Learning a Limited Text Space for Cross-Media Retrieval

Zheng Yu, Wenmin Wang *, and Mengdi Fan

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,
fanmengdi@sz.pku.edu.cn

Abstract. In this paper, we propose a novel model for cross-media retrieval which relies on a limited text space rather than a common space or an image space. More specifically, the model consists of three parts: A visual part that consists of a convolutional neural network and an image understanding network; A language model part that achieves sentence understanding by recurrent neural network; An embedding part that contains a fusion layer to capture both visual label information and semantic correlations between images and sentences, as well as learn the final limited text space by optimizing pairwise ranking loss. Experimental results on three benchmark datasets show that our proposed model gains promising improvement in accuracy for cross-media retrieval especially on sentence retrieval compared with the related state-of-the-art methods.

Keywords: Cross-media retrieval, Limited text space, Fusion layer, Image understanding network, Recurrent neural network

1 Introduction

Along with the popularization of the Internet, there has been a rapid growth of multimedia data such as images, texts, videos and audios which always appear together. As a result, single-media retrieval can not meet people's daily needs since some people want to search sentences that can best describe a given image or show images that can best depict a given sentence. Therefore, cross-media retrieval has been proposed which comprises two problems. The first problem is how to efficiently represent multimedia data. Traditional methods such as the bag-of-words for sentences and SIFT for images transformed multimedia data into low-level features. In order to learn more abstractive representations, deep neural network was proposed to represent data at a higher level which was proven to be more efficient than low-level features. However, the semantic gap among heterogeneous data features still exists.

Hence, the second problem is how to embed heterogeneous multimedia data such as images and sentences into a homogeneous space so that their similarity can be measured directly. Since we only focus on the retrieval between images and sentences,

This project was supported by Shenzhen Peacock Plan (20130408-183003656).

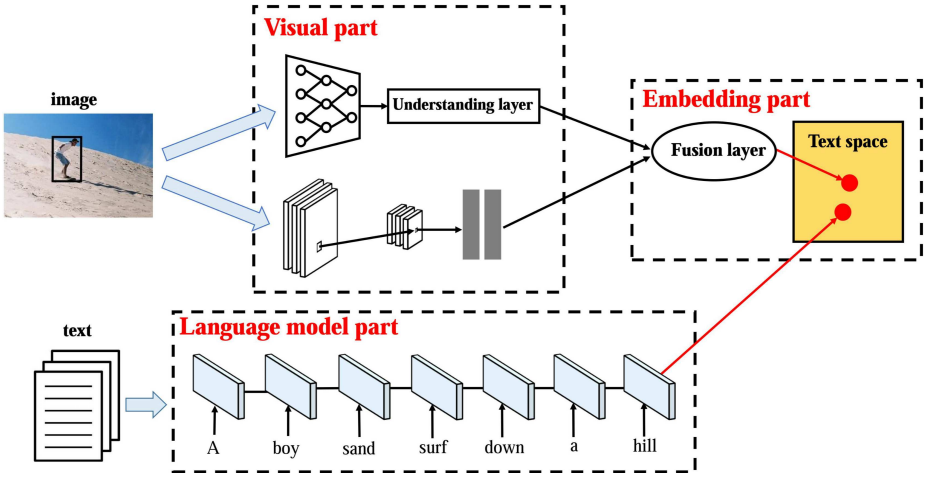

Fig. 1. A detailed illustration of our proposed model which consists of three parts: Visual part, Language model part and Embedding part.

cross-media retrieval can be achieved based on a common space [4][19][7][8][21], a text space [18][9], or an image space [10]. When performing cross-media retrieval, from the human point of view, we always try to understand the images and sentences sufficiently before retrieval. It is simple and intuitive for brains to understand the sentences but a little bit complicated to understand the images. Given an image, we first caption the image subconsciously by nature language and then understand it. In order to make the model behave as similar as human, we hope it is able to understand images and sentences sufficiently before retrieval. Therefore, we aim to perform cross-media retrieval in a text space. Current methods based on a text space mostly employed the Word2Vec space. Image understanding was achieved by a convex combination of the word embedding vectors of the visual labels predicted to be the most relevant to the image. However, the visual labels only reflect the objects contained in an image but ignore how these objects relate to each other as well as their attributes and the activities they are involved in. Thus, the Word2Vec space is not an effective text space for cross-media retrieval.

