GCT634-HW2

20244418 임주은

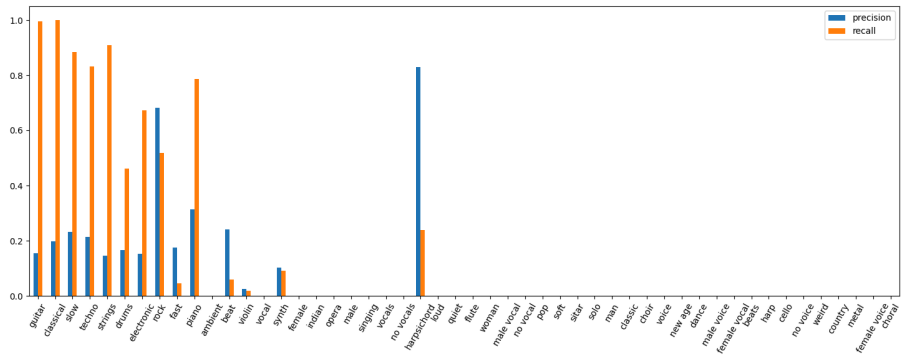
[Table of contents]

1. Data Analysis

Pre-analysis on baseline 1D CNN.

A. Tag-wise Precision and Recall

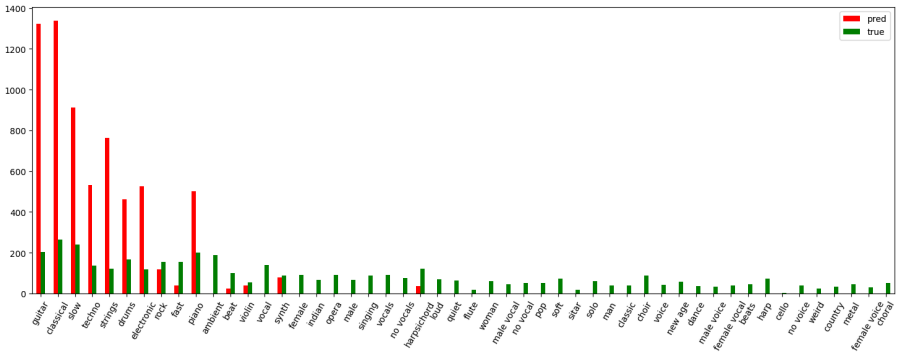
To identify which tags are easier or harder to label, I first plotted the precision and recall values for each tag:



The plot reveals that there are large imbalance of the result of tagging, which means the model tend to produce similar labels and logits regardless of the input.

B. Tag-wise Labeling Count

To verify this observation, I plotted the count of ground-truth and predicted labels for each tag:



This confirms that the tagging results were indeed imbalanced. Furthermore, the model seems to have learned only the bias of the whole dataset, resulting exaggeration of the difference in the distribution, rather than reasoning based on the given audio features itself. The reason could be just due to model underfitting, or label imbalance.

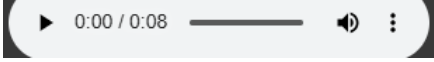
C. Label Inspection

Next, I examined every first sample corresponding to each tag without repetition. Some samples were predicted with contradicting labels simultaneously, for example, ‘slow’ and ‘fast’ appearing together,

```
annotation tag:  ['electronic', 'ambient', 'synth', 'new age']
model predict tags:  ['classical', 'guitar', 'slow', 'piano', 'fast']
```

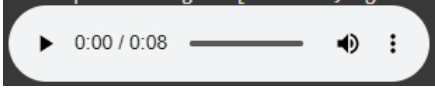
or some genres like ‘classical’ and ‘rock’, which are conceptually far apart, yet coexisting.

```
annotation tag:  ['no vocals', 'loud', 'no vocal']
model predict tags:  ['classical', 'slow', 'guitar', 'electronic', 'rock']
```



Meanwhile some tags had similar semantic meanings, such as, ‘female’ and ‘female vocal’.

```
annotation tag:  ['female', 'female vocal']
model predict tags:  ['techno', 'guitar', 'beat', 'electronic', 'slow']
```



These linguistic relationship among tags are not considered but regarded as just uniform components. Thus, incorporating a loss which ensures those relationships is expected to yield better results.

