PROJECT

FAKE SPEECH DETECTION

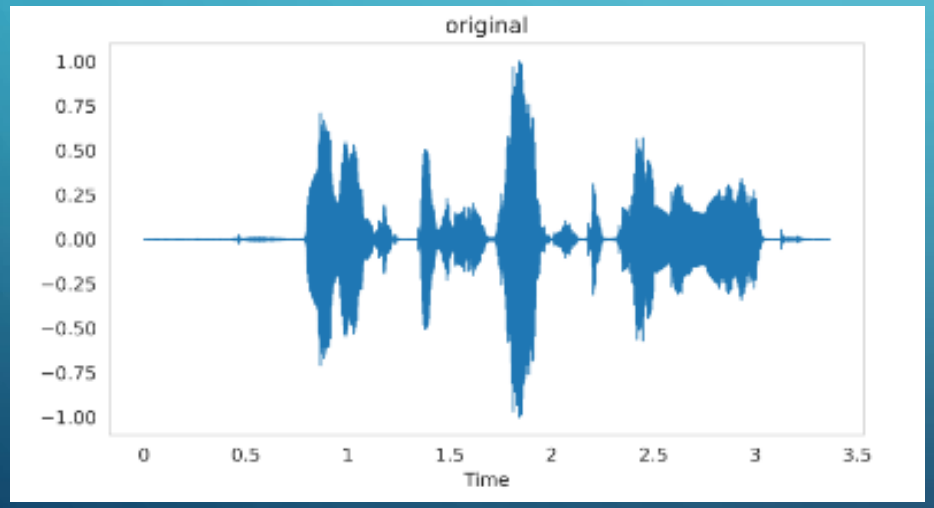
(CONFORMER)

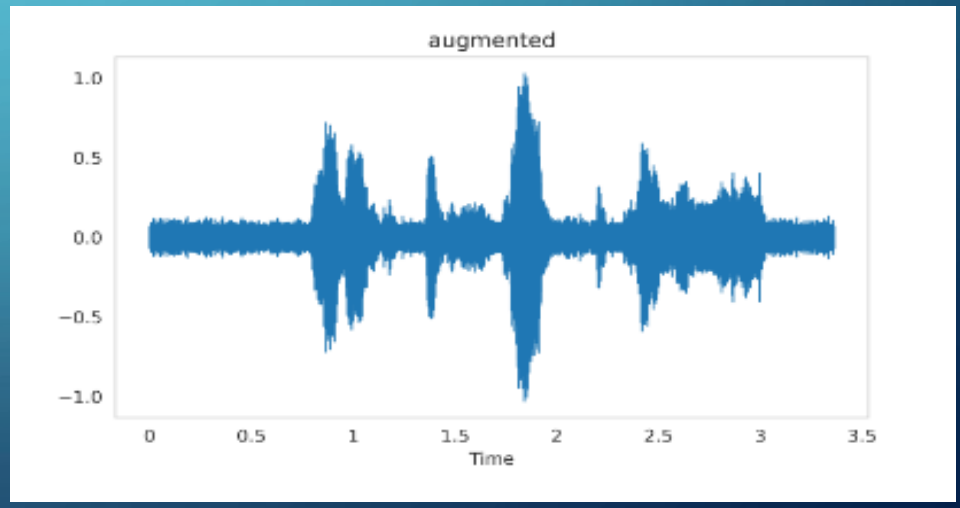
DATASET

- ***ASVspoof 2019 Dataset :-*** **The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database.**
  - **It Contains a 363k files of .flac files**
- ***ASVspoof 2019 tfrecord Dataset:-*** **It is simple Subset of the dataset.**
  - **It contains a 30 files of .tfrec format.**
  - (http://www.asvspoof.org)

