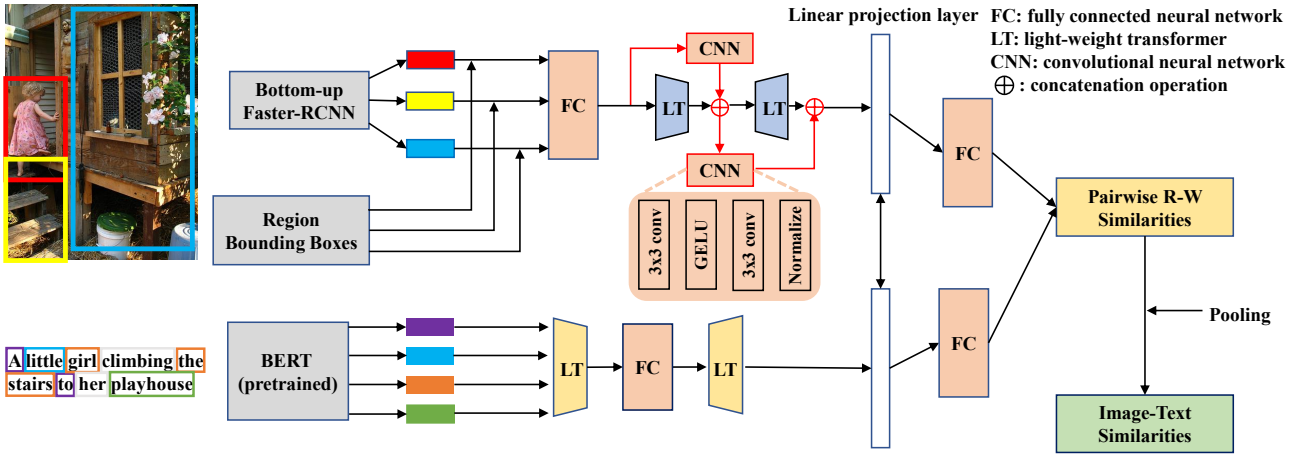

Fig. 1. The overall architecture of LTAN. Visual features and textual features are processed by Faster-RCNN and pre-trained BERT model respectively. In visual module, we construct enhancement paths in defferent LT layers to shorten the distance between low-level and high-level features, and use the precise positional signals which stored in low-level features to improve model accuracy. Concatenation is used for feature fusion to avoid the side effects of feature superposition on information. As for textual features, we add fully connected neural network between LT layers to make model learn more extensively. Finally, the calculated region-word (R-W) similarity matrix obtains the image-text (I-T) similarity by pooling functions.

former (LT) layer are passed through linear projection layer to make them have the same dimension. The enhancement path is indicated in red lines in Fig. 1. A symmetrical structure is constructed to fuse the precise location signals stored in low-level features with the semantic information extracted from high-level features. The CNN unit consists of two convolutional layers, a Gaussian Error Linear Units (GELU) as activation function and a normalization layer. The upper CNN unit receives visual features from Faster-RCNN, and the lower CNN unit receives the features after first fusion. Then, two different fully connected layers are used to process and generate visual and textual intrinsic nature. LTAN matches the basic components of image and text while preserving the information richness of both modalities.

The visual and textual processing pipelines in LTAN are built with LT layers extracted from DeLighT architecture [12]. The overall architecture of LT layer is shown in Fig. 2. LT layer uses single-head attention and light-weight feed forward network to replace multi-head attention and feed forward network. In particular, single-head attention is computed by using d_0-dimensional outputs through the scaled dot-product which is formulated as

$$Attention(K, Q, V) = softmax(\frac{QK^T}{\sqrt{d_0}})V, \qquad (1)$$

where Q, K and V are feature outputs, K^T represents the transpose of K. Transformations use group linear transformations to learn local representations by deriving the output from a specific part of the input.

