Cross-model retrieval with deep learning for business application

Yufei Wang 1, Huanting Wang 2, 3, Jiating Yang 2, *, Jianbo Chen 3

1 Simon Fraser University, 8888 University Dr, Burnaby, BC, V5A 1S6, Canada
2 Xian Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China
3 University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author e-mail: yangjiating@opt.ac.cn

Abstract. Cross-modal retrieval has been used in many fields, such as business and search engines. Most search engines for business are text-based, but text-based search engines are limited by equipment and the strict requirement for knowledge. Text-based search needs keyboards to finish the search process, which requires users to have the knowledge of using keyboards. Compared to the text-based search, audio-based search has advantages. First, it avoids the traditional ways of inputting information. And it gets rid of the gap in time between inputting information for searching and getting useful information. In this paper, we propose a way to use audio to search images for business applications. We use deep learning to implement cross-modal retrieval systems between images and audio. We first extract features from images and audio respectively. And then we implement a neural network with two identical networks to learn the correspondence between images and audio. The first network extracts the features from images and audio further for calculation, and the second network learns whether two features from different modalities are related. This research provides a new way for business applications to search for information more instantly.

Key words: Cross-modal retrieval; Audio features; Deep hashing; Useful information.

1. Introduction

Nowadays, image retrieval systems are often used in shopping sites such as Amazon and Taobao. An image retrieval system is a system for users to browsing, searching, and retrieving images from a large database. These systems that are used in shopping sites are meant to improve the shopping experience of users so that they can get desired products rapidly and accurately. An image retrieval system has two basic methods including text-based image retrieval and content-based image retrieval. Text-based image retrieval systems require users to input keywords, tags, and descriptions related to products through keyboards. However, this method requires users to have access to keyboards and to input accurate keywords. Although this method has been used widely, the lack of effectiveness and limitation of inputting equipment are the main two problems with this method. Content-based image retrieval systems are systems that retrieval similar images from a database based on the content of sample images. This
method has an obvious defect, which is that it requires users to have some sample images as inputs. However, users often do not have such images for searching or retrieval in practice.

Speech is one of the main communication ways in daily life. Audio-based image retrieval systems are still a field that lacks too much focus. There are some difficulties related to this kind of retrieval systems. First, there is not too much audio information related to the corresponding images. Thus, the construction of the dataset for training is difficult. Second, it is difficult to learn the correspondence between audio features and image features. Unlike content-based image retrieval systems, audio-based image retrieval systems can not simply compare the distance between features of images and audio. However, an audio-based image retrieval system does not need sample images for inputs as content-based image retrieval and it also gets rid of the limitation of inputting equipment like keyboards. The main advantage of an audio-based image retrieval system is that it is instantaneous. Without the process of inputting keywords and tags, the effectiveness of image retrieval can be improved by using speech to search specific images from a large database.

In this paper, in order to resolve the difficulties of figuring out whether images and audio are correlated, we use deep learning to learn the correspondence between images and audio. We use a Siamese neural network to train the model to figure out the correspondence between different modalities. A Siamese neural network [1] is often used to learn whether two features from the same modal are similar or not. In a Siamese neural network, it includes two networks. The first neural network is used for extracting features from images or audio, and the second one is used to figure out whether two features are correlated or not. In this paper, we prove the feasibility of audio-based image retrieval systems and point out the difficulties and limitations of this method.

The remaining parts of this paper are organized as follows: in Section II, we will introduce some previous work related to this method including the method for extracting audio features and other cross-modal research about images and audio; in Section III, we will introduce the details of the neural networks we use for extracting features and learning the correspondence between images and audio; In Section IV, we will introduce the datasets we use for this research and demonstrate the details about experiments and the results of experiments; In Section V, we will conclude this paper and discuss the future work and the future applications about this retrieval system.