DATA AUGMENTATION

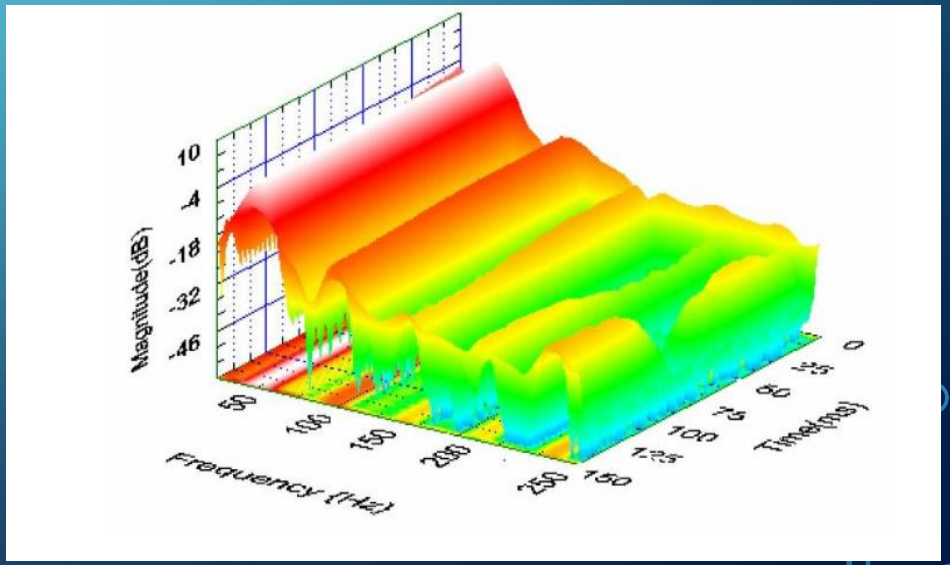
- Two types of Augmentations are used here
  - AudioAug
  - SpecAug





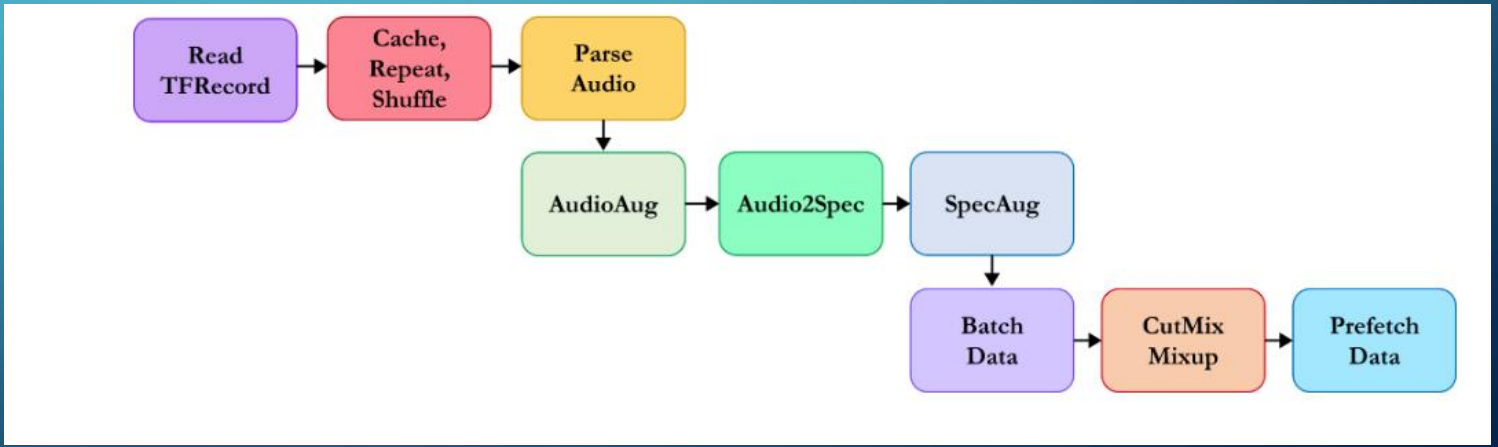
FEATURE EXTRACTION

- We'll feed Mel-Spectrogram feature to our model as input.
- Spectrogram feature is extracted from audio sample using Short Time Fourier Transform (STFT).
- The fast Fourier transform is a powerful tool that allows us to analyze the frequency content of a signal.



