Two-Stream Convolutional Neural Network for Multimodal Matching

Youcai Zhang, Yiwei Gu, and Xiaodong Gu(✉)

Department of Electronic Engineering, Fudan University, Shanghai 200433, China
xdgu@fudan.edu.cn

Abstract. Mulitimudal matching aims to establish relationship across different modalities such as image and text. Existing works mainly focus on maximizing the correlation between feature vectors extracted from the off-the-shelf models. The feature extraction and the matching are two-stage learning process. This paper presents a novel two-stream convolutional neural network that integrates the feature extraction and the matching under an end-to-end manner. Visual and textual stream are designed for feature extraction and then are concatenated with multiple shared layers for multimodal matching. The network is trained using an extreme multiclass classification loss by viewing each multimodal data as a class. Then a finetuning step is performed by a ranking constraint. Experimental results on Flickr30k datasets demonstrate the effectiveness of the proposed network for multimodal matching.

Keywords: Multimodal matching · Two-stream network
Convolutional neural network

1 Introduction

Multimodal analysis has received ever-increasing research focus due to the explosive growth of multimodal data such as image, text, video and audio. A core problem for multimodal analysis is to mine the internal correlation across different modalities. In this paper, we focus on the image-text matching. For example, given a query image, our aim is to retrieve the relevant texts in the database that best illustrate the image. There are two major challenges in multimodal matching: (1) effectively extracting the feature from the multimodal data; (2) inherently correlating the feature across different modalities.

Previous works for multimodal matching prefered to adopt off-the-shelf models to extract the features rather than learn modality-specific features. For the image, some well-known hand-crafted feature extraction techniques such as SIFT [1], GIST [2] were widely used. Inspired by recent breakthroughs of convolutional neural network (CNN) in visual recognition, CNN visual features were also introduced to multimodal matching [14]. For the text, latent Dirichlet allocation (LDA) [3] and *word2vec* [18] models were two typical choices for vectorization. Despite their contributions to the multimodal matching, off-the-shelf

V. Kůrková et al. (Eds.): ICANN 2018, LNCS 11139, pp. 14–21, 2018.
https://doi.org/10.1007/978-3-030-01418-6_2

models suffer from some weaknesses. They are not specific designed for the task of multimodal matching. That is, these features are not discriminative enough, which limits the final matching performance.

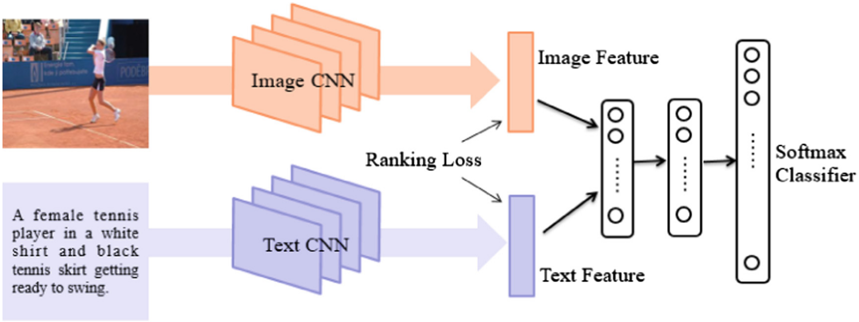
A female tennis player in a white shirt and black tennis skirt getting ready to swing.
Image CNN
Text CNN
Ranking Loss
Image Feature
Text Feature
Softmax Classifier

Fig. 1. Overview of the proposed two-stream convolutional neural network.

Another challenge is to correlate these multimodal features. Most deep learning based methods [4,5] are highly dependent on the categorical information for network training. However, such high-level semantic information is absent in most scenarios and requires much manual labels. Furthermore, the explosive increase of data makes it unrealistic to label each data with a certain category. Luckily, co-occurred data usually delivers correlated information (i.e. image-text pair information). The pair information is relatively easy to be obtained via the web crawler and should be fully explored for multimodal matching.

To address above issues, we propose a novel two-stream convolutional neural network as shown in Fig. 1, which extracts visual and textual representations and simultaneously performs the task of multimodal matching. Thus the similarity between images and texts can be measured directly according to the learned representations. More specifically, CNN is the backbone to extract the feature from the raw images and texts respectively. The outputs of the two stream are concatenated and followed by several shared fully connected layers. The final output of the network is the class probabilities after a softmax regression. To train the network, we adopt an extreme multiclass classification loss and a ranking loss both based on the pair information.