2. Hypotheses

Based on these observations, hypotheses are formulated as following:

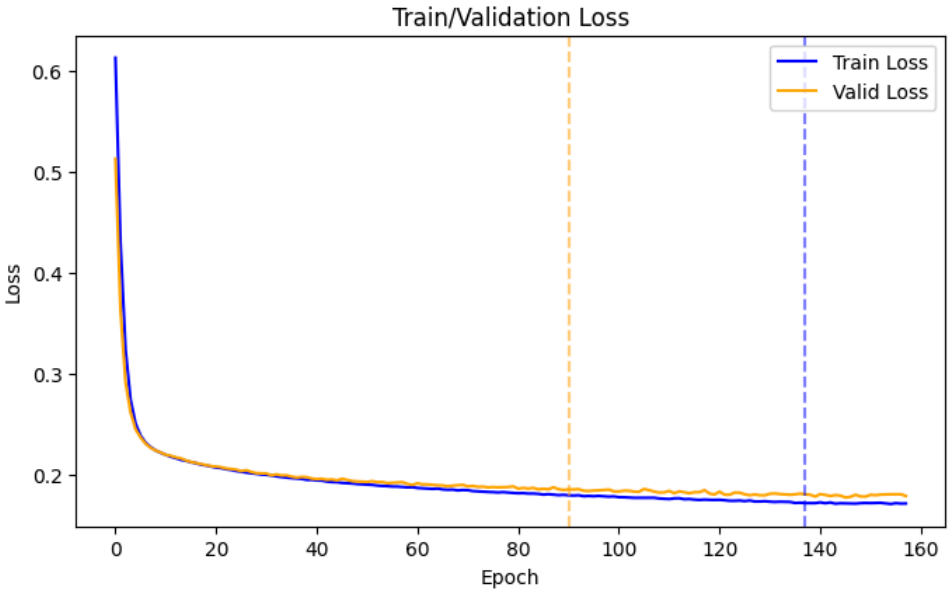
1. The model is underfitted. Increasing training epochs or using higher complexity models will improve performance.
2. The data labels are imbalanced. Using tag-wise uniform sampling or focal loss will increase the performance.
3. The label representations are entangled. Introducing a triplet loss will increase the performance.

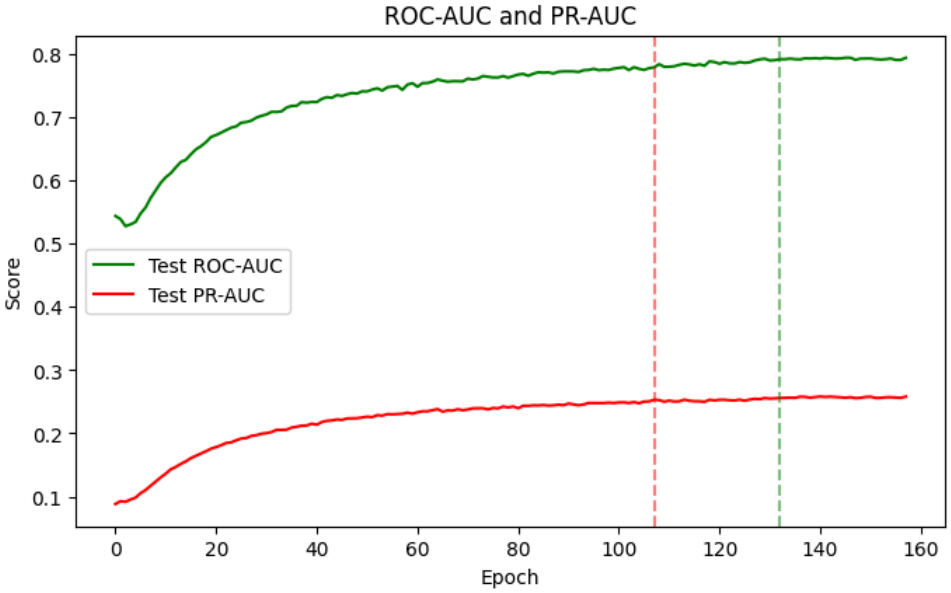
3. Experiments and Results

A. Effect of Training Epochs

The first experiment investigates whether increasing the number of training epochs improves model performance.

