AUGMENTED VISUAL-SEMANTIC EMBEDDINGS FOR IMAGE AND SENTENCE MATCHING

Zerui Chen[1,3] *Yan Huang*[1,3] *Liang Wang*[1,2,3,4]

[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
[4]Chinese Academy of Sciences Artificial Intelligence Research (CAS-AIR)
zerui.chen@cripac.ia.ac.cn, {yhuang, wangliang}@nlpr.ia.ac.cn

ABSTRACT

The task of image and sentence matching has witnessed significant progress recently, but it is still challenging arising from the tremendous semantic gap between a pixel-level image and its matched sentences. Due to limited training data, it is rather challenging to optimize the visual-semantic embeddings. In this work, we propose to augment visual-semantic embeddings via enlarging the training dataset. With more data, models can learn discriminative features with high-quality semantic concepts. More specifically, we augment data by generating sentences for given images. Our method consists of two steps. At first, to enlarge the training dataset, given an image, we perform image captioning. Instead of introducing redundancy to our augmented dataset, we hope that our generated sentences are in diverse style and maintain its fidelity at the same time. Therefore, we consult to generative adversarial networks (GANs) which can produce more flexible expressions compared to methods based on the maximum likelihood principle. Then, we augment visual-semantic embeddings with the augmented training dataset and obtain the model for the task of image and sentence matching. Experiments on the popular benchmark demonstrate the effectiveness of our method by achieving superior results compared to our baseline.

Index Terms— Generative Adversarial Networks, Image and Sentence Matching, Visual-Semantic Embeddings

1. INTRODUCTION

The task of image and sentence matching aims to find a way to measure the semantic similarity between an image and a sentence. It plays a vital role in many applications, e.g., given an image query to find similar sentences, namely image annotation, and given a sentence query to retrieve matched images, namely sentence based image search. While much progress [1, 2] has been witnessed in this area recently, narrowing the semantic gap between a pixel-level image and its matched sentences remains a challenge. MSCOCO dataset, the most significant benchmark for this task, contains less than 500000 sentences in training subset. Limited data has always been the bottleneck for improving model performance further, making it hard to optimize the visual-semantic embeddings. In this work, what we focus is not to propose a new pipeline for the task of image and text matching, we intend to augment data with GANs [3] and learn more discriminative semantic concepts for this task. Our method is not only limited to methods based on visual-semantic embeddings but also can generalize to different kinds of approaches in this field for further performance improvement.

Our method can be mainly summarized into two steps. The first step is to apply conditional GANs to perform image captioning and to enlarge the training dataset. The quality of visual-semantic embedding has been dramatically improved recently. The embedding features can capture fruitful semantic information for images and sentences. However, it is still a challenge to learn discriminative visual-semantic embeddings under the circumstance of limited training data and limited model complexity. In this work, instead of increasing the complexity of the model for performance improvement, we intend to enlarge the training dataset with generated data. Augmenting training data via generating sentences is not a trivial task. We cannot merely apply traditional image caption framework based on the maximum likelihood principle or the distributions of our generated data and original training data will be roughly the same. Considering that we hope to generate diverse and natural sentences for a given image, we consult to GANs equipped with reinforcement learning for sequence generation as in [4]. Different from [4] which focuses on producing natural and accurate image captions, we attach attention to learning better visual-semantic embeddings for the task of image and sentence matching. By uti-

lizing augmented training dataset with generated data, we can lift the model performance. The second step is to learn visual-semantic embeddings with our enlarged training dataset. In our experiment, we adopt VSE++ [5] model as our baseline, which proposes a triplet ranking loss designed for image and sentence matching. To demonstrate the effectiveness of our method, we conduct several experiments on the large-scale MSCOCO dataset and achieve superior results than our baseline.

The proposed method has two merits. First, enlarging the training dataset can benefit the retrieval model to learn more solid visual-semantic embedding under the condition of limited training data and limited model complexity. Second, our method can generalize to nearly all the methods in the task of image and sentence matching. Our practice lifts the performance of VSE++ baseline and achieves competitive results. Beyond that, compared with models with complex structures, models with clear and simple structures can achieve comparable performance when equipped with our method.

2. BACKGROUND

In this section, we review recent progress on the task of image and sentence matching and the task of image captioning. In our method, we perform image captioning to enlarge the training dataset and then augment visual-semantic embeddings for image and sentence matching.