The remainder of this paper is structured as follows. Section 2 reviews the related work. Section 3 presents our two-stream network for multimodal matching and its learning process, followed by experimental results in Sect. 4. Section 5 draws an overall conclusion.

2 Related Work

The core issue for multimodal matching is to learn discriminative and joint image-text representations. Canonical correlation analysis (CCA) [7] and cross-modal factor analysis (CFA) [8] were two classic methods. They linearly projected vectors from the two views into a shared correlation maximum space. Andrew *et al.* proposed deep CCA [12] to learn the nonlinear transformation through two deep networks, whose outputs are maximally correlated. Yan *et al.* [13] further introduced DCCA into image-text matching.

Inspired by recent breakthroughs in visual recognition, CNN was also widely employed in multimodal matching. Wei *et al.* [14] provided a new baseline for cross-modal retrieval with CNN visual features instead of traditional SIFT [1] and GIST [2] features. CNN has also shown its powerful abilities in natural language processing. Hu *et al.* [10] proposed a sentence matching model based on CNN that represented the sentence and captured the matching relation simultaneously. In [9], convolutional architectures were first employed to learn the correlation between image and sentence by encoding their separate representations into a joint one.

There are also some deep models related to our work. In [6], a three-stream deep convolutional network was proposed to generate a shared representation across image, text, and sound modality. Wang *et al.* [15] presented a two-branch network to learn the image-text joint embedding. The network was trained by an extended ranking constraint and only received the input of feature vectors. Mao *et al.* [16] proposed a multimodal Recurrent Neural Network (m-RNN) model for image captioning and cross-modal retrieval. [17] presented a selective multimodal network that incorporated attention and recurrent selection mechanism based on long short term memory.

3 Two-Stream CNN

3.1 Network Architecture

Overall Architecture. As exhibited in Fig. 1, the overall architecture of the proposed network contains two parts. The color part with two streams focuses on the feature extraction from the raw image and text. The gray one integrates the feature vectors from different modalities with shared weights and fully connected layers for further multimodal matching. In general, to generate a joint representation, the color part is specific to modality but gray one is shared across modalities.

Image Stream. We adopt a 50-layer ResNet model [11] pretrained on ImageNet classification tasks as the visual CNN. We discard the top fully connected layer designed for ImageNet. Thus, given a raw image resized to 224 × 224, a 2048-*dim* vector considered as the image representation is produced by the model after average pooling.

Text Stream. Since each image can be represented by a fixed-length vector with CNN, we also design a textual CNN with three convolutional layers to vectorize the text as shown in Fig. 2. Text is first encoded into a $1 \times n \times d$

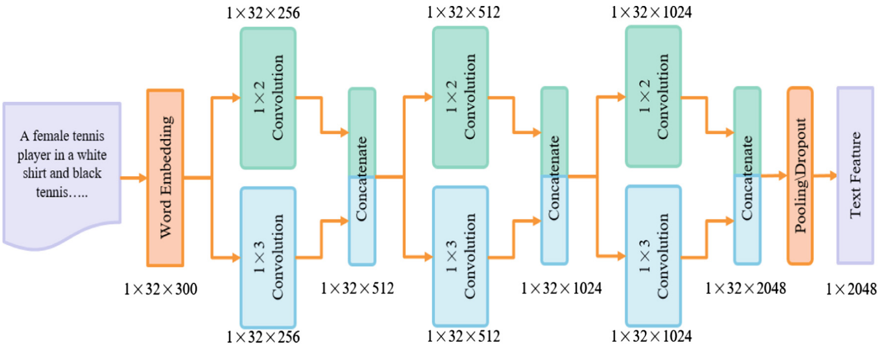

Fig. 2. Overview of the textual CNN stream.

numerical matrix $\mathbf{T}$, where n is the length of the sentence and d is the size of the vocabulary. The vocabulary contains all tokens appeared in the corpus. Let w_i be the *i-th* word in the vocabulary, thus w_i can be converted into a one-hot high-dimensional sparse vector $\mathbf{v_i}$ where the *i-th* element is set to be 1 and rests to be 0. Then the embedding layer turns each $\mathbf{v_i}$ into a low-dimensional dense word embedding $\mathbf{e_i}$ with the length of k via a lookup table. Thus, each sentence is encoded into a $1 \times n \times k$ matrix.

