VISUALIZATION AND INTERPRETATION OF SIAMESE STYLE CONVOLUTIONAL NEURAL NETWORKS FOR SOUND SEARCH BY VOCAL IMITATION

Yichi Zhang, Student Member, IEEE, Zhiyao Duan, Member, IEEE

Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA

ABSTRACT

Designing systems that allow users to search sounds through vocal imitation augments the current text-based search engines and advances human-computer interaction. Previously we proposed a Siamese style convolutional network called IMINET for sound search by vocal imitation, which jointly addresses feature extraction by Convolutional Neural Network (CNN) and similarity calculation by Fully Connected Network (FCN), and is currently the state of the art. However, how such architecture works is still a mystery. In this paper, we try to answer this question. First, we visualize the input patterns that maximize the activation of different neurons in each CNN tower; this helps us understand what features are extracted from vocal imitations and sound candidates. Second, we visualize the imitation-sound input pairs that maximize the activation of different neurons in the FCN layers; this helps us understand what kind of input pattern pairs are recognized during the similarity calculation. Interesting patterns are found to reveal the local-to-global and simple-to-conceptual learning mechanism of TL-IMINET. Experiments also show how transfer learning helps to improve TL-IMINET performance from the visualization aspect.

Index Terms— Neural network visualization, CNN, Siamese style networks, vocal imitation, transfer learning

1. INTRODUCTION

Vocal imitation is a common behavior that people use voice to mimic sounds. It is an effective way to convey ideas that are difficult to describe by language. Designing computational systems that allow users to search sounds through vocal imitation [1, 2] goes beyond the current text-based search and enables novel human-computer interactions. It can also be applied to many applications including movie and music production, multimedia retrieval, and security and surveillance.

There are two main challenges in the design of vocal-imitation-based sound search systems: 1) feature extraction: what features are appropriate to capture the similarity/dissimilarity between vocal imitations and sound candidates, given that humans tend to imitate different acoustic aspects for different sounds; 2) similarity calculation: how to measure the similarity between imitation queries and sound candidates, given that imitations and general sounds are produced by drastically different sound sources.

In [3], we proposed a Siamese style neural network called IMINET for sound search by vocal imitation. It consists of 1) two Convolutional Neural Network (CNN) towers for feature extraction for vocal imitations (query) and sound candidates (recording) respectively; and 2) a Fully Connected Network (FCN) for metric learning that predicts the similarity between the imitation and the candidate. This Siamese architecture integrating feature learning and metric learning outperforms existing systems that rely on hand-crafted features [4] and that only perform feature learning [5].

However, the mystery of how such Siamese style architecture works for sound search by vocal imitation is still unknown. Questions like what features are extracted from the vocal imitations and sound recordings and what patterns in imitation-recording pairs are recognized during metric learning need to be answered to give us deeper insights.

In this paper, we answer these questions by visualizing the input patterns that maximize the activation of different neurons of the network. From the visualization, we interpret the patterns in imitations and recordings captured by their feature extractors, as well as patterns in imitation-recording pairs captured by the FCN. There has been little work on network visualization and interpretation, especially in the audio domain. In addition, to our best knowledge, no work exists on visualizing networks that have multiple inputs.

Instead of the original IMINET, however, we visualize a transfer learning boosted version named TL-IMINET [6]. TL-IMINET adopts a similar Siamese style framework, but uses domain-specific networks for feature extraction. Specifically, the imitation CNN tower uses a model designed for spoken language classification [7], while the recording CNN tower uses a model designed for environmental sound classification [8]. Both towers are pre-trained on their own tasks. They are then fine-tuned together with the FCN on the target sound search task. This transfer learning process significantly improves the performance and makes TL-IMINET the new state of the art. The visualization techniques and findings presented in this paper, however, can be well generalized to other types of Siamese networks including IMINET.

This work was funded by the National Science Foundation grant No. 1617107. We also acknowledge NVIDIA's GPU donation for this research.

2. RELATED WORK

Query by vocal imitation belongs to the task of Query by Example (QEB) [9]. QEB has been applied to sound related applications like query by humming [10], query by beat boxing [11], cover song recognition [12], and spoken document retrieval [13]. However, little work has been reported on general sound search using vocal imitation queries.

