Cross-modal retrieval by an end to end way

Bang Hu Yin* and Xiu Li
Tsinghua University, China

*Email: ybh16@mails.tsinghua.edu.cn

Abstract. Cross-modal retrieval has attracted most attention in the recent years. For the image and text, how to measure the semantic similarity among them is still a challenging problem in cross-modal retrieval task. In our paper, we propose a cross-modal retrieval method that uses a CNN to obtain the semantic similarity between image and text. Most methods to solve this problem employ two separate parts for each modality to obtain the semantic similarity between them. However, in our work we just use a CNN to get the semantic similarity without having to use separate networks. In addition, we are aiming to solve the problem between long text and image, so we use the topic model to process the text. We evaluate our approach on Wikipedia dataset.

1. Introduction

Convolutional Neural Networks has already made great success in image classification[1][2][3], object detection[4][5]. And now it has started to pivot from classical image classification and object detection problems to associating text descriptions to visual data. For example, Convolutional Neural Networks has made progress in image caption[6][7] which is aimed to produce the sentences describing the images. From these works, we can know that the deep Convolutional Neural Network can extract state-of-the-art image features and can obtain a better experiment result than other classical methods. It uses an image of the size W×H×3 as input, and then put it into a convolutional layer which is consisting of C-channel filters. These filters will scan the input and then output a feature map with the size of W1×H1×C. Certainly, The result will be the input of the next convolutional layer.

The task of cross-modal retrieval is that according to queries in one modality, users would search for semantically relevant content in other modalities. In image-text cross modal retrieval, we need find images that best illustrate the topic of a textual query, or textual descriptions that best explain the content of a visual query. Therefore, it is a hard task in cross-modal retrieval that obtain the semantic similarity between queries and database entries from different modalities.

The standard solution to cross-modal retrieval is to explore a shared low dimensional latent feature space for these modalities using data pairs[8][9]. And then we can obtain the semantic similarity by processing feature vectors in the feature space. The efficient and effective distance measurements are Cosine distance[9][10] or Euclidean distance[11], both methods can be used to measure the semantic similarity. Besides, we can take other methods to get the semantic similarity. In[12], they train a network that can learn to judge video frame is matching with short audio clip. The positives are those frames and audios from one video at the same time, while those extracted from different videos are negatives. By this method, they get the best results at that time.

In our paper we propose an approach that obtain the semantic similarity of text and image by one network. We utilize the great power of Convolutional Neural Network to realize it. We aim to let the
network output the similarity, not just get the feature vector of the image. In addition, we experiment on long texts not several sentences and transform them into vectors which are as inputs for network.

2. Related work
Many methods in the field of cross-modal retrieval have been proposed over recent years. Even if they are different from each other, they have a common feature that they try to search a shared subspace where the item of one modality can directly compare with others. And then these methods will transform each modality into vectors of common subspace. As a result, they can obtain the semantic similarity by computing the distance of two vectors.

One of the classical approaches towards is named Canonical Correlation Analysis (CCA) [10] which aims to search the common feature space that can maximizes the relevance between two types of data. Kernel canonical correlation analysis (KCCA) [13] uses the kernel trick based on CCA. And there are many other extensions of CCA. Just like CCA, Graph-based methods [14][15] also learn linear projections to discover a shared feature space by maximizing the cross-modal pairwise item correlation. Triplet ranking [16][17] is another effective approach. This method impose a constraint that make the distance between positive pairs be closer to the negative pairs. The cross-modal hashing [18][19] aims at searching a shared binary Hamming space where we can compare them directly. In TextTopicNet [20], they employ the topic level feature vectors of long texts to supervise the image feature learning of CaffeNet [28]. All these approaches could be divided into two parts: supervised and unsupervised approaches. The difference among them is if they utilize class or categorical information related to every image-document pair, of course, supervised methods use.

The method presented in our paper is the category of unsupervised methods. Our method is related to these cross-modal retrieval methods to some extent that we adopt LDA [21] to extract features of texts and use convolutional neural network to extract image feature vectors. However, our method differs from these methods in that we use the features of texts as one of inputs of the convolutional neural network with the image, besides, the output is the semantic similarity between the images and texts. Therefore, we needn’t calculate the similarity with the features of images and texts, and we can get the similarity directly through the network.