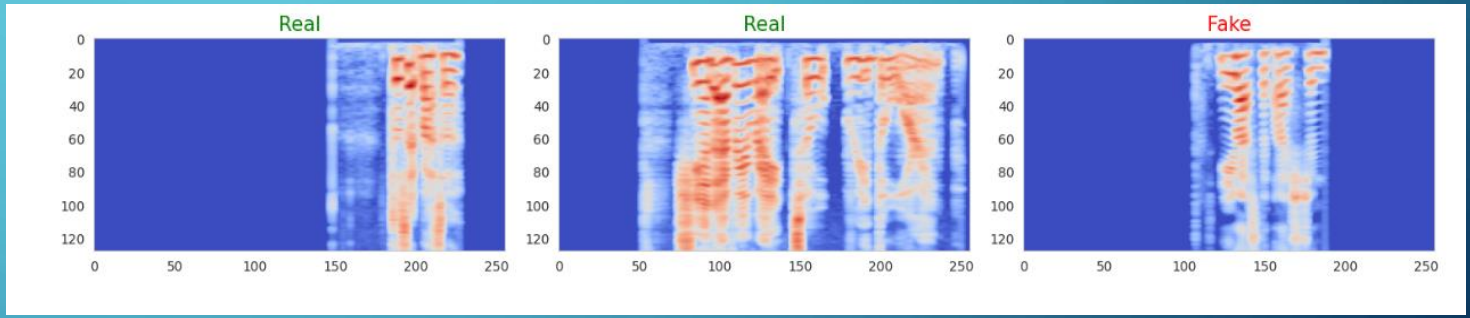
DATA PIPELINE

- We can build complex input pipelines from simple, reusable pieces using tf.data API.
- https://www.tensorflow.org/guide/data

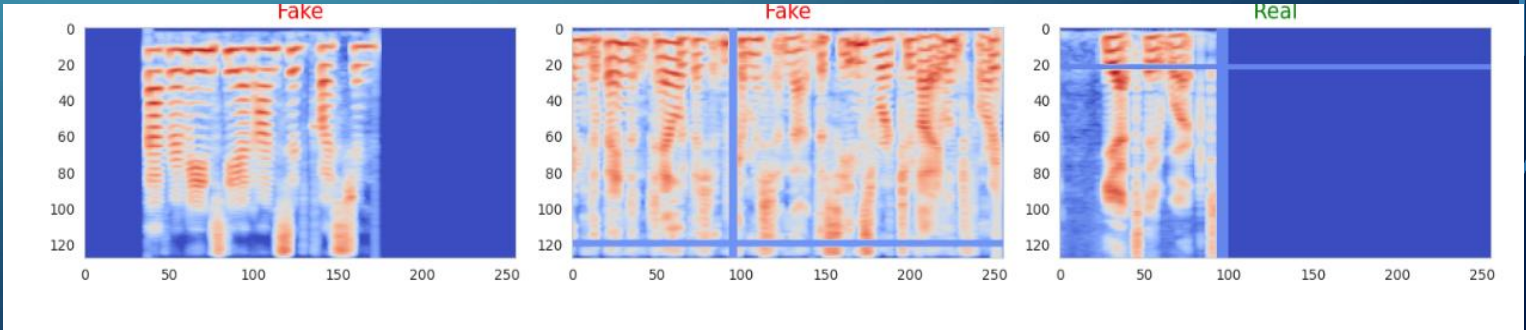


VISUALIZATION

- spectrogram and its associate label.
  - Without Augmentation



  - With Augmentation



MODEL (CONFORMER)

- A Neural net for speech recognition that was published by Google Brain in 2020.
- I used as a transfer learning concept to train on the data.
- Here is the Model Summary as well as Number of trainable parameters

```
#########################################################
#### IMAGE_SIZE: (256, 128) | BATCH_SIZE: 32 | EPOCHS: 12
#### MODEL: Conformer | LOSS: binary_crossentropy
#### NUM_TRAIN: 5,000 | NUM_VALID: 4,000
#########################################################
```

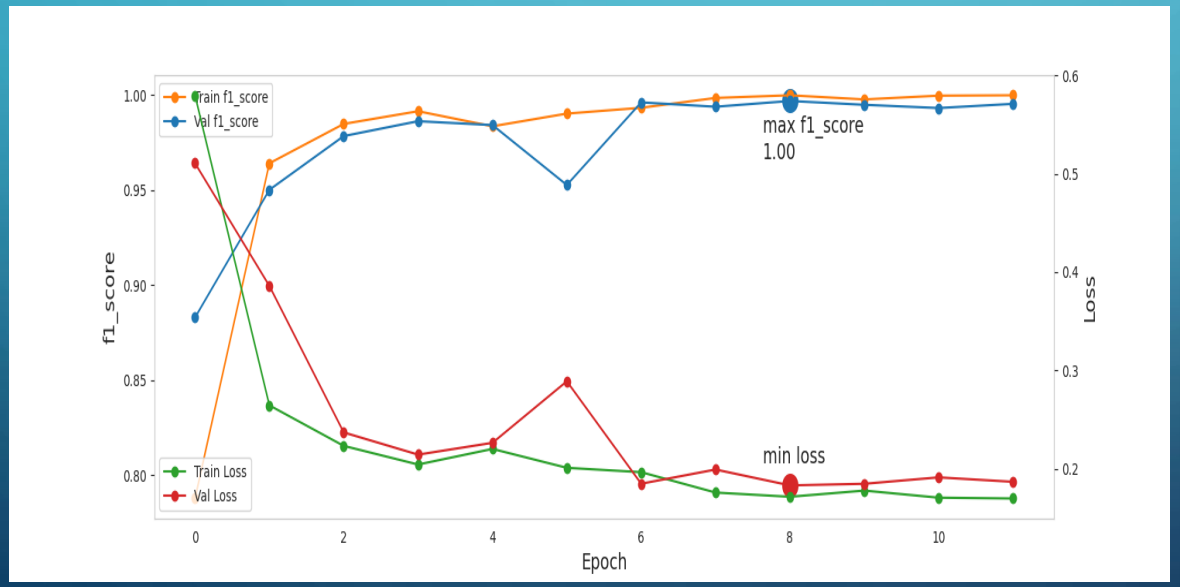
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 256, 128, 1)]	0
conformer_encoder (ConformerEncoder)	(None, None, 144)	8959680
global_average_pooling1d (GlobalAveragePooling1D)	(None, 144)	0
dense (Dense)	(None, 1)	145

