Audio Classification using Spectrograms with transfer learning

Mayank Gulati: gulati@kth.se

Federico Favia: favia@kth.se

2020.01.13

Problem Statement

Widespread use of audio classification for focusing on speech recognition, crime detection, music production.
Exploit image classification CNNs architectures to analyze spectrograms of NSynth Audio dataset.
Research the use of transfer learning with state-of-the-art infrastructures along with custom layers in the end of neural networks.
Our focus is to use deep CNN for classification of 10 musical instruments.

Background and related work

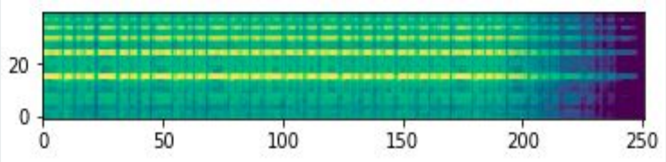
- Based on idea of CNN for audio classification.
- Until now not satisfied results as deep learning with visual data.
- Need to take in account serial and temporal dependence of audio when feeding spectrograms in CNN.

Background and related work

- Paradigm shift in computer vision from hand-engineered features to CNNs.
- ImageNet project disrupting in visual object recognition.
- Two strategies in transfer learning approach: (1) fine-tuning transferred layers.
  (2) **Keep frozen transferred layers**.

Methods

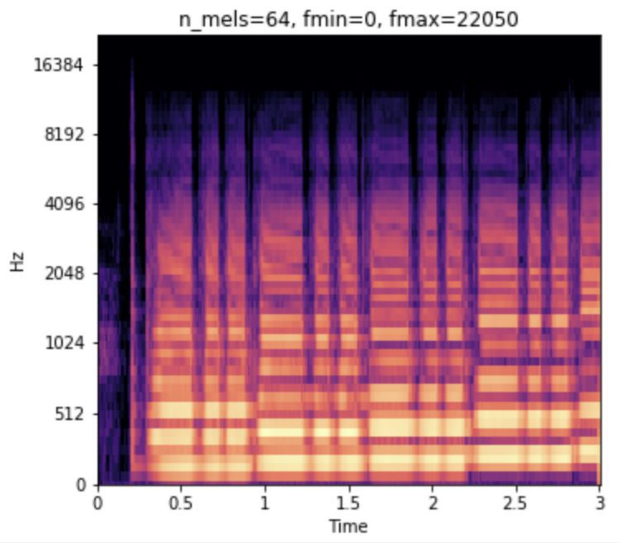
- From NSynth dataset build *mel* spectrograms dataset using *librosa*.
- Spectrograms: frequency content in the audio over time.
- Deep neural networks for recognize critical parts and classify different musical instruments after training.



frequencies (on a Mel frequency scale) vs time

Framework

- Audio files in *.wav* format with single channel
- Every audio file has associate sample rate
- Apply N-point Fast Fourier Transform on small overlapping chunks to move from time domain to frequency domain (STFT): usually N = 256, 512
- *Mel* spectrogram because of human perception



Architecture

- Using different architecture: adapted AlexNet (approximate 62M trainable parameters), adapted GoogLeNet (over 6M trainable parameters) and adapted ResNet18 (over 11M learnable parameters).

- Accuracy evaluation using PyTorch native performance metric conducted over batch size of 64 examples on Microsoft Azure's GPU (Tesla K80) for 500 epochs.

Results and Analysis

Predicted Actual	bass	brass	flute	guitar	keyboard	mallet	organ	reed	string	vocal	All
bass	807	19	42	20	9	8	8	15	47	25	1000
brass	20	886	0	12	0	5	13	4	10	50	1000
flute	40	5	675	52	48	67	77	26	8	2	1000
guitar	6	6	38	665	32	126	81	24	10	12	1000
keyboard	15	3	46	25	751	26	76	49	6	3	1000
mallet	3	2	33	73	26	796	34	26	5	2	1000
organ	5	2	18	27	14	14	918	2	0	0	1000
reed	8	4	28	37	34	52	16	814	0	7	1000
string	49	33	9	15	12	7	2	2	850	21	1000
vocal	20	49	3	19	5	4	13	17	14	856	1000
All	973	1009	892	945	931	1105	1238	979	950	978	10000

Performance metric : Accuracy (cumulative correct matches for all classes by total no. of examples)

Conclusion and Future work

- Our idea of spectrogram image based audio classification has potential application. Satisfied with baseline result.

