Project Report: Fast and Memory-Efficient Coarse-Grained Audio Classification using Deep Neural Network

Md Shamim Hussain
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY, USA
hussam4@rpi.edu

Nafis Neehal
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY, USA
neehan@cs.rpi.edu

Abstract

Deep neural networks are notorious for their high computational cost and memory requirements. Not only that, often they need specialized hardware for efficient execution. However, for simple tasks such as a coarse-grained classification of sound into superclasses such as speech, music, and noise, one may not be able to spare a lot of computational resources. In this work, we devise a new convolutional network - SwishNetV2 which is fast and memory-efficient, but gets close to big convolutional networks such as VGG16 in terms of classification performance. This architecture improves upon the previous SwishNet architecture and performs well on the CPU with limited memory requirements, similar to its predecessor. We verified the performance of this model on the diverse Audioset dataset, with manually verified gold standard labels. We also devised a method for including data with unchecked noisy labels from AudioSet which improves classification performance. Our annotated dataset is available publicly at `https://www.kaggle.com/snirjhar/audioset-speech-music-noise`. The code for our implimentation can be found at `https://github.com/shamim-hussain/audioset_coarse_grained_classification`.

1 Introduction

Music, speech and noise classification and segmentation is an important task because these three types of signals are inherently different in nature [1] and require different types of processing and/or coding schemes. For example, a good compression scheme for speech may not be good for compressing music. Also, the signal processing steps used for these two types of media are likely to be very different. So, correctly identifying the type of media before further processing should be considered a very important task.

With the advent of widespread use of the internet, it is now possible to build large media databases from user-contributed data. However, labels for the collected data are rarely available. Manual media labeling can be both expensive and time-consuming. So, a reliable method is required to automate the indexing of large media databases. It is intuitive to first classify/ segment the media into broad categories such as speech, music and noise which may be followed by further fine-grained classification/segmentation procedure (for example, into different musical genres).

The above discussion suggests that it is desirable to bring the classification/segmentation of media into speech, music and noise under a single framework. In recent years deep learning [2] algorithms have achieved unprecedented success in numerous classification problems, without any need for

Preprint. Under review.