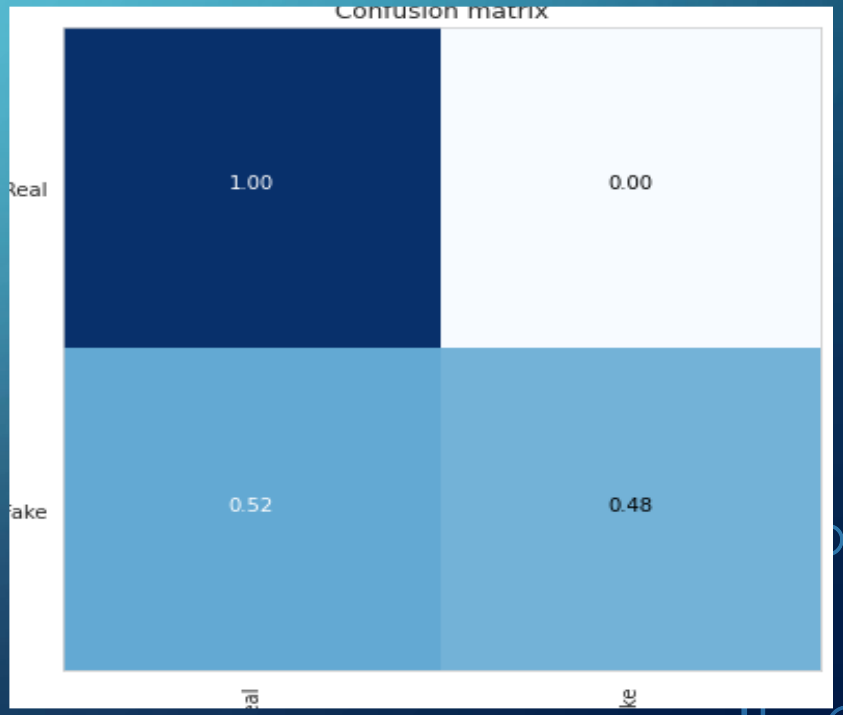
Total params: 8959825 (34.18 MB)
Trainable params: 8955217 (34.16 MB)
Non-trainable params: 4608 (18.00 KB)

TRANING AND MATRIX

Here is the training Status and Confusion Matrix.

We used Wandb for the looking of current Status.





CONCLUSION AND FUTURE SCOPE

- In the battle against audio manipulation, Conformers have become heroes.
- Studies on datasets like ASVspoof showcase their impressive accuracy.
- They can be used as a Ensumble Learning as well as other techniques for the Betterment of result.
- Currenlty, It is in research field and widely used in Domestic Stations.

REFERENCES

- View Project at https://wandb.ai/anony-moose-21470533058967856/fake-speech-detection?apiKey=55191f42f262408d6d22cec5d9c08fc9e9af0fd4
- View Run at https://wandb.ai/anony-moose-21470533058967856/fake-speech-detection/runs/u5htw2se?apiKey=55191f42f262408d6d22cec5d9c08fc9e9af0fd4
- Dataset at http://www.asvspoof.org
- Papers:
  - **Conformer: Convolution-augmented Transformer for Speech Recognition https://arxiv.org/abs/2005.08100**
  - **Synthetic Voice Detection and Audio Splicing Detection using SE-Res2Net-Conformer Architecture https://ieeexplore.ieee.org/abstract/document/10037999**
  - **SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection https://ieeexplore.ieee.org/abstract/document/10063734**