MULTI-STEP SELF-ATTENTION NETWORK FOR CROSS-MODAL RETRIEVAL BASED ON A LIMITED TEXT SPACE

*Zheng Yu, Wenmin Wang *, Ge Li*

School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
yuzheng@pku.edu.cn, wangwm@ece.pku.edu.cn, geli@ece.pku.edu.cn

ABSTRACT

Cross-modal retrieval has been recently proposed to find an appropriate subspace where the similarity among different modalities, such as image and text, can be directly measured. In this paper, we propose **M**ulti-step **S**elf-**A**ttention **N**etwork (MSAN) to perform cross-modal retrieval in a limited text space with multiple attention steps, that can selectively attend to partial shared information at each step and aggregate useful information over multiple steps to measure the final similarity. In order to achieve better retrieval results with faster training speed, we introduce global prior knowledge as the global reference information. Extensive experiments on Flickr30K and MSCOCO, show that MSAN achieves new state-of-the-art results in accuracy for cross-modal retrieval.

Index Terms— Cross-modal retrieval, Multi-step self-attention, Limited text space, Global prior knowledge

1. INTRODUCTION

Due to the intrinsical heterogeneity of multimedia data, the main challenge in cross-modal retrieval is how to embed heterogeneous multimedia data into a homogeneous space, so that their similarity can be measured directly. Focusing on the retrieval between image and text, we address two problems here.

The first problem is how to learn efficient features. Most traditional methods [1, 2, 3, 4] simply extract global features for both image and text by CNN [5, 6] or RNN. However, they ignore the fact that global features always contain massive redundant information, that is, modality-specific information. Modality-specific information is unique, which may not exist in any other modalities in addition to itself. Recently, some attention-based methods [7, 8, 9] try to extract a list of features for image regions and words. However, they only consider the object-level alignments between image regions and words but ignore the rich relation information which may play an indispensable role in cross-modal retrieval. As mentioned in [10], image captioning models [11, 3, 12] can be used to learn image features with rich relation information. Given an input image, we can get sensible descriptive sentences which contain nouns and verbs. That is, image captioning models are able to not only recognize the objects in the image (nouns), but also preserve rich relation information among different objects (verbs). Therefore, we adopt image captioning models to make up for the shortcomings of the traditional CNN features.

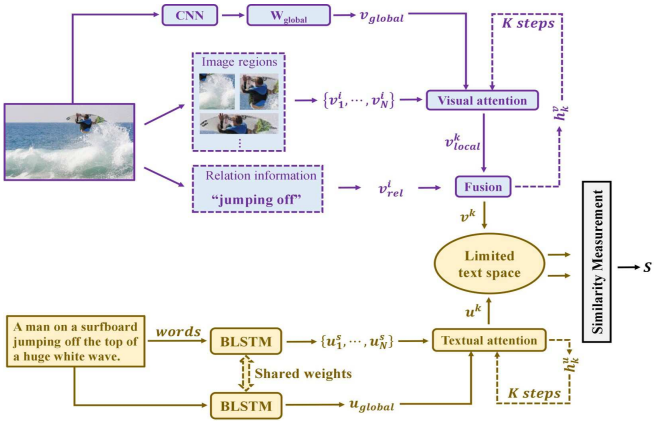

Fig. 1. An overview of our proposed Multi-step Self-Attention Network (MSAN). Paths in purple and gold represent the network of visual attention and textual attention respectively.

The second problem is how to find a homogeneous space. Since we only focus on the retrieval between image and text, cross-modal retrieval can be achieved by a common space [13, 14, 1, 15], a text space [16, 17, 10] or an image space [2]. For the human brain, textual features are closer to human understanding (and language) than the pixel-based features [18]. Thus a text space can better simulate the human cognitive behavior during retrieval. In most cases, people only need to remember some of the commonly used words to meet their daily needs. Accordingly, we aim to explore the possibility of performing cross-modal retrieval in a limited text space [10]. The ability for the text space to understand is limited by the size of the vocabulary. The bigger the vocabulary, the stronger

This project was supported by Shenzhen Peacock Plan (20130408-183003656), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467), and National Natural Science Foundation of China (NSFC, No.U1613209).

