Residual Audio Neural Networks with Multiple Features for Sound Classification

Verbitskiy Sergey

Novosibirsk State University
*THE REAL SCIENCE

Some subtasks of audio pattern recognition

- Environmental sound and an acoustic scene classification tasks (DCASE, ESC-50, MSoS)
- Musical genre classification tasks (GTZAN)
- Identifying and detecting species of animals (Cornell Birdcall Identification and Rainforest Connection Species Audio Detection on kaggle)
- Emotion recognition tasks (RAVDESS, EmoDB)
- Speaker recognition tasks (VoxCeleb)
- Some applications in medicine. For example, classification of lung diseases using sound recordings which are recorded by electronic stethoscopes (Respiratory Sound Database on kaggle)

* My approaches

- A new architecture of CNN based on WideResNet [11]
- Applying several data augmentation techniques
- Using several 2D audio features as input to CNNs and using different ensemble methods (the weighted average and D-S theory)

* Data augmentation techniques for audio signals

- temporal cropping (“t_c” = the temporal cropping length. Segments duration)
- speed stretching
- pitch shifting
- white noise (Gaussian)
- SpecAugment [1]
- mixup for audio signals [2]

* Residual Audio Neural Network and Optimization

- WideResNet as a based model with basic blocks as in ResNet-v2 [12]
- Changing of stride, kernel and padding sizes (t_m is the temporal decreasing parameter)

For example:

sampling rate is 44100 Hz, duration is 8 sec, hop size is 320 and mel bins is 128 (the best choice for model as a trade-off between computational complexity and system performance), then input tensor shape is:

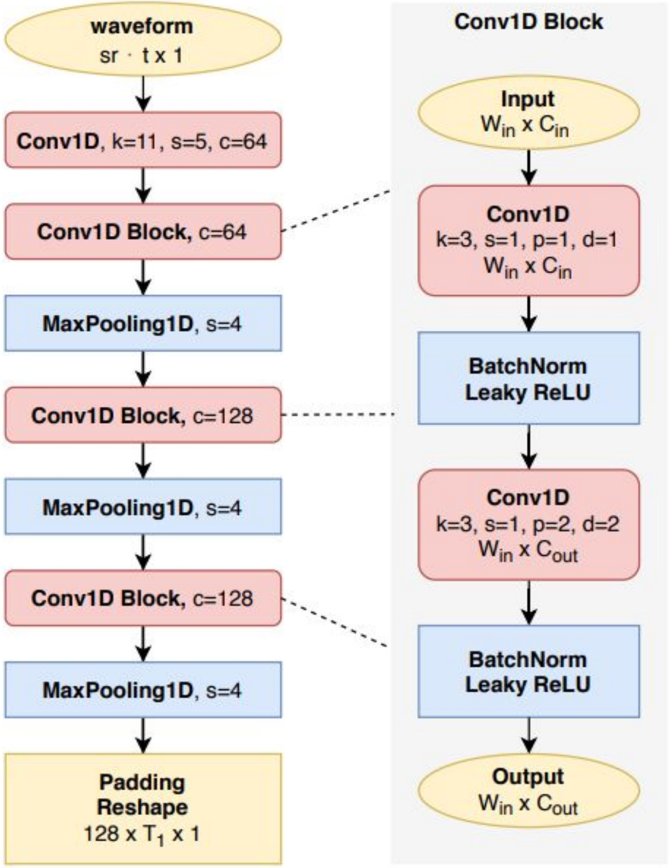
(128, [44100 * 8 / 320] + 1) = (128, 1103). The width of input tensor to model is about 9 times higher than height!

- Leaky ReLU with 0.01 [13]
- Adam optimizer [14], One Cyclic Learning Rate Scheduler [15] and EMA of model parameters [16]

* Features

- Log Mel Spectrogram [3]
- Mel-Frequency Cepstral Coefficients (MFCC) [4]
- Gammatone Frequency Cepstral Coefficients (GFCC) [5]
- Chromagram [6]
- Constant-Q Transform (CQT) [7]
- Tempogram [8]
- Wavegram [3]

Wavegram



sr - sampling rate

t - duration (sec)