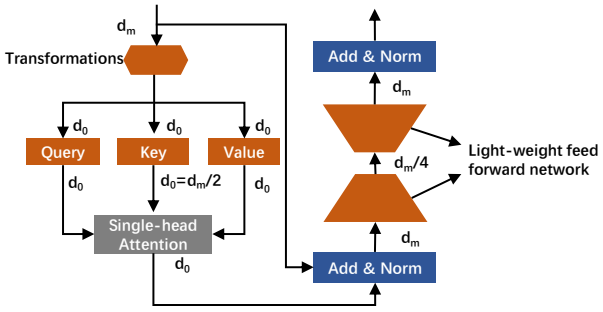

Fig. 2. The architecture of light-weight transformer layer

2.1. Visual Features Processing Pipeline

LTAN uses Faster-RCNN, a state-of-the-art object detector, to extract visual features [6]. Many downstream tasks utilize it to obtain salient object regions extracted from images. The visual features extracted from Faster-RCNN are conditioned with the information related to geometry bounding-boxes using fully connected neural network. In visual module, we construct enhancement paths in defferent LT layers to shorten the distance between low-level and high-level features, use precisely positional signals which stored in low-level features to improve model accuracy. Concatenation is used for feature fusion to avoid side effects of feature superposition on information. With CNN choosing features to fuse, the effectiveness of the model is improved.

Considering the generalization ability of the model, activation function used between CNN is Gaussian Error Linear

Units (GELUS) as

$$GELU(x) = 0.5x(1 + \tanh[\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)]), \quad (2)$$

where x is the input feature. LTAN uses LT layers and enhancement paths to mine the high-level semantic information and fuse early interaction in raw image pixels respectively. The linear projection layer makes the dimension of concatenated features same with model embedding space. Finally, the fused features are reasoned by fully connected neural network.

2.2. Textual Features Processing Pipeline

LTAN uses pretrained BERT to extract textual features [7]. BERT uses multi-layer transformer to capture the relationships between words in sentence through the powerful self-attention mechanism. BERT embeddings are trained on some general natural language processing tasks such as sentence prediction or sentence classification and demonstrate state-of-the-art results in many downstream natural language tasks.

In our BERT, we replaced transformer encoder layer with the LT layer using single-head attention mechanism. Although shuffle features can learn global information, the location information of original features would be disrupted, so we don't use it. Instead of feature mixers in DeLighT [12], we adopte fully connected neural network to fuse global information in different LT layers. Similarly, the linear projection layer makes the dimension of processed features same with model embedding space. The processed features are finally reasoned by fully connected neural network.

2.3. Similarity Between Different Modalities

In LTAN, the processed features through the two fully connected layers will be used to compute region-word (R-W) similarity matrix $RW \in \mathbb{R}^{|gk|\times|gl|}$, where gk is the set of indexes of region features from the k-th image and gl is the set of indexes of words from the l-th sentence. Using cosine similarity to calculate affinity between i-th region and j-th word. Let v_i and s_j be feature vector sets extracted from k-th image and l-th sentence respectively.

The R-W similarity matrix is used to get global similarity of k-th image and l-th sentence by pooling function: max-sum [11], which measures maximum value of each row of RW and sum it up. Max-sum pooling function as

$$P_{kl} = \sum_{j\in gl} \max_{i\in gk} \frac{v_i^T s_j}{\|v_i\| \, \|s_j\|}, \quad (3)$$

where $i \in gk, j \in gl$, $\|\cdot\|$ is the ℓ_2 norm. Our learning method is hinge-based triplet ranking loss [7], focusing on hard negatives $l^{'}$ and $k^{'}$. The alignment loss function as

$$L_{kl} = \max_{l^{'}}[\alpha + P_{kl^{'}} - P_{kl}]_+ + \max_{k^{'}}[\alpha + P_{k^{'}l} - P_{kl}]_+, \quad (4)$$

where α is margin parameter that defines the minimum separation between negative pairs and truly matching word-region embeddings, $[x]_+ \equiv \max(x, 0)$. The hard negatives are sampled from mini-batch but not globally.

3. EXPERIMENT

The single-head attention mechanism is used with $dropout$=0.1, normalization is $LayerNorm$ and activation function is GELU. The visual bottom-up features are extracted as the work of [13] and then pass through LT layer. The dimension of visual and textual features processed by LT layer is mapped to 1024-D which is the same as dimension of model embedding space. We trained for 30 epochs using $Adam$ optimizer with a batch size of 30. The $warmup$ method is adopted for training acceleration and the hard negative margin of triplet ranking loss is set as 0.2.

