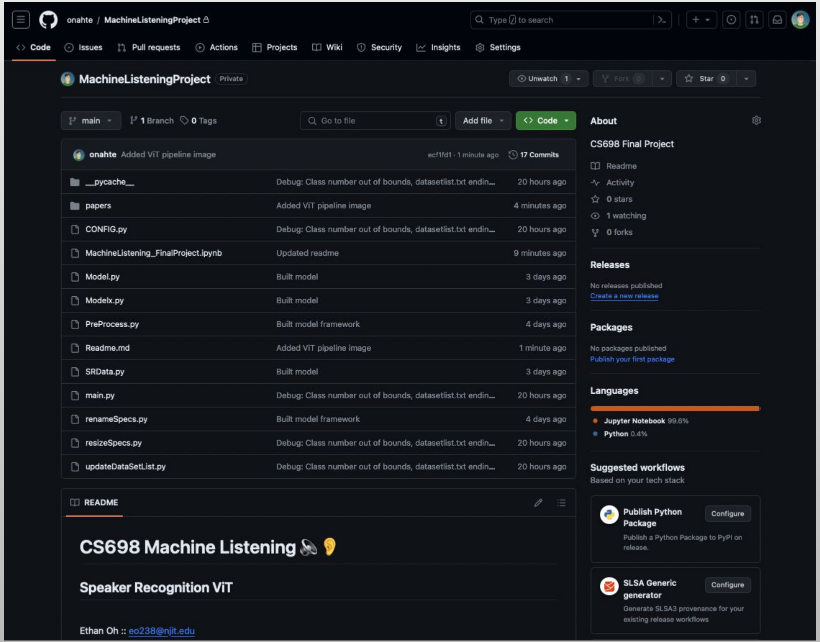
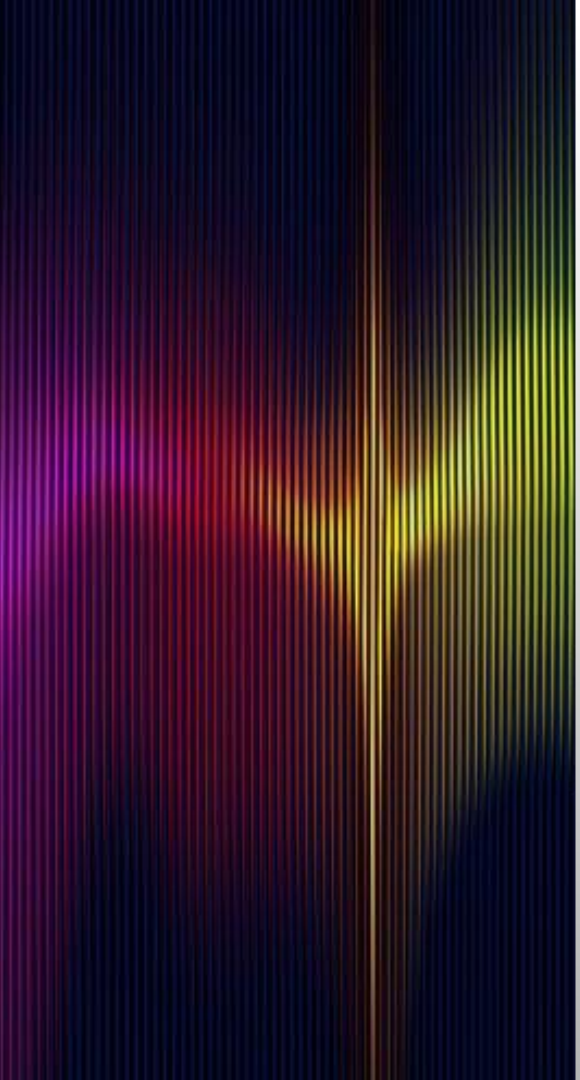
SPEAKER RECOGNITION via ViT