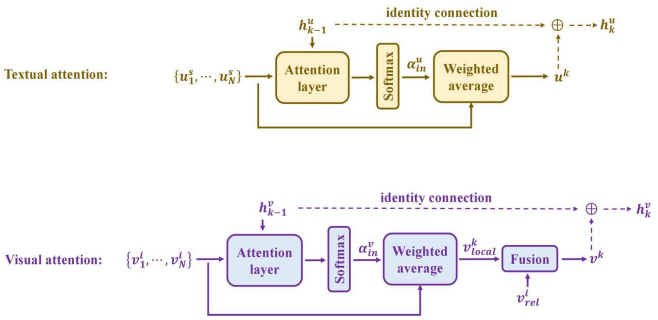

Fig. 2. Detailed Illustration of the multi-step self-attention network.

the understanding ability. Increasing the number of words blindly will not improve the retrieval performance but increase the complexity of the method in time and space.

To address the two problems mentioned above, we propose Multi-step Self-Attention Network (MSAN). Given an image-text pair, MSAN first extracts a list of features from image regions and key words. Then, MSAN adopts self-attention strategy to learn object-level alignments between image and text, and obtain the local features by weighted average. Following [10], we add a fusion layer on top to fuse the image local features and relation features, as well as embed them into a limited text space. Considering the step-by-step nature of human cognition during retrieval, MSAN adopts multiple attention steps to recurrently attend to partial shared information at each step, and then aggregate useful information over multiple steps to distill the shared information as much as possible. Finally, in order to achieve better retrieval results with faster training speed, we introduce global prior knowledge as the global reference information. The final similarity is obtained by the summation of the similarity at each step.

Our core contributions are summarised as follows:

- In addition to object-level alignments, MSAN is able to capture rich relation information ignored by prior attention-based methods, which plays an indispensable role in cross-modal retrieval.
- We introduce global prior knowledge as the global reference information, which is able to achieve better retrieval results with faster training speed.

2. PROPOSED METHOD

2.1. Feature Extraction

Image representation As shown in the purple path in Fig. 1, features for image regions are extracted from the last pooling layer of VGG19 (pool5) [5]. The pool5 layer consists of 512 feature maps and the size of each feature map is 7×7, which means that the number of image regions is 49 and each is represented by a 512 dimensional vector. Given an input image i, we can extract a list of features $\{v_1^i, ..., v_N^i\}$, where N denotes the total number of image regions and v_n^i is a 512 dimensional vector for the n-th region. As for relation feature, following [10], we regard the 512-dimensional NIC feature as the relation feature v_{rel}^i, which contains rich relation information.

Text representation As shown in the golden path in Fig. 1, we employ bidirectional LSTM to extract d dimensional features for each word in the text, which takes the form:

$$\begin{aligned} h_t^{fw} &= \mathrm{LSTM}_{\mathrm{fw}}\left(x_t, h_{t-1}^{fw}\right), \\ h_t^{bw} &= \mathrm{LSTM}_{\mathrm{bw}}\left(x_t, h_{t-1}^{bw}\right), \\ u_n^s &= \frac{\left(h_t^{fw} + h_t^{bw}\right)}{2}, \end{aligned} \tag{1}$$

where x_t represent the input word at time t. h_t^{fw} and h_t^{bw} represent the hidden states at time t from the forward LSTM and backward LSTM respectively. The feature u_n^s for each word is obtained by averaging h_t^{fw} and h_t^{bw}. Finally, we can extract a list of features $\{u_1^s, ..., u_N^s\}$ from each word in the text s, where u_n^s denotes the d-dimensional feature vector for the n-th word. Specially, d is also the dimensionality of the limited text space.

2.2. Multi-Step Self-Attention Network

Fig. 2 shows the detailed self-attention network employed at each step k, which contains two separate paths for image and text respectively. Accordingly, self-attention network is able to make the image or text learn to attend to itself without any image-text pair information.