2. Related work

2.1. Mel-Frequency Cepstral Coefficients

Mel-frequency Cepstral is a representation of the power spectrum of a sound, and it is a cepstral analysis synthesis on the Mel-Frequency scale [2]. Mel-frequency Cepstral Coefficient is the coefficients that make up an MFC. MFCC is usually used as a technique that extracts features from audio files. This process including windowing the signal, applying Discrete Fourier Transform, and other processes including taking the log spectrum on the Mel frequency scale [3]. MFCC is used to approximates the human auditory system. Since the human perception of sounds for speech signal is not on a linear scale, applying sound signals in the Mel Scale is a good way to imitate the human auditory system. In order to imitate the human auditory system as well as possible and to get the best performance of the MFCC, Zheng et al. [4] have compared different performances of different implementations of MFCC. There is a lot of researches related to the application of MFCC in speech recognition [5][6][7][8]. Ittichaichareon el al. [5] used MFCC and Machine learning to accomplish speech recognition. Tiwari [6] created a speaker recognition system and made a comparison among the performances of different implementations of MFCC including changing the number of filters and the types of window. Since speech recognition using MFCC is feasible and there are many researches that improve the performances of MFCC, we use MFCC as the technique to extract features from audio files.
2.2. Cross-modal learning
Cross-modal retrieval provides a way of retrieving information among different modalities. For retrieving images, text-based retrieval systems are popular since there are many researches related to this method and the applications are easy to use in business. Many works related to cross-modal including exploring novel neural networks and new approaches for cross-modal multimedia retrieval [9][10][11][12][13]. In recent years, the cross-modal between images and audio become more popular [13][14][15][16][17]. Mao et al. [13] proved the feasibility of cross-modal retrieval between images and audio. They extracted features from images and audio, and then made a fusion of two features. They conducted linear calculation on the fusion and passed the results into an activation function. As a result, the method that they proposed in that paper is available for retrieving images using audio. In this paper, we use the Siamese network instead of making a fusion between the features of images and audio to train the model for retrieving images.

3. Methodology
The process includes three main parts: (1) extracting features from images, (2) extracting features from audio, (3) training the neural network for retrieving images through audio. The image feature extraction part introduces how we pre-process images and how we extract features from images. The audio feature extraction part introduces what technique we use to extract features from audio files. And the Siamese network part introduces the structure of the network we use for training and how we use that network to retrieve images through audio features.

3.1. Image feature extraction
There are many datasets for studying cross-modal retrieval, but the datasets we use for this paper have problems with the sizes of images. Images from the same dataset do not have the same size, so it is difficult to extract features of all the images at the same time. Therefore, we decide to resize all the images from datasets to have the same size, like 224 x 224 or 500 x 500. Resizing images is available for this study since we only care about the features of images and the plentiful details resulted from high dimensions are not helpful in this study. After resizing all the images, every image has the same size, which is appropriate for extracting features from multiple images at the same time. The small size of images can reduce the time of calculation. The features of images are extracted by GoogLeNet [18] that is available on Matconvnet. As a result, there is a 4096-dimension feature vector for each image.

3.2. Audio feature extraction
MFCC as introduced above is a technique that imitates the human auditory system, so we use MFCC [7] to extract features from audio files. We use 44100 as the sampling rate of audio time series, 20 as the number of MFCCs to return. We also limit the maximum length of MFCCs. For those MFCCs with length larger than 170, we truncate the rest of those MFCCs. For those MFCCs with length less than 170, we pad zero into those MFCCs. These processes reduce the amount of calculation while keeping the important features from audio files. However, the results from MFCC calculation have too much redundancy, and the dimension of the results is not available for the Siamese network, which requires both features have the same dimension. Therefore, we proposed a convolutional neural network to normalize the dimension of the features extracted from audio files. The convolutional neural network is simple, and it contains two convolutional layers, one average pool layer, and one linear layer since we want to keep useful information as much as possible. As a result, all the features of audio files have 4096-dimension feature vectors, which is compatible with the dimension of features of images.
3.3. Siamese network

Cross-modal retrieval systems use distances across different modalities to retrieve related images or audio. Thus, it is important to learn whether two features are related or not and the distances between their features. Siamese network is used to classify whether two images are the same based on the distance between features. In this paper, we also use this neural network to figure out the correspondence across different modalities. Siamese network [1] is a neural network that has two identical neural networks, named that share the same weights. First, two features from the audio-image pair are inputted into the two identical neural networks. In this paper, \( \text{\textsc{Net}} \) contains two linear layers, which is used to reduce the dimensions of features further and then reduce the time for calculation. After inputting two features into the two networks respectively, the two outputs from two networks are calculated as 

\[(\text{output}_1 - \text{output}_2)^2.\]

And another neural network for classification, only containing one linear layer, is used to guide the training process of the Siamese network. We use Adam optimizer for this neural network.

When we test the performance of the model, we abandon the last linear layer to have a larger dimension for features so that we can have more features to calculate the Euclidean distance between two features from different modalities to have a better result.

4. Dataset, Experiments and Results

4.1. Dataset

We conduct experiments on the PASCAL sentence dataset [19] and the Wikipedia dataset [20]. Currently, most datasets for studying cross-modal only have image-text pair, so we extract useful texts from either annotations or descriptions. And then we use text-to-speech software to transfer texts to audio so that we can get image-audio pairs for this experiment.