Roma and Serra [14] designed a system that allows users to search sounds on freesound.org by recording their own audio, but no formal evaluation was reported. Blancas et al. [1] built a supervised system using hand-crafted features [4] and an SVM classifier to classify a vocal imitation query and retrieve sounds in that class. The major limitation of supervised systems, however, is that they cannot retrieve sounds not having training imitations. Helén and Virtanen [15] extracted hand-crafted features from both query and sound samples and measured the query-sample pairwise similarity on their feature distributions. In our previous work [2], we first proposed a supervised system using a Stacked Auto-Encoder (SAE) for automatic feature learning and an SVM for imitation classification. Considering the limitation of supervised systems, we then proposed an unsupervised system called IMISOUND [16], using the SAE for feature extraction and various kinds of distances for query-sample similarity measure. However, feature extraction and similarity calculation are independent, suggesting that the extracted features may not work the best with the distance measure. Therefore, we further proposed a Semi-Siamese Convolutional Network called IMINET [3] to tune feature extraction and metric learning together.

On another aspect, several methods for visualizing what a network learns have been developed [17]. The most straightforward method is to visualize the activations of each layer [18]. Another method is activation maximization [19], which generates an input that maximally activate a certain neuron by performing gradient ascent of the neuron activation w.r.t. the input while keeping the filter fixed. A related technique is to search for the inputs within a dataset that maximally activate a neuron [20], which requires a large scale dataset including extensive input patterns. These techniques, however, have not been applied to audio signals, nor was it adapted for multiple-input neural network structures like Siamese style networks.

3. THE TL-IMINET MODEL

TL-IMINET is improved from IMINET through transfer learning. Its overall structure is shown in Figure 1. Different from IMINET, the two towers for feature extraction do not share the same structure. Instead, they adopt structures that are specially designed for spoken language recognition (vocal imitation tower) and environmental sound classification (sound recording tower), respectively. Their input dimensions are also different, accounting for the different types of sounds they receive. Network weights of the two tower

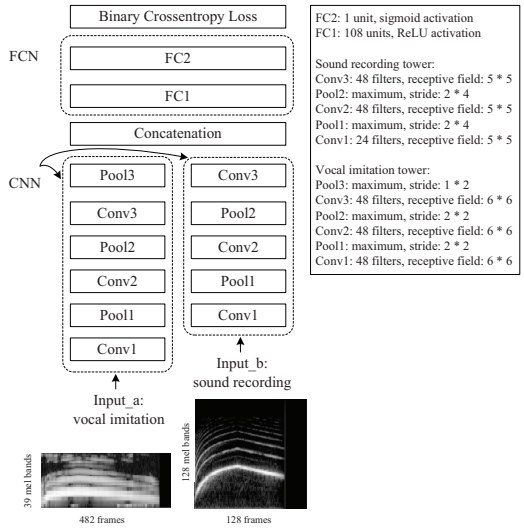

Fig. 1. TL-IMINET structure. The imitation input is a log-mel spectrogram with frequency range of 0-5 kHz, 4 seconds long, frame and hop size both of 8.33 ms. The sound input is also a log-mel spectrogram with frequency range of 0-22,050 Hz, 3 seconds long, frame and hop size both of 23 ms.

are pre-trained on their own source tasks, then fined tuned together with the FCN on the sound search task.

3.1. Pre-training CNN Towers

Imitation Tower: Vocal imitations and speech utterances are all human voices, therefore, we pre-train the imitation CNN on a spoken language recognition task. We adopt a network structure proposed in [7] and slightly modify it: we use three convolutional layers, each followed by a pooling layer. The number and size of filters and the pooling parameters are specified in Figure 1. We pre-train the network on a 7-class (Dutch, English, French, German, Italian, Russian, and Spanish) spoken language recognition task, using data from VoxForge [21]. For each language, we use 8,000 speech clips contributed by various users, 70% for training and 30% for testing. Each speech clip is truncated to 4 seconds and converted to a 39-band log-mel spectrogram with 8.33 ms for both the frame size and hop size. Therefore, the log-mel spectrogram has a dimensionality of 39 * 482. The 7-class classification model achieves 69.8% accuracy.

Recording Tower: A sound search database may contain various kinds of sounds, and a large portion are everyday sounds. Therefore, we adopt an environmental sound classification network structure [8] for the sound recording tower. The structure is described in Figure 1. We train the network on the same 10-class sound classification task using the UrbanSound8K dataset [8] without data augmentation. The dataset contains 8,732 labeled sound excerpts from 10 classes. On the same 10-fold cross validation setup, we achieve 70.2%

Table 1. MRR comparisons of TL-IMINET with the baseline.