Visual attention Given a list of features $\{v_1^i, ..., v_N^i\}$ for each region in image i, the image local feature v_{local}^k at step k is given by:

$$\begin{aligned} v_{local}^k &= \sum_{n=1}^{N} \alpha_{in}^v v_n^i, \\ \alpha_{in}^v &= \mathrm{softmax}\left(f_{att}\left(v_n^i, h_{k-1}^v\right)\right), \\ f_{att}(v_n^i, h_{k-1}^v) &= \tanh(W_v^k v_n^i) \odot \tanh(W_{vh}^k h_{k-1}^v), \end{aligned} \tag{2}$$

where h_{k-1}^v denotes the previous context vector for image. α_{in}^v represents the attention weight corresponding to the n-th image region. The image local feature v_{local}^k at step k is computed as an average of the image region features weighted with attention weight α_{in}^v. $f_{att}(v_n^i, h_{k-1}^v)$ denotes the visual self-attention function which computes the unnormalized attention weight for the n-th image region. W_v^k and W_{vh}^k represent the trainable parameters of the visual attention layer in Fig. 2.

Then, following [10], we add a fusion layer to fuse v_{local}^k and v_{rel}^i, as well as embed them into a limited text space:

$$\begin{aligned} v_{local}^k &= \mathrm{BN}\left(W^k v_{local}^k\right), \\ v^k &= \mathrm{Relu}\left(\mathrm{BN}\left(v_{local}^k + v_{rel}^i\right)\right), \end{aligned} \tag{3}$$

where v^k denotes the limited text space feature for image i, and W^k embeds v_{local}^k into a limited text space.

Textual attention We have obtained a list features $\{u_1^s, ..., u_N^s\}$ for each word in the text s. At step k, similar to visual attention, the text local feature u^k is given by:

$$\begin{aligned} u^k &= \sum_{n=1}^{N} \alpha_{in}^u u_n^s, \\ \alpha_{in}^u &= \text{softmax}\left(f_{att}\left(u_n^s, h_{k-1}^u\right)\right), \\ f_{att}(u_n^s, h_{k-1}^u) &= \tanh(W_u^k u_n^s) \odot \tanh(W_{uh}^k h_{k-1}^u), \end{aligned} \tag{4}$$

where h_{k-1}^u denotes the previous context vector for text. α_{in}^u represents the attention weight corresponding to the n-th word. The text local feature u_{local}^k is obtained from an weighted average of the word features. $f_{att}(u_n^s, h_{k-1}^u)$ denotes the self-attention function for the text. W_u^k and W_{uh}^k represent the trainable parameters of the textual attention layer in Fig. 2.

Context vector In order to encode the context information that has already been attended to, we employ an extra identity connection to obtain the context vector h_k^v and h_k^u at the next step inspired by ResNet [6], which takes the form:

$$\begin{aligned} h_k^v &= \text{V\_att}\left(v^k, h_{k-1}^v\right) + h_{k-1}^v, \\ h_k^u &= \text{T\_att}\left(u^k, h_{k-1}^u\right) + h_{k-1}^u, \end{aligned} \tag{5}$$

where $k \in \{1, \cdots, K\}$ and V_att and T_att represent the procedure of visual attention and textual attention respectively. The identity connection is able to control the flow of information and pass on information that needs to be preserved to the next step.

Instead of initializing h_0^v and h_0^u by mean vectors, we introduce global prior knowledge which acts as a "mentor". Global prior knowledge is able to help the self-attention network locate key information quickly and thus lead to faster convergence and better accuracy. The initialization is give by:

$$\begin{aligned} h_0^v &= v_{global} = W_{global} \cdot f_{vgg}, \\ h_0^u &= u_{global}, \end{aligned} \tag{6}$$

where v_{global} and u_{global} represent the global features for image and text respectively, which can be regarded as the global reference information as well. W_{global} embeds 4096-dimensional VGG feature f_{vgg} into the limited text space. u_{global} is the averaged hidden state of BiLSTM at the last time step.

Finally, we perform K steps to recurrently attend to partial shared information, and the self-attention mechanism employed at each step k stays the same.

2.3. Similarity Measurement

Since we have obtained the image and text local features at each step k, the next step is to compare their similarity respectively. We define a scoring function $s(v,\ u) = v \cdot u$, where v and u represent the image and text local features respectively. To make s equivalent to cosine similarity, v and t are first scaled to have unit norm by the L2Norm layer. Accordingly, the similarity s^k at step k is give by:

$$s^k = v^k \cdot u^k, \tag{7}$$

Table 2. The effect of different numbers of attention steps on Flickr30K. K denotes the total number of attention steps.