Accordingly, we propose a novel model to learn the text space effectively which is capable of understanding images like human. For the first problem mentioned above, we propose a visual part to learn deep representations for images. Meanwhile, recurrent neural network is used to learn dense representations for sentences. For the second problem, we propose an embedding part to learn a limited text space. More specifically, the whole model contains a language model part, a visual part and an embedding part. The language model part learns a dense limited text space representation for each sentence which contains rich semantic information. The visual part contains a deep convolutional neural network and an image understanding network. Deep CNN can be used to generate deep convolutional representations containing rich visual label information, such as the objects contained in an image. The image understanding network represents im-

ages in a pre-trained limited text space which can capture strong semantic correlations between images and sentences, such as the attributes and the activities these objects are involved in. For the embedding part, a single fusion layer is added on top of the visual part to capture both visual label information and semantic correlations between images and sentences, as well as transform the image representations into the final limited text space by optimizing the pairwise ranking loss. About the word "limited" in the paper title, it means that the text space is spanned by a set of base vectors which are also known as different words in a vocabulary. Therefore, the ability for the text space to understand is limited due to the limited number of words in the vocabulary.

Our core contributions are: 1) We propose a novel model to do the retrieval of images and sentences humanly in a limited text space. Pairwise ranking loss function is exploited as the objective function to be optimized. 2) The image understanding network is capable of modeling strong semantic correlations between images and sentences. 3) A single fusion layer is added on top of the visual part in order to capture both visual label information and semantic correlations between images and sentences, as well as transform the image representations into a limited text space.

The rest of the paper is organized as follows. Section 2 reviews the related work for cross-media retrieval. Section 3 describes details of our proposed model. Section 4 presents the experimental results on three datasets. Finally, we make a summary of the paper in Section 5.

2 RELATED WORK

There are a lot of methods that have been proposed to handle the aforementioned two problems. For the first problem of learning efficient image and sentence representations, Sharif et al. [1] argued that a pre-trained deep convolutional neural network (CNN) was an effective image feature extractor which had achieved the state-of-the-art performances on many image processing tasks. Simonyan et al. [2] investigated how the depth of convolutional neural networks affected their performance and proposed VGG which had won the first and the second places in the localization and classification tasks respectively. For sentence, traditional methods including Word2Vec [3], LDA [17], or FV [5] were used to learn low-level representations for sentences without concerning the rich contextual information. Recently, with the great progress on machine translation [6], recurrent neural network is found to be a more powerful tool for language modeling which is able to take advantage of the contextual information of the whole sentence. Wang et al. [23] proposed a deep alternative neural network (DANN) to extract contextual information for action recognition in video.

For the second problem of learning a homogeneous space, the mainstream approach is to learn a common space by affine or deep transformation on both sentence and image sides. Canonical correlation analysis [4] learned a common space by maximizing the correlations between relevant sentences and images. Karpathy et al. [19] broke down both images and sentences into fragments and embedded them into a common multimodal space. Fan et al. [21] performed coupled feature mapping and correlation mining successively for cross-media retrieval. Coupled feature mapping learned two projection matrices to map the multimodal features into a common category space and then corre-

lation mining was used to take advantage of the semantic category information. Yan et al. [7] stacked fully connected layers together to represent the sentences and used deep canonical correlation analysis for matching images and sentences. Ma et al. [8] proposed multimodal convolutional neural networks (m-CNNs) which captured relations between images and sentences at different level. In addition to a common space, in the DeViSE model developed by Frome et al. [18], the text space was formed by a pre-trained Word2Vec model. In a follow-up work, Norouzi et al. [9] employed Word2Vec for both sentence and image embeddings. The text space vector of an image was obtained by a convex combination of the word embedding vectors of the visual labels predicted to be the most relevant to the image. Recently, a distributional visual embedding space provided by Word2VisualVec [10] was found to be an effective space to perform cross-media retrieval by embedding sentences into a visual feature space.

Apart from those models designed for cross-media retrieval, image caption models can also be used to learn an appropriate homogeneous space. For example, multimodal recurrent neural network (m-RNN) [11], neural image caption (NIC) [13], deep visual-semantic alignments (DVSA) [14], Unifying Visual-Semantic Embeddings (VSE) [15] were used to learn the relations between images and sentences and generate the captions for a given image. Before translating the image representations to descriptive sentences, those models first transformed the image representations into a limited text space. Thus, the limited text space vectors for images contain rich semantic correlations between images and sentences.

3 PROPOSED METHOD

The architecture of our proposed model is shown in Fig. 1 which contains a language model part, a visual part and an embedding part.