The training and validation losses, along with the test AUC values of the baseline model, were tracked across epochs until early stopping was triggered:

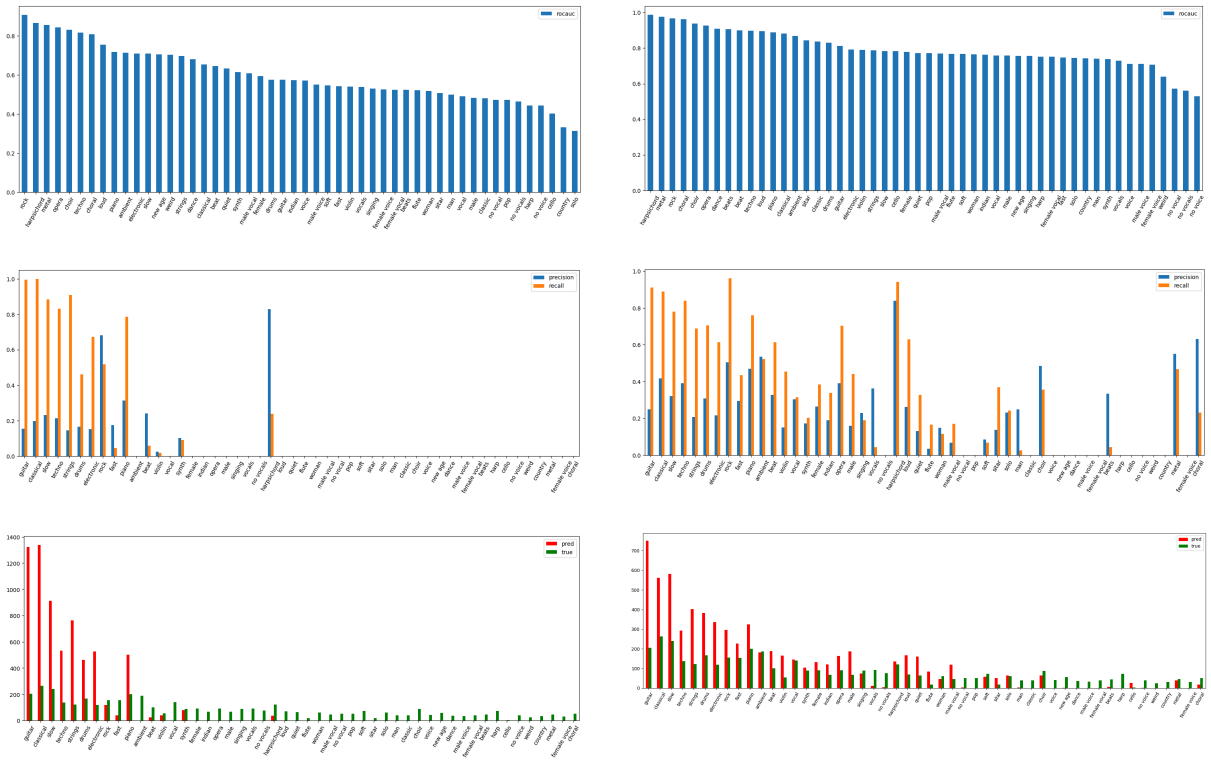




The model was clearly underfitted at epoch 10 and continued to improve until approximately epoch 120, around where the trend of the moving average reversed.

To confirm the performance increasement, compare the tag-wise metric plots:

[Epoch=10] **[Epoch=159]**



All label-wise ROC AUC resulted over 0.5, scores exceeded 0.5, which is the minimum threshold for a model to be considered functionally effective. Furthermore, predictions across labels became more diverse compared to earlier epochs.

