INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Open-Set Audio Classification with Limited Training Resources based on Augmentation Enhanced Variational Auto-Encoder GAN with Detection-Classification Joint Training

Kah Kuan Teh, Huy Dat Tran

Aural and Language Intelligence Department, Institute for Infocomm Research, A*STAR, Singapore

tehkk@i2r.a-star.edu.sg, hdtran@i2r.a-star.edu.sg

Abstract

In this paper, we propose a novel method to address practical problems when deploying audio classification systems in operations that are the presence of unseen sound classes (open-set) and the limitation of training resources. To solve it, a novel method which embeds variational auto-encoder (VAE), data augmentation and detection-classification joint training into conventional GAN networks is proposed. The VAE input to GAN-generator helps to generate realistic outlier samples which are not too far from in-distribution class and hence improve the open-set discrimination capabilities of classifiers. Next, the augmentation enhanced GAN scheme developed in our previous work [4] for close-set audio classification, will help to address the limited training resources by in cooperating the physical data augmentation to work together with traditional GAN produced samples to prevent overfitting and improve the optimization convergences. The detection-classification joint training further steps on advantages of VAE and Augmentation GAN to further improving the performances of detection and classification tasks. The experiments carried out on Google Speech Command database show great improvements of open-set classification accuracy from 62.41% to 88.29% when using only 10% amount of training data.

Index Terms: Audio Classification, Limited Training, Variational Autoencoder, Generative Adversarial Networks, Open set classification, Sinkhorn divergence

1. Introduction

It is well known that audio classification has received enormous interest from the research community [1]-[3], [5]-[7] and current deep learning networks have demonstrated state-of-the-art results in many challenging tasks. However, the majority of current tasks are limited by closed-set scenarios, where audio samples are classified into a fixed set of classes, or sufficient data is normally assumed available for training [7]. Those conditions, however, are not typical for real-life applications, where the system should be operating under continuously changing conditions. Hence, open-set and limited training resources should be the key problems of the task. Secondly, the business models require the system to operate under various unseen acoustic conditions. The idea of having universal models to work across multiple deployment conditions may thus face big challenges due to the uncontrolled mismatches. Therefore, solutions with actual limited labeled training data are more practically valuable. In summary, open-set and limited training resource are the two key problems of audio classification systems in deployments. In our previous work [4], we have proposed an approach to address the limited training resources under close-set classification conditions. The method embeds physical augmentations and wavelet scattering transform into a generative adversarial network (GAN) to improves its stability and generalization which translates into great improvements in GAN-based classification accuracy. In this paper, we are going to extend this method [4] for open-set situations. The open-set classification problem [8] is highly related to the outlier detection problem where a test sample is classified into in-of-distribution (i.e. available in training data) or out-of-distribution (i.e. not available in training data). Despite the importance, open-set audio classification has received little attention from the research community. There have been recent developments utilizing a Deep Neural Network (DNN) classifier and applying a threshold on the output probability [9]-[11] for open set classification problems. For each input, it measures a maximum value of the predictive distribution based on a pre-trained classifier and compares the score to some pre-define threshold. However, those methods were not designed to work with limited training resources and hence not so effective for investigating the problems.

In this paper, we integrate the outlier detection into the previously developed augmentation-embedded GAN classification method. The basic idea here is to have a multi-task joint training of both detection and classification under augmentation-embedded GAN method. To achieve that purpose, the random input to the GAN-generator is replaced by a Variational Auto-Encoder (VAE) [12]-[14] with reparameterization tricks. The reparametrized VAE input to GAN-generator helps to generate realistic outlier samples which are not too far from in-distribution class. This would improve the open-set discrimination capabilities of classifiers. We evaluate the proposed method on a real-life task of developing a low-cost low-powered (non-ASR type) speech command (hot-words) recognition through experiments using Google speech command dataset. The experiment will emphasize on the practical issue of current smart speaker developments that requiring huge amounts of data to train hot words recognition. This results in inflexibility in introducing new sets of speech commands. The experiments in this paper evaluate the methods using only 10% of the original training resources. The organization of the paper is as follows: next, in Sec. 2, we will give an overview of our system. Sec.3 then reports experimental results on Google Speech Command data [15]. Finally, Sec.4 concludes the work.

