EMBEDDING PHYSICAL AUGMENTATION AND WAVELET SCATTERING TRANSFORM TO GENERATIVE ADVERSARIAL NETWORKS FOR AUDIO CLASSIFICATION WITH LIMITED TRAINING RESOURCES

Teh Kah Kuan and Tran Huy Dat

Acoustic, Speech and Language Department, Institute for Infocomm Research, A*STAR Singapore

ABSTRACT

This paper addresses audio classification with limited training resources. We first investigate different types of data augmentation including physical modeling, wavelet scattering transform and Generative Adversarial Networks (GAN). We than propose a novel GAN method to embed physical augmentation and wavelet scattering transform in processing. The experimental results on Google Speech Command show significant improvements of the proposed method when training with limited resources. It could lift up classification accuracy from the best baselines of 62.06% and 77.29% on ResNet, to as far as 91.96% and 93.38%, when training with 10% and 25% training data, respectively.

Index Terms— Audio Classification, Limited Training, Augmentation, Generative Adversarial Networks, Wavelet Scattering Transform

1. INTRODUCTION

It is well known that, current state-of-the-art deep learning [1]-[2] requires huge amount of labeled data that comes with enormous costs, especially in audio classification where more training data is needed to cover the variations caused by the uncontrollable nature of audio sources. Therefore, building a robust audio classification engine with limited training resources is an important and practical problem that we are going to address in this paper.

Recent developments in related areas such as speech recognition and image classification suggest practical engineering solutions by employing data augmentation on top of available training data [3]-[6]. To improve the environmental robustness against noise and reverberation, speech samples were mixed with noises [3] to tackle with environments , or convolved to simulated or measured room impulse responses (RIRs) [4] to model the multi-paths and reverberations, or shifted in frequency or pitch scales to model the variations of vocal tracts [5]-[6]. We call these methods physical augmentation and that could be the first idea to be adopted in audio classification with limited training resources. Wavelet scattering transform is another great idea to combine augmentation with deep learning, particularly the Convolutional Neural Network (CNN) which is proven in image classification [7]-[8]. It is done by scattering 2-D image localized paths from wavelet transforms, to create more data variations before feeding into CNN layers of the classifier. In this paper, we also adopt this concept in the audio classification task.

Using deep learning for augmentation is a next logical idea that we applied in our previous work [9] on through-the-wall audio classification. When the audio recording involves many non-linear effects caused by modulation effects, physical augmentation and scattering transform are not applicable but deep learning can come to model the observations to supply samples to training. More advanced technology in this direction is the Generative Adversarial Networks (GAN), which simultaneously generates synthesized data for training and classifies the samples. The joint optimization of generator/discriminator could elegantly balance the over fitting and improve classification. It has achieved amazing results in image classification [10], speech synthesis [11], and recently being applied with a little success in noisy speech recognition [12], but not yet to be adopted in audio classification.

In this paper, we first systematically investigate the augmentation methods, particularly: physical augmentation, wavelet scattering transform and GAN for audio classification with limited training resources. We than propose a new GAN design which allows embedding of the former two to significantly improve the classification accuracy. We note that besides augmentation and GAN methods, there are other methodologies to address the task such as model adaptation which has been applied in speech recognition [13], transfer learning of the model borrowed from other large-scale classification tasks [14] but those methods need relatively good starting points and that is not always trivial in audio classification. Using unsupervised data is also another effective way to address the task but we will leave it for future works and just focus on developing augmentation-based GAN approaches to solve the problem without help from additional unsupervised data.

The organization of the paper is as follows: next, in Sec. 2, we give the details of physical augmentation (PA), wavelet scattering transform (WST) and Generative Adversarial

Networks (GAN), before proposing novel schemes to embed the first two into GAN training. Sec.3 then reports experimental results on Google Speech Command data [15]. Finally, Sec.4 concludes the work.

2. PHYSICAL AUGMENTATION AND WAVELET SCATTERING TRANSFORM EMEBDDED GAN

In this section, we review the methods to address audio classification with limited training resources before proposing a novel GAN scheme to integrate the physical augmentation and wavelet scattering transform.