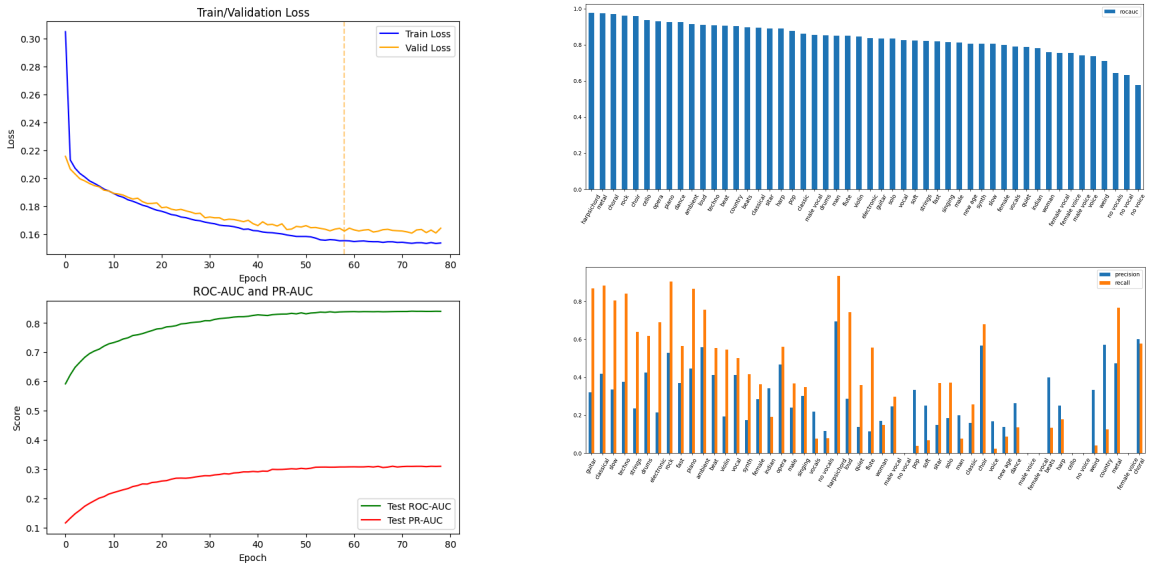
While the performance showed clear improvement, the issue of label imbalance still remained not entirely solved.

B. Model Architecture Comparison

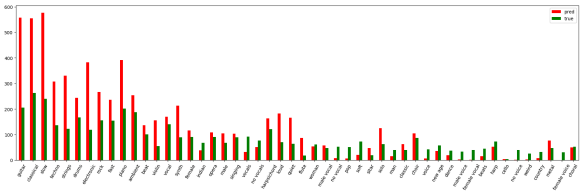
This experiment investigates whether more complex model architectures achieve better performance.

First, compare the baseline 1D CNN model with the implemented 2D CNN model:

[2D CNN]



	ROC AUC	PR AUC
1D CNN (baseline)	0.794	0.258
2D CNN	**0.840**	**0.310**



As shown in the table, the 2D CNN achieves higher ROC AUC and PR AUC scores, suggesting that increased dimensional complexity improves performance. Additionally, the predicted label distribution becomes more balanced across classes.

Next, compare three models from https://github.com/minzwon/sota-music-tagging-models:

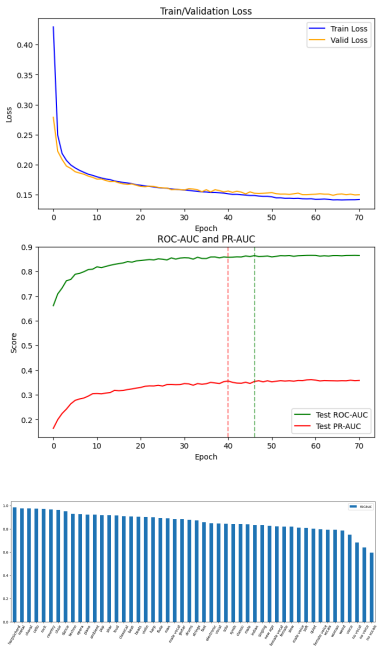
```
!git clone https://github.com/minzwon/sota-music-tagging-models
path = 'sota-music-tagging-models/training'
if path not in sys.path:
  sys.path.append(os.path.abspath(repo_path))

from model import ShortChunkCNN, ShortChunkCNN_Res, CNNSA
```

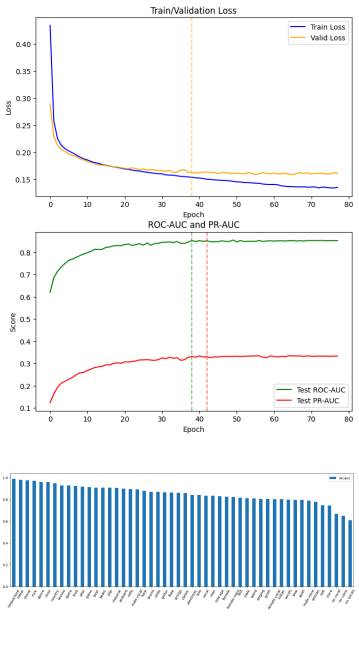
- 1D kernel : Baseline
- 2D kernel : **Short-chunk CNN, Short-chunk CNN + Residual**
- unfixed-kernel : **CNN with Self-Attention (CNNSA)**

The goal of this comparison is to examine how performance varies with the degree of kernel flexibility. In particular, the target is to assess whether loosening kernel constraints leads to further improvements.