3. Proposed method
Our goal is to learn the semantic similarity between the images and texts by the deep neural network. To achieve this goal, we will extract the texts features at the first through the LDA [21] topic model, and then we will accomplish the fusion of image features and text features in the neural network. Finally, we can get the semantic similarity between the image and text.

3.1. LDA model
For the text, we suppose that several hidden topics can generate the textual information by some ways. Therefore, we take advantage of the latent dirichlet allocation model to obtain these latent topics and represent the texts associated with images as a probability distribution over the set of discovered topics. LDA, a three-level hierarchical Bayesian model, is a generative probabilistic model of texts. The core of it is that texts can be seen as different mixtures over latent topics, and each topic is distinguished by a probability distribution over all words of texts. We can see the structure of the model in figure 1.

For a text corpus including M documents and a dictionary containing N words, we can form a new text consisting of several topics by these steps:
• Select parameter $\theta$ which is generated by Dirichlet distribution
• For word $\omega_n$ in the new text $d$:
  - Select a topic $z_n$ which is generated by multinomial distribution with the parameter $\theta$
  - Select a word $\omega_n$ from multinomial distribution $p(\omega_n | z_n, \beta)$

So the LDA model can be expressed as (K is the number of topics):
\begin{align}
p(d \mid \alpha, \beta) &= \int \prod_{n=1}^{N} p(z_k \mid \theta) p(\omega_n \mid z_k, \beta) \ d \theta 
\end{align}

We can know that LDA obtains a more compact representation of text, so it reduces the dimensionality of data to a great extent. And we can easily interpret the meaning of the text vector. In our work, we need to extract the features of texts in advance which are inputs of the neural network. What we process are long texts, therefore, we need take some measures to present the texts in a compact representation.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{lda_schema.png}
\caption{Graphical model of LDA}
\end{figure}

3.2. Network architecture

To achieve our goals, we design the network structure shown in figure 2. In our experiments we do with LDA topic model to obtain the compact representation of texts, and process the images by convolutional layers to get the feature vectors of image. Then the two 40-D image and text features are concatenated into a 80-D vector which is used to generate a 2-way classification result by two fully connected layers. We can judge whether the images correlate to the texts according to the results.

In experiment, we add some other layers behind each the convolutional layers for the purpose of improving the accuracy of retrieval, such as max-pooling, Relu, Local Response Normalization layers. After we get the trained model, we can calculate the semantic similarity of the new image and text through inputting them into the network. By this method, we achieve the goals that to learn cross-modality similarity by an end to end ways and we make no use of class and categorical information about each image-text pair.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{network_architecture.png}
\caption{Network architecture.}
\end{figure}

4. Experiment and results
In this part we will describe our dataset, implementation details, and compare our method to other approaches so as to show the quality of the semantic similarity learned by our method. In our experiment, we utilize the texts of the dataset to train the LDA topic model so as to get the features of texts, then we input the images and text features into the network to obtain the semantic similarity, finally, we use the semantic similarity to finish the cross-modal retrieval task.

4.1. Dataset

We aim to solve the problem of text-image cross-modal retrieval by an end to end way. So we train our LDA and CNN models using Wikipedia articles contained in the Wikepedia ImageCLEF dataset[22]. We use the dataset which is the same as to TextTopicNet[20]. It consists of 35582 different articles which describe 100785 images. In order to test the quality of our models for cross-modal retrieval, we compare with other methods on a subset of Wikipedia dataset[10] which includes 693 pairs.

4.2. Implementation details

We do several experiments to choose the proper number of topics of LDA by varying it from 10 to 100, and finally we discover that 40 topics can obtain the best performance. One of the core ideas of this paper is that the fusion of image and text features, when we fuse the features of images and texts, we try to use the product of them, but it cannot reach the level ofconcating them. Finally, we use the softmax as our loss function. When we use it to calculate the semantic similarity, it will get a value between 0 and 1, which can be seen as the possibility that the image and text are correlated to each other.