2.1. Visual-semantic Embedding in Image and Sentence Matching

Frome et al. [6] propose the first visual-semantic embedding framework in which CNN [7] and Skip-Gram [8] are set as the optimization objective. Under the similar framework, Kiros et al. [9] apply LSTM [10] to learn feature representations for sentences. Vendrov et al. [11] try to encode visual-semantic hierarchy in the objective function. Additionally, Wang et al. [12] intend to learn structure-aware representations under within-view constraints. Yan and Mikolajczyk [13] apply deep canonical correlation analysis to associate the image and sentence. Huang et al. [14] reason the semantic order for more discriminative feature representations for both image and sentence. Lee et al. [2] use the stacked cross attention mechanism to achieve very competitive results.

2.2. Image Captioning

In recent years, the Encoder-and-Decoder paradigm proposed in [15] is the mainstream framework. Many state-of-art methods [16, 17, 18] utilize the maximum likelihood principle for learning. However, as mentioned above, we hope our generated data is different from data in the training dataset. Therefore, it is unreasonable to follow the classic framework based on the maximum likelihood principle directly. Compared with traditional framework, methods based on GANs [4] can generate diverse and natural sentences for a given image. With the maximum likelihood principle, the model aims to create similar word patterns existing in training samples but often overlooks words with low-frequency in training data. For sequence generation methods based on GANs, low frequency words can be emphasized in reinforcement learning, leading to more diverse expressions. Therefore, in order to generate varied sentences while maintaining fidelity, we consult to methods based on conditional GANs for augmenting training data.

3. ALGORITHM

In this section, we summarize our method for the task of image and sentence matching. Section 3.1 describes how to utilize conditional GANs to enlarge the training dataset. Section 3.2 describes how to conduct image and sentence matching via visual-semantic embeddings.

3.1. Data Augmentation with Conditional GANs

Given an image, first, we perform image captioning to enlarge the training dataset. We consult to conditional generative adversarial networks for this task. The entire pipeline consists of two components: a Generator (G) and a Discriminator (D). Their structures are depicted in Figure 1 (a) and (b) respectively.

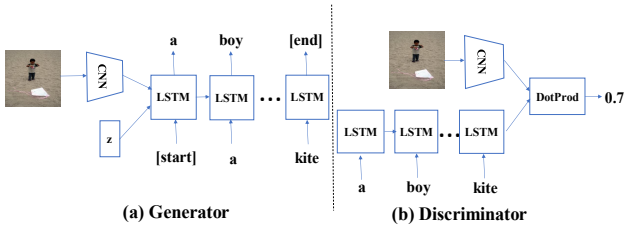

Fig. 1. The overall structure of our image captioning model.

In the framework, the Generator (G) takes into input image and a noise vector which controls the diversity of generated sentences. We adopt VGG16 [19] as our backbone network to extract feature representation for a given image. After extracting the image feature, LSTM network [10] is to model words' distribution and outputs a sentence word by word. Specifically, LSTM contains a sequence of latent states $(s_0, s_1, ...s_n)$ encoding continuous feature for previous words. Then a word w_t drawn from the latent distributions $p(w|s_t)$ step by step, forming a complete sentence.

The Discriminator (D) aims to measure the semantic similarity between the image and sentence. It consists of two branches. The first branch is to encode the image feature with CNN, specifically the VGG16 in our setting. The second branch is to encode the sentence feature with LSTM net-

work. Then the Discriminator decides how well the sentence describes a given image by calculating cosine distance between the image feature vector and sentence feature vector and the semantic similarity is measured as:

$$Similarity(I, S) = \sigma(f(I, \theta_I), g(S, \theta_S)) \tag{1}$$

Where (θ_I, θ_S) denotes parameters in image and sentence encoder network respectively. $f(I, \theta_I)$ and $g(S, \theta_S)$ represent normalized encoded feature vectors for image and sentence respectively. σ is dot product function that measures similarity between image and sentence and casts value into $[0, 1]$.

Compared with image generation, it is challenging to apply GANs for sentence generation. The difficulty lies in the word generation procedure when we need to sample from the hidden distribution in LSTM network in the Generator, and this procedure is unpredictable, making it hard to optimize. Following [4], we consult to policy gradient, a classic method in reinforcement learning [20]. The overall optimization procedure can be mainly divided into two parts, one for the Generator and the other for the Discriminator.

