AURALISATION OF DEEP CONVOLUTIONAL NEURAL NETWORKS: LISTENING TO LEARNED FEATURES

Keunwoo Choi, George Fazekas, Mark Sandler
Queen Mary University of London
keunwoo.choi@qmul.ac.uk,

Jeonghee Kim
Naver Labs
jeonghee.kim@navercorp.com

ABSTRACT

Deep learning has been actively adopted in the field of music information retrieval, e.g. genre classification, mood detection, and chord recognition. Deep convolutional neural networks (CNNs), one of the most popular deep learning approach, also have been used for these tasks. However, the process of learning and prediction is little understood, particularly when it is applied to spectrograms. We introduce auralisation of CNNs to understand its underlying mechanism.

1. INTRODUCTION

In the field of computer vision, deep learning approaches become *de facto standard* after convolutional neural networks (CNNs) showed break-through results in ImageNet competition in 2012 [3]. It rapidly became popular while the reason of success was not completely understood.

One effective way to understand and explain the CNNs was introduced in [5], where the features in deeper levels are visualised by a method called *deconvolution*. By deconvolving and un-pooling layers, it enables people to see which part of the input image are focused on by each filter.

However, spectrograms have not been analysed by this approach. Moreover, it is not clear what can be understood by deconvolving spectrograms. The information from 'seeing' a part of spectrograms can be extended by auralising the convolutional filters.

In this paper, we introduce the procedure and results of deconvolution of spectrograms. Furthermore, we propose auralisastion of filters by extending it to time-domain reconstruction. In Section 2, the background of CNNs and deconvolution are explained. The proposed auralisation method is introduced in Section 3. The results are discussed in Section 4.

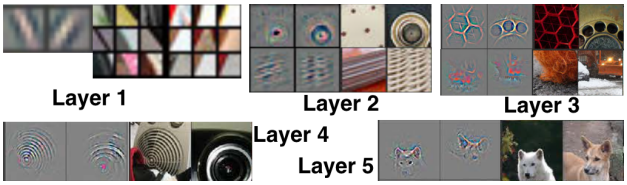

Figure 1. Deconvolution of CNNs trained for image classification. The filters' response and a corresponding part in the input images are shown respectively on the left and right for each layer. Image courtesy of [5].

2. BACKGROUND

The majority of research of CNNs on audio signal uses 2D time-frequency representation as input data, considering it as an image. Various types of representations have been used including Short-time Fourier transform (STFT), Mel-spectrogram and constant-Q transform (CQT). In [4], for example, 80-by-15 (mel-band-by-frame) mel-spectrogram is used with 7-by-3 and 3-by-3 convolutions for onset detection and CQT is used with 5-by-25 and 5-by-13 convolutions for chord recognition in [2].

Visualisation of CNNs was introduced in [5], which showed how high-level features (postures/objects) are combined from low-level features (lines/curves), as illustrated in Figure 1. Visualisation of CNNs helps not only to understand the process inside the black box model, but also to decide hyper-parameters of the networks. For example, redundancy or deficiency of the capacity of the networks, which is limited by hyper-parameters such as the number of layers and filters, can be judged by inspecting the learned filters. Network visualisation provides useful information since fine tuning hyper-parameters is among the most crucial factors in designing CNNs, while it is a computationally expensive process.

3. AURALISATION OF FILTERS

The spectrograms used in CNNs can be also deconvolved as well. Unlike visual images however, a deconvolved spectrogram does not generally facilitate an intuitive explanation. This is because, first, seeing a spectrogram does not necessarily provide clear intuition that is comparable to observing an image. Second, detecting edges of a spectro-

This paper has been supported by EPSRC Grant EP/L019981/1, Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption.

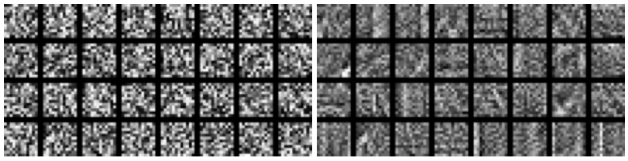

Figure 2. Filters at the first layer of CNNs trained for genre classification, initial (on the left) and learned filters (on the right).

gram, which is known to happen at the first layer in CNNs, results in removing components of the spectrograms.