http://dx.doi.org/10.21437/Interspeech.2021-1142

2. System Overview

In this section, we describe the proposed GAN scheme for audio classification under open-set and limited training resource conditions. Fig.1 illustrates the block diagram of the proposed method. The upper part is similar to previous augmentation-embedded GAN method [4], where the physic modeling and wavelet scattering module are added to a typical multi-class GAN scheme to improve the stability and accuracy of the classification task. To tackle the open-set situations two novelties have been adopted in this work: (1) the replacement of random input to GAN-generator by a VAE network which encodes the existing real data via a statistically driven embedding layer. A re-parametrization trick is introduced by adding Gaussian noise to the embedding to produce slightly

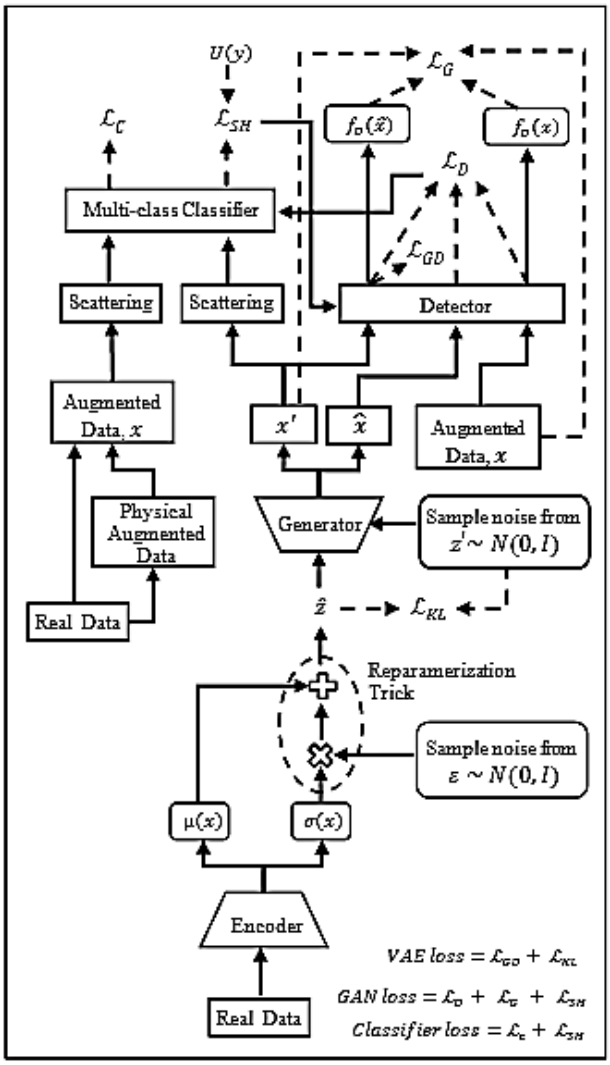

Figure 1: *The architecture of the proposed method. Network structure contains four parts: 1) The Encoder network (VAE); 2) The Generative network (GAN). VAE-GAN used to generate new data by regularizing the latent space in an unsupervised manner; 3) Detector network, used to discriminate the seen and unseen data; and 4) Classifier network. Dash lines are the loss functions as described in Equation [1-8]*

different samples. The aim is for the generator to produce challenging outlier samples so that the discriminator (detector) can learn better; (2) the joint training between multi-class classification and detection allowing the GAN to perform both outlier detection and classification. This can be done by simultaneously optimizing the cross-entropy for the classification and Sinkhorn divergence [16]-[17] for outlier detection. The latter measures distance between empirical distribution between in and out-of-distributions which separates the distributions better than classical KL distance based on Gaussian assumptions. In the next section, we provide details about each of our proposed framework's modules.

