tranSLATor

Milan Ilic

Problem description

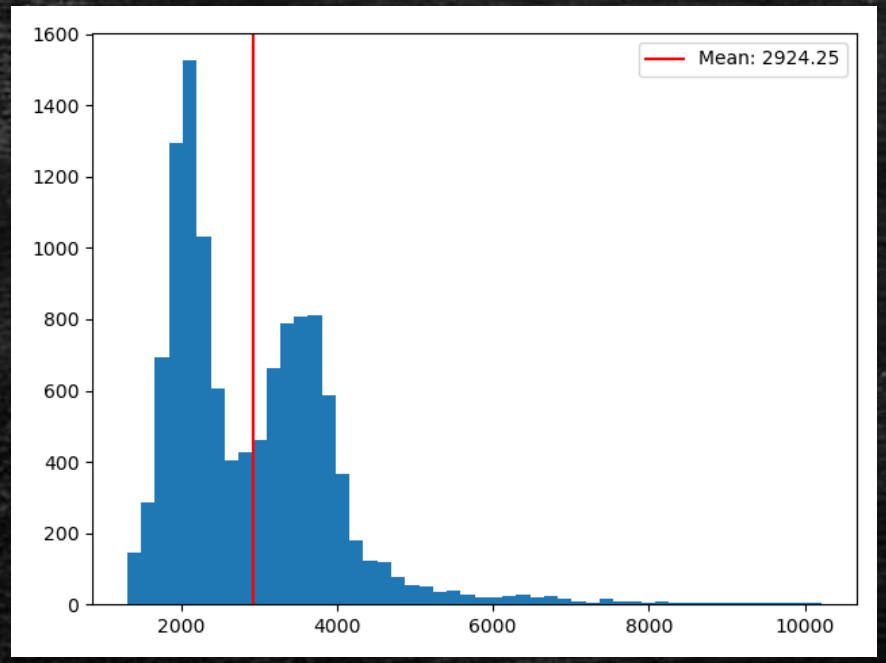
- The task of this project was music classification, specifically classifying the song to the correct rap artist.
- The biggest challenge of this project is the similarity of songs between different artists.
- Most of the songs used for the project can be catagorized as mumble rap which also increases the difficulty of this problem.

Dataset

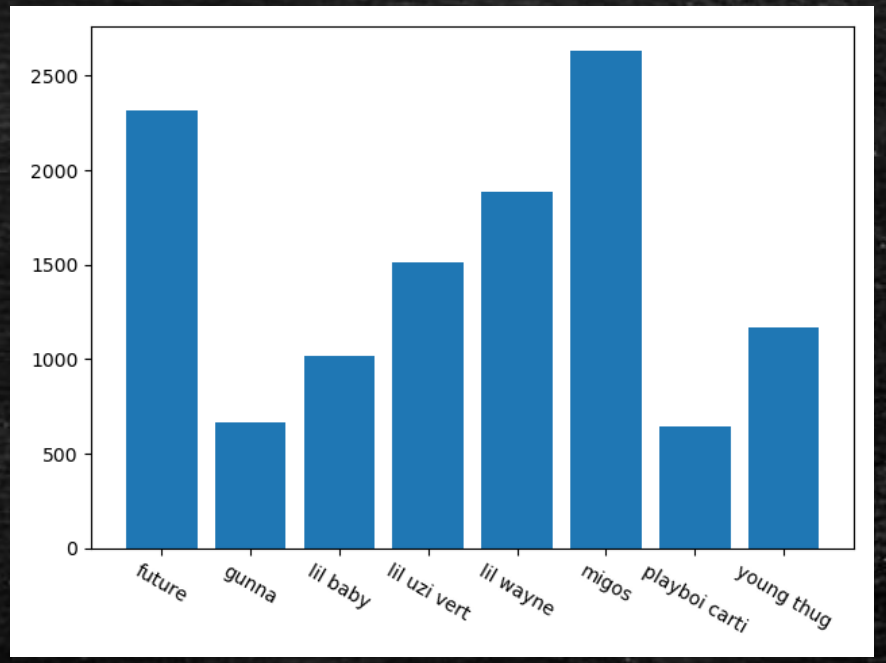
- Dataset consisted out of 437 different songs collected from 8 artists.
- These songs are split into 32166 lines using timestamp information of each line.
- The songs were filtered to contain only solo songs and then further filtered so there is no duplicate line transcripts in the dataset.
- After all preprocessing the number of lines was 11867.

Dataset statistics

Histogram of slice length in milliseconds.

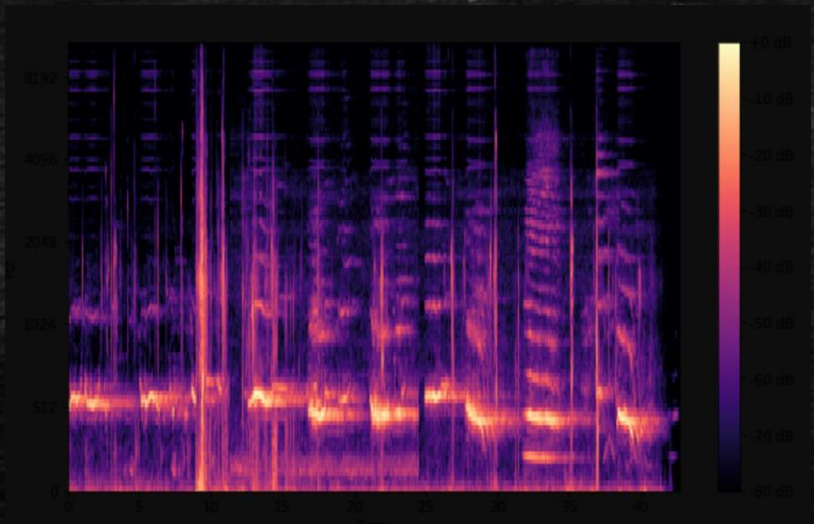


Barplot of number of slices per artist.



