Audio Plugin for SAGE

Enrique Nueve

enriquenueve9@gmail.com

1 Summary

This article documents a plugin example made for SAGE. This particular plugin demonstrates the usage of audio data. To demonstrate the use of a plugin with audio data, a series of models that classify audio were made and then packaged. The design of the models and used datasets were based on the experiments from the paper [PSY20]. To build upon the work of [PSY20], we explore the performance trade-off of using more edge efficient models such as MobileNetV3[How+19] and EfficientNetV2[TL21]. We also compare the computational performance of the different model configurations over an Jetson Xavier NX (an edge device). We also document the process needed to make the constructed models plugin compatible with SAGE. Both links to the code for the models (**INSERT CODE URL FOR MODEL**) and the plugin (**INSERT URL FOR PLUGIN**) are provided.

2 *Rethinking CNN Models for Audio Classification*

To optimize the performance of CNNs, it is common to perform transfer learning as with a model trained on large image datasets. In classifying audio, it is possible to convert audio data to images through spectrograms, making audio classification possible with CNNs. Thus, transfer learning with CNNs trained on large audio datasets such as AudioSet in the form of spectrograms has been tried. Uniquely, in the paper[PSY20], they tried transfer learning for the audio datasets UrbanSounds8k and ESC-50 with CNNs trained on the large image dataset ImageNet and outperformed state-of-art methods. This procedure is much more feasible for implementation given many pre-trained CNN exists for ImageNet, unlike large audio datasets such as AudioSet. Through inspiration from [PSY20], this project replicates the best configuration from the experiment, DenseNet ensemble of five, and compares against efficient models such as MobileNetV3[How+19] and EfficientNetV2[TL21] given our interest in performing audio classification on the edge.

3 Datasets

- UrbanSounds8k: https://urbansounddataset.weebly.com/urbansound8k.html
- ECS-50: https://github.com/karolpiczak/ESC-50

4 Data Pre-Processing

For consistency, we follow the data pre-processing documented from the [PSY20]. Used data augmentation methods from [SB17].

UrbanSounds8k

1. Resample at 22.5 kHz
2. Apply data augment: time-shift (.81,.93,1.07,1.23), pitch-shift one (-2,-1,1,2), pitch-shift two (-3.5,-2.5,2.5,3.5), and background noise (street workers, street traffic, street people, park) with function $z = (1-w)*x + w*y$ where x is audio, y is background noise, and w is a value from a uniform between (.1,.5).
3. Convert audio to Mel-Spectrogram with 128 mel-bins and log scaled. Do for three channels with (window size, hop length) of (25ms,10ms), (50ms,25ms), and (100ms,50ms) on each channel. Input image size of (128,50).
4. Make into tfRecords

ECS-50

1. step 1
2. step 2

5 Experiments

Redo from *Rethinking CNN Models for Audio Classification* test with DenseNet on ImageNet for ensemble of $n = 5$. Do same test but with EfficientNet and MobileNet for ensembles of $n = 1, 3, 5$.

5.1 Accuracy

Compare accuracy from paper and my results.

5.2 Inference Time

Compare inference time on Jetson Xavier NX with batch size of (1,2,4,8,16,32). Take into account data will not be batched in tfRecord file.

6 Plugin

7 Future Work

Audio source seperate!

References

[SB17] Justin Salamon and Juan Pablo Bello. "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification". In: *IEEE Signal Processing Letters* 24.3 (Mar. 2017), pp. 279–283. ISSN: 1558-2361. DOI: `10.1109/lsp.2017.2657381`. URL: `http://dx.doi.org/10.1109/LSP.2017.2657381`.

[How+19] Andrew Howard et al. *Searching for MobileNetV3*. 2019. arXiv: `1905.02244 [cs.CV]`.

[PSY20] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. *Rethinking CNN Models for Audio Classification*. 2020. arXiv: `2007.11154 [cs.CV]`.

[TL21] Mingxing Tan and Quoc V Le. "EfficientNetV2: Smaller Models and Faster Training". In: *arXiv preprint arXiv:2104.00298* (2021).