2.1. Classifiers

The key difference of the GAN scheme proposed in this work is that the discriminator includes both multi-class classification to classify the samples and detection to reject the outliers. We propose to formulate the classifier in a multi-task learning framework. The first learning task of the classifier is the closed-set classification and is trained to produce correct labels for samples from in-distribution. The objective function cross-entropy can be formulated as:

$$L_C = -E_{x \sim P_{real}}[logP_\theta(x)] \quad (1)$$

The second learning task is to force predictive distribution on out-of-distribution samples $q_\theta(x)$to be closer to the uniform distribution $U(y)$. To achieve this, the Sinkhorn divergence [16]-[17] between $q_\theta(x)$ and $U(y)$ is minimized. In other words, we need to minimize:

$$L_{SH} = SH\big(q_\theta(x), U(y)\big) \quad (2)$$

we refer L_{SH} as a confident loss. Sinkhorn divergences rely on an idea of blurring the transport plan through the addition of entropy regularization, in order to approximate the optimal transport between two distributions Q and P respectively, at a low computation cost. Let γ be the transport matrix and the entropy be $\Omega(\gamma)$. The regularization has the following expression:

$$\Omega(\gamma) = -\sum_{i,j} \gamma_{i,j} log\ (\gamma_{i,j}) \quad (3)$$

The Sinkhorn distance denoted by:

$$SH(Q, P) = argarg \sum_{i,j} \gamma_{i,j} M_{i,j} + \varepsilon\Omega(\gamma) \quad (4)$$

where $M_{i,j}$ is the metric cost, which is defined as the cost to move mass from a_i to b_j and ε is a blurring parameter.

2.2. Reparametrized Variational Auto Encoder (VAE)

A novel feature of the proposed method compared to the previous work [4] is having a VAE module in the input part. The key for this method is that the encoder maps the training data around the standard normal distribution via a statically driven embedding layer. Therefore, when we sample from the standard normal, it should be similar to what can be retrieved from the training data. Reparameterization trick [18]-[19] is then applied by changing the embedding representation to get the generator to produce out-of-class samples that look like it is from the training data, which is called the prior predictive sample. The cost function of a VAE is the combination of two terms:

1. Expected log-likelihood, negative binary cross-entropy between input data x and the reconstruction x'.

$$L_{GD} = -E_{z\sim P_z}\left[logD\left(G(z)\right)\right] \quad (5)$$

2. KL-divergence between the posterior and prior.

$$L_{KL} = \frac{1}{2}\mu^T\mu + \sum e^{\varepsilon} - \varepsilon - 1 \quad (6)$$

Expected log-likelihood is responsible for the reconstruction penalty, and KL divergence is responsible for the regularization penalty.

2.3. Generator & Detector

The detector is used for adversarial training and is trained to discriminate generated samples from the real samples. Formally, the objective for the detector is as follows:

$$L_D = -E_{x\sim P_{real}}[logD(x)] - E_{z\sim P_z}\left[log\left(1 - D\left(G(z)\right)\right)\right] \quad (7)$$

The generator is used to generate unseen-in-training samples. By using the l_2 reconstruction loss and discriminator feature matching [20]-[21], the generative model is enforced to emit diverse samples and generate structure-preserving samples which improves the stability of GAN learning. The objective for the generator is as follows:

$$L_G = \frac{1}{2}(\alpha\|x - \hat{x}\|_2^2 + \beta\|f_D(x) - f_D(\hat{x})\|_2^2) \quad (8)$$

where x is real samples, $\hat{x}$ is synthesized samples and f_D is the features of an intermediate layer of discriminator, α and β are weight coefficients.