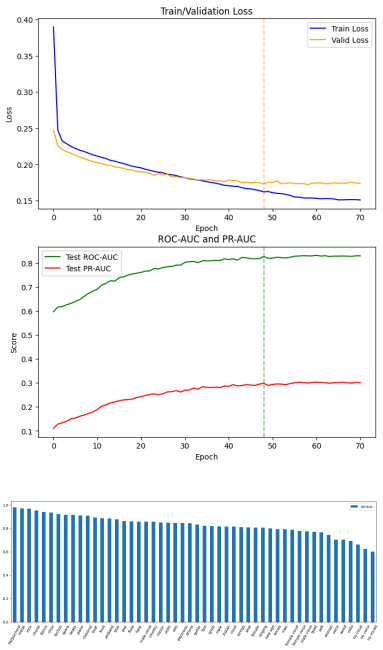
[Short-chunk CNN]

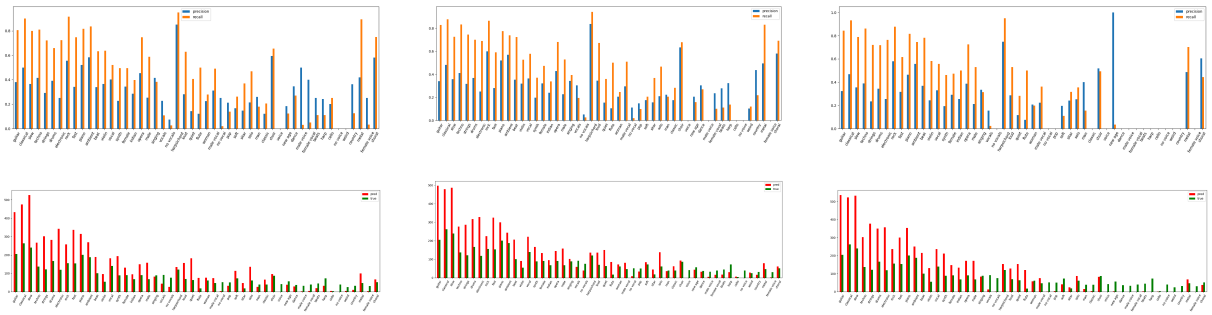


[Short-chunk CNN + Res]



[Self-attention]





	ROC AUC	PR AUC
baseline (1D)	0.794	0.258
Short-chunk CNN	**0.865**	**0.358**
Short-chunk CNN + Residual	0.853	0.333
Self-attention	0.831	0.301

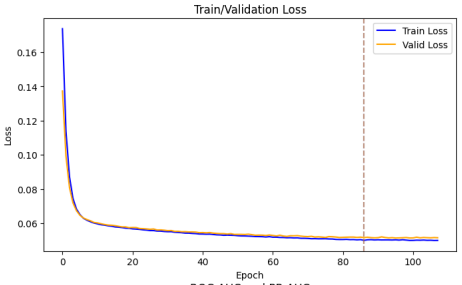
Results suggest that relaxing kernel constraints enhances performance up to a point, but deeper or residual structures may yield diminishing returns.

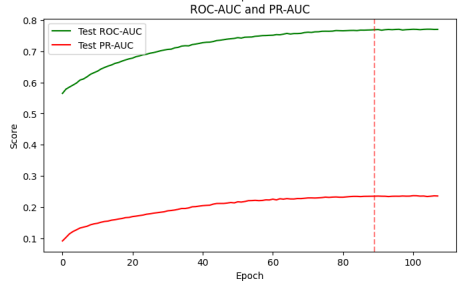
C. Loss Function Comparison

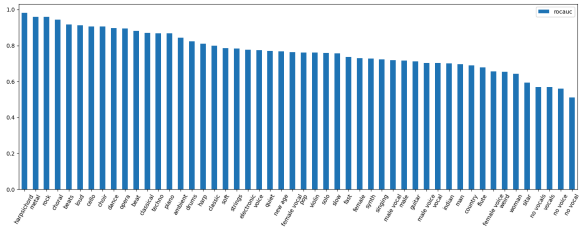
This section explores the impact of using focal loss and triplet loss on model performance.