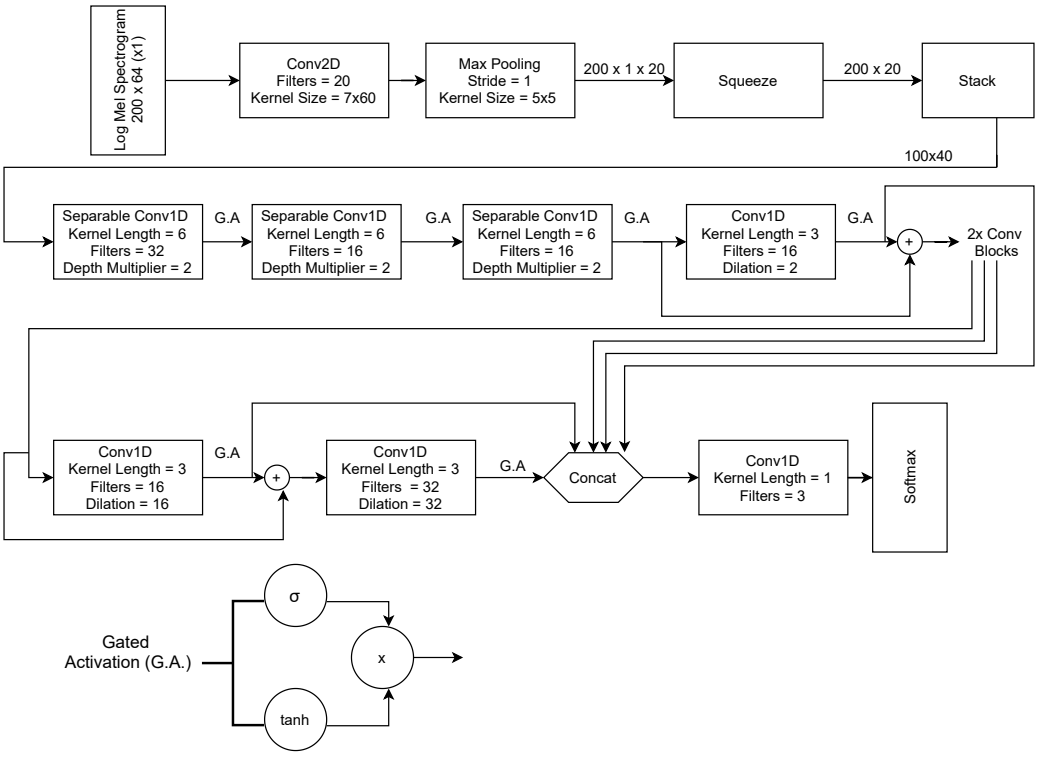

Figure 1: Network Architecture of SwishNetV2

careful feature selection. Especially, deep convolutional neural networks have set new frontiers in many fields such as image and audio classification and segmentation. However, deep neural networks are generally known to be more computationally expensive and slower than other more conventional models. These models often need large-scale parallelization in GPUs for faster implementation. Since classification/segmentation is often set as a pre-processing step in an audio processing pipeline, it needs to be fast enough to prevent the delay in further processing. Also, the computational resources are limited in many scenarios such as in mobile devices and embedded systems where it is unreasonable to waste too much resource in the preprocessing step. This is where we put forth our contribution.

Recently, it has been shown that for coarse grained audio classification tasks a fast and lean 1D CNN applied on derived features can show performance close to that of a 2D CNN - the SwishNet architecture [3]. Although, this architecture has shown good performance on a clean and less diverse dataset such as the MUSAN [4] dataset, it's performance on a more difficult and diversified realistic setting has not been verified. In this work, we further explored the use of lightweight CNN architecutres for audio classification task but in a more diverse and realistic setting - on the AudioSet dataset [5], which is a dataset of 2.1 million annotated audio clips derived from youtube videos. The AudioSet ontology contains 527 different classes, but it does not directly annotate clips into suprerclasses such as speech, music and noise. Instead, it labels each clip with a set of subclasses. We defined these superclasses in the ontology which gives us plausible superclass labels for the clips. However, due to mistakes/missing annotations the actual superclass maybe wrong/ambiguous. So we, manually annotated a set of gold standard labels for around 12000 clips. We devised a new network architecture SwishNetV2 which is an improvement on the SwishNet architecture and directly applies on log mel spectrogram. We verified the performance of this model on the AudioSet dataset with manually verified gold standard labels. We also developed a method for including data with unchecked noisy labels from AudioSet which improves classification performance.

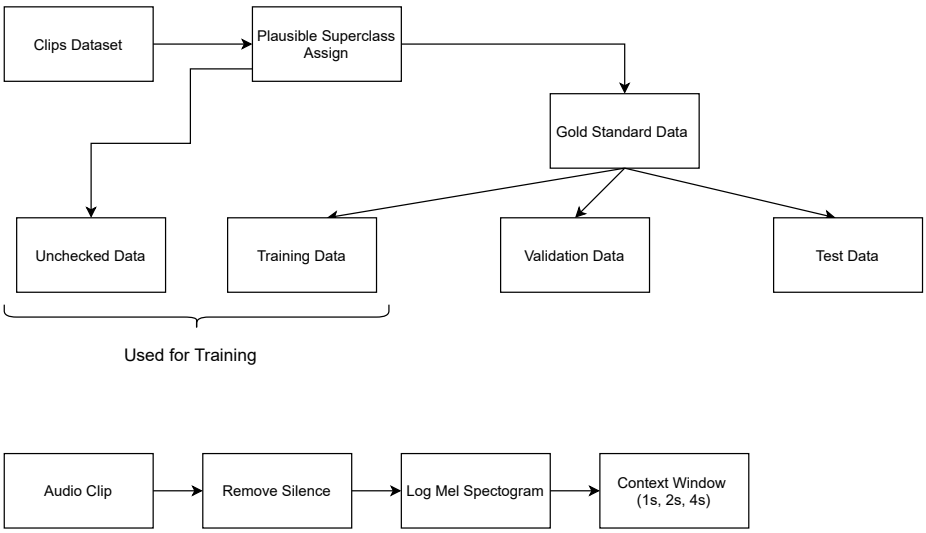

Figure 2: Network Architecture of SwishNet-slim

2 Related Works

In general, audio content analysis in video parsing can be considered in two directions [6, 7]. One is to discriminate audio streams into different classes such as speech, music, noise etc., the other is to classify audio streams into segments of different speakers. In this paper, our research work in the first direction will be presented.

