Project Progress Report

Project Title: Fast and memory-efficient coarse-grained audio classification/segmentation using a deep neural network

Team Member Names:

1. Md Shamim Hussain (RIN: 661989634)
2. Nafis Neehal (RIN: 661990881)

Data Collection:

1. The Audioset [1] ontology was used to identify fine-grained sound event classes.
   https://research.google.com/audioset/ontology/index.html
2. 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.
3. We downloaded the annotations of the youtube video segments.
   https://research.google.com/audioset/download.html
   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.
4. 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 also the unbalanced training set. We downloaded about 81,000, 10s clips from youtube.
5. 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, it was tagged as "bad", otherwise it was tagged as "good". We performed these manual annotations until we reached 4,000 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 at -
   https://www.kaggle.com/snirjhar/audioset-speech-music-noise

Feature extraction:

We extracted log mel spectrogram features and energy for each clip. Before that, silence portions of the audio clips were removed, since we do not consider silence in our classification task. Our derived features can be found at -
https://www.kaggle.com/snirjhar/audioset-derived-features

The MFCC features used in [2] can be easily derived from these features.

Train-validation-test split:

We randomly split the gold standard dataset into a training and a test set in a stratified manner. 35% of the clips were selected for the test set. Note that, we will use the unchecked clips with noisy labels in the future either for pretraining, or by refining their labels by some means.

We randomly split the gold standard training dataset into 5 equally sized, stratified folds. For now, we assigned only 1 of the folds as the validation set. But in the future, these will be used for 5-fold cross-validation.

Primary model design and results:

We have trained a variant of the SwishNet architecture [2] on the gold standard dataset, taking 2 second long clips. We are getting around 91% accuracy on the validation set, which is lower than typical 2D CNNs 93-95%. But this model is lightweight and faster than 2D CNNs.

Future Direction:

1. We intend to use the unchecked clips with noisy labels, either as a pre-training stage, or by automatically refining their labels by some means.
2. We intend to add clips from other datasets into training. Possible candidates are
   (i) FSD50k [3]
   (ii) MUSAN [4]
3. We intend to try and combine different features, including MFCC, log Mel spectrogram, and also NMF based features.
4. We intend to experiment with different architectures to produce best results, while keeping the framework lightweight and fast.

References:

[1] Gemmeke, Jort F., Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. "Audio set: An ontology and human-labeled dataset for audio events." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776-780. IEEE, 2017.

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

[3] Fonseca, Eduardo, et al. "FSD50k: an open dataset of human-labeled sound events." *arXiv preprint arXiv:2010.00475* (2020).

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