Project Repo

https://github.com/onahte/MachineListeningProject

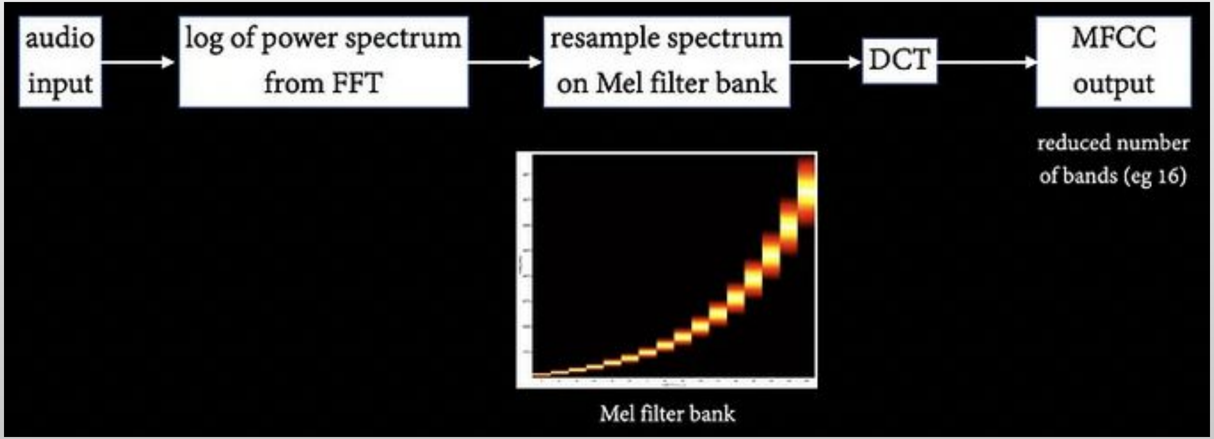


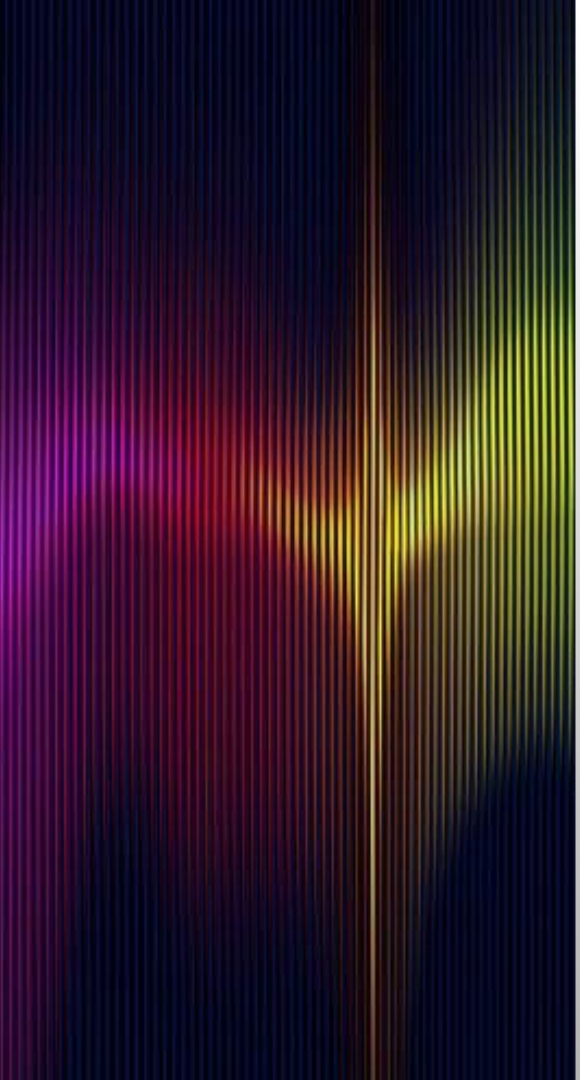


Speaker Recognition

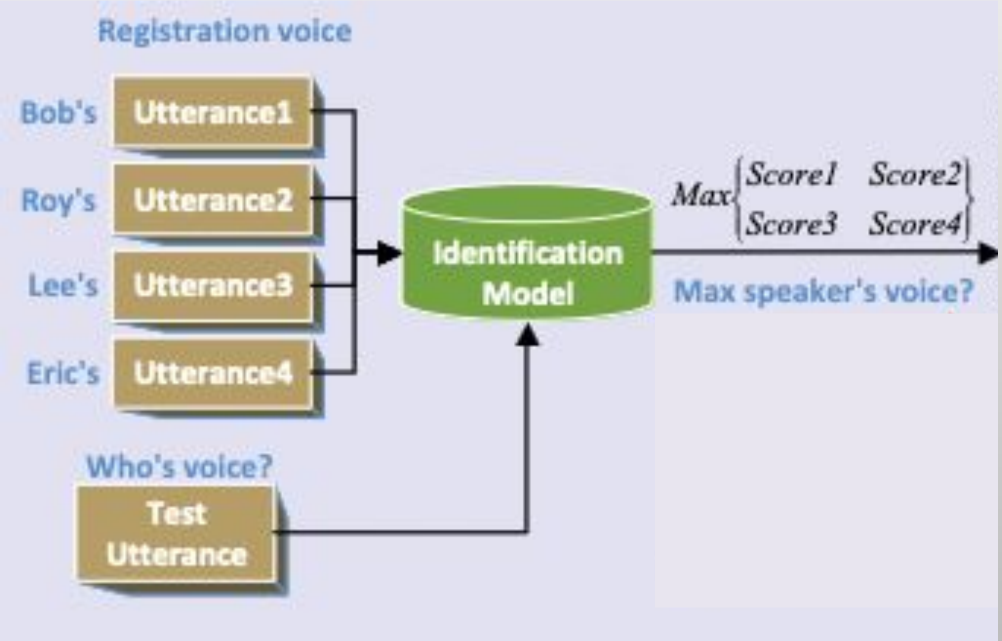
- Automatic Speaker Recognition (ASR)
- Label utterance with speaker ID
- Deep Learning
    - Mel Frequency Cepstral Coefficient
        - Speech Recognition
        - Captures timbre

Speaker Recognition





Speaker Recognition

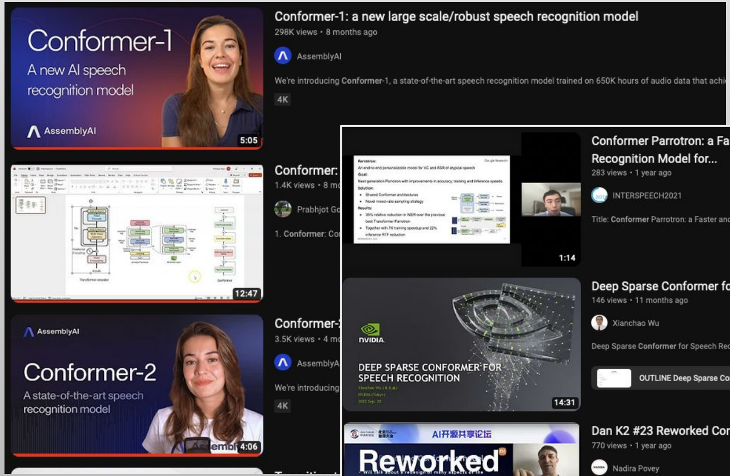


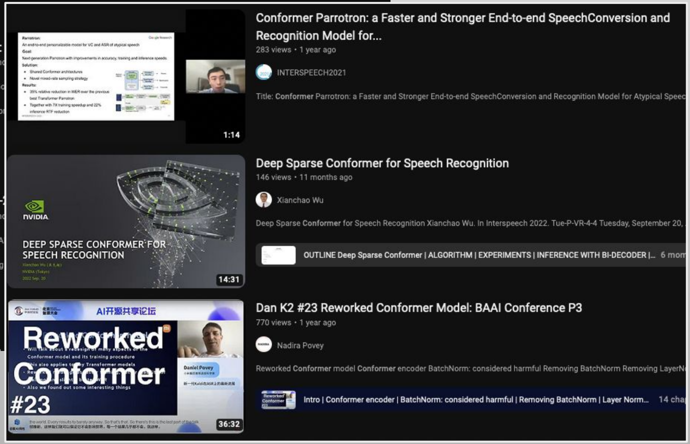
Speaker Recognition

Inputs	CNN	LSTM	Hybrid structures
Wave	Others [52, 53].	—	CNN-LSTM [54, 55]; CNN-GRU [56, 57].
Spectrogram	ResNet [58, 59, 60, 61]; VGGNet [15, 24]; Inception-resnet-v1 [62, 63].	—	CNN-GRU [64]
F-bank	TDNN [14, 65, 66, 67]; ResNet [68, 69, 70, 71]; VGGNet [72]; Inception-resnet-v1 [63, 73, 74]; Others [75, 76].	[77, 78, 79].	BLSTM-ResNet [80], TDNN-LSTM [81]
MFCC	TDNN [82, 51, 83, 84, 85, 86, 87, 88, 67, 89, 90, 91]; ResNet [92]; Others [93, 94].	—	TDNN-LSTM [95]

