Troy Sabia

C964

**Task D – Testing, Revision, Optimization**

Initial testing was conducted with the GTZAN dataset, but it was quickly determined that the dataset was too small. A catalog of commercial music files was instead acquired to extract features from.

**Setup**

Music files were transcoded to mono WAV format, 22.050 kHz samplerate. Three 15 second clips were taken from each file and MFCC features extracted using the following parameters:

num of MFCC = 15

FFT bins = 512

Hop length = 128

We trained a CNN on MFCC features using the cached datasets (train: 3**933**, val: **569**, test: 290). Metrics were logged per epoch and final models evaluated on the **test** set only for unbiased comparison.

**Baseline (unoptimized)**

Some assumptions were made for the baseline model. Three layers, doubling of parameters at each layer, Leaky ReLu (slope = -0.2), and dropout of 0.1 at each layer and in the classifier.

**Configuration:**

**Total parameters = 1577**

LR **1e-4**

batch **16**

weight\_decay **1e-4**

**LR Scheduler (step size 5, gamma 0.6)**

Trained 4**0** epochs

**Results:**

**Accuracy 0.907**

F1 **0.905**

AUC 0.960

**Test:**

Accuracy **0.879**

Precision **0.848**

Recall **0.914**

F1 **0.879**

AUC **0.967**

**Confusion matrix (testset, n = 290, threshold = 0.5):**

[[TN **127** FP **23**],

[FN **12** TP **128**]]

**Observations**

**The results are already quite promising, and it seems that meeting the minimum requirement of 90% accuracy should be doable with some tweaking.**

**The learning curves are relatively smooth, but the accuracy curve has a lot of jitter. This suggests that increasing parameter count and regularization may be required.**

**Optimizations**

Dropout increased across later layers, increasing regularization. Overall parameter amount increased. Added a second Linear() layer to the classifier and increased dropout before the final classification neuron.

**Optimized Results**

**Configuration:**

**Total parameters = 25,633**

LR **1e-4**

batch **16**

weight\_decay **1e-4**

**LR Scheduler (step size 5, gamma 0.6)**

Trained 4**0** epochs

**Results:**

**Accuracy 0.928**

F1 **0.926**

AUC 0.984

**Test:**

Accuracy **0.946**

Precision **0.957**

Recall **0.925**

F1 **0.941**

AUC **0.987**

**Confusion matrix (testset, n = 290, threshold = 0.5):**

[[TN **144** FP **6**],

[FN **7** TP **133**]]

**Improvements**

**Accuracy increased by 0.067** absolute on test (from **0.879 to 0.946**)

**AUC increased by 0.02 absolute**, showing better ranking quality

Precision increased, False Positives decreased

Increased parameter count should add to robustness of outcomes

**Business Impact**

The optimized model meets the success criterion (≥**90%** test accuracy) and reduces false approvals, directly supporting usage of the automated algorithmic sorting. We accept the hypothesis that an MFCC-based CNN can achieve the required accuracy on this task.