Scenery Soundscaping using Cross-Modal Information Retrieval

[1]Abheetha Pradhan, [1]Vikram C, [2]Dr. Manjula K

[1]Student, [2]Associate Professor
[1,2]Department of Computer Science & Engineering,
[1,2]Global Academy of Technology, Bengaluru, India

Abstract: Human beings associate certain sceneries with certain ambient sounds. A picture of the ocean can make us think of the sound of ocean waves, or a picture of a busy market can make us imagine the sounds of people talking and the sound of vehicles. But computers do not have this innate ability to associate audio to a visual image. If you separate the audio from its corresponding video, the computer would not be able to associate these with each other, unless programmed to do so explicitly. Cross-Modal Information Retrieval (CMIR) is a technique that measures similarities between different types of data, usually media, which here would be audio in the form of mp3 files, and images in the form of jpg files. The Scenery Soundscaping System accurately predicts the ambient sounds that you would hear in the scenery of a particular image using this technique. The project also creates a web app for the user to provide image inputs to the Scenery Soundscaping System and to listen to the soundscape predicted by it.

***Index Terms*—Deep Learning, Cross-Modal Information Retrieval**

I. Introduction

Humans naturally have the innate ability to look at something, and to imagine what that specific thing would sound like, without explicitly listening to it. This is true for many kinds of imagery that are available around us. For example, a human could look at the image of a dog and imagine it barking loudly. Or they could see a river and imagine the sound of water crashing against the rocks, as it flows downstream. Or they could be shown the faces of different people they know, and they can correctly associate each of them to what their voices sound like. Cross-modal information retrieval aims to achieve something similar, where data of different modalities are associated with each other to enable flexible retrieval. This method transforms the data from different modes into a common representational space so that the similarity metric between them can directly be measured.

Computers do not have this ability. Computers store information in many different forms, such as audio, images, text, etc. If you separate the audio from its corresponding video, the computer would not be able to associate these with each other, unless programmed to do so explicitly. For artificial intelligence to learn the way humans do, it should be able to perform cross-modal information retrieval efficiently.

II. Literature Survey

The paper by Jing Yu, Yuhang Lu, Zengchang Qin, Yanbing Liu, Jianlong Tan, Li Guo, and Weifeng Zhang [1] uses a two path neural network that learns representations of different modes of data, as well as a metric to compare similarities at the same time. It aims to retrieve text from images, by mapping them to a similar representational space and using a similarity metric. The text path uses a Graph Convolutional Network to model text based on its representation in a graph. The image path uses a hand-crafted fully connected neural network where the last layer maps the image features to the same representational space as the other path. Then, an objective distance metric is used to minimize the error to make the model learn. It is important to study these kinds of models, because even though our aim is slightly different, to correlate audio and images, the outline of the model will remain roughly the same, with mapping both onto the same space, and then comparing distances.

The paper by Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, Liang Wang [2] explores different methods of approaching Cross-Modal Information Retrieval. It showcases the general framework of such networks, stating how extracting the features from any modes of data into a common representation space is always the first step to performing the task of cross-modal retrieval. In these representations, there are two kinds: real valued, and binary valued representations. Real-valued representation uses real values for the different modalities of data. As this might be slow, binary-valued representations map them onto a common Hamming space. Moreover, there are four methods to actually perform this mapping: unsupervised, pairwise based, rank based, and supervised methods. In our case, we will be using a supervised method on both ends that are not related.

The paper by Liangli Zhen, Peng Hu, Xu Wang, Dezhong Peng [3] showcases a novel method to perform Cross-Modal retrieval, called Deep Supervised Cross-Modal Retrieval (DSCMR). This method finds a similar representation space between the different forms of data, where they can be compared with each other in that form. In DSCMR, the loss is optimised both in the labelling space to ensure proximity, and in the common representation space to make sure that the closeness can be measured easily. It uses two subnetworks: one for images and one for text modality. The image network uses the VGG-19, and the text network uses the sentence CNN. Then two fully connected ReLU networks compare the two. The representations that are derived work quite well to correlate data and they are also independent of which modes of data are being used. But sometimes, some of the data from different categories in the dataset, are confused for each other and overlap, which makes using DSCMR not relevant.