3.1 Language Model Part

We first review GRU which is used for learning dense limited text space representations for sentences. Cho et al. [16] proposed Gated Recurrent Unit as a simpler alternative to the LSTM. The GRU uses reset gates r and update gates z to control the flow of information inside the unit. h represents the activation of the GRU and $\hat{h}$ is the previous computed activation.

Let $S = (s_0, s_1, \cdots, s_t), t \in \{0 \cdots T\}$ be an input sentence, where we represent each word as a one-hot vector s_t of dimension equals to the size of the dictionary. Note that we denote by s_T a special end word which designates the end of the sentence. Before fed into the GRU, s_t should be embedded into a more dense space such as the Word2Vec space:

$$x_t = W_e s_t, t \in \{0 \cdots T\}, \tag{1}$$

The word embedding matrix W_e maps the one-hot vectors to a more dense text space. As mentioned in [16], the GRU takes the form:

$$\begin{aligned} h_t^j &= (1 - z_t^j)h_{t-1}^j + z_t^j \hat{h}_t^j \\ z_t^j &= \sigma(W_z x_t + U_z h_{t-1})^j \\ \hat{h}_t^j &= tanh(W x_t + U(r_t \odot h_{t-1}))^j \\ r_t^j &= \sigma(W_r x_t + U_r h_{t-1})^j \end{aligned} \tag{2}$$

As shown in equation (2), the activation h_t^j of the GRU at time t is a linear interpolation between the previous activation h_{t-1}^j and the candidate activation $\hat{h}_t^j$. An update gate z_t^j decides how much the unit updates its activation. The reset gate r_t^j controls the unit whether to forget the previous computed states or not. Finally, the representation of a sentence S is the hidden state of the GRU at time T.

3.2 Visual Part

The visual part contains a deep convolutional neural network and an image understanding network to achieve image understanding. For convolutional neural network, we employ VGG [2] to extract 4096-dimensional image representations X_{vgg} which contain rich visual label information.

For the image understanding network, inspired by the idea of automatic image captioning, we propose a novel method to map image pixels to pre-trained limited text space representations which contain strong semantic correlations between images and sentences similar to NIC [13]. The understanding process can be divided into two subprocesses:

(1) Learning image representations. We choose $Inception\ v3$ image recognition model pre-trained on the ILSVRC-2012-CLS image classification dataset as an image feature extractor:

$$X_{img} = Inception\_v3(Image) \tag{3}$$

where X_{img} is a 1024-dimensional vector in an image space.

(2) Embedding image representations into a pre-trained limited text space. A single linear embedding layer is added on top of the $Inception\ v3$ model to transform the image representations into a pre-trained limited text space:

$$X_{txt} = X_{img} W_{img} \tag{4}$$

where W_{img} maps X_{img} to a 512-dimensional vector.

3.3 Embedding part

As mentioned in the previous subsection, X_{vgg} is able to capture rich visual label information but ignores semantic correlations between images and sentences. X_{txt} is particularly good at modeling semantic correlations between images and sentences which

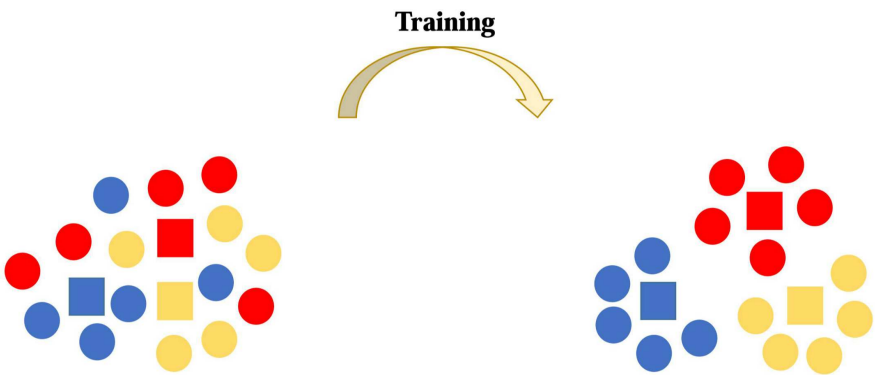

Fig. 2. Illustration of the pairwise ranking loss for learning the limited text space. Rectangles represent images and circles represent sentences. Matching image-sentence pairs are denoted in the same color.

is complementary to X_{vgg}. According to it, we add a linear fusion layer on top of the visual part to combine X_{txt} with X_{vgg} as well as embed them into a limited text space:

$$X_{final} = X_{vgg}W_{vgg\_fuse} + X_{txt}W_{txt\_fuse}, \tag{5}$$

where W_{vgg_fuse} and W_{txt_fuse} are embedding matrices for X_{vgg} and X_{txt} respectively.

