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

Teh Kah Kuan & Tran Huy Dat

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

BACKGROUND / MOTIVATION

- Current state-of-the-art deep learning requires huge amount of labeled data that comes with enormous costs.
- Problem is more serious with audio classification amid uncontrolled environment conditions.
- Building a robust audio classification engine with limited training resources is essential.

METHODOLOGY

- The key idea is to train with augmented data.
- Physical augmentation (PA) - vocal tract length variations, speaking rate variations and far-field noisy simulations.
- Wavelet scattering transform (WST)- improve translation invariant and deformation stability of audio spectrogram images
- GAN – deep learning data generation: not stable with few training data.

RESULTS/BENCHMARKING

Method	Accuracy	
	10% (200 samples) training data	25% (450 samples) 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 Experimental results comparing the classification accuracy of the baseline and proposed GAN system with limited training resource.

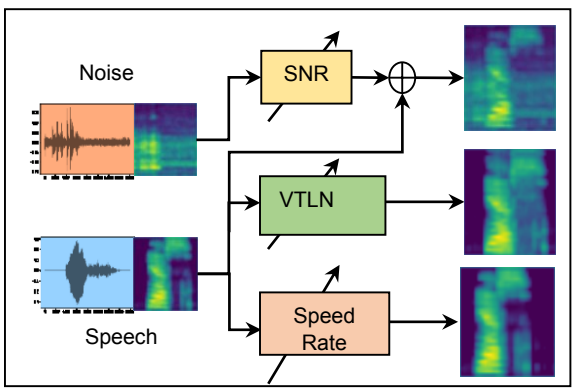

Fig.1 Physical augmentation generation.

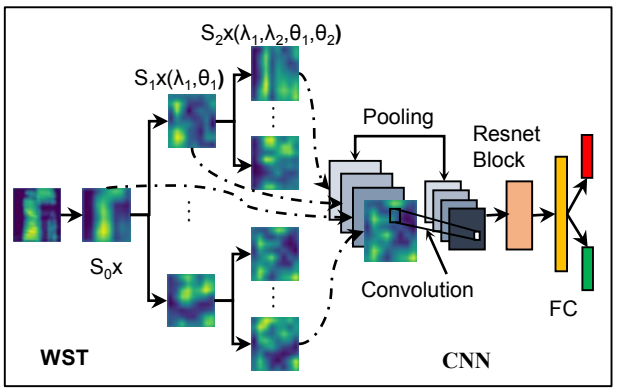

Fig.2 The basic diagram of a 2-layers wavelet scattering transform as a first layer of the CNN classifier.

Approach

- Embed physical modeling and wavelet scattering transform to GAN to improve its stability and generalization.
- Multi-class labels are fed to the generator together with random noises to simulate training samples.
- Both augmented & generated data input to GAN's discriminator.
- GAN is simultaneously optimized by two objective functions (binary-fake/no-fake & multiclass – classification task)
- Improve the discriminator's translation invariance and deformation stability by replacing its first layers by wavelet filters.

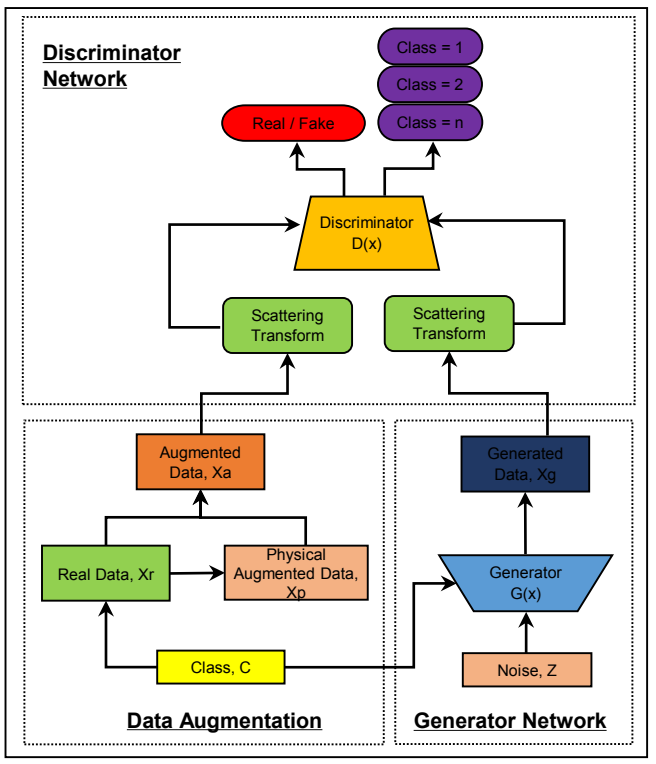

Fig.3 Overview of the proposed GAN integrated with physical augmentation and wavelet scattering transform.

Experiments

- Using Google Speech Commands Dataset
  - 10 core command words were selected
  - Limited training resources (only use 10% or 25% of original, respectively)
  - Full testing set of 2700 samples.
  - No additional unlabeled data.
- Network design
  - Audio samples are transformed into 32x32 normalized Mel-spectrogram.
  - Both generator (G) and discriminator (D) adopt the ResNet-18 architecture.
  - Morlet wavelet filters are used in WST.
- Results & discussions
  - Proposed method significantly improve classification accuracy with limited training resource.
  - Same accuracy level could be achieved with only 10-25% training data.
  - Physical augmentation & wavelet scattering greatly improve the GAN stability and performance.

SUMMARY

- A novel GAN design which embeds physical augmentation and wavelet scattering transform is proposed and it shows very promising results of audio classification with limited training resources.

Institute for Infocomm Research
A*STAR