Speaker Recognition

- CNN + Transformer = Conformer
  - Speech Recognition





VISION TRANSFORMER

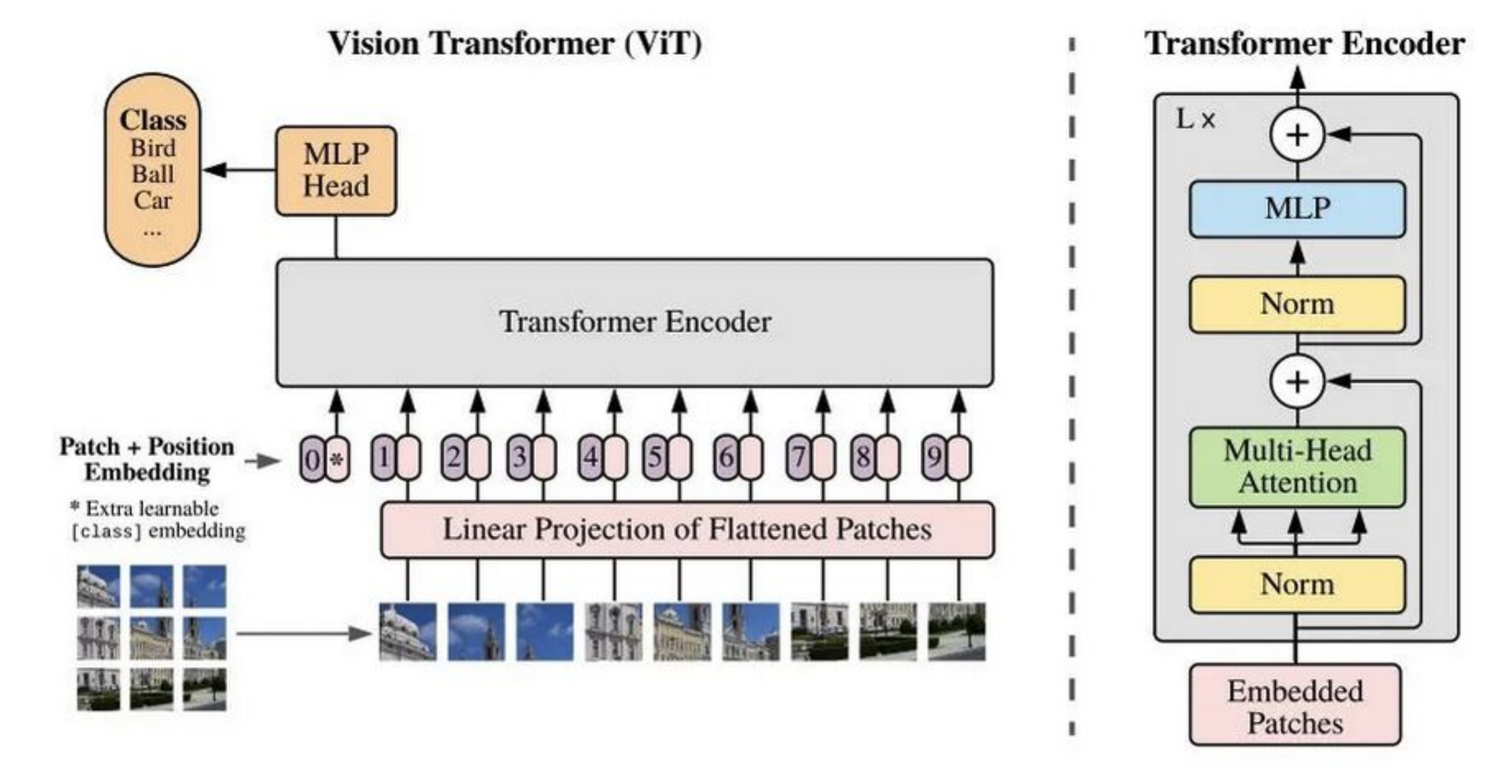
VisionTransformer

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

- Transformer reimagined for images
- Competitive with SOTA CNNs
- Patching

VisionTransformer



VisionTransformer

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

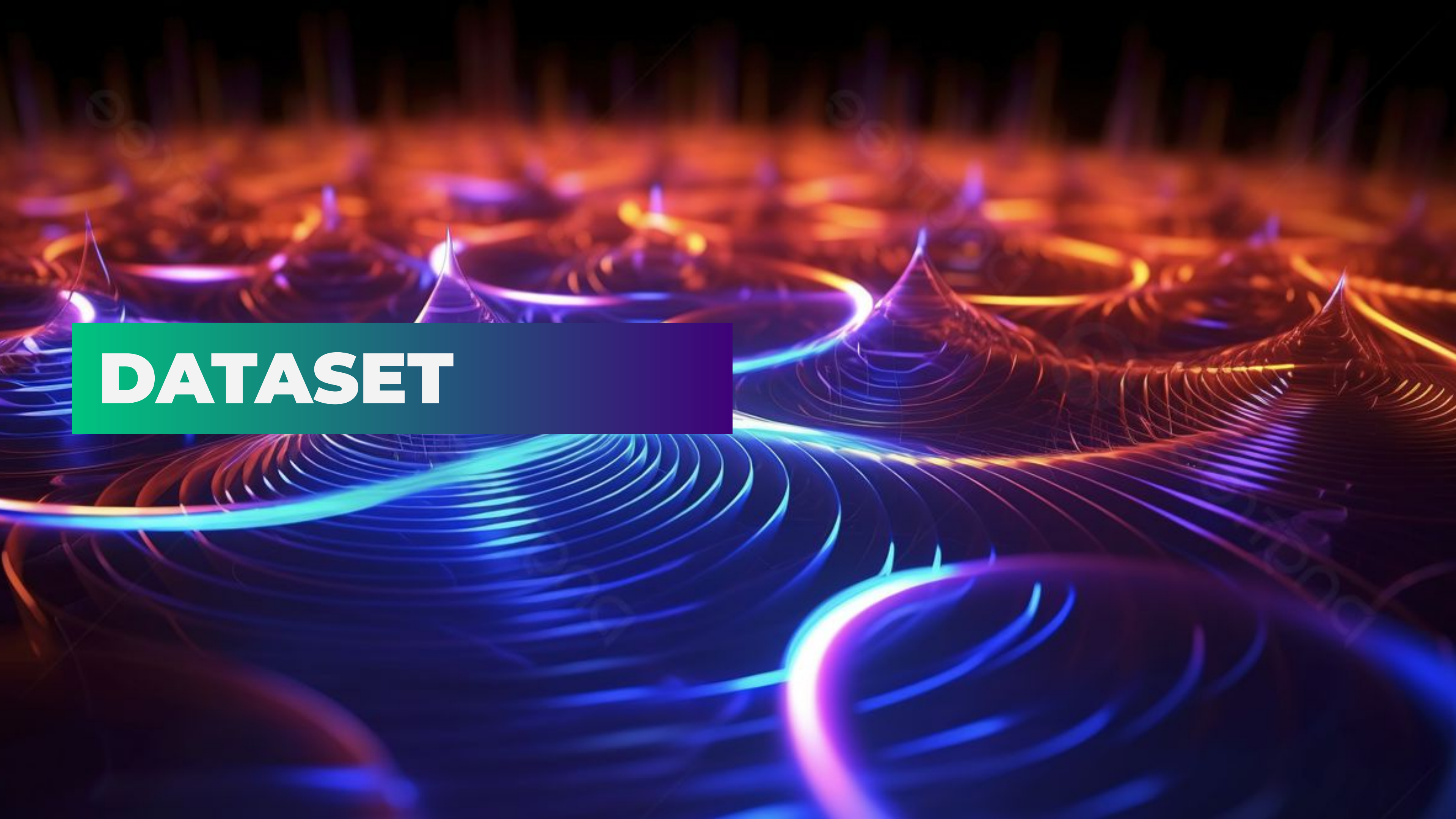
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

All models were trained on TPUv3 hardware. Report of the number of TPUv3-core-days taken to pre-train each of them: number of TPU v3 cores (2 per chip) used for training multiplied by the training time in days

VisionTransformer

- No inductive bias
    - CNNs have strong inductive bias
- Global attention
    - CNNs use growing receptive field
- Data hungry
    - CNNs are not so data hungry
- Lighter than Transformer

DATASET

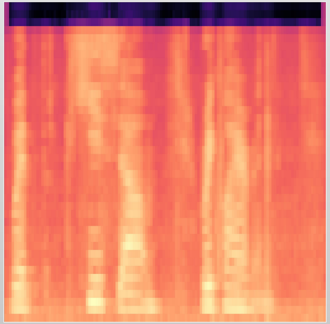


Dataset

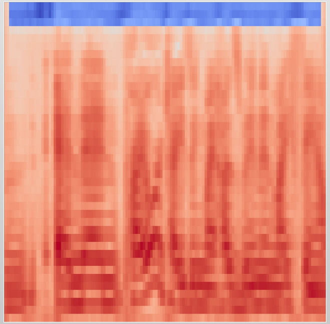
- VoxCeleb1
    - 113_985 clips
    - YouTube audio
    - 932 classes