There have been many studies on audio content analysis, using different features and different methods [8] - Pfeiffer et al [9], presented a theoretic framework and application of automatic audio content analysis using some perceptual features. Saunders [10], presented a speech/music classifier based on simple features such as zero crossing rate and short time energy for radio broadcast. When a window size of 2.4s was used, the reported accuracy rate would be 98%. Scheirer et al [11] introduced many more features into audio classification and performed experiments with different classification models including GMM (Gaussian Mixture Model), BP-ANN (Back Propagation Artificial Neural Network) and KNN (K-Nearest Neighbor). When using a window of the same size (2.4s), the reported error rate would be 1.4%. However, it is found that such simple feature-based methods cannot work well when a smaller window is used or more audio classes such as environment sounds are taken into consideration.

Many other works have been done to enhance audio classification algorithms. In [12], audio recordings are classified into speech, silence, laughter and non-speech sounds, to segment discussion recordings in meetings. The accuracy of the segmentation result using his method varies considerably for different types of recording. In the work by Zhang and Kuo [13], pitch tracking methods are introduced to discriminate audio recordings into more classes, such as songs, speeches over music, with a heuristic-based model. Accuracy of above 90% is reported. Srinivasan et al [5], try to detect and classify audio that consists of mixed classes, such as combinations of speech and music together with background sound. The accuracy of classification is over 80%.

3 Methodology

3.1 Data Collection

The Audioset [5] ontology was used to identify fine-grained sound event classes (source). We combined the fine-grained event classes into 3 major superclasses - speech, music and noise. We define each superclass uniquely by saying that a particular audio segment must not contain events from other superclasses, so that there is no ambiguity in classification.

Table 1: Network Sizes and Prediction Times per Sample

Network	No. of Parameters	Prediction Time (ms) per Sample for Different Sample Lengths						Weight file size
		1.0s		2.0s		4.0s		
		CPU	GPU	CPU	GPU	CPU	GPU	
SwishNet V2	15,127	2.15	1.85	2.52	1.87	2.86	1.87	59KB
VGG16	14.75M	31.4	2.92	51.6	3.97	89.7	5.88	56MB

Table 2: Overall Classification Accuracy (%) for Clips of Different Lengths. (GST = Trained on Gold Standard Training Set, GST + UT = Trained on Gold Standard and Unchecked Training Set, GST + UT + FT = Trained on Gold Standard and Unchecked Training Set, Fine Tuned on Gold Standard Training Set)

Network	1.0s	2.0s	4.0s
SwishNet (GST)	89.69%	91.50%	91.09%
SwishNet (GST + UT)	91.90%	93.42%	92.85%
SwishNet (GST + UT + FT)	92.45%	94.41%	94.54%
VGG16 (GST + UT + FT)	94.72%	96.01%	96.57%

We downloaded the annotations of the YouTube video segments. Each segment is assigned multiple sound event tags. We defined plausible superclasses, based on these tags. If a clip contains only the subclasses of a given superclass (speech/music/noise) it is said to belong to that plausible superclass. If the superclass is ambiguous, we ignore that clip. But due to mistakes in annotations, missing annotations and human errors, the actual superclass may still be ambiguous or even wrong. So, the labels assigned at this point are noisy.

Next, we downloaded YouTube clips in uncompressed wav format which was later compressed by the free lossless audio codec (FLAC). Clips were downloaded from the evaluation set, the balanced training set and the unbalanced training set. We downloaded about 83,600, 10s clips from YouTube.

To generate ground truth (gold standard) annotations for the clips we manually checked each clip and verified their plausible superclass. If we found that a clip's superclass was ambiguous/wrong, it was tagged as "bad", otherwise it was tagged as "good". We performed these manual annotations until we reached 4,000 "good" clips for each superclass (speech, music, and noise). Thus, the gold standard ("good") dataset contains 12,000 annotated unambiguous ground truth labeled clips. Also, the ambiguous ("bad") labels are also kept for possible future use in automated label clean up. Our annotated dataset, along with unchecked clips is available publicly at `https://www.kaggle.com/snirjhar/audioset-speech-music-noise`.

3.2 Feature Extraction

3.3 Network Architecture

4 Results and Discussions

5 Conclusion and Future Work

References

[1] Joe Wolfe. "Speech and music, acoustics and coding, and what music might be 'for'". In: *Proc. 7th International Conference on Music Perception and Cognition*. 2002, pp. 10–13.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), p. 436. ISSN: 1476-4687.