Approach

- For solving this problem, the chosen approach is as follows:
  1. Take the line slice given as *.wav* file and transform it into mel spectogram.
  2. Resize the mel spectogram to fixed size and perform some data augmentation techniques.
  3. Use a CNN backbone and MLP head to perform classification.



Training

- Several neural network architecutres were used for training the dataset:
  - MobileNetV2, InceptionV3, Resnet50V2, DenseNet121
- Training was done with the next parameters:

Parameter	Train/Val	Input shape	Epochs	Batch size	Optimizer	Scheduler	Loss
Value	80/20	(128, 256, 3)	200	4	RAdam	Cosine Annealing	CCE

Evaluation Metrics

- Metrics used to evaluate the performance of models are:
  - ***Slice accuracy*** (samplewise accurace)
  - ***Song accuracy*** (song prediction is calculated by getting the majority vote of song slices predictions)

Results

- Results were recorded for each network architecture alone, and for their ensemble.

Validation

Architecture	Augmentations	Slice Accuracy	Song Accuracy*
MobileNetV2	SpecAugment	86.93%	**100%**
InceptionV3	SpecAugment	88.17%	**100%**
Resnet50V2	SpecAugment	89.27%	**100%**
DenseNet121	SpecAugment	**89.62%**	97.78%
Ensemble (4 nets)	-	**91.87%**	**100%**

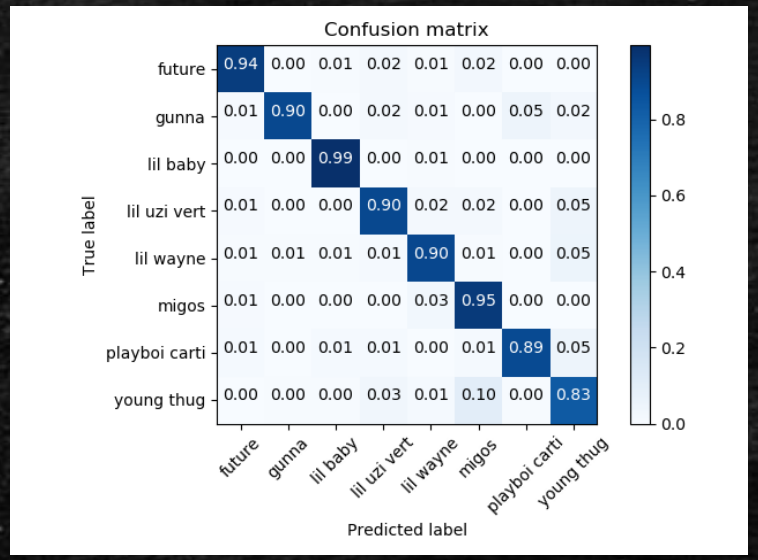
Test

Architecture	Augmentations	Slice Accuracy	Song Accuracy*
MobileNetV2	SpecAugment	72.48%	88.89%
InceptionV3	SpecAugment	**77.23%**	88.89%
Resnet50V2	SpecAugment	71.23%	83.33%
DenseNet121	SpecAugment	75.65%	**100%**
Ensemble (4 nets)	-	**78.65%**	88.89%

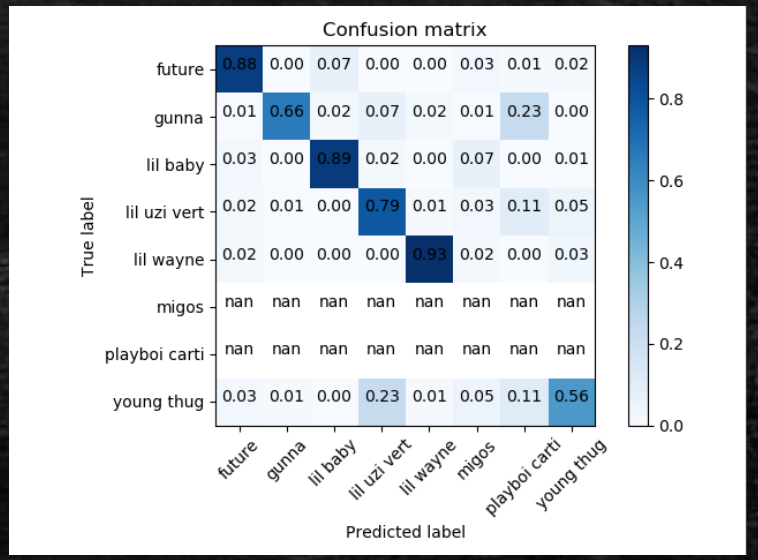
Confusion Matrix

- Confusion matrix is plotted only for ensemble method. (No new test songs for two artists)

Validation



Test



References

1. **SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition (https://arxiv.org/abs/1904.08779)**