For the Generator (G), we cast the word generation procedure as a reinforcement learning problem. The LSTM network is an agent interacting with the outside environment and input of the agent at every step is the hidden state vector. This agent defines policy, and the result of this policy can be viewed as the generated word at every time step. After the agent finishes the task of sequence generation, it will get a reward. In our work, the reward is obtained from the Discriminator. Our optimization objective is to update the parameters in the Generator and to maximize the expected reward. We can define loss function as follows:

$$\begin{aligned} L_\theta &= -\sum_{w_1^g,...w_T^g} p_\theta(w_1^g, ..., w_T^g) r(w_1^g, ..., w_T^g) \\ &= -E_{(w_1^g,...,w_T^g)\sim p_\theta} r(w_1^g, ..., w_T^g) \\ &= -E_{w_1 \sim p_\theta(w_1)} ... E_{w_T \sim p_\theta(w_T|w_{1:T-1})} \sum_{t=1}^{T} r(w_{1:t}) \end{aligned} \tag{2}$$

Where, w_n^g is the chosen word at step n, and r denotes the reward for the generated sentence. p_θ is a parameterized policy and here is the LSTM network. Then we go further to derive the gradient of our optimization objective as:

$$\begin{aligned} \nabla_\theta L(\theta) &= -\sum_{w_1^g:w_T^g} \nabla_\theta p_\theta(w_1^g : w_T^g) r(w_1^g : w_T^g) \\ &= -\sum_{w_1^g:w_T^g} p_\theta(w_1^g : w_T^g) \nabla_\theta log(p_\theta(w_1^g : w_T^g)) r(w_1^g, ..., w_T^g) \\ &\approx -\sum_{t=1}^{T} \nabla_\theta log(p_\theta(w_t^g | w_{1:(t-1)}^g)) \sum_{t=1}^{T} r(w_{1:t}^g) \end{aligned} \tag{3}$$

Thus, we can update the parameters in the Generator (G) according to rewards obtained from the Discriminator (D) with the technique of standard gradient descent.

For the Discriminator (D), its function is to evaluate how well pairs of image and sentence matched and the naturalness of generated sentences. We design three types of loss functions to discriminate the generated sentence, which is defined as follows:

$$\begin{aligned} minL &= \frac{1}{N} \sum_{i=1}^{N} L(I, S) \\ &= \frac{1}{N} \sum_{i=1}^{N} (L_{S \in S_T}(I, S) + L_{S \in S_F}(I, S) + L_{S \in S_M}(I, S)) \end{aligned} \tag{4}$$

Where N is the number of training images. I and S represent sets of image and sentence respectively. S_T and S_F are ground truth sentences and generated sentences for given images respectively. S_M contains semantic-matched sentences for given images. The Discriminator (D) is a classification network, and our three criteria enable the network not only to discriminate fake or real sentences for a given image but also to evaluate the semantic similarity between two instances.

3.2. Visual-Semantic Embeddings

After we enlarge the training dataset, what we do next is to match pairs of image and sentence. For the task of image and sentence retrieval, it contains two sub-tasks. When the query is a caption, the task is to retrieve the most relevant image in the database. When the query is an image, the task is to retrieve the most relevant sentence in the database. The optimization goal is to improve the scores for well-matched pairs and suppress scores for irrelevant pairs, which is known as recall at *K* ($R@K$).

We intend to learn visual-semantic embeddings, mapping image features and sentence features into the joint embedding space which is convenient to measure the semantic similarity. We use the VGG19 network and LSTM network to extract feature vectors from image and sentence respectively. Then we use linear projections to map them into a joint embedding space in which we apply the cosine distance as a measurement for the semantic similarity between images and sentences. Then, in order to push the network to separate well-matched pairs away from irrelevant pairs, we adopt the ranking loss defined as follows:

$$\begin{aligned} L(i, c) = &\max_{c'}[\alpha + s(i, c') - s(i, c)] \\ &+ \max_{i'}[\alpha + s(i', c) - s(i, c)] \end{aligned} \tag{5}$$

Where this loss consists of two terms, the first term is to optimize the image retrieval sub-task and the other for caption retrieval. (i, c) is a positive pair and the hardest negatives are defined as $i' = argmax_{j \neq i} s(j, c)$ or $c' = argmax_{d \neq c} s(i, d)$. The loss emphasizes on eliminating hard negatives to dig out well-matched pairs.

Table 1. Comparison results of image and sentence matching on the MSCOCO (1000 testing) dataset.