2.4. Joint Training of Classifier and Detector

Finally, we proposed a joint training scheme for both classification and generative neural networks for detecting and generating out-of-distribution by minimizing their losses alternatively. We suggest the following joint objective function:

$$E_{x\sim P_{real}}[-logP_\theta(x)] + \lambda SH\left(q_\theta(x), U(y)\right) + E_{x\sim P_{real}}[logD(x)] - E_{z\sim P_z}\left[log\left(1 - D\left(G(z)\right)\right)\right] \quad (9)$$

The confident loss is weighted by coefficients λ that balance the quality of generation and sampling. In our experiments, we empirically set $\lambda = 0.1$.

3. Experiment and Results

In this section, we evaluate and compare methods for multi-class open-sets audio classification using Google Speech Command datasets [15]. The actual task tackles on a real-life task of developments of a low-cost low-powered (non-ASR type) speech command (hot-words) recognition for smart speakers. The experiment will be emphasized on the practical issue of current developments that requires huge amount of training data and hence not flexible with a new set of commands. The experiments in this paper evaluate the methods using only 10% of original training resources.

3.1. Data Description

The dataset has 65,000 one-second long utterances of 30 short voice commands and are spoken by a variety of speakers. We split the dataset into 2 main classes, known classes: 10-classes with distinctly labeled positive training samples from command words "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", and unknown classes: these are command words unseen in training. The unknown classes serve to simulate the conditions of an open set scenario. Training data consists of the 10 known classes and from there, we only used 10% of its resources. Testing dataset includes both known and unknown classes, consisting of all command words.

3.2. Methods

The aim of this evaluation is to benchmark the proposed methods to conventional approaches for a combined task of open set and limited training resources. Each system includes of an outlier detector and multi-class classifier. Since the previous work [4] has provided comprehensive benchmark of the audio classification methods for limited training resource problem, this paper will focus on comparing the detections in combination to AugGAN the best method for close-set found in [4]. We also compare it to the current approaches adopted in conventional smart speakers, which are based on comprehensive deep learning close-set classification schemes trained on large collected datasets. The implemented methods can be summarized as follows:

- Conventional smart speaker with full training: applying posterior thresholding [10] on closet-trained ResNet [25]-[26], using the full training set.
- Conventional smart speaker with limited training: applying posterior thresholding [10] on closet-trained ResNet [25]-[26], using only 10% of the full training data.
- Posterior-AugGAN (baseline): Posterior thresholding [10] as detector followed by the best proven classifier AugGAN classification [4], using only 10% of original training data.
- OpVAEAugGAN: Proposed joint detection-classification training with reparametrized VAE embedded into AugGAN, using only 10% of original training data.

For posterior thresholding, the threshold is empirically optimized on development data.

The details of OpVAEAugGAN processing is as follow: Unique signal processing is performed by 32x32 segmented Mel-spectrogram and scaling to the range [-1, 1]. We train OpVAEAugGAN with mini batches of 100 with a learning rate of 0.0002 for 150 epochs and Adam optimization was adopted.

VAE Architecture: The Encoder network takes input spectrograms of size 1x32x32, followed by two-layers convolution network with 4x4 kernel size and ReLU activation function. The output consists of two separated fully connected layers (latent representation space z with output dimensions 20).

The generator and discriminator have similar network architecture. both uses three-layers convolution network with 4x4 kernel size and ReLU activation function with slope of the leak of 0.2 applied to all layers.

Classifier Architecture: We take Resnet-18 [22], as a reference and construct a similar architecture, 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:

$$1 + JL + \frac{1}{2}J(J-1)L^2 \quad (10)$$

where $J = 2$ and $L = 8$, and the original size is down sampled by a factor 2^J. Average pooling is applied to the last ResNet block with a kernel size of 4 and stride of 4. The output consists of fully connected layers with a softmax output distribution. All the methods were implemented on PyTorch [23] and Sinkhorn divergence calculation using GeomLoss library [24].