3.1. Results

The LTAN is compared with the following baseline: M3A-Net [3], GraDual [14], TIMAM [15], CASC [4], MMCA [10], CAMERA [16], VSE++ [9], SCAN [1], VSRN [17], Full-IMRAM [2], TERAN$_{MrSw}$ [11]. Many of the listed methods report their results using an ensemble of two models having different training initialization parameters, where the final similarity is obtained by averaging the scores in output from each model. But we believe that the ensemble method introduces complexity among different models. For purpose of unify standards, we only evaluate basic Model without ensemble.

Results on MS-COCO. The results are shown in Table 1. For this test, LTAN achieves the highest performance in all indicators. By concatenating features obtained by LT layer and CNN, intrinsic nature between features is further explored. LTAN outperforms previous state-of-the-art method (TERAN$_{MrSw}$) by 3.9%, 1.1% and 1% in image retrieval and 2.5%, 2.1% and 1.1% in sentence retrieval on Recall@1, Recall@5 and Recall@10, respectively.

Results on Flickr30k. Results are reported in Table 2. In image retrieval, LTAN improves results on Recall@1, Recall@10 and Recall@5 by 4.9%, 0.7% (GraDual) and 4.5% (MMCA) respectively. For sentence retrieval, our LTAN still performs very well. It obtains 3.2% of improvement on Recall@1 over Full-IMRAM, while the latter obtains the best result, slightly better than LTAN on Recall@10.

Algorithm characteristics. In Fig. 3, we show alignment loss and sum of Recall@K (rsum) of LTAN model on Flickr30k. Obviously, LTAN's loss function fluctuates greatly in process of convergence, which we believe the network is searching for better results positively, and it reflects our model's strong ability of learning. This may be related to random regularization introduced by GELU activation function we adopte. By observing the changes of our evaluation

Table 1. Results on MS-COCO dataset

	Image Retrieval			Sentence Retrieval		
model	R@1	R@5	R@10	R@1	R@5	R@10
M3A-Net [3]	58.4	87.1	94.0	70.4	91.7	96.8
CASC [4]	58.9	89.8	96.0	72.3	96.0	99.0
MMCA [10]	61.6	89.8	95.2	74.8	95.6	97.7
CAMERA [16]	62.3	90.1	95.2	75.9	95.5	98.6
VSE++ [9]	52.0	84.3	92.0	64.6	90.0	95.7
SCAN [1]	58.8	88.4	94.8	72.7	94.8	98.4
VSRN [17]	60.8	88.4	94.1	74.0	94.3	97.8
Full-IMRAM [2]	61.7	89.1	95.0	76.7	95.6	98.5
TERAN$_{MrSw}$ [11]	65.0	91.2	96.4	77.7	95.9	98.6
GraDual [14]	63.7	90.8	95.6	76.8	95.9	98.3
LTAN(ours)	**67.5**	**92.2**	**97.4**	**79.6**	**97.9**	**99.7**

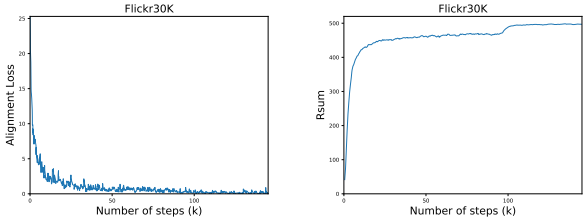

Fig. 3. Changes in various evaluation indicators

indicators, it is obvious that LTAN can quickly reach a stable optimal results in training process, and has potential to break through the bottleneck bulit in the first period of learning. Our parameter size is 181M, which is 65.8% and 83.8% of that of VILBERT [18] and TERAN respectively.

3.2. Ablation Study

Whether to share the weights in the last fully connected neural network. According to the results in 1st and 2nd rows in Table 3, it can be found that results of not sharing weights in fully connected neural network has a small gain. However, sharing the weights in fully connected neural network can reduce model parameters, make model more simply and difficult to overfit, increase stability of model. On the other hand, not sharing the weights can make model have higher potential, increase diversity of model, and make model learn better.