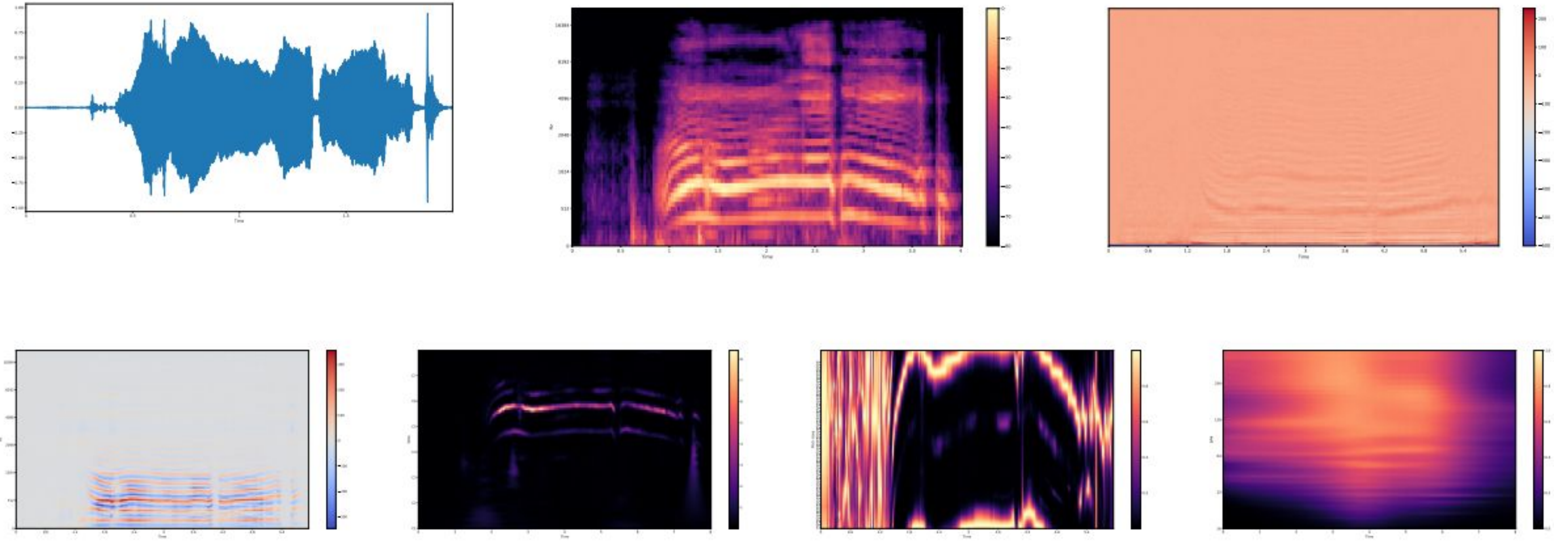
k - kernel size

s - stride size

p - padding size

d - dilation

T_1 - the number of temporal frames



CPU:

- librosa
- spafe
- pywavelets

GPU, PyTorch:

- torch.fft
- torchlibrosa
- torchaudio

How to combine models with different features?

- weighted average method
- D-S Evidence Fusion method [9]
- Improved D-S Evidence Fusion method [9]

$$m(\emptyset) = 0$$

$$0 \leq m(A) \leq 1, \ \ \forall A \subset \Theta$$

$$\sum_{A \subset \Theta} m(A) = 1$$

$$(m_1 \oplus ... \oplus m_n)(x \in A) = \frac{1}{1-k} \prod_{i=1}^{n} m_i(x \in A)$$

$$k = 1 - \sum_{A \subset \Theta} \prod_{i=1}^{n} m_i(x \in A)$$

$$m_a(x) = \sum_{i=1}^{n} w_i \cdot m_i(x)$$

$$\sum_{i=1}^{n} w_i = 1$$

$$m_a(x \in A) = (m_1 \oplus ... \oplus m_n)(x \in A)$$

$$pred_a(x) = \max_{A} m_a(x \in A)$$

Results, Audioset

Comparison of the computational complexity and the performance of models with different hyper-parameters (only Mel Spectrogram)

TABLE VII

COMPARISON OF RANNS WITH DIFFERENT VALUES FOR PAIRS OF t_c AND t_m FOR THE AUDIOSET TAGGING

System	$F \times T$	mAP	mAUC
RANN-4x4-6	8×8	0.407	0.974
RANN-8x1-6	8×64	0.428	0.974
RANN-8x2-6	8×32	0.435	0.975
RANN-8x4-6	8×16	**0.443**	**0.975**
RANN-8x8-6	8×8	0.432	0.974

TABLE XII

COMPARISON OF THE COMPUTATIONAL COMPLEXITY AND THE PERFORMANCE OF DIFFERENT SYSTEMS

System	mAP	Parameters	Multi-Adds
CNN14 [3]	0.431	80,753,615	42.220×10^9
ResNet38 [3]	0.434	73,783,247	48.962×10^9
Wavegram-Logmel-CNN [3]	0.439	81,065,487	53.510×10^9
RANN-4x4-6	0.407	54,919,313	23.569×10^9
RANN-8x1-6	0.428	54,435,473	101.231×10^9
RANN-8x2-6	0.435	54,532,241	61.745×10^9
RANN-8x4-6	**0.443**	**54,919,313**	**47.137**$\times 10^9$
RANN-8x8-6	0.432	56,467,601	42.399×10^9
RANN-8x4-5	0.424	38,198,545	32.743×10^9
RANN-8x4-4	0.410	24,504,849	20.964×10^9

The previous best score: **0.439** [3]

My best score (from scratch): **0.443**

Results, ESC-50

Comparison of the performance of models with different audio features

	Accuracy	mAP	F1
Mel Spectrogram	**0.878**	**0.945**	**0.876**
MFCC	0.853	0.916	0.850
CQT	0.823	0.889	0.819
GFCC	0.813	0.896	0.809
Wavegram	0.813	0.889	0.806
Chromagram	0.708	0.720	0.707
Tempogram	0.455	0.467	0.453