PASCAL sentence dataset has 20 categories including aeroplane, bicycle, and so on and each category includes 50 images. There are 5 sentences for each image, and we select one representative sentence for each image. For the PASCAL sentence dataset, we use 800 image-audio pairs for creating the training set and 200 image-audio for creating the test set.

Wikipedia dataset has 10 categories including art, biology, and so on. We extract descriptions from annotations for every image and select representative sentences for text-to-speech. Since each category
has a different number of images, we use 80 percent image-audio pairs to create the training set and the rest 20 percent to create the test set, which results in 2289 image-audio pairs for creating the training set and 577 image-audio pairs for creating the test set.

To create the training set, we create pairs that images and audio are correlated and pairs that images and audio are not correlated. First, we combine images and their own corresponding audio into an 8192-dimension feature vector and label them as ‘correspond’. Since we are creating an image retrieval system, images, and audio from the same category can be also seen as correlated. Then we randomly combine images and audio from the same category into an 8192-dimension feature vector and label them as ‘correspond’. Finally, we combine images and audio from different categories into an 8192-dimension feature vector and label them as ‘not correspond’. We do the same process for the test set.

4.2. Evaluation Metric

We use mean average precision (MAP) to evaluate the performance of our model in this paper.

4.2.1. Mean average precision. MAP is the mean value of average precision (AP). AP is the average precision of top N retrieved instances for each audio feature (we use N=10 in this paper). The formula for calculating AP is defined as:

\[
AP = \frac{1}{R} \sum_{r=1}^{N} \text{Precision}(r) \times \text{corr}(r)
\]

Where r is the rank after sorting by their distances between features of images and audio, R is the number of related images, Precision(r) is the precision of the current image, corr(r) indicates whether the audio for searching and the current image are correlated or not.

4.2.2. Precision. Precision is the fraction of relevant instances among the retrieved instances. Precision is calculated as:

\[
\text{Precision}(r) = \frac{\text{the number of relevant instances}}{\text{the current rank } r}
\]

4.3. Experiment and Analysis

For proving the feasibility of this method, we test the performance of the image retrieval system by using different epochs, different batch sizes, and different learning rates to training the model.

Table 1. The MAPs of using different parameters for the PASCAL sentence dataset

| Epochs | Batch size | Learning rate | MAP       |
|--------|------------|---------------|-----------|
|        | Different numbers of epochs |               |           |
| 100    | 50         | 0.001         | 0.439958  |
| 500    | 50         | 0.001         | 0.46236   |
| 1000   | 50         | 0.001         | **0.485697** |
| 2000   | 50         | 0.001         | 0.429291  |
|        | Different batch sizes |               |           |
| 1000   | 25         | 0.001         | **0.529254** |
| 1000   | 50         | 0.001         | 0.485697  |
| 1000   | 100        | 0.001         | 0.444057  |
| 1000   | 200        | 0.001         | 0.482942  |
|        | Different learning rates |               |           |
| 1000   | 25         | 0.00001       | 0.498152  |
| 1000   | 25         | **0.0001**    | **0.562212** |
| 1000   | 25         | 0.0005        | 0.48659   |
| 1000   | 25         | 0.001         | 0.529254  |
| 1000   | 25         | 0.005         | 0.444521  |
| 1000   | 25         | 0.01          | 0.34503   |
Table 1 shows the MAPs with different parameters for the PASCAL sentence dataset. First, we try different epochs for training, we use 50 as the batch size and 0.001 as the learning rate since we have not tested the performance with different batch sizes and learning rates. As the results show, when the number of epochs is 1000, MAP has the highest number, which means that the performance for the image retrieval system is the best. Then, we use 1000 as the number of epochs for the rest of the experiments. We try different batch sizes for the performance of the retrieval system and keep using 0.001 as the learning rate. The results indicate that the performance of the retrieval system is the best when the batch size is 25. We do not try smaller batch sizes because if the batch size is too small, the effectiveness of vectorization will lose. Finally, we test the performance with different learning rates. And we use 1000 as the number of epochs and 25 as the batch size since the performances with those parameters are the best from the previous experiments. During different learning rates, we can see that when the learning rate is 0.0001, the MAP is 0.562212, which is much higher than other MAPs. In conclusion, when we choose 1000 as the number of epochs, 25 as the batch size, and 0.0001 as the learning rate, the model has the best performance and the MAP for the image retrieval system is 0.562212.