In order to optimize the model parameters, pairwise ranking loss function is exploited to be the objective function. That is, as shown in Fig. 2, given an image query x, we want the distance between x and its matching sentences to be smaller than the distance between x and its non-matching sentences by a margin, and vice versa for a sentence query. Thus, we optimize the following loss function:

$$\begin{aligned} L = & \min_{\Theta} \sum_{x} \sum_{k} max\left\{0, margin + d(x,s) - d(x,s_k)\right\} \\ & + \sum_{v} \sum_{k} max\left\{0, margin + d(s,x) - d(s,x_k)\right\} \end{aligned} \tag{6}$$

where s_k is a negative sentence for a given image x and x_k is a negative image for a given sentence s. In order to obtain the non-matching terms, we choose them randomly from the training set and re-sampled every epoch.

4 EXPERIMENTS

In order to evaluate the effectiveness of our proposed model on cross-media retrieval, we have performed an extensive set of experiments on three benchmark datasets. We follow the evaluation metrics adopted in [15] for a fair comparison using Recall@K and Med r. The R@K (with K = 1, 5, 10) computes the mean number of images for which the correct caption is ranked within the top-K retrieved results and vice versa for sentences. Med r is the median rank of the first correct result in the ranking list. Higher R@K and lower Med r thus mean better performance.

Table 1. Bidirectional image and sentence retrieval results on Flickr8K

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE[18]	4.8	16.5	27.3	28.0	5.9	20.1	29.6	29
m-RNN[11]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
Deep Fragment[19]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
DCCA[7]	17.9	40.3	51.9	9	12.7	31.2	44.4	13
m-CNN$_{wd}$[8]	15.6	40.1	55.7	8	14.5	38.2	52.6	9
m-CNN$_{phs}$[8]	18.0	43.5	57.2	8	14.6	39.5	53.8	9
m-CNN$_{phl}$[8]	16.7	43.0	56.7	7	14.4	38.6	52.2	9
m-CNN$_{st}$[8]	18.1	44.1	57.9	7	14.6	38.5	53.5	9
m-CNN$_{ENS}$[8]	24.8	53.7	67.1	5	20.3	47.6	61.7	5
FV(GMM+HGLMM)[5]	**31.0**	**59.3**	**73.7**	**4**	**21.3**	**50.0**	**64.8**	**5**
VSE[15]	18.0	40.9	55.0	8	12.5	37.0	51.5	10
Ours_single	21.8	50.2	64.5	5	13.8	37.5	52.2	10
Ours_fusion	21.5	50.3	66.2	5	15.4	40.5	54	9

4.1 Datasets

For evaluation we use three benchmark datasets consisting of images and their corresponding descriptive sentences. The statistics of the datasets are as follows:
Flickr8K [24] : This dataset consists of 8,000 images. Each image is annotated with 5 sentences describing the content. We use 6,000 images for training, 1,000 images for validation and 1,000 images for testing.
Flickr30K [22] : This dataset consists of 31,783 images. Each image is also annotated with 5 sentences describing the content of the image. We use 29,000 images for training, 1,000 images for validation and 1,000 images for testing.
MSCOCO [12] : This dataset consists of 82,783 training ,40,504 validation, and 40,775 testing images. Each image is also annotated with 5 sentences describing the content of the image. We reserve 1,000 random images from the MSCOCO validation set as test and use it to report results.

4.2 Experimental Configurations

For the visual part, we adopt the similar architecture as NIC [13] to embed image pixels into a pre-trained limited text space which was pre-trained on MSCOCO. More specifically, we use 512 dimensions for the word embeddings. For the language model part, we randomly initialize the word embeddings W_e to be 1024-dimensional vectors. Similar to [15], our GRU uses one layer with 1024 units and weights initialized uniformly from [-0.08, 0.08]. For the embedding part, Xavier initialization [20] is used to initialize the embedding matrices W_{vgg_fuse} and W_{txt_fuse}. In order to choose an appropriate $margin$ that achieves the best performance on all datasets, we set $margin = [0.1, 0.2, ..., 0.9]$ for validation and finally select $margin$ = 0.7. The