4.3. Comparison to other methods

We evaluate our models by two kinds of cross-modal retrieval tasks: (1) given image retrieve text, (2) given text retrieve image. For the process of retrieval, we input images and texts features into the network, and obtain the semantic similarity in the end of the network. We compare our method with other supervised and unsupervised approaches in the table 1.

| Method       | Image query | Text query | Average |
|--------------|-------------|------------|---------|
| CCA[13]      | 19.70       | 17.84      | 18.77   |
| PLS[23]      | 30.55       | 28.03      | 29.29   |
| TextTopicNet[20] | 39.58   | 38.16      | 38.87   |
| SCM[10]      | 37.13       | 28.23      | 32.68   |
| GMMFA[24]    | 38.74       | 31.09      | 34.91   |
| CCA-3V[25]  | 40.49       | 36.51      | 38.50   |
| GMLDA[24]    | 40.84       | 36.93      | 38.88   |
| LCFS[26]     | 41.32       | 38.45      | 39.88   |
| JFSSL[27]    | 42.79       | 39.57      | 41.18   |
| ours         | 36.44       | 29.02      | 32.73   |

Table 1. MAP comparison with other methods.
Our method outperforms some supervised methods and unsupervised approaches. Most importantly, our method achieves the goals of learning the semantic similarity by an end to end way. It simplifies the process of calculating the semantic similarity.

5. Conclusion
In our experiments, we train our models to learn cross-modality similarity by an end to end way without using two networks like other methods and our performance is better than some classical ways.

References
[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[3] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[4] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
[5] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Springer, Cham, 2016.
[6] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 6. 2017.
[7] Chen, Long, et al. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning." arXiv preprint arXiv:1611.05594 (2016).
[8] Kuang, Zhanghui, and Kwan-Yee K. Wong. "Relatively-Paired Spce Anlysis: Lerning Ltent Common Spce From Relitively-Pired Observtions." International Journal of Computer Vision113.3 (2015): 176-192.
[9] Sharma, Abhishek, et al. "Generalized multiview analysis: A discriminative latent space." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
[10] Rasiwasia, Nikhil, et al. "A new approach to cross-modal multimedia retrieval." Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010.
[11] Hua, Yan, et al. "Cross-Modal Correlation Learning by Adaptive Hierarchical Semantic Aggregation." IEEE Trans. Multimedia 18.6 (2016): 1201-216.
[12] Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.
[13] Hardoon, David R., Sandor Szedmak, and John Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods." Neural computation 16.12 (2004): 2639-664.
[14] Zhai, Xiaohua, Yuxin Peng, and Jianguo Xiao. "Learning cross-media joint representation with sparse and semisupervised regularization." IEEE Transactions on Circuits and Systems for Video Technology 24.6 (2014): 965-78.
[15] Zhuang, Yueting, et al. "Supervised coupled dictionary learning with group structures for multimodal retrieval." AAAI. 2013.
[16] Frome, Andrea, et al. "Derive: A deep visual-semantic embedding model." Advances in neural information processing systems. 2013.
[17] Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[18] Xu, Xing, et al. "Learning discriminative binary codes for large-scale cross-modal retrieval." IEEE Transactions on Image Processing 26.5 (2017): 2494-507.
[19] Zhu, Xiaofeng, et al. "Linear cross-modal hashing for efficient multimedia search." Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013.

[20] Gomez, Lluis, et al. "Self-supervised learning of visual features through embedding images into text topic spaces." arXiv preprint arXiv:1705.08631 (2017).

[21] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research3.Jan (2003): 993-1022.

[22] Tsikrika, Theodora, Adrian Popescu, and Jana Kludas. "Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011." CLEF (Notebook Papers/Labs/Workshop). Vol. 4. 2011.

[23] Rosipal, Roman, and Nicole Krämer. "Overview and recent advances in partial least squares." International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection". Springer, Berlin, Heidelberg, 2005.

[24] Sharma, Abhishek, et al. "Generalized multiview analysis: A discriminative latent space." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[25] Gong, Yunchao, et al. "A multi-view embedding space for modeling internet images, tags, and their semantics." International journal of computer vision 106.2 (2014): 210-33.

[26] Wang, Kaiye, et al. "Learning coupled feature spaces for cross-modal matching." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[27] Wang, Kaiye, et al. "Joint feature selection and subspace learning for cross-modal retrieval." IEEE transactions on pattern analysis and machine intelligence 38.10 (2016): 2010-023.

[28] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.