Table 2 shows the MAPs with different parameters for the Wikipedia dataset. We use the same method as mentioned above for testing the performances of the model trained with different values of parameters. As a result, when we choose 500 as the number of epochs, 25 as the batch size, and 0.001 as the learning rate, the model has the best performance and the MAP for the image retrieval system is 0.553293.

| Epochs | Batch size | Learning rate | MAP     |
|--------|------------|---------------|---------|
| **500** | 25         | 0.001         | **0.553293** |
| 500    | 50         | 0.001         | 0.45576 |
| 500    | 100        | 0.001         | 0.37017 |
| 500    | 200        | 0.001         | 0.307737 |

Table 2. The MAPs of using different parameters for the Wikipedia dataset

| Epochs | Batch size | Learning rate | MAP     |
|--------|------------|---------------|---------|
| 500    | 25         | 0.000001      | 0.48175 |
| 500    | 25         | 0.0001        | 0.51441 |
| 500    | 25         | 0.0005        | 0.4832 |
| 500    | 25         | **0.001**     | **0.553293** |
| 500    | 25         | 0.005         | 0.37332 |
| 500    | 25         | 0.01          | 0.41424 |
Figure 2 shows the top 10 retrieved images using audio from the category “bird”. As we can see in the figure, there are some images that do not belong to the category “bird”, but most of them are birds. This proves that using the Siamese network to train the model for the image retrieval system is feasible, and we can get acceptable results.

5. Conclusion
In this paper, we propose a method, which is using the Siamese network to train the model for the image retrieving system, to solve cross-modal retrieval problems between images and audio. The results have proved the feasibility of this method. And from the results, we can see that the cross-modal retrieval between images and audio is available. We think that the results can be improved with larger datasets, more complex Siamese networks, and creating more available image-audio pairs. For example, currently, we have multiple corresponding sentences for each image, and we can use all of them to make image-audio pairs for the training and test sets. Moreover, we will add more layers to the neural networks. This study provides a new way to search images in business applications. When the image retrieval system using audio becomes more efficient, users can get rid of the limitation of equipment and get the results instantly and accurately.

References
[1] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In Advances in neural information processing systems (pp. 737-744).
[2] Imai, S. (1983, April). Cepstral analysis synthesis on the mel frequency scale. In ICASSP’83. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 8, pp. 93-96). IEEE.
[3] Rao, K. S., & Manjunath, K. E. (2017). Speech recognition using articulatory and excitation source features. Springer.
[4] Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. Journal of Computer science and Technology, 16(6), 582-589.
[5] Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012, July). Speech recognition using MFCC. In International Conference on Computer Graphics, Simulation and Modeling (pp. 135-138).
[6] Tiwari, V. (2010). MFCC and its applications in speaker recognition. International journal on emerging technologies, 1(1), 19-22.
[7] Han, W., Chan, C. F., Choy, C. S., & Pun, K. P. (2006, May). An efficient MFCC extraction method in speech recognition. In 2006 IEEE international symposium on circuits and systems (pp. 4-pp). IEEE.
[8] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6), 1-4.

[9] Karpathy, A., Joulin, A., & Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In Advances in neural information processing systems (pp. 1889-1897).

[10] Chen, Y., Lu X., Feng, Y. (2019, October). Deep Voice-Visual Cross-Modal Retrieval with Deep Feature Similarity Learning. In Proceedings of Chinese Conference on Pattern Recognition and Computer Vision (pp. 454-465).

[11] Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017, October). Adversarial cross-modal retrieval. In Proceedings of the 25th ACM international conference on Multimedia (pp. 154-162).

[12] Feng, F., Wang, X., & Li, R. (2014, November). Cross-modal retrieval with correspondence autoencoder. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 7-16).

[13] Zhen, L., Hu, P., Wang, X., & Peng, D. (2019). Deep supervised cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10394-10403).

[14] Mao, G., Yuan, Y., & Xiaoqiang, L. (2018, August). Deep cross-modal retrieval for remote sensing image and audio. In 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS) (pp. 1-7). IEEE.

[15] Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In Advances in Neural Information Processing Systems (pp. 1858-1866).

[16] Nagrani, A., Albanie, S., & Zisserman, A. (2018). Seeing voices and hearing faces: Cross-modal biometric matching. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8427-8436).

[17] Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. In Proceedings of the IEEE International Conference on Computer Vision (pp. 609-617).

[18] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[19] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting Image Annotations Using Amazon’s Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

[20] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010, October). A new approach to cross-modal multimedia retrieval. In Proceedings of the 18th ACM international conference on Multimedia (pp. 251-260).