The impact of using enhancement paths. Better results (1st and 3nd rows in Table 3) are obtained with using enhancement paths (2.2% and 3.8% improvement in image retrieval and sentence retrieval on Recall@1 respectively). It is likely that the visual features reasoned through deep LT layer may loss accuracy for the location of features. By concatenation of low-level and high-level information, precisely feature location information in low-level is combined with semantic information in high-level, and the visual feature extraction

Table 2. Results on Flickr30k dataset

	Image Retrieval			Sentence Retrieval		
model	R@1	R@5	R@10	R@1	R@5	R@10
TIMAM [15]	42.6	71.6	81.9	53.1	78.8	87.6
M3A-Net [3]	44.7	72.4	81.1	58.1	82.8	90.1
MMCA [10]	54.8	81.4	87.8	74.2	92.8	96.4
VSE++ [9]	39.6	70.1	79.5	52.9	80.5	87.2
SCAN [1]	48.6	77.7	85.2	67.4	90.3	95.8
VSRN [17]	53.0	77.9	85.7	70.4	89.2	93.7
Full-IMRAM [2]	53.9	79.4	87.2	74.1	93.0	96.6
TERAN$_{MrSw}$ [11]	59.5	84.9	90.6	75.8	93.2	96.7
GraDual [14]	57.7	84.1	90.5	76.1	94.7	**97.7**
LTAN(ours)	**60.5**	**85.1**	**91.1**	**76.5**	**93.5**	96.5

Table 3. Results from ablation study

	Image Retrieval			Sentence Retrieval		
model	R@1	R@5	R@10	R@1	R@5	R@10
LTAN	60.5	85.1	91.1	76.5	93.5	96.5
LTAN$_{shareW}$	59.4	84.9	90.7	75.7	92.5	95.8
LTAN$_{NoEnPath}$	59.2	84.9	90.7	73.7	91.4	96.1
LTAN$_{Relu}$	58.5	84.2	90.5	74.0	92.4	96.0

module pays more attention to global information and locates the features more accurately.

Differences between activation functions Relu, Sigmoid and GELU. According to the results in 1st and 4th rows in Table 3, It is shown that Relu activation function is slightly inferior to GELU activation function on Recall@K, but sigmoid activation function has very slow convergence speed and poor effect. It is obvious that power operation in sigmoid activation function is time-consuming. Due to sigmoid function characteristics, the convergence speed of model is slow. Convergence of Relu activation function is stable and results are good. The GELU activation function introduces the idea of random regularization, intuitively better understanding of nature, and has greater potential and best effect in results.

4. CONCLUSION

We consturct the LTAN based on light-weight transformer and convolutional neural network. It outperforms the previous state-of-the-art methods on MS-COCO and Flickr30K datasets. In order to study early interaction in raw image pixels, we add enhancement paths in LTAN to shorten distance between low-level and high-level features. Using light-weight transformer to reduce the model size. Experiments show that our LTAN is more effective and stable than other methods. In our future work, we will give priority to how to use unpaired datasets on internet to study semi-supervised learning algorithms to improve efficiency of image-text retrieval.

5. REFERENCES

[1] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Zhong Ji, Zhigang Lin, Haoran Wang, and Yuqing He, "Multi-modal memory enhancement attention network for image-text matching," *IEEE Access*, vol. 8, pp. 38438–38447, 2020.

[4] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen, "Cross-modal attention with semantic consistence for image–text matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5412–5425, 2020.

[5] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*, 2019.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "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)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 6000–6010, Curran Associates Inc.

[9] F. Faghri, D.J. Fleet, J.R. Kiros, and S. Fidler, "VSE++: improved visual-semantic embeddings with hard negatives," in *British Machine Vision Conference (BMVC)*, Newcastle, 2018.

[10] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet, "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. abs/2008.05231, 2020.

[12] Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi, "Delight: Deep and light-weight transformer," in *International Conference on Learning Representations*, 2021.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon, "Gradual: Graph-based dual-modal representation for image-text matching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 3459–3468.

[15] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[16] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian, "Context-aware multi-view summarization network for image-text matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, MM '20, p. 1047–1055, Association for Computing Machinery.

[17] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 13–23.