	Img2Txt			Txt2Img		
	R@1	R@5	R@10	R@1	R@5	R@10
K = 1	**44.0**	73.0	82.6	33.2	64.0	75.4
K = 2	43.0	**73.7**	**83.7**	**33.5**	**64.5**	**75.4**
K = 3	40.5	71.1	80.1	32.2	62.3	73.0

we add up the similarity at each step to obtain the final similarity S:

$$S = \sum_{k=0}^{K} s^k. \tag{8}$$

Then, pairwise ranking loss function is exploited to optimize the model.

3. EXPERIMENTS

In this section, we perform extensive experiments on Flickr30K [20] and MSCOCO [21] following the dataset splits in [14]. Evaluation is performed using Recall@K (with K = 1, 5, 10), which computes the mean number of images (texts) for which the correct texts (images) is ranked within the top-K retrieved results. Higher Recall@K indicates better results.

3.1. Implementation Details

To demonstrate the efficiency of self-attention mechanism and relation information in MSAN, we design the following variants: MSAN-obj abandons the relation information v_{rel}^i and only considers the object-level alignments between image regions and key words; MSAN-glob removes the multi-step self-attention network and simply use the global features for both image and text; MSAN is our full model with self-attention mechanism and relation information.

We set d to 1024 and W_{global} is a 4096×1024 embedding matrix. The dimension of each attention layer is set to 512 with dropout ratio 0.5. The number of attention steps K is set to 2, which empirically shows the best experimental results. And the margin m is set to 0.3 in all our experiments. During training, we adopt Adam optimizer to optimize the model with learning rate 0.0002 for the first 10 epochs and then decay the learning rate by 0.1 for the remaining 10 epochs. We use a mini-batch size of 128 in all our experiments.

3.2. Comparison with the State-of-the-art

First, we compare MSAN with several current state-of-the-art methods on Flickr30K and MSCOCO in Table 1. Img2Txt and Txt2Img denote image-to-text retrieval and text-to-image retrieval respectively. From Table 1, we can observe that MSAN achieves new state-of-the-art results in accuracy for cross-modal retrieval on all datasets, which demonstrates the

Table 1. Image-to-text and text-to-image retrieval results on Flickr30K and MSCOCO.

	Flickr30K						MSCOCO					
	Img2Txt			Txt2Img			Img2Txt			Txt2Img		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DSPE [1]	40.3	68.9	79.9	29.7	60.1	72.1	50.1	79.7	89.2	39.6	75.2	86.9
sm-LSTM [7]	42.4	67.5	79.9	28.2	57.0	68.4	52.4	81.7	90.8	38.6	73.4	84.6
DAN (VGG) [8]	41.4	73.5	82.5	31.8	61.7	72.5						
LTS [10]	31.2	62.5	75.8	21.5	48.9	61.5	45.5	78.7	88.8	30.2	66.0	80.5
VSE++ [19]	31.9	-	68.0	23.1	-	60.7	43.6	-	84.6	33.7	-	81.0
HM-LSTM [15]	38.1	-	76.5	27.7	-	68.8	43.9	-	87.8	36.1	-	86.7
MSAN-obj	37.6	64.6	75.8	27.3	56.2	67.8	46.1	80.0	89.5	36.6	71.3	85.1
MSAN-glob	35.4	66.7	76.5	26.9	57.3	70	46.2	80.4	89.1	38.1	73.5	86.1
MSAN	**43.0**	**73.7**	**83.7**	**33.5**	**64.5**	**75.4**	**52.9**	**86.5**	**94.4**	**43.0**	**79.0**	**89.4**

Table 3. The effect of global prior knowledge on Flickr30K. "MSAN with prior" is trained with global prior knowledge and "MSAN w/o prior" is trained without global prior knowledge.

	Img2Txt			Txt2Img		
	R@1	R@5	R@10	R@1	R@5	R@10
MSAN with prior	**43.0**	**73.7**	**83.7**	**33.5**	**64.5**	**75.4**
MSAN w/o prior	42.0	72.1	81.9	32.9	63.6	74.6

efficiency of MSAN. For a fair comparison, the results of VSE++ are based on 1-crop VGG image features without fine-tuning. And sm-LSTM represents the best single model without ensemble. Better results can be observed in [19] and [7] respectively.