3.3. Performance Metrics

We measure the following metrics:

- Area under the receiver operation characteristic curve (AUROC): The ROC curve is a graph plotting true positive rate = TP/(TP+FN) against the false positive rate = FP/(FP+TN).
- Area under the precision-recall curve (AUPR): The PR curve is a graph plotting the precision = TP/(TP+FP) against recall = TP/(TP+FN). AUPR_IN (or _OUT) is AUPR where in-(or out-of-) distribution samples are specified as positive.

3.4. Overview of Results

Table 1 shows the evaluation results of methods in open-set setup when around 50% of testing samples are from unseen classes to the training. We can see that conventional approaches currently adopted in commercial smart speakers, which includes close-set deep learning classifier (aka ResNet) [25]-[26] and posterior thresholding detector, performs poorly with limited training as the original classifier is not designed to handle limited training resources. Multi-class GAN scheme, particularly our augmentation enhanced GAN (AugGAN) [4], was shown to be effective in solving limited training resources problem. It can be extended into open-set scenario by adopting conventional outlier detectors such as posterior thresholding or VAE. The former yields more than 80% overall accuracy on open-set test set when using only 10% of training data. The proposed joint training AugGAN with reparametrized VAE input could significantly improve the outlier detection,

Table 1: *Overall detection and classification results over methods in open-set setup.*

Method	AUROC	AUPR_IN	AUPR_OUT
Conventional smart speaker (full training)	85.15%	86.77%	80.51%
Conventional smart speaker (10% training data)	62.41%	64.35%	60.12%
Posterior-AugGAN (baseline) (10% training data)	81.53%	82.13%	76.13%
OpVAEAugGAN (proposed) (10% training data)	88.29%	89.41%	86.27%

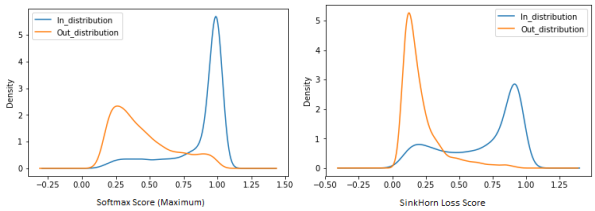

Figure 2: *In and out-of-distribution of the maximum prediction value in Softmax score and Sinkhorn score.*

resulting in close to 90% overall accuracy for the open-set classification task. The proposed method outperformed the conventional smart speaker approach using the full training set and that is a very important result as collecting training data is very expensive. This approach opens opportunities for low-cost developments of voice control features in smart speaker or IoT devices for new and low-resource languages.

3.5 Sinkhorn scores vs Softmax scores

An interesting question is to analyze the contribution of Sinkhorn distance in the proposed solution. Figure 2 compares the in- and out-of-training distributions of conventional Softmax and introduced Sinkhorn scores. It can be seen that the latter compressed the in-class distribution and significantly reduced the intersection area between the two distributions. This translates into better ROC curves of OpVAEAugGAN which yields equal error rate (EER) at 20% better than state-of-the-art methods.

4. Conclusions

This paper proposes a novel method to address open-set audio classification with limited training resources. The method extends from our previous work [4], where an augmentation embedded GAN scheme was proposed and found superior to others on limited training resources, when applied to close-set multi-class classification tasks. In this work, Variational Auto-Encoder (VAE) with reparametrized trick was introduced as the input module to the previous AugGAN scheme, on top of joint training between outlier detector and classification. The reparametrized VAE improves the quality of generated input to GAN generator and subsequently improves the outlier detection as well as stability of GAN classification module. This method outperformed commercial solutions using only 10% of original training data on non-ASR low-powered speech command recognition task. The technology can be applied for any open-set limited AI training tasks.

5. References

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: tasks, datasets and baseline system," In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), pp. 85–92. November 2017.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "A multi-device dataset for urban acoustic scene classification," In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), pp. 9–13. November 2018.

[3] Zhu, Houwei, et al. "DCASE 2019 challenge task1 technical report." Tech. Rep., DCASE2019 Challenge, Tech. Rep (2019).