Dataset

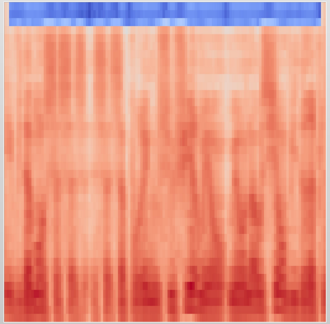
Mel Spectrogram



0



639

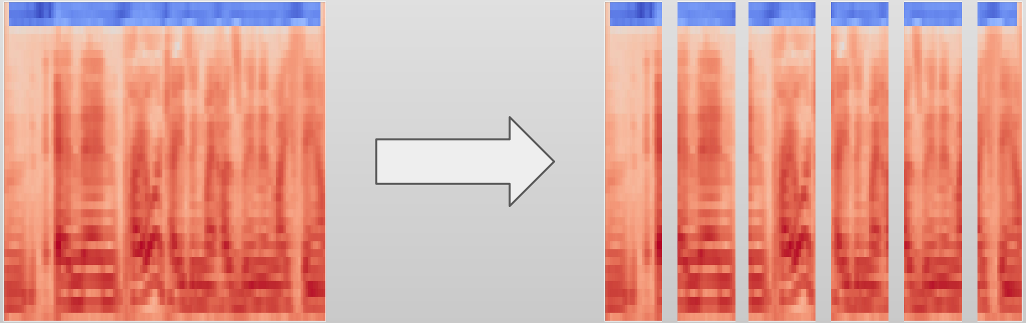


924

PUTTING IT ALL TOGETHER

Putting It All Together

- Spectrograms patchified



Putting It All Together

- Batch size: 16
- Encoder layer: 8
- Embedding size: 932
- Attention heads: 4
- Learning rate: 3e-3

Parameter count: 19,423,738

Putting It All Together

Baseline Model: ECAPA-TDNN

- Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network
- Hybrid Model
  - CNN block (ResNet)
  - Attentive Statistics Pooling

Putting It All Together

Metric: Equal Error Rate (EER)

- Percentage of FAR=FRR

$$EER = \frac{FAR + FRR}{2}$$

FAR is the false acceptance rate and FRR is false recognition rate and they are defined:

$$FAR = \frac{number\ of\ false\ positives}{number\ of\ false\ positives + number\ of\ true\ negatives}\ x\ 100$$

$$FRR = \frac{number\ of\ false\ negatives}{number\ of\ false\ negatives + number\ of\ true\ positives}\ x\ 100$$

Putting It All Together

Model	Parameters	EER
ECAPA-TDNN	20.8M	0.82
SR-ViT	19.4	0.4796

```
EER Score: 0.47967687249183655
```

Bibliography

[1] A. Gulati *et al.*, “Conformer: Convolution-augmented Transformer for Speech Recognition,” *arXiv:2005.08100 [cs, eess]*, May 2020, Available: https://arxiv.org/abs/2005.08100

[2] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, “Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges,” *Expert Systems with Applications*, vol. 171, p. 114591, Jun. 2021, doi: https://doi.org/10.1016/j.eswa.2021.114591.

[3] Y. Zhang *et al.*, “MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification,” Mar. 2022, doi: https://doi.org/10.48550/arxiv.2203.15249.

[4] A. Gulati *et al.*, “Conformer: Convolution-augmented Transformer for Speech Recognition,” *arXiv:2005.08100 [cs, eess]*, May 2020, Available: https://arxiv.org/abs/2005.08100

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” *Interspeech 2020*, pp. 3830–3834, Oct. 2020, doi: https://doi.org/10.21437/Interspeech.2020-2650.

[6] Z. Bai and X.-L. Zhang, “Speaker recognition based on deep learning: An overview,” *Neural Networks*, vol. 140, pp. 65–99, Aug. 2021, doi: https://doi.org/10.1016/j.neunet.2021.03.004.

[7] A. Dosovitskiy *et al.*, “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE,” Jun. 2021. Available: https://arxiv.org/pdf/2010.11929.pdf