#	Model	Trainset	Caption retrieval		Image retrieval	
			R@1	R@10	R@1	R@10
		1K Test Images				
1	VSE[9]	1C(1 fold)	43.4	85.8	31.0	79.9
2	VSE++[5]	1C(1 fold)	42.5	85.4	31.1	79.4
3	VSE++[5]	RC	49.0	88.4	37.1	83.8
4	Ours	1C(1 fold)	43.7	86.8	33.6	81.7
5	Ours	RC	**50.4**	**90.3**	**39.7**	**85.1**

4. EXPERIMENTS

In this section, to demonstrate the effectiveness of our method, we conduct several experiments in terms of image and sentence matching on MSCOCO dataset [21].

4.1. Dataset and Protocol

The MSCOCO dataset and its protocol are described as follows. MSCOCO consists of 82783 training and 40504 validation images, each of which is associated with five sentences. We follow [9] to split the dataset, with 82783, 4000 and 1000 images for training, validation and testing.

4.2. Implementation Details

In the procedure of extending the training dataset, we generate another three captions for a given image. We pre train the Generator (G) based on the standard maximum likelihood principle for 20 epochs. The Discriminator (D) is supervised by loss function in Eq. 4 for five epochs. Then we conduct adversarial training, in which one iteration consists of one step of updating parameters in the Discriminator after one stage of updating parameters in the Generator. The batch size is 64, and the learning rate is 0.0001.

4.3. Quantitative Results

We conduct the training procedure on our augmented training dataset and evaluate the performance of our model on the standard test dataset. The commonly used evaluation criteria for image and sentence matching are 'R@1', 'R@10', i.e., recall rates at the top one and ten results. From the results illustrated in Table 1, we can see that our model with augmented training data improves the performance of both caption retrieval and image retrieval based on VSE++ baseline in random crop (RC) setting or centre crop (1C) setting. For caption retrieval, since more sentences have been used to consolidate the visual-semantic embeddings, better sentence feature representations lead to advance in performance. For image retrieval, the model benefits from the larger scale of data to discriminate semantic similarity between matched pairs and unmatched ones, leading to a more significant gain in performance even compared to image retrieval.

4.4. Qualitative Results

We show the qualitative results of our image captioning model in Figure 2. Our generator can produce diverse and natural image captions without losing fidelity.

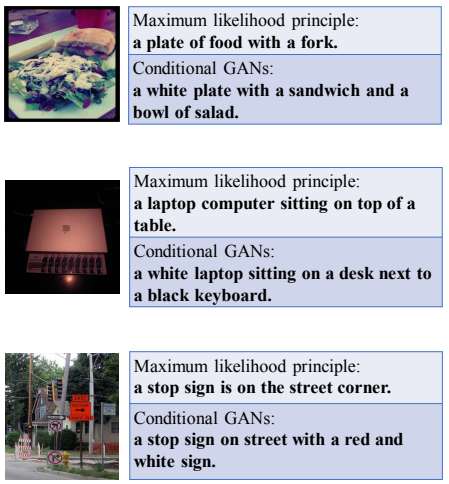

Fig. 2. Comparison results of image captioning on the MSCOCO dataset

5. CONCLUSION

This paper has proposed a method to further improve the performance in the task of image and sentence matching. With the help of GANs and policy gradient in reinforcement learning, we can produce diverse and natural descriptions for a given image, which can efficiently enlarge the training dataset. More training data benefits the retrieval network to learn solid visual-semantic embeddings, narrowing the semantic gap between images and sentences. As it is shown in the experiments, compared with the baseline VSE++, our model achieves superior results. Our method can benefit different kinds of retrieval methods in this field. In the future, we will equip state-of-art methods with our approach for further performance improvement.

Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015, 61806194), Capital Science and Technology Leading Talent Training Project (Z181100006318030), Beijing Science and Technology Project (Z181100008918010), and CAS-AIR.

6. REFERENCES

[1] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7181–7189.

[2] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching," *arXiv preprint arXiv:1803.08024*, 2018.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[4] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin, "Towards diverse and natural image descriptions via a conditional gan," *arXiv preprint arXiv:1703.06029*, 2017.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "Vse++: improved visual-semantic embeddings," *arXiv preprint arXiv:1707.05612*, 2017.

[6] 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.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Computer Science*, 2014.

[10] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun, "Order-embeddings of images and language," *arXiv preprint arXiv:1511.06361*, 2015.

[12] 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.

[13] Fei Yan and Krystian Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3441–3450.

[14] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang, "Learning semantic concepts and order for image and sentence matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6163–6171.

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2017.

[16] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J Corso, "Image caption generation with text-conditional semantic attention. arxiv preprint," *arXiv preprint arXiv:1606.04621*, vol. 2, 2016.

[17] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 3242–3250.

[18] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

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

[20] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[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.