Though embedding layer encodes the semantic information of each word into vectors, simply concatenating word vectors ignores many subtleties of a possible good representation, e.g. consideration of word ordering. Therefore, following convolutional layers are employed to extract the word sequence information of the words. In each convolutional layer, the context in the sentence is modeled using two convolution kernels of size 1×2 and 1×3, respectively. And the outputs of two convolutional operations are concatenated directly, fed into following layers. At the end of network, a pooling layer with dropout is used to produce final output, which matches the size of image features. Convolutional layers combined with word embedding ensure that the output feature contains most necessary information to effectively represent sentences for further multimodal matching.

3.2 Network Learning

Objective Function. Supervised semantic labels usually play an important role in deep neural network learning. However, the lack of labels poses a unique challenge to multimodal matching: how to effectively utilize the only image-text pair information. In this paper, we transform the multimodal matching into an extreme multiclass classification task where the matching becomes accurately classifying a specific data among tens of thousands classes. Here, each multimodal document including an image and corresponding text is viewed as a pseudo class. Given an instance x^i, we apply the *softmax function* to the output of the network $\mathbf{z} \in \Re^{1 \times n}$ (n is the number of multimodal document). Thus, we can obtain the posterior probability of the instance being classified into the right category c. It can be formally written as Eq. (1).

$$P(c|x^i) = softmax(\mathbf{z}) = \frac{e^{\mathbf{z}_c}}{\sum_{j=1}^{n} e^{\mathbf{z}_j}}. \tag{1}$$

Then we minimize the negative log-likelihood $P(c|x^i)$, defined as Eq. (2).

$$L_{cls} = -log(P(c|x^i)). \tag{2}$$

To obtain more discriminative representations, we also performed a metric learning based on a ranking constraint. Pair of distances in the feature space between x^p and x^n against the anchor x^a should be pulled apart up to a margin α ($\alpha = 0.1$ in our case) as $d(x^a, x^p) + \alpha < d(x^a, x^n)$. Instances sharing the same pseudo class with x^a are defined as x^p, otherwise, x^n. We compute the cosine distance between the feature vectors $(\mathbf{v}_i, \mathbf{v}_j)$ of two instances (x^i, x^j) as $d(x^i, x^j) = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}$. We further define the bi-directional ranking constraint with a hinge loss for the given image reference $(x^a_{img}, x^p_{txt}, x^n_{txt})$ and the text reference $(x^a_{txt}, x^p_{img}, x^n_{img})$ respectively as Eq. (3).

$$\begin{aligned} L_{rank} &= max\{0, d(x^a_{img}, x^p_{txt}) - d(x^a_{img}, x^n_{txt}) + \alpha\} \\ &+ max\{0, d(x^a_{txt}, x^p_{img}) - d(x^a_{txt}, x^n_{img}) + \alpha\}. \end{aligned} \tag{3}$$

The final objective function is a weighted combination of the classification loss and ranking loss as Eq. (4).

$$L = \lambda_1 L_{cls} + \lambda_2 L_{rank}. \tag{4}$$

Training Scheme. Network training is done in three steps. Firstly, we fix the image stream and train the remaining part using the classification loss ($\lambda_2 = 0$, only text data is used). The reason behind is that pre-trained weights on Imagenet can be used for image stream but weights of the remaining part have to be learned from scratch. Secondly, we update the weights of the entire network after step 1 converges ($\lambda_2 = 0$, both text and image data are used). Considering that ranking loss usually converges very slowly or even does not converge especially in two-stream network learning, we fine-tune the entire network using the combination of the classification loss and ranking loss ($\lambda_1 = 1, \lambda_2 = 1$) only in the last step.

4 Experiment

4.1 Datasets and Evaluation Metrics

We choose widely-used Flickr30k [19] for experiments. Flickr30k contains 31,783 images collected from website Flickr. Each image is described with five sentences. We follow the partition scheme in [16,17], where 29,783, 1,000, and 1,000 images are used for training, validation, and test respectively. *R@k* and *Med r* are adopted as evaluation metrics. *R@k* is the average recall rate over all queries in the test set. Specifically, given a query, the recall rate will be 1 if at least one ground truth occurs in the *top-k* returned results and 0 otherwise. *Med r* is the median rank of the closest ground truth in the ranking list.

4.2 Implementation Details

For Flickr30k, the vocabulary size d is 20,074, and each word is encoded into a 300-*dim* dense vector. To ensure that each input sentence has the same length of 32, we use 0 vectors as paddings for those short sentenses. And we use the pre-trained vectors of the *word2vec* [18] model to initialize our embedding layer. The network is optimized by backpropagation and mini-batch stochastic gradient descent with the momentum fixed to 0.9. For the three training steps, learning rate is set to 0.001. 0.0001 and 0.00005 respectively. The maximum epochs are set to 180, 60 and 20 accordingly. In our experiments, we observe convergence within 150, 30, 10 epochs.