Config.	AI	CS	ED	SS
Tied-IMINET	0.401	0.327	0.158	0.380
TL-IMINET (w/o pretrain)	0.397	0.309	0.225	0.377
TL-IMINET (w/ pretrain)	**0.462**	**0.349**	**0.246**	**0.390**

average accuracy, similar to the reported performance in [8].

3.2. Fine-tuning Entire Network

After pre-training the two CNN towers on their own source tasks, we fine-tune them together with the FCN for our target task: sound search by vocal imitation. To do so, we use the VocalSketch Data Set v1.0.4 [22], which contains 240 sound recordings with different sound concetps falling into 4 categories, namely Acoustic Instruments (AI), Commercial Synthesizers (CS), Everyday (ED), and Single Synthesizer (SS). with different concepts in total. Each recording has 10 vocal imitations. We create 840 positive pairs and 840 negative pairs of imitations and recordings from 120 sound concepts. In a positive pair, the imitation was produced to imitate the recording. In a negative pair, the imitation and the recording have no relationship.

We employ the Stochastic Gradient Descent (SGD) algorithm to minimize the binary cross-entropy between the predicted similarity (between 0 and 1) and the ground-truth label (binary), where 1 and 0 denotes positive and negative pairs, respectively. The learning rate is 0.01, learning rate decay is 0.0001, and momentum is 0.9. The batch size is 128. Training is terminated after 30 epochs.

3.3. State-of-the-art Performance

We use the other 120 sound recordings and their vocal imitations of the VocalSketch Data Set v1.0.4 [22] to evaluate TL-IMINET against existing models. We employ the same experimental setup described in our previous work [3]. Mean Reciprocal Rank (MRR) is adopted to evaluate the retrieval performance in each category.

MRR ranges from 0 to 1 with a higher value indicating a better sound retrieval performance. We report the average MRR across 10 runs of the system. In Table 1 we compare it with our previous Tied-IMINET system [3], which was the best model on this dataset. It can be seen that without pre-training the CNN towers, TL-IMINET achieves a similar MRR as Tied-IMINET on AI, CS and SS categories, and outperforms Tied-IMINET on ED category. With pre-training, the MRR value is significantly improved across all categories, showing the benefit of transfer learning and that TL-IMINET is the new state of the art based on our previous work [2, 16, 5, 3].

4. VISUALIZATION AND INTERPRETATION

4.1. Dual-Input Activation Maximization

We generate and visualize the input spectrograms that maximally activate each neuron [19]. This can be done by taking the gradient of that neuron's activation w.r.t. the input while keeping the trained weights unchanged and updating the input by gradient ascent from a random initialization. After convergence, the updated input spectrogram can be interpreted as what that neuron learns. For better visualization purposes, ReLU activations in TL-IMINET are replaced by leaky ReLU with a slope coefficient of 0.3 for negative inputs. This is to prevent the zero gradient issue when the input value to the ReLU activation is negative, which will cause the optimization of the input data for certain neurons to be trapped.

1) CNN Neurons: Each CNN neuron receives the signal from either the vocal imitation input or the sound recording input, but not both. Therefore, we adopt the following objective function for gradient ascent:

$$\underset{x}{\text{argmax}}\left(\frac{\partial(A_{c_{ij}} - \lambda \left\|x\right\|^2)}{\partial x}\right. \tag{1}$$

where $A_{c_{ij}}$ is the activation of the i-th neuron in the j-th convolutional layer from imitation or recording tower, x is the input pattern for either the imitation or recording tower. $\lambda = 0.1$ is a weighting factor to prevent x from being arbitrarily large.

2) FC Neurons: Each FC neuron receives the signal from both the vocal imitation and sound recording inputs. Therefore, we should actually visualize a pair of imitation and recording inputs. This is different from the visualization for Single-Input-Single-Output (SISO) networks. We adopt the following objective function:

$$\underset{x_{imi},x_{rec}}{\text{argmax}}\left(\frac{\partial(A_{f_{ij}} - \lambda \left\|x_{imi}\right\|^2)}{\partial x_{imi}} + \frac{\partial(A_{f_{ij}} - \lambda \left\|x_{rec}\right\|^2)}{\partial x_{rec}}\right) \tag{2}$$

where A_{fij} is the activation of the i-th neuron in the j-th dense layer, x_{imi} and x_{rec} are the input patterns for imitation and recording towers.

4.2. Visualization of Convolutional Neurons

We visualize the input patterns that maximize convolutional layer activations for the recording tower in Figure 2. The imitation tower visualization shares the similar patterns, and can be accessed via: `https://goo.gl/Y5ytv6`.