The paper by Meinard Müller, Andreas Arzt, Stefan Balke, Matthias Dorfer, Gerhard Widmer [4] explores cross-modal information retrieval in a similar domain as ours, between images and audio. More specifically, it aims to look at sheet music (images) and produce corresponding audio. As with the previous papers, it converts both of them into a similar representation space and then compares them. It introduces different methods to represent music in a form to perform cross-modal information retrieval, by trying to do this regardless of the way in which the music has been represented. This makes it easy for music streaming services and applications to find and browse music to provide relevant results, making it more relevant educationally and commercially. But the main hurdle to overcome is to optimise the transformation techniques, which is the main step of this method, that creates a bottleneck, as we cannot optimise.

This paper by Yusuf Aytar, Carl Vondrick, Antonio Torralba [5] studies unlabelled sound data from different sources on the internet, namely video streaming services on the internet. They have used over two million videos downloaded from Flickr. The choice to use Flickr was made because it is home to videos from the internet that have not been edited. This usually means that a video has its natural audio. These videos can be used as labelled data to correlate images and audio because of their inherent correlation. This data is used to train a model that has learnt the correlation between these models, named SoundNet. Also, an advantage of using unlabelled videos from an online source is that there are massive amounts of available data to form a dataset. This results in a diverse and rich dataset that is being used to train the network.

In the paper by Alex Krizhevsky, Ilya Sutzkever, and Geoffrey E. Hinton [6], a large deep convolutional neural network was trained to classify 1.2 million high-resolution images into 1000 different classes in the ImageNet LSVRC-2010 contest. It achieved superior test results compared to its predecessors, with a top-1 error rate of 37.5% and a top-5 error rate of 17.0%. The deep neural network they use consists of over 60 million parameters and 650,000 neurons across 5 convolutional layers, 3 fully connected layers and a softmax function with 1000 possibilities for the different classes of images. They also optimise GPU usage for the convolution tasks to speed up training. A relatively new regularization method called “dropout” has also been used to make sure that the neural network does not suffer from overfitting. This resulted in accuracy scores that were better than any supervised learning models used before it. And it also proved that this can be used for very complex data. We make use of this data and model to make sure that the representation learning for the images is performed to great accuracy.

III. Proposed Work

In the last decade or so, cross-modal information retrieval has been the subject of extensive research both by academics, as well as by people working in the industry. A method that is used commonly to overcome the representation gap across different modes like images, text, and so on is representation learning. This method transforms the data from different modes into a common representational space so that the similarity metric between them can directly be measured.

As deep learning is being used to great success in the industry to solve a wide array of problems, deep learning has also seen an insurgence in being used to solve the problem of representation learning, which is to map data of different modalities to a common representation space. The term Deep Supervised Cross-Modal Retrieval (DSCMR) is used for these methods.

The deep learning model will consist of two different neural networks: one that will convert the image into a feature vector, and another that will convert the audio files in the dataset to a list of feature vectors. A different algorithm will compare the image feature vectors with each of the audio feature vectors and find the most suitable ones. The web app will take an image and feed it to the trained model to output corresponding audio.

IV. Methodology

A. Dataset

The dataset is taken from the SoundNet archive, which is a dataset of audio files that has been created by using videos from the online video streaming service, Flickr. Flickr has been chosen here because it is home to videos that are more natural and have not been professionally edited unlike some of their counterparts such as YouTube. Over two million videos have been used and the audio has been separated from them, and stored in the mp3 format.

The same dataset has also been used to train the SoundNet model and is available for download on their website, which has been used in the model architecture explained below.

B. Architecture

As seen in Fig. 4.1, we use the neural network for images to process the image into an image feature vector. On the other side, we use the SoundNet model to process all the audio files available into feature vectors. The image feature vectors are then compared with all the audio feature vectors to determine the closest ones, using the entropy statistic.

The audio feature vector with the least entropy compared to the image feature vector will be chosen as the best audio corresponding to the particular image. To process the images, we choose MobileNet, a deep learning model trained on the ImageNet classification database optimized for embedded vision applications. Additionally, we precompute all the feature vectors for the audio files in the database to save on computational power which would be spent on performing the same computations every time otherwise.