Table 3: Normalized Confusion Matrices for Clips of Different Lengths (Rows: True Labels, Columns: Predicted labels, Ordering: Speech, Music, and Noise. GST = Trained on Gold Standard Training Set, GST + UT = Trained on Gold Standard and Unchecked Training Set, GST + UT + FT = Trained on Gold Standard and Unchecked Training Set, Fine Tuned on Gold Standard Training Set)

Network	1.0s	2.0s	4.0s
SwishNet(GST)	$\begin{bmatrix} .90 & .05 & .05 \\ .02 & .92 & .06 \\ .02 & .11 & .86 \end{bmatrix}$	$\begin{bmatrix} .93 & .03 & .03 \\ .01 & .94 & .05 \\ .02 & .11 & .87 \end{bmatrix}$	$\begin{bmatrix} .93 & .03 & .03 \\ .01 & .94 & .05 \\ .03 & .11 & .86 \end{bmatrix}$
SwishNet(GST + UT)	$\begin{bmatrix} .97 & .01 & .02 \\ .03 & .91 & .06 \\ .06 & .05 & .89 \end{bmatrix}$	$\begin{bmatrix} .98 & .01 & .01 \\ .01 & .95 & .04 \\ .06 & .06 & 0.88 \end{bmatrix}$	$\begin{bmatrix} .96 & .03 & .01 \\ .00 & .97 & .03 \\ .06 & .08 & .86 \end{bmatrix}$
SwishNet(GST + UT + FT)	$\begin{bmatrix} .91 & .03 & .05 \\ .01 & .92 & .07 \\ .01 & .05 & .93 \end{bmatrix}$	$\begin{bmatrix} .96 & .01 & .03 \\ .01 & .95 & .03 \\ .02 & .06 & .92 \end{bmatrix}$	$\begin{bmatrix} .95 & .01 & .04 \\ .01 & .95 & .04 \\ .01 & .05 & .94 \end{bmatrix}$
VGG16 (GST + UT + FT)	$\begin{bmatrix} .94 & .01 & .05 \\ .01 & .94 & .05 \\ .01 & .03 & .95 \end{bmatrix}$	$\begin{bmatrix} .97 & .01 & .02 \\ .00 & .98 & .02 \\ .02 & .05 & .93 \end{bmatrix}$	$\begin{bmatrix} .98 & .00 & .02 \\ .00 & .97 & .03 \\ .01 & .03 & .96 \end{bmatrix}$

[3] Md Hussain, Mohammad Ariful Haque, et al. "Swishnet: A fast convolutional neural network for speech, music and noise classification and segmentation". In: *arXiv preprint arXiv:1812.00149* (2018).

[4] David Snyder, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus". In: *arXiv preprint arXiv:1510.08484* (2015).

[5] Jort F Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[6] Savitha Srinivasan, Dragutin Petkovic, and Dulce Ponceleon. "Towards robust features for classifying audio in the CueVideo system". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. 1999, pp. 393–400.

[7] Zhu Liu, Yao Wang, and Tsuhan Chen. "Audio feature extraction and analysis for scene segmentation and classification". In: *Journal of VLSI signal processing systems for signal, image and video technology* 20.1 (1998), pp. 61–79.

[8] Lie Lu, Hao Jiang, and HongJiang Zhang. "A robust audio classification and segmentation method". In: *Proceedings of the ninth ACM international conference on Multimedia*. 2001, pp. 203–211.

[9] Silvia Pfeiffer, Stephan Fischer, and Wolfgang Effelsberg. "Automatic audio content analysis". In: *Proceedings of the fourth ACM international conference on Multimedia*. 1997, pp. 21–30.

[10] John Saunders. "Real-time discrimination of broadcast speech/music". In: *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*. Vol. 2. IEEE. 1996, pp. 993–996.

[11] Eric Scheirer and Malcolm Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator". In: *1997 IEEE international conference on acoustics, speech, and signal processing*. Vol. 2. IEEE. 1997, pp. 1331–1334.

[12] Don Kimber, Lynn Wilcox, et al. "Acoustic segmentation for audio browsers". In: *Computing Science and Statistics* (1997), pp. 295–304.

[13] Tong Zhang and C-C Jay Kuo. "Video content parsing based on combined audio and visual information". In: *Multimedia Storage and Archiving Systems IV*. Vol. 3846. International Society for Optics and Photonics. 1999, pp. 78–89.