Table 2. Bidirectional image and sentence retrieval results on Flickr30K

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE[18]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
Deep Fragment[19]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
m-RNN-vgg[11]	**35.4**	63.8	73.7	3	22.8	50.7	63.1	5
DCCA[7]	16.7	39.3	52.9	8	12.6	31.0	43.0	15
m-CNN$_{wd}$[8]	21.3	53.2	66.1	5	18.2	47.2	60.9	6
m-CNN$_{phs}$[8]	25.0	54.8	66.8	4.5	19.7	48.2	62.2	6
m-CNN$_{phl}$[8]	23.9	54.2	66.0	5	19.4	49.3	62.4	6
m-CNN$_{st}$[8]	27.0	56.4	70.1	4	19.7	48.4	62.3	6
m-CNN$_{ENS}$[8]	33.6	**64.1**	74.9	**3**	**26.2**	**56.3**	**69.6**	**4**
FV(GMM+HGLMM)[5]	35.4	62.0	73.8	3	25.0	52.7	66.0	5
VSE[15]	23.0	50.7	62.9	5	16.8	42.0	56.5	8
Ours_single	24.5	54.2	69.3	5	17.7	43.6	55.9	8
Ours_fusion	31.2	62.5	**75.8**	3	21.5	48.9	61.5	6

whole model is implemented in TensorFlow and Theano based on NVIDIA Tesla K80 GPU. We use minibatches of 40 on Flickr8K, 100 on Flickr30K and MSCOCO in the training procedure.

4.3 Experimental Results

We aim to show the experimental results on two aspects. Firstly, in order to emphasize the importance of the fusion layer, we design three contrast models: 1) VSE is the baseline model which uses VGG features to represent images; 2) $Ours\_single$ removes the fusion layer and uses the image understanding network to learn image representations; 3) $Ours\_fusion$ reserves the fusion layer. Secondly, we compare the three models with the related state-of-the-art methods so as to verify the effectiveness of the limited text space. The experimental results on Flickr8K, Flickr30K and MSCOCO are illustrated in Table 1, 2, and 3 where the best performance of each evaluation metric has been hightlighted.

The experimental results among the three contrast models show that our proposed fusion model $Ours\_fusion$ outperforms VSE and $Ours\_single$ on all datasets. It demonstrates that the fusion layer is able to capture both visual label information and semantic correlations between images and sentences which is beneficial to the performance of cross-media retrieval compared with VGG features and the pre-trained limited text space representations. Moreover, $Ours\_single$ outperforms VSE as well especially on sentence retrieval task due to the sufficient understanding of images by the visual part.

The contrast experiments between the fusion model $Ours\_fusion$ and the related state-of-the-art methods are shown as follows. On Flickr8K, FV achieves the best performance. $Ours\_fusion$ performs inferiorly to FV and m-CNN$_{ENS}$. Due to the lack

Table 3. Bidirectional image and sentence retrieval results on MSCOCO

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
m-RNN-vgg[11]	41.0	73.0	83.5	2	29.0	42.2	77.0	3
DVSA[14]	38.4	69.9	80.5	**1**	27.4	60.2	74.8	3
m-CNN$_{wd}$[8]	34.1	66.9	79.7	3	27.9	64.7	80.4	3
m-CNN$_{phs}$[8]	34.6	67.5	81.4	3	27.6	64.4	79.5	3
m-CNN$_{phl}$[8]	35.1	67.3	81.6	5	27.1	62.8	79.3	3
m-CNN$_{st}$[8]	38.3	69.6	81.0	2	27.4	63.4	79.5	3
m-CNN$_{ENS}$[8]	42.8	73.1	84.1	2	**32.6**	**68.6**	**82.8**	**3**
FV(GMM+HGLMM)[5]	39.4	67.9	80.9	2	25.1	59.8	76.6	4
VSE[15]	43.4	75.7	85.8	2	31	66.7	79.9	3
Ours_single	34.6	68.5	82.9	3	17.8	49.9	66.9	6
Ours_fusion	**45.5**	**78.7**	**88.8**	2	30.2	66	80.5	3

of training samples, Fisher vector is proven to be the most powerful method on modeling sentences. However, recurrent neural network is essentially a kind of temporally deep neural network and thus needs sufficient data to tune the parameters adequately. Except FV, m-CNN$_{ENS}$ performs better than us due to the integration of four separate models. It is worth mentioning that $Ours\_fusion$ outperforms the four models on sentence retrieval and matches their results on image retrieval.