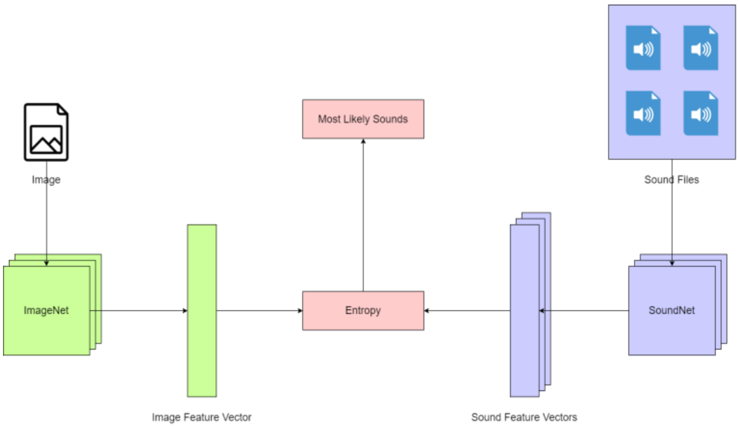

Fig. 1. Architecture Diagram

V. Results

A. Testing Procedure

Consider the example where we are finding the best audio that is associated with an image of a waterfall. Even if that particular waterfall sounds a certain way, the sound of any waterfall would still be a correct match for the image. The sound of a river flowing would be decently accurate match too, but the sound of a bustling city would clearly be wrong.

Because of this ambiguity where a single image can have multiple correct audios, and many others could be close, it is not possible for us to automatically label the results of the choice of audio by just comparing it with the labels present in the dataset. Instead, we need a human labeller(s) who is aware of sounds of all varieties and what their origin would look like. The human labeller(s) should individually look at each image and compare it with the audio file selected and make a decision on whether it is a good match.

As this would be quite expensive and time-consuming on a large sample of images, we resort to using a smaller sample of images and labelling it ourselves.

B. Experiments

We will now proceed to showcase some images and describe the audio that was selected by the model and make decisions on whether the selection was a good match or not.

The first image we consider is that of an open crop field on a cloudy day, with a downpour sighted far off in the distance, shown in Fig. 2. The audio selected for this image is that of a heavy breeze blowing strongly. This is the kind of sound you would expect to hear when seeing this sight; therefore, we consider this to be a good match.

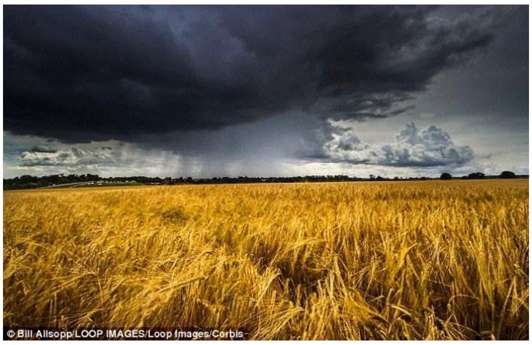
© Bill Allsopp/LOOP IMAGES/Loop Images/Corbis

Fig. 2. Image of a Crop Field

Next, we select the image of a beach with a lot of people on it playing and sunbathing, shown in Fig. 3. The audio selected contains sounds of waves crashing, and people shouting, cheering, and laughing.

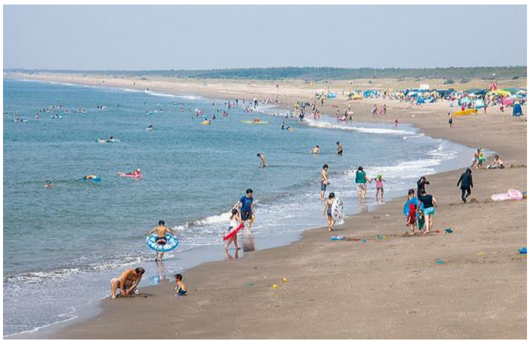

Fig. 3. Image of a beach with people on it

Next, we select the image of a wooded forest with sunlight falling through the gaps in the trees, shown in Fig. 4. The audio selected contains sounds of birds chirping and light winds, but also people walking and talking somewhere outside about camping. Although this seems like a close match in terms of setting, the lack of humans whatsoever in the image makes it a bad match.

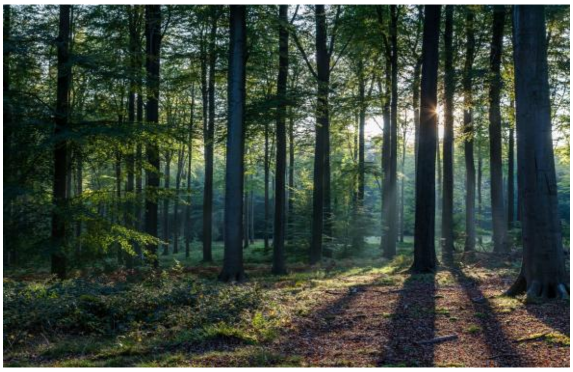

Fig. 4. Image of a Wooded Forest

Next, we select the image of a crowded city square, shown in Fig. 5. The audio selected contains sounds of the bustle of people walking and talking, and also bagpipes playing. These sorts of sounds are exactly what you would expect when looking at such a scenery, so we label it as a good match.