For figure arrangement, columns from left to right represent the learned input patterns in Conv1, Conv2, and Conv3, respectively. Rows show different network configurations. The first row shows input patterns before network fine-tuning, i.e., the network only trained on the UrbanSound8K for environmental sound classification. The second row shows the learned patterns of the network trained from scratch only using the VocalSketch dataset without pre-training. The last

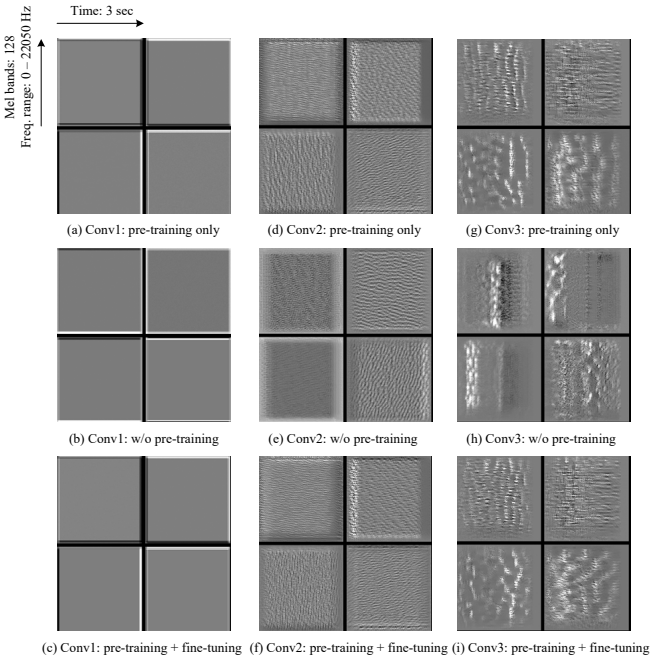

Fig. 2. Recording tower input pattern visualization. Brighter color represents higher energy.

row shows learned patterns after the complete process of pre-training and fine-tuning. We only show 4 input patterns in each layer due to space constraints.

The first column shows that input patterns with local edge information activates Conv1 neurons the most. Only a single stripe is observed in each input pattern. The second column displays more texture-like features with horizontal, vertical, and inclined stripes. For the third column, complicated temporal-frequency patterns are observed, which present global structures. Both Conv2 and Conv3 input patterns have periodic stripes, but patterns in Conv3 are coarser. For example, we observe vertical stripes in bottom left pattern of (d) and bottom right pattern of (g), the period in Conv3 is roughly 4 times larger than Conv2. This may be due to the 2*4 (4 in time) pooling effect after Conv2. Also interestingly, in Conv3 some patterns represent harmonic structure, some represent vertical stripes showing rhythmic patterns.

Besides, we find that pattern visualization with fine-tuning is shaper and clearer compared with the ones without fine-tuning, as comparing the first and third rows. For those layers without pre-training, interesting patterns appear locally. Also much finer patterns can be obtained with pre-training by comparing patterns in the first/third row with the second row. This suggests that prior knowledge from the original UrbanSound8k dataset provides essential information about various sound events, which is lacking from the learning from VocalSketch data set itself. Further more, by comparing the first and third row, after fine-tuning (the model used for the third row is initialized by the weights learned in

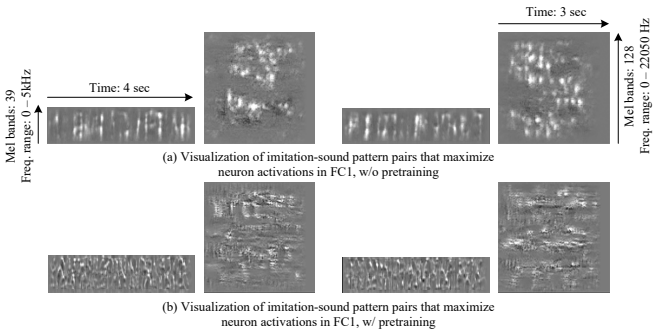

Fig. 3. Imitation-recording input pair pattern visualization in FC1. Brighter color represents higher energy.

the first row), the input patterns are derived from pre-training, but more variations and detailed structures can be observed.