[4] K. K. Teh, Tran Huy Dat, “Embedding Physical Augmentation and Wavelet Scattering Transform to Generative Adversarial Networks for Audio Classification with Limited Training Resources,” In ICASSP, 2019.
[5] Wilkinghoff, Kevin, and Frank Kurth. "Open-set acoustic scene classification with deep convolutional autoencoders." (2019).
[6] Rakowski, Alexander, and Michal Kosmider. Frequency-aware CNN for open set acoustic scene classification. DCASE2019 Challenge, Tech. Rep, 2019.
[7] Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. “Acoustic scene classification in DCASE 2019 challenge: Closed and open set classification and data mismatch setups.” In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), pp. 164-168, 2019.
[8] Walter J Scheirer, Lalit P Jain, and Terrance E Boult, “Probability models for open set recognition,” IEEE transactions on pattern analysis and machine intelligence, 36(11):2317-2324, 2014.
[9] Abhijit Bendale and Terrance E Boult, “Towards open set deep networks,” CVPR, pp. 1563-1572, 2016.
[10] Lalit P Jain, Walter J Scheirer, Terrance E Boult, “Multi-Class Open Set Recognition Using Probability of Inclusion,” ECCV, september, 2016.
[11] L. Shu, H. Xu, and B. Liu, “DOC: Deep open classification of text documents,” EMNLP, 2017
[12] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, Ole Winther, “Autoencoding beyond pixels using a learned similarity metric,” PMLR 48:1558-1566, 2016
[13] C. Wan, T. Probst, L. V. Gool and A. Yao, “Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation,” CVPR, pp. 1196-1205, 2017.
[14] Yu, Xianwen & Zhang, Xiaoning & Cao, Yang & Xia, Min, “VAEGAN: A Collaborative Filtering Framework based on Adversarial Variational Autoencoders,” IJCAI, pp. 4206-4212, 2019.
[15] Warden P, "Speech Commands: A public dataset for single-word speech recognition," 2017.
[16] Marco Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” Advances in Neural Information Processing Systems 26, pp. 2292–2300, 2013.
[17] Aude Genevay, Gabriel Peyré, and Marco Cuturi, “Learning generative models with sinkhorn divergences,” In International Conference on Artificial Intelligence and Statistics, pp. 1608–1617, 2018.
[18] Salimans, T. and Knowles, D. A., “Fixed-form variational posterior approximation through stochastic linear regression,” Bayesian Analysis, pp. 8(4):837–882, 2013.
[19] Kingma, D. P. and Welling, M., “Auto-encoding variational Bayes,” In International Conference on Learning Representations 2014.
[20] Lars Mescheder, Sebastian Nowozin and Andreas Geiger, “Which Training Methods for GANs do actually Converge?,” ICML 2018.
[21] Youssef Mroueh, Tom Sercu, Vaibhava Goel, “McGan: Mean and Covariance Feature Matching GAN,” PMLR 70:2527-2535, 2017.
[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition.” arXiv preprint, pp. arXiv:1512.03385, 2015.
[23] Paszke, Adam and Gross, Sam and Chintala, Soumith and Chanan, Gregory and Yang, Edward and DeVito, Zachary and Lin, Zeming and Desmaison, Alban and Antiga, Luca and Lerer, Adam, "Automatic differentiation in PyTorch,” 2017.
[24] Jean Feydy, Thibault Sejourne, Francois-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyre, “Interpolating between Optimal Transport and MMD using Sinkhorn Divergences,” AISTATS, pp. 2681-2690, 2019.
[25] Tara N. Sainath and Carolina Parada, “Convolutional neural networks for small-footprint keyword spotting,” in Interspeech, pp. 1478–1482, 2015.
[26] Tang, Raphael and Jimmy Lin. “Deep Residual Learning for Small-Footprint Keyword Spotting.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5484-5488, 2018.