2.1. Physical augmentation

Physical augmentation is the simplest way of increasing amount of training data in audio and speech classification tasks. It has been successfully applied in far-field noisy speech recognition tasks, first with HMM-GMM [3] and then deep learning frame works [4]-[5]. Thanks to the linear property of sound propagation, far-field speech $x(t)$ can be simply modeled as a convolution of clean speech $s(t)$ and a room impulse $h(t)$ response and further added to different noise type and levels $n(t)$ to create noisy speech samples, denoted by

$$x(t) = s(t) * h(t) + n(t) \tag{1}$$

The second type of physical augmentation is vocal tract length normalization (VTLN) [6], where the basic idea comes from the observation of frequency shifting of vocal sounds due to the vocal tract length change. It can be described by adding a linear (or bi-linear) modulation $\varphi_{VTLN}(\omega)$ [6] to the signal spectrum denoted by

$$X(\omega) \rightarrow X[\varphi_{VTLN}(\omega)] \tag{2}$$

The third type of augmentation is the effect of speaking rate in vocal sound production. It creates a wrapped time signal which translates into linear scaling in frequency domain

$$x(\alpha t) \rightarrow \frac{1}{\alpha} X\left(\frac{\omega}{\alpha}\right) \tag{3}$$

But unlike VTLN, speaking rate augmentation changes the signal duration in time domain. In this work, we combine all three physical augmentations above which transform one set of training data into three equal data sets with varying parameters of SNR, VTL, speaking rate, as shown in Fig. 1.

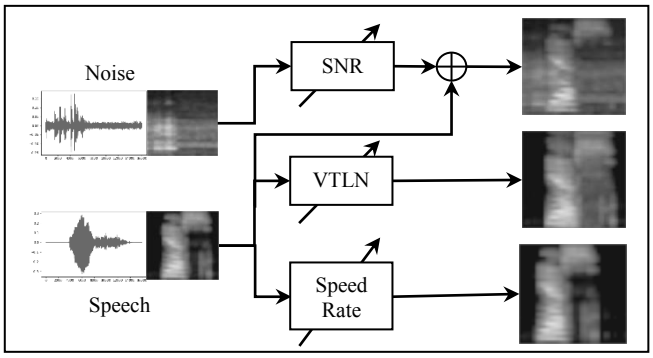

Figure 1: Physical augmentation generation

2.2. Wavelet scattering transform

The physical nature of vocal and non-vocal sound production leads to the effects of time, pitch or frequency shifting and fluctuation of modulation curves, hence translation invariance and deformation stability are the key factors to deliver a robust classifier. Basically, CNN, particularly its pooling operation, is understood as a main engine for invariance for small shifts and distortions [16], although the latest works have shown the limitation of its effectiveness [17]. Wavelet scattering transform (WST) [18] is viewed as a physic-driven CNN layer to allow invariance in translations and deformation stability. It was successfully applied in computer vision [19] and in this paper we adopt this concept for audio classification of spectrograms. Given an audio spectrogram x, WST exploits multi-scale analysis using a cascade of wavelet filter $\psi_{\lambda_k,\theta_k}$, and convolving with local averaging filter ϕ_J with a spatial window of scale 2^J, where λ is the scale of the wavelet and θ is angular sector. Zeroth-order scattering coefficient denoted by

$$S_0 \mathrm{x} = \mathrm{x} * \phi_J \tag{4}$$

Next order of the scattering coefficients can be obtained as

$$S_k \mathrm{x}(\lambda_1, .., \lambda_k, \theta_1, .. \theta_k) = \{|\, \mathrm{x} * \psi_{\lambda_1,\theta_1}| * \psi_{\lambda_2,\theta_2}| \ldots * \psi_{\lambda_k,\theta_k}| * \phi_J\}_{\lambda_k \le J, \theta_k = 2\pi\frac{l}{L}, 1 \le l \le L} \tag{5}$$

Fig.2 shows the concept of 2-layers WST with CNN network. It can be considered as a type of data augmentation which increases the variation of audio spectrogram in the training.

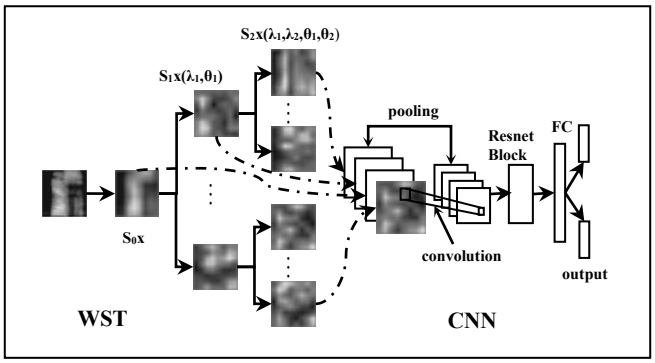

Figure 2: The basic diagram of a 2-layers wavelet scattering transform as a first layer of the CNN classifier.