When comparing between MSAN-glob and MSAN, we can observe that our multi-step self-attention network is very effective, since MSAN outperforms MSAN-glob on all datasets. Meanwhile, MSAN achieves significant improvement in accuracy compared with MSAN-obj, which reveals the importance of relation information.

3.3. Effect of the Number of Attention Steps

In Table 2, we show the experimental results of our full model MSAN with different numbers of attention steps (K = 1,2,3) on Flickr30K. We can observe that MSAN achieves the best results when $K = 2$ for Flickr30K. Note that when K grows bigger, the performance degrades obviously due to the potential over-fitting problem. Therefore, we set $K = 2$ for Flickr30K and MSCOCO, which empirically shows the best results.

3.4. Effect of Global Prior Knowledge

To demonstrate the importance of global prior knowledge, we show the experimental results for two variants of our full model: "MSAN with prior" and "MSAN w/o prior". As shown in Table 3, "MSAN with prior" performs better than " MSAN w/o prior", especially on image-to-text retrieval. So global prior knowledge is able to improve the retrieval accuracy.

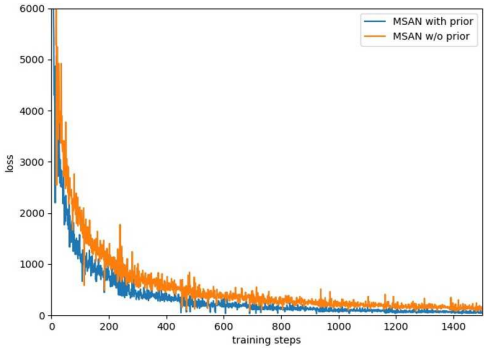

Fig. 3. Illustration of two curves that each depicts the change of losses during training. The orange and blue curve reflect the change of losses for "MSAN w/o prior" and "MSAN with prior" respectively.

Moreover, global prior knowledge is able to accelerate convergence during training. As shown in Fig. 3 the orange curve ("MSAN w/o prior") lies above the blue curve ("MSAN with prior"), which demonstrates that "MSAN with prior" trains more faster than "MSAN w/o prior".

Therefore, due to the use of global prior knowledge as global reference information, we can achieve better retrieval results with faster training speed.

4. CONCLUSIONS

In this paper, we propose a novel model MSAN to simulate the procedure of human cognitive behaviour during retrieval, aiming to perform cross-modal retrieval in a limited text space with multiple self-attention steps. Extensive experiments on three benchmark datasets demonstrate the efficiency of our proposed model. In the future, we will explore the efficiency of relation features more deeply and try some stronger CNNs, such as ResNet.

5. REFERENCES

[1] Liwei Wang, Yin Li, and Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5005–5013.

[2] Jianfeng Dong, Xirong Li, and Cees GM Snoek, "Word2visualvec: Cross-media retrieval by visual feature prediction," *arXiv preprint arXiv:1604.06838*, 2016.

[3] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[4] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen, "Dual-path convolutional image-text embedding," *arXiv preprint arXiv:1711.05535*, 2017.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Yan Huang, Wei Wang, and Liang Wang, "Instance-aware image and sentence matching with selective multimodal lstm," *arXiv preprint arXiv:1611.05588*, 2016.

[8] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim, "Dual attention networks for multimodal reasoning and matching," *arXiv preprint arXiv:1611.00471*, 2016.

[9] Yuxin Peng, Jinwei Qi, and Yuxin Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *arXiv preprint arXiv:1708.04776*, 2017.

[10] Zheng Yu, Wenmin Wang, and Mengdi Fan, "Learning a limited text space for cross-media retrieval," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2017, pp. 292–303.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3156–3164.

[12] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[13] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[14] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[15] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua, "Hierarchical multimodal lstm for dense visual-semantic embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1881–1889.

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al., "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[17] Ines Chami, Youssef Tamaazousti, and Hervé Le Borgne, "Amecon: Abstract meta-concept features for text-illustration," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017, pp. 347–355.

[18] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[19] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," 2017.

[20] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[21] 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 *European conference on computer vision*. Springer, 2014, pp. 740–755.