4.3 Experimental Results

We consider two basic multimodal tasks: Img2Txt (an image query to retrieve texts) and Txt2Img (a text query to retrieve images). Table 1 presents the experimental results of different methods in terms of *R@k* and *Med r*. The proposed network outperforms other methods in the Img2Txt task with the highest *R@1* of 48.4%. In the Txt2Img task, *R@1* obtained by our method is only 0.7% lower than the best method RBF-Net [20]. The results indicate that the learned features are effective for multimodal matching. The superiority of our network can be explained by the following two aspects: (1) We simultaneously perform feature extraction and multimodal matching. Compared with off-the-Shelf models, the learned features are more targeted for the matching task instead of previous generic representations; (2) We fully explore the image-text pair information via the classification and ranking loss to generate more discriminative representations.

We also conduct experiments to analyze the effect of the training scheme. Step 1 only trains the text stream using the classification loss and directly adopts the image features extracted from pre-trained ResNet-50. Step2 trains the entire network using the classification loss, which encourages instance from the same document to fall into one category. Thus, results obtained from step 2 gains a great increase of about 10%, 6% on *R@1* in the bidirectional retrieval respectively. Step 3 combines ranking constraints to further finetune the network, which provides a higher performance for the final model.

Another issue to be noticed is that the improvement brought by step 2 is not as impressive as that by step 3. On the one hand, that illustrates the effectiveness of posing multimodal matching as a classification problem. On the other hand, considering the effectiveness of ranking loss in previous works, there could be space for improvement in our network especially the weakness of *R@5* and *R@10.*

Table 1. Bidirectional image and text retrieval results on Flickr30K.

Methods	Img2Txt				Txt2Img			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
DCCA [13]	16.7	39.3	52.9	8	12.6	31.0	43.0	15
m-CNN [9]	33.6	64.1	74.9	3	26.2	56.3	69.6	4
m-RNN [16]	35.4	63.8	73.7	3	22.8	50.7	63.1	5
2-branch [15]	40.3	68.9	79.9	-	29.7	60.1	72.1	-
sm-LSTM [17]	42.5	71.9	81.5	2	30.2	60.4	72.3	3
RBF-Net [20]	47.6	**77.4**	**87.1**	-	**35.4**	**68.3**	**79.9**	-
Ours (step 1)	38.4	68.4	79.3	2	28.4	56.1	68.2	4
Ours (step 2)	46.8	75.7	85.6	2	33.5	63.0	74.9	3
Ours (step 3)	**48.4**	77.2	85.9	2	34.7	64.9	76.4	3

Ranking loss requires a careful triplet sampling strategy from the extremely unbalanced positive and negative ones, which points out the direction of our future work.

5 Conclusion

This paper mainly addresses the issue of multimodal matching via a novel two-stream convolutional neural network. The proposed network can extract the features from the raw image and text. To guarantee the features shared between different modalities, a classifier and ranking constraint are adopted for network learning by utilizing the pair information. Experimental results on Flickr30k datasets demonstrate the effectiveness of viewing each multimodal document as a discrete class. For further research, the ranking constraint will be polished to perform a more effective metric learning. Also, more detailed experiments on the Microsoft COCO datasets will be conducted to further validate the validity of our network.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants No. 61771145 and No. 61371148.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
2. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

4. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H. T.: Adversarial cross-modal retrieval. In: ACM International Conference on Multimedia Conference, pp. 154–162 (2017)
5. Huang, X., Peng, Y.: Cross-modal deep metric learning with multi-task regularization. In: IEEE International Conference on Multimedia and Expo, pp. 943–948 (2017)
6. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
7. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
8. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: ACM International Conference on Multimedia, pp. 604–611 (2003)
9. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: IEEE International Conference on Computer Vision, pp. 2623–2631 (2015)
10. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: Advances in Neural Information Processing Systems, pp. 2042–2050 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255 (2013)
13. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3450 (2015)
14. Wei, Y., et al.: Cross-modal retrieval with CNN visual features: a new baseline. IEEE Trans. Cybern. **47**(2), 449–460 (2017)
15. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
16. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint arXiv:1412.6632 (2014)
17. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2310–2318 (2017)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
19. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for Richer image-to-sentence models. In: IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)
20. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4107–4116 (2017)