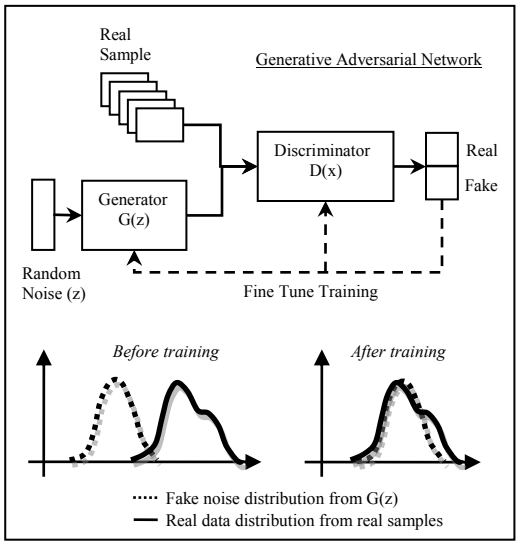

Figure 3: Generator learned how to sample from real data distribution p_{data}

2.3. Generative Adversarial Networks (GAN)

In our previous work [9], we applied LSTM for data augmentation in a task of through-the-wall audio classification. However, a serious problem of deep learning augmentation is the over-fitting situation. GAN [20] is elegantly and effectively designed to solve this problem. The original idea here is to employ a pair of networks, generator and discriminator, to bring closer the distributions of real and fake (simulated) data in order to improve the original classification problem, as seen in Fig.3.

AC-GAN (Auxiliary Classifier Generative Adversarial Network) is a newer GAN method [21] to address the multi-class classification. The main difference here is that that the multi-class training label is fed to the generator to create specific label oriented fake samples while the discriminator simultaneously and auxiliary optimize both multi-class and fake/non-fake objectives and by doing so could prevent the occurrence of over-fitting. We note that GAN has not been yet applied in similar tasks for audio classification.

2.4. Augmentation GAN

In this paper, we propose a novel GAN scheme to allow embedding of the physical augmentation and wavelet scattering transform into GAN. We call our GAN augmentation-GAN (augGAN). The block diagram of the augGAN is shown in Fig.4. Similarly to AC-GAN [21], the multi-class labels are fed to the generator together with random noises to simulate training samples. But unlike AC-GAN, both of the real, the physically augmented and the faked data join to the discriminator. The discriminator will simultaneously optimizes two objective functions for binary (real/fake) and multi-class classification but the

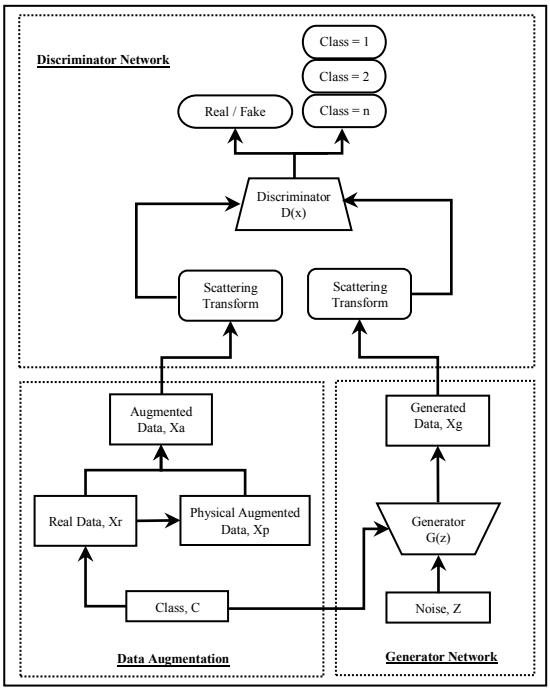

Figure 4: The block diagram of GAN integrated with physical augmentation and wavelet scattering transform.

optimization is performed on generated, real and physically augmented data. The wavelet scattering transform is embedded in the first layer of the discriminator to improve the translation invariance and deformation stability of the classifier. Another difference to the original AC-GAN is that both generator (G) and discriminator (D) adopt the ResNet architecture [23]. Cross-entropy criteria are used in both G and D with ResNet architecture. The experiments in next section will show the effectiveness of augGAN when training with limited resources. Note that there are more recent GAN schemes which allow usage of unsupervised, unlabeled data to boost the classification accuracy but that is out of the scope of this paper which focuses on the problem of limited resources.

3. EXPERIMENTS AND RESULTS

In this section, we evaluate and compare methods for audio classification with limited training resources.