Mel Spect	MFCC	CQT	GFCC	Wave	Acc. WA	Acc. (D-S)	Acc. (ID-S)
✓	✓				**0.896**	**0.890**	**0.887**
✓		✓			0.886	0.882	0.883
✓			✓		0.892	0.881	0.882
✓				✓	0.892	0.882	0.886
	✓	✓			0.881	0.879	0.880
	✓		✓		0.875	0.873	0.874
	✓			✓	0.872	0.870	0.870
		✓	✓		0.863	0.866	0.860
		✓		✓	0.856	0.869	0.856
			✓	✓	0.851	0.849	0.850
✓	✓	✓			0.904	0.897	0.897
✓	✓		✓		0.901	0.894	0.895
✓	✓			✓	**0.904**	**0.897**	**0.899**
✓		✓	✓		0.897	0.890	0.886
✓		✓		✓	0.900	0.893	0.895
✓			✓	✓	0.898	0.886	0.883
	✓	✓	✓		0.891	0.885	0.885
	✓	✓		✓	0.891	0.889	0.887
	✓		✓	✓	0.887	0.883	0.884
		✓	✓	✓	0.876	0.877	0.869
✓	✓	✓	✓		0.900	0.897	0.898
✓		✓	✓	✓	0.898	0.893	0.891
✓	✓		✓	✓	0.901	0.895	0.895
✓	✓	✓		✓	**0.910**	**0.904**	**0.904**
	✓	✓	✓	✓	0.891	0.891	0.890
✓	✓	✓	✓	✓	**0.910**	**0.901**	**0.898**

The previous best score (from scratch, 2020): **0.89** [10]

My best score (from scratch): **0.91**

Results, RAVDESS

Comparison of the performance of models with different audio features

System	Accuracy
Mel Spectrogram	0.739
Mel Spectrogram + MFCC + CQT, WA	0.772
Mel Spectrogram + MFCC + CQT, scratch,(our), D-S	**0.768**
Mel Spectrogram + MFCC + CQT + GFCC + Wave, scratch, (our), WA	**0.774**

The previous best score (fine-tune, 2020): **0.721** [3]

My best score (from scratch): **0.774** (the weighted average) and **0.768** (D-S theory)

* Future work....

- Training models with another features as input and using ensembles methods for the AudioSet dataset. Already an mAP of **0.443** have been achieved with only one model with Mel Spectrogram as input. I expect an mAP of ~ 0.5 with several features... Previous best mAP on the AudioSet dataset is 0.439 [3].
- Transfer system pretrained on AudioSet to other task and achieve best score on DCASE 2019 Task 1A, DCASE 2020 Task 1A, MSoS, ESC-50 (previous best accuracy with fine-tuning is 0.945 [3], i expect ~ 0.97)
- Kaggle Competition (Rainforest Connection Species Audio Detection)
- Paper publication (~february of 2021) and participating in INTERSPEECH 2021 and in DCASE 2021.

References

[1]: D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *in INTERSPEECH*, 2019.

[2]: Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017

[3]: Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880-2894, 2020.

[4]: S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, August 1980.

[5]: S. Chachada, and C. . J. Kuo, "Environmental sound recognition: A survey," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9, 2013.

[6]: Ayush Shah, Manasi Kattel, Araju Nepal and D. Shrestha, "Chroma Feature Extraction,', 2019.

[7]: Christian Schorkhuber, "Constant-q transform toolbox for music process," in *The 7th Sound and Music Computing Conference*, Barcelona, Spain, 2010.

[8]: P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram—A mid-level tempo representation for musics ignals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, pp. 5522-552.

[9]: Shaobo Li, Guokai Liu, Xianghong Tang, Jianguang Lu, and Jianjun Hu, "An Ensemble Deep Convolutional Neural Network Model with Improved D-S Evidence Fusion for Bearing Fault Diagnosis," in *Sensors*, 17(8), 2017.

References

[10]: Jinbae Park, Teerath Kumar, and Sung-Ho Bae, "Search of an Optimal Sound Augmentation Policy for Environmental Sound Classification with Deep Neural Networks," in *Proceedings of the Korean Society of Broadcast Engineers Conference*, 2020.
[11]: Sergey Zagoruyko, and Nikos Komodakis, "Wide Residual Networks," *arXiv preprint arXiv:1605.07146*, 2016.
[12]: Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity Mappings in Deep Residual Networks," *arXiv preprint arXiv:1603.05027*, 2016.
[13]: Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR, abs/1505.00853*, 2015.
[14]: D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[15]: Leslie N. Smith, Nicholay Topin, "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates, *arXiv preprint* arXiv:1708.07120, 2017
[16]: David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel, "MixMatch: A Holistic Approach to Semi-Supervised Learning," *arXiv preprint arXiv:1905.02249*, 2019.

Novosibirsk State University
*THE REAL SCIENCE

RUSSIA, 630090, NOVOSIBIRSK, PIROGOVA STR., 2

nsuniversity.official

nsu24

@nsuniversity

WWW.NSU.RU/N/