Focal loss assigns greater weight to harder-to-classify labels, while triplet loss adjusts embedding distances based on label similarity measured by jaccard equation. Triplet loss is added to the baseline loss, acting as an auxiliary component.

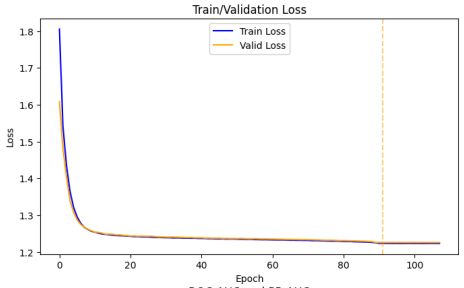
[Focal loss]

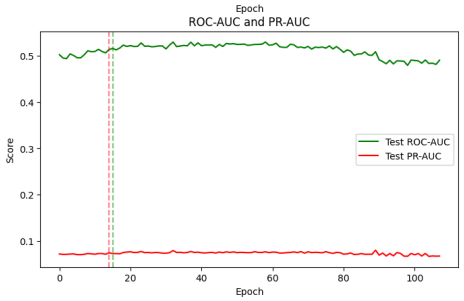


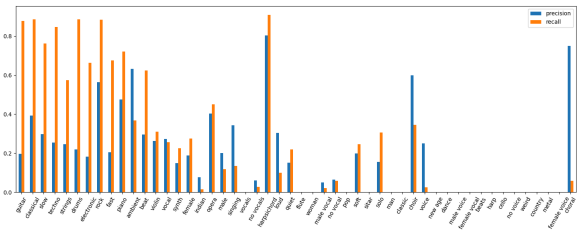


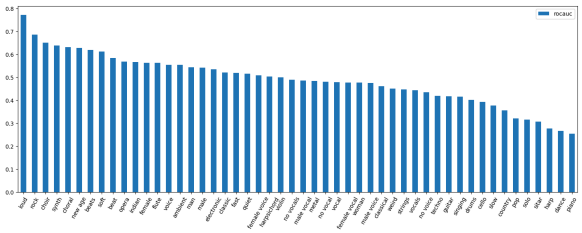


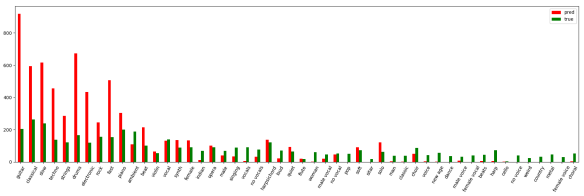
[BCE loss+Triplet loss]

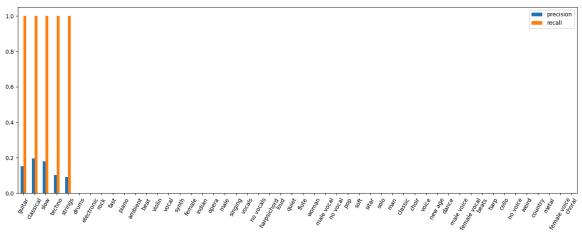


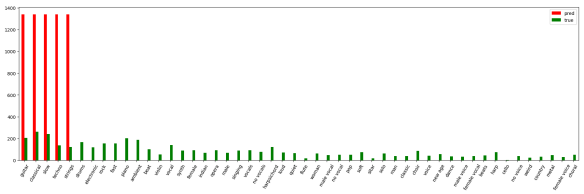












	ROC AUC	PR AUC
baseline (BCE loss)	**0.794**	**0.258**
Focal loss	0.770	0.236
baseline + Triplet loss	0.490	0.068

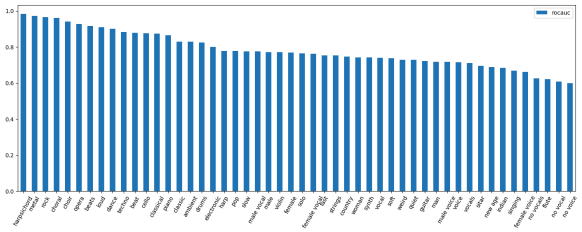
Focal loss led to marginal performance drops but improved balance for rare tags. Triplet loss significantly reduced performance, implying conflict with the multi-label classification task.