3.1. Data Description

The Google Speech Commands Dataset [15] is used in our experiments. The dataset has 65,000 one-second long utterances of 30 short voice commands and are spoken by a variety of speakers. Ten core command words were selected, they are "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go". To emphasize the limited training data situations, we investigate two scenarios: using 10% (200 samples per class) and 25% (450 samples per class) of the original training data. We note that

using full training data would achieve 96%-97% overall classification accuracy with the ResNet design (compared to 95% baseline in [24])

3.2. Methods

The following methods have been implemented and evaluated with the noted two training situations.

- **Baseline ResNet-18***:* The best fine-tuned baseline.
- **WST + ResNet**: WST is added into ResNet
- **GAN_ResNet:** AC-GAN with ResNet designs in both G and D
- **WST + GAN_ResNet***:* WST is added into GAN's discriminator
- **Physical augmentation + ResNet***:* Physical augmentation on top of baseline ResNet
- **Physical augmentation + WST + ResNet:** Augmentation followed by WST added ResNet
- **Physical augmentation + GAN_ResNet:** Physical augmentation followed by GAN
- **Physical augmentation + WST + GAN_ResNet:** Full version of proposed AugGAN

Unique signal processing is performed by 32x32 segmented Mel-spectrogram and scaling to the range [-1, 1]. We train augGAN with mini-batches of 64 with a learning rate of 0.0002 for 250 epochs. To improved GAN stability, we used embedding layer [22] acts as look-up table, LeakyReLU, average pooling and Adam optimization was adopted.

Generator Architecture: We take the Resnet-18 [23], as a reference and construct a similar architecture. The generator network takes input as a 100-dimension random vector drawn from a Gaussian distribution and output spectrogram images of size 1x32x32. Class labels from training data are embedded in the fully connected layer through a look-up table. The output reshape to 512x4x4 and 4 residual stages of 2 blocks each with a 3x3 kernel size, 1x1 stride, 1x1 padding and up-samples the spectrogram image with scale 2. The ReLU activation function with slope of the leak of 0.2 is applied to all layers except the output layer which uses a tanh activation function.

Discriminator Architecture: The discriminator network has similar architecture as the generator except for the first ResNet blocks which are replaced with wavelet scattering transform coefficients, as shown in figure 2. In this work, we used a two-layer wavelet scattering transform and Morlet wavelets. The final scattering coefficients, have a size equal to $1 + JL + \frac{1}{2}J(J-1)L^2$ where J=2 and L =8, and the original size is down-sampled by a factor 2^J. Average pooling is applied to last ResNet block with a kernel size of 4 and stride of 4. The output consists of two separated fully connected layers, one with a sigmoid (adversarial classifier) and another has a softmax output distribution (auxiliary classifier). All the methods were implemented on PyTorch.

3.3. Overview of Results

Method	Accuracy	
	10% training data	25% training data
No physical augmentation		
ResNet-18 (baseline)	62.06%	77.29%
WST + ResNet	78.42%	85.66%
GAN_ResNet	74.68%	83.95%
WST + GAN_ResNet	88.55%	90.30%
With physical augmentation		
ResNet-18	80.25%	89.44%
WST + ResNet	84.61%	89.91%
GAN_ResNet	85.39%	91.86%
WST + GAN_ResNet	**91.96%**	**93.38%**

Table 1: Overall classification accuracies over methods

yes	No	up	down	left	right	on	off	stop	go
240	1	0	4	5	0	1	0	2	3
0	**216**	0	17	2	1	0	1	1	14
1	1	**253**	1	4	0	1	3	6	2
1	11	0	**228**	0	0	1	1	1	10
3	0	3	4	**254**	1	2	0	0	0
1	0	0	2	1	**251**	1	0	0	3
0	0	0	3	0	1	**240**	2	0	0
0	1	7	2	0	0	13	**229**	3	5
0	0	4	6	3	0	0	0	**232**	4
0	17	3	10	0	1	0	0	2	**218**

Table 2: Confusion matrix – train with 200 samples/class

Table 1-2 reports overall classification accuracies and confusion matrix over augmentation methods. We can see that for a "single" augmentation, physical augmentation is still the most effective compared to WST or GAN. WST is also very effective as it increases the data size by 81 times by performing scattering. All the methods are complementary. GAN looks less effective when used alone but greatly improves single augmentation, particularly with WST. It can be explained that by introducing a pair of generator/discriminator, the sample generation is well balanced hence making better use of huge multiplication of sample by scattering transforms resulting a boost over its normal data fitting capability.