To solve this problem, we propose to reconstruct the audio signal using a process called *auralisation*. This requires an additional stage for inverse-transformation of a deconvolved spectrogram. The phase information is provided by the phase of the original STFT, following the generic approach in spectrogram-based sound source separation algorithms. STFT is therefore recommended which allows us to later obtain a time-domain signal easily.

4. EXPERIMENTS AND DISCUSSION

We implemented a CNNs-based genre classification algorithm using a dataset obtained from Naver Music [1]. Three genres (ballade, dance, and hiphop) were classified using 8,000 songs in total. 10 clips of 4 seconds were extracted for each song, generating 80,000 data samples by STFT with 512-point using windowed Fast Fourier Transform and 50% hop size. 6,600/700/700 songs were designated as training/validation/test sets respectively. As a reference of performance, CNNs with 5 layers, 3-by-3 filters, max-pooling with size and stride of 2 was used. This system showed 75% of accuracy with a loss of 0.62.

4.1 Visualisation of First-layer Filters

A 4-layer CNNs was built with larger filter sizes at the first layer. Since max-pooling layers are involved with in deeper layers, filters at deeper layers can be only illustrated when input data is given. To obtain a more intuitive visualisation, large convolutional filters were used, as shown in [1]. This system showed 76% of accuracy.

Figure 2 shows the visualisation of the first layer, which consists of 12-by-12 filters. Filters are initialised with uniformly distributed random values, which resemble white noise as shown on the left. After training, some of the filters develop patterns that can detect certain shapes. For image classification tasks, the detectors for edges with different orientations usually are observed since the outlines of objects are a useful cue for the task. Here however, several vertical edge detectors are observed as shown on the right. This may be because the networks learn to extract not only edges but more complex patterns from spectrograms.

An example code of the whole deconvolution procedure is open to public at https://github.com/gnuchoi/CNNauralisation. The results are demonstrated at author's soundcloud https://soundcloud.com/kchoi-research

[1] http://music.naver.com, a Korean music streaming service

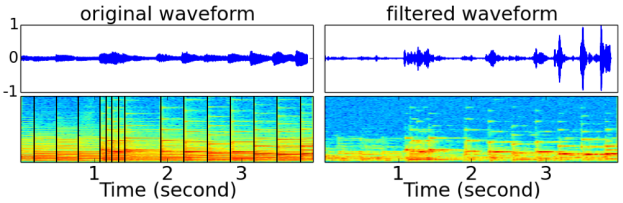

Figure 3. Original and deconvolved signals and spectrograms with ground-truth onsets annotated (on the left)

4.2 Auralisation of Filters

Fig 3 shows original and deconvolved signals and spectrograms of *Bach's English Suite No. 1: Prelude* from 7th feature at the 1st layer. By listening to these deconvolved signals, it turns out that this feature provides an onset detector-like filter. It can also be explained by visualising the filter. Vertical edge detectors can work as a crude onset detector when it is applied to spectrograms, since rapid change along time-axis will pass the edge detector while the sustain parts will be filtered out.

There was a different case at deeper layers, where high-level information can be expressed in deconvolved spectrograms. One of the features at the deepest layer was mostly deactivated by hiphop music. However, not all the features can be easily interpreted by listening to the signal.

5. CONCLUSIONS

We introduce auralisation of CNNs, which is an extension to CNNs visualisation. This is done by inverse-transformation of a deconvolved spectrogram to obtain a time-domain audio signal. Listening the audio signal enables researchers to understand the mechanism of CNNs when they are applied to spectrograms. Further research will include more in depth interpretation of the learned features.

6. REFERENCES

[1] Luiz Gustavo Hafemann. An analysis of deep neural networks for texture classification. 2014.

[2] Eric J Humphrey and Juan P Bello. Rethinking automatic chord recognition with convolutional neural networks. In *Machine Learning and Applications (ICMLA), International Conference on*. IEEE, 2012.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[4] Jan Schluter and Sebastian Bock. Improved musical onset detection with convolutional neural networks. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014.

[5] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.