4.3. Visualization of FC Neurons

The neuron in fully connected layer receives a pair of inputs, and the receptive field of each neuron covers the entire input ranges of both the imitation and recording. So we need to find a certain imitation-recording pair to maximally activate the neuron. In Figure 3, by selecting 2 neurons in FC1 that the rest neurons can be well represented by, the corresponding imitation-recording input pattern pairs are shown without and with pre-training TL-IMINET respectively. Complete results can be found on our webpage. By pre-training TL-IMINET, much detailed structures from the pairs can be observed compared with the configuration of without pre-training. In both Figure 3(a) and (b), imitation and recording are alike to form pairs that maximally activate the FC neurons.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we applied the activation maximization technique to visualize input patterns that maximize the activation of different neurons of a Siamese style network. This network, called TL-IMINET, was designed for sound search by vocal imitation. It uses two towers of specialized Convolutional Neural Networks (CNN) to extract features from vocal imitations and sound recordings, which are pretrained by two source task datasets respectively, then uses a Fully Connected Network (FCN) to predict the similarity between the imitation and the sound. Through visualization, we attempted to gain insights about how such architecture works. Interesting patterns are found to reveal the local-to-global and simple-to-conceptual learning mechanism of TL-IMINET. Experiments also show how transfer learning helps to improve TL-IMINET performance from the visualization aspect. For future work, we would explore sonifications that transform generated spectrogram-like input patterns to audible sounds. Also we would conduct subjective studies for our system.

6. REFERENCES

[1] David Sanchez. Blancas and Jordi Janer, "Sound retrieval from voice imitation queries in collaborative databases," in *Proc. Audio Engineering Society 53rd International Conference on Semantic Audio*, 2014, pp. 1–6.

[2] Yichi Zhang and Zhiyao Duan, "Retrieving sounds by vocal imitation recognition," in *Proc. Machine Learning for Signal Processing (MLSP), 2015 IEEE International Workshop on*, 2015, pp. 1–6.

[3] Yichi Zhang and Zhiyao Duan, "IMINET: Convolutional semi-Siamese networks for sound search by vocal imitation," in *Proc. Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*, 2017, pp. 304–308.

[4] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams, "The timbre toolbox: Extracting audio descriptors from musical signal," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[5] Yichi Zhang and Zhiyao Duan, "Supervised and unsupervised sound retrieval by vocal imitation," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 533–543, 2016.

[6] Yichi Zhang, Bryan Pardo, and Zhiyao Duan, "Convolutional siamese style neural networks for sound search by vocal imitation," (in preparation).

[7] Gregoire Montavon, "Deep learning for spoken language identification," in *Proc. NIPS Workshop on deep learning for Speech Recognition and Related Applications*, 2009, pp. 1–4.

[8] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[9] Moshe M. Zloof, "Query-by-example: A data base language," *IBM Systems Journal*, vol. 16, no. 4, pp. 324–343, 1977.

[10] Asif Ghias, Jonathan Logan, David. Chamberlin, and Brian C. Smith, "Query by humming: musical information retrieval in an audio database," in *Proc. the 3rd ACM International Conference on Multimedia*, 1995, pp. 231–236.

[11] Ajay Kapur, Manj Benning, and George Tzanetakis, "Query-by-beating-boxing: Music retrieval for the DJ," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2004, pp. 170–177.

[12] Thierry Bertin-Mahieux and Daniel PW Ellis, "Large-scale cover song recognition using hashed chroma landmarks," in *Proc. Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, 2011, pp. 117–120.

[13] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng, "A lattice-based approach to query-by-example spoken document retrieval," in *Proc. the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 363–370.

[14] Gerard Roma and Xavier Serra, "Querying freesound with a microphone," in *Proc. the 1st Web Audio Conference (WAC)*, 2015.

[15] Marko Helén and Tuomas Virtanen, "Audio query by example using similarity measures between probability density functions of features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 179303, 2009.

[16] Yichi Zhang and Zhiyao Duan, "IMISOUND: an unsupervised system for sound query by vocal imitation," in *Proc. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 2269–2273.

[17] "Understanding and visualizing convolutional neural networks," `http://cs231n.github.io/understanding-cnn/`, Accessed: 2017-09-30.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[19] Dumitru Erhan, Aaron Courville, and Yoshua Bengio, "Understanding representations learned in deep architectures," *Department d'Informatique et Recherche Operationnelle, University of Montreal, QC, Canada, Tech. Rep 1355*, pp. 1–25, 2010.

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 580–587.

[21] "VoxForge," `http://www.voxforge.org/`, Accessed: 2017-09-30.

[22] Mark Cartwright and Bryan Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proc. the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 43–46.