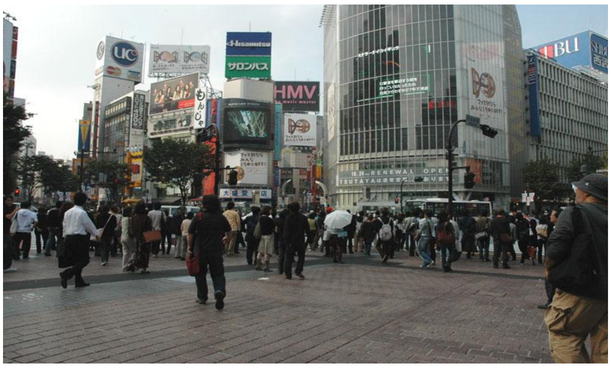

Fig. 2. Image of a Crop Field

Next, we select the image of a rain pouring over an empty road, shown in Fig. 6. The audio selected contains sounds of the pitter-patter of raindrops, as well as human talking. Although there are no humans in the picture, you would usually expect that people would be present on a road, so we label this as a decent match.

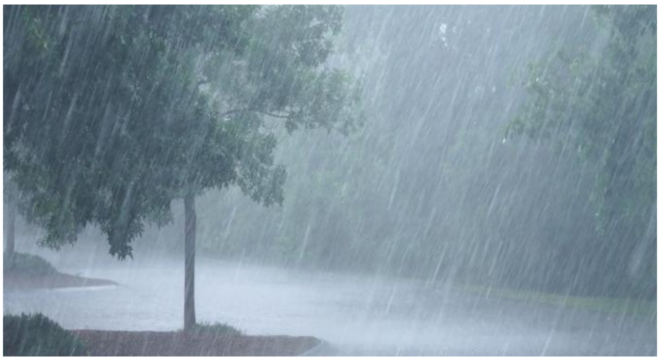

Fig. 2. Image of a Crop Field

C. Discussion and Future Trends

By experimenting with a number of images, it becomes apparent that the model performs exceptionally well with images that are more natural, while the accuracy falls when we come to sceneries that contain humans, or sceneries where you would usually find humans.

This is because the Scenery Soundscaping model is limited to the dataset it operates on, i.e., SoundNet’s rich natural sound database. As most of the audio files have been extracted from videos shot by humans spontaneously, even many sceneries without humans contain human voices in their selected audio. If the dataset were to be expanded to contain more sounds from other environments that are not abundant in the dataset currently, the system would be able to accurately predict ambient sounds for an even wider array of images. This could also be expanded to simulators and video games, which could automatically generate appropriate sounds for any environment it operates upon. As more data would get added, the models also would have to be tweaked and tuned to improve performance.

VI. CONCLUSION

The Scenery Soundscaping System bridges the gap in digital knowledge where computers can use it to associate imagery with appropriate sounds that would be heard in the scenery present in the image, using deep learning models for cross-modal information retrieval.

The Scenery Soundscaping System enables users to imagine what any environment in the world would sound like. People who have lived in mountainous areas all their lives could hear what a day at the beach would sound like. People can experience the auditory delights of the world in their homes for anything that they can visualise.

REFERENCES

1. Yu, Jing, Yuhang Lu, Zengchang Qin, Weifeng Zhang, Yanbing Liu, Jianlong Tan, and Li Guo. "Modeling text with graph convolutional network for cross-modal information retrieval." In Pacific Rim Conference on Multimedia, pp. 223-234. Springer, Cham, 2018.
2. Wang, Kaiye, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. "A comprehensive survey on cross-modal retrieval." arXiv preprint arXiv:1607.06215 (2016).
3. Zhen, Liangli, Peng Hu, Xu Wang, and Dezhong Peng. "Deep supervised cross-modal retrieval." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10394-10403. 2019.
4. Müller, Meinard, Andreas Arzt, Stefan Balke, Matthias Dorfer, and Gerhard Widmer. "Cross-modal music retrieval and applications: An overview of key methodologies." IEEE Signal Processing Magazine 36, no. 1 (2018): 52-62.
5. Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." Advances in neural information processing systems 29 (2016): 892-900
6. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105