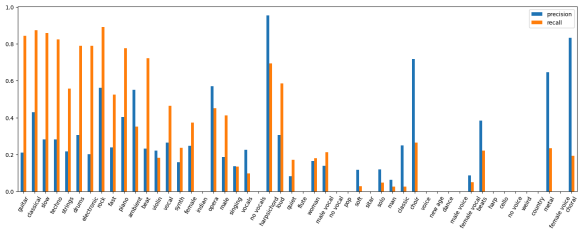
D. Sampling Strategy Analysis

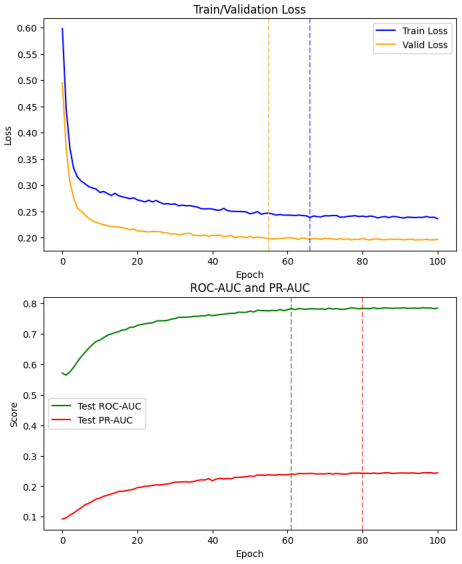
This section analyzes the effect of tag-wise uniform sampling in addressing label imbalance issue.

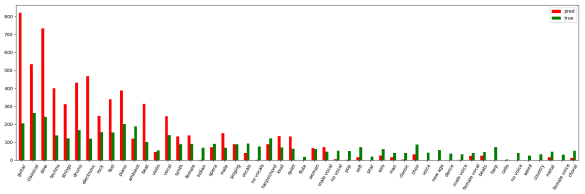
Uniform sampler balances label distribution by first randomly selecting a label, and then sampling a corresponding data. This results in oversampling minority labels and undersampling majority ones.

[Uniform Sampler]









	ROC AUC	PR AUC
baseline (no sampler)	**0.794**	**0.258**
Uniform sampler	0.784	0.244

Although performance gains were limited, predictions became slightly more balanced across tags.

4. Discussions

Revisit the hypotheses:

1. **The model is underfitted. Increasing training epochs or using higher complexity models will improve performance.**

   → *Partially confirmed.*

   The baseline model showed clear underfitting, which improved significantly with more training epochs and higher-capacity architectures. 2D CNNs and attention-based models achieved higher ROC AUC and PR AUC scores, validating the benefit of increased model complexity. However, the residual variant, although deeper, showed slightly lower performance, possibly due to over-parameterization and insufficient regularization on a relatively small dataset. Furthermore, self-attention models, while flexible, may have failed to capture time-frequency locality effectively compared to convolutional counterparts, which may explain their weaker performance in this context.

2. **The data labels are imbalanced. Using tag-wise uniform sampling or focal loss will increase the performance.**

   → *Partially confirmed.*

   Tag-wise uniform sampling and focal loss slightly improved performance on rare tags. However, gains were limited, and imbalance remained a challenge. ㄹocal loss may have overemphasized difficult samples, leading to reduced confidence in predictions and potential overfitting to ambiguous examples. More adaptive sampling or reweighting strategies may be needed for substantial improvement.

3. **The label representations are entangled. Introducing a triplet loss will increase the performance.**

   → *Not supported.*

Triplet loss degraded overall performance. This suggests that contrastive objectives may conflict with multi-label classification. Unlike single-label classification, where positive and negative pairs are well-defined, multi-label tasks introduce ambiguity in defining ‘similar’ vs ‘dissimilar’ pairs. This may explain the poor performance of triplet loss. Alternative approaches like supervised contrastive loss could be explored in future work.