Sr. No.	Model architecture	Accuracy	No. of Parameters
1	AlexNet	0.723	61.8M
2	GoogLeNet	0.762	6.13M
3	ResNet18	0.814	11.24M

- Future work:
  (1) possible dataset augmentation.
  (2) detection of musical notes from any instrument.

Questions?

References

[1] S. Subha and P. Kannan, “Speaker identification techniques – a survey,” International Journal of Advanced and Innovative Research (2278-7844), vol. 4 Issue 10, 2015. [Online]. Available: https://www.academia.edu/18551460/Speakeridentification-Asurvey

[2] V. Passricha and R. K. Aggarwal, “Convolutional neural networks for raw speech recognition,” in From Natural to Artificial Intelligence, R. Lopez-Ruiz, Ed. Rijeka: IntechOpen, 2018, ch. 2. [Online]. Available: https://doi.org/10.5772/intechopen.80026

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 , ser. NIPS’12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” CoRR , vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[6] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, “Neural audio synthesis of musical notes with wavenet autoencoders,” 2017.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in CVPR09, 2009.

References

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015. doi: 10.1007/s11263-015-0816-y

[9] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, “CNN architectures for large-scale audio classification,”

CoRR, vol. abs/1609.09430, 2016. [Online]. Available: http://arxiv.org/abs/1609.09430

[10] J. B. Allen, “How do humans process and recognize speech?” IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 567–577, Oct. 1994.

[11] L. Wyse, “Audio spectrogram representations for processing with convolutional neural networks,” CoRR, vol. abs/1706.09559, 2017. [Online]. Available: http://arxiv.org/abs/1706.09559

[12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3320–3328. [Online]. Available:http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf

[13] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” CoRR, vol. abs/1310.1531, 2013. [Online]. Available: http://arxiv.org/abs/1310.1531

References

[14] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” CoRR, vol. abs/1311.2901, 2013. [Online]. Available: http://arxiv.org/abs/1311.2901

[15] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” 2013.

[16] P. Herrera-Boyer, G. Peeters, and S. Dubnov, “Automatic classification of musical instrument sounds,” Journal of New Music Research, vol. 32, no. 1, pp. 3–21, 2003. doi: 10.1076/jnmr.32.1.3.16798. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1076/jnmr.32.1.3.16798

[17] B. McFee, M. McVicar, S. Balke, V. Lostanlen, C. Thomé, C. Raffel, D. Lee, Kyungyun Lee, O. Nieto, F. Zalkow, D. Ellis, E. Battenberg, R. Yamamoto, J. Moore, Ziyao Wei, R. Bittner, Keunwoo Choi, Nullmightybofo, P. Friesch, Fabian-Robert Stöter, , Thassilo, M. Vollrath, Siddhartha Kumar Golu, Nehz, S. Waloschek, , Seth, R. Naktinis, D. Repetto, C. Hawthorne, and CJ Carr, “librosa/librosa: 0.6.3,” 2019. [Online]. Available: https://zenodo.org/record/2564164

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017

Fourier Formulas

- Audio files in *.wav* format with single channel
- Every audio file has associate sample rate
- Apply N-point Fast Fourier Transform on small overlapping chunks to move from time domain to frequency domain (STFT): usually N = 256, 512
- *Mel* spectrogram because of human perception

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn}$$

$$P = \frac{|FFT(x_i)|^2}{N}$$

$$m = \frac{1000}{log2} log(1 + \frac{f}{1000})$$