On Flickr30K, with more training samples than Flickr8K, $Ours\_fusion$ gains a significant improvement on sentence retrieval task which shows competitive experimental results compared with FV and m-CNN$_{ENS}$. However, the model performs poor on image retrieval task. The most probable cause may be the insufficient understanding of sentences by RNN which may lead to the ambiguity during the retrieval.

On MSCOCO, with the largest number of training samples, the performance of $Ours\_fusion$ on sentence retrieval has been significantly improved, compared with all the other methods. Only DVSA outperforms $Ours\_fusion$ in terms of Med r. It demonstrates that with enough training samples, the parameters of GRU and the embedding matrices can be more adequately tuned. On image retrieval task, $Ours\_fusion$ performs inferiorly to m-CNN$_{ENS}$ but still superior to the other methods.

Table 4 shows three examples of sentence retrieval. It can be observed that our model finds the closest results for a given image query. For example, the groundtruth descriptive sentences for the first image query are: 1) A wooden ball on top of a wooden stick. 2) The table is full of wooden spoons and utensils. 3) A wood table holding an assortment of wood cooking utensils. 4) A selection of wooden kitchen tools on a counter. 5) Wooden spoons are lined up on a table. Although retrieved sentences "Wooden spoons and forks are all over a table" and "Multiple wooden spoons are shown on a table top" are regarded as irrelevant to the image, they can describe the content more accurately than "A wooden ball on top of a wooden stick" and "A selection of wooden kitchen tools on a counter" from a subjective perspective of human. However, there are

Table 4. Three examples of sentence retrieval. The first column contains the image queries and the second column shows the top five retrieved sentences on MSCOCO dataset. The correctly retrieved sentences for each image query are denoted in blue. The incorrectly identified objects are marked in red.

Image queries	Top five retrieved sentences
	**1. Wooden spoons are lined up on a table**. 2. Wooden spoons and forks are all over a table. **3. The table is full of wooden spoons and utensils.** 4. Multiple wooden spoons are shown on a table top. **5. A wood table holding an assortment of wood cooking utensils.**
	**1. A person is riding a bicycle but there is a train in the background.** **2. A man on a bicycle riding next to a train.** **3. A guy that is riding his bike next to a train.** **4. A red and white train and a man riding a bicycle.** **5. A man riding a bike past a train traveling along tracks.**
	1. **A tall woman** is standing in a small kitchen. 2. **A woman** observing something on a kitchen stove. **3. A kitchen with two windows and two metal sinks.** **4. A kitchen has the windows open and plaid curtains.** 5. The view of a kitchen from across the room.

some unreasonable results for the third image query. As shown by the words marked in red, our model identifies visual concept "woman" incorrectly which is nonexistent in the image.

5 Conclusion

In this paper, we propose a novel model to perform cross-media retrieval in a limited text space which aims to learn limited text space representations for both images and sentences. Firstly, the visual part learns image representations including deep convolutional features and pre-trained limited text space representations. The language model part learns a dense limited text space representation for each sentence. Secondly, the embedding part captures both visual label information and semantic correlations between images and sentences, as well as learns the final limited text space. Experimental results on three benchmark datasets demonstrate the importance of the fusion layer and the effectiveness of the limited text space. Our proposed fusion model achieves promising improvement compared with the related state-of-the-art methods. In the future, we

will pay more attention to the image retrieval task and further improve the accuracy of it.

References

1. A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.
2. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
3. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
4. B. Thompson, "Canonical correlation analysis," *Encyclopedia of statistics in behavioral science*, 2005.
5. B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation," *arXiv preprint arXiv:1411.7399*, 2014.
6. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
7. F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3441–3450.
8. L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2623–2631.
9. M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013.
10. J. Dong, X. Li, and C. G. Snoek, "Word2visualvec: Cross-media retrieval by visual feature prediction," *arXiv preprint arXiv:1604.06838*, 2016.
11. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.
12. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
13. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, 2016.
14. A. Karpathy and L. 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. R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.
16. K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.
17. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
18. A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

19. A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.
20. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.
21. M. Fan, W. Wang, and R. Wang, "Coupled feature mapping and correlation mining for cross-media retrieval," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.
22. P. Young, A. Lai, M. Hodosh, and J. 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.
23. J. Wang, W. Wang, R. Wang, W. Gao *et al.*, "Deep alternative neural network: Exploring contexts as early as possible for action recognition," in *Advances in Neural Information Processing Systems*, 2016, pp. 811–819.
24. M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.