Finally, the full version of augmentation GAN, taking advantage of physical augmentation, and wavelet scattering transform into GAN could significantly improve audio classification with limited training resources. It achieved 91.96% and 93.38% classification accuracies using only 10% and 25% training data, respectively.

4. CONCLUSIONS

This paper proposes a novel augmentation GAN which allows embedding of physical augmentation and wavelet scattering transforms into GAN scheme to address the task of audio classification with limited training resources. The experimental results show that the proposed method could greatly improve the classification accuracy coming close to full training range with just 25% of its resources.

REFERENCES

[1] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal and Marvin Ritter, “Audio Set: An ontology and human-labeled dataset for audio events,” ICASSP, pp. 776-780, 2017.

[2]Aren Jansen, Jort F. Gemmeke, Daniel P. W. Ellis, Xiaofeng Liu, Wade Lawrence and Dylan Freedman, “Large-scale audio event discovery in one million YouTube videos,” ICASSP, pp. 786-790, 2017.

[3] H.G.Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” Proceedings of the ISCA workshop ASR2000, Paris, France, 2000.

[4] Jonathan William Dennis and Tran Huy Dat, “Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2R'S system description for the ASpIRE challenge,” ASRU, pp. 518-524, 2015.

[5] Tom Ko, Vijayaditya Peddinti, Daniel Povey and Sanjeev Khudanpur, “Audio augmentation for speech recognition,” INTERSPEECH, 2015.

[6] Puming Zhan and Alex Waibel, “Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition,” CMU COMPUTER SCIENCE TECHNICAL REPORTS, 1997.

[7] Stephane Mallat, “Group Invariant Scattering,” Communications on Pure and Applied Mathematics, vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

[8] Edouard Oyallon, Stéphane Mallat and Laurent Sifre, “Generic Deep Networks with Wavelet Scattering.” CoRR, pp. abs/1312.5940, 2013.

[9] Tran Huy Dat, Wen Zheng Terence Ng and Yi Ren Leng, “Data Augmentation, Missing Feature Mask and Kernel Classification for Through-the-Wall Acoustic Surveillance,” INTERSPEECH, pp. 3807-3811, 2017.

[10] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford and Xi Chen, “Improved Techniques for Training GANs,” NIPS, pp. 2226-2234, 2016.

[11] Saito, Yuki, Shinnosuke Takamichi, and Hiroshi Saruwatari, “Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017.

[12] Ke Wang, Junbo Zhang, Sining Sun, Yujun Wang, Fei Xiang and Lei Xie, “Investigating Generative Adversarial Networks based Speech Dereverberation for Robust Speech Recognition,” arXiv preprint, pp. arXiv:1803.10132, 2018.

[13] Yajie Miao, Hao Zhang and Florian Metze, “Towards Speaker Adaptive Training of Deep Neural Network Acoustic Models,” INTERSPEECH, 2014.

[14] Yusuf Aytar, Carl Vondrick and Antonio Torralba, “SoundNet: Learning Sound Representations from Unlabeled Video,” NIPS, 2016.

[15] Warden P, "Speech Commands: A public dataset for single-word speech recognition," 2017.

[16] Avraham Ruderman, Neil C. Rabinowitz, Ari S. Morcos and Daniel Zoran, “Learned Deformation Stability in Convolutional Neural Networks,” CoRR, pp. abs/1804.04438, 2018.

[17] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel, “Handwritten digit recognition with a backpropagation network,” In Advances in neural information processing systems, pp. 396–404, 1990.

[18] Edouard Oyallon, Eugene Belilovsky and Sergey Zagoruyko, “Scaling the Scattering Transform: Deep Hybrid Networks,” CoRR, pp. abs/1703.08961, 2017.

[19] Edouard Oyallon, Sergey Zagoruyko, Gabriel Huang, Nikos Komodakis, Simon Lacoste-Julien, Matthew B. Blaschko and Eugene Belilovsky, “Scattering Networks for Hybrid Representation Learning,” CoRR, pp. abs/1809.06367, 2018.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” ArXiv e-prints, pp. arXiv:1406.2261, June 2014.

[21] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” arXiv preprint, pp. arXiv: 1610.09585, 2016.

[22] Yoon Kim, “Convolution Neural Networks for Sentence Classification.” arXiv preprint, pp. arXiv:1408.5882, 2014.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition.” arXiv preprint, pp. arXiv:1512.03385, 2015.

[24] Raphael Tang and Jimmy Lin, “Deep Residual Learning for Small-Footprint Keyword Spotting.” arXiv:1710.10361v2, 2018.