

SPEAKER

RECOGNITION

via ViT

Project Repo

<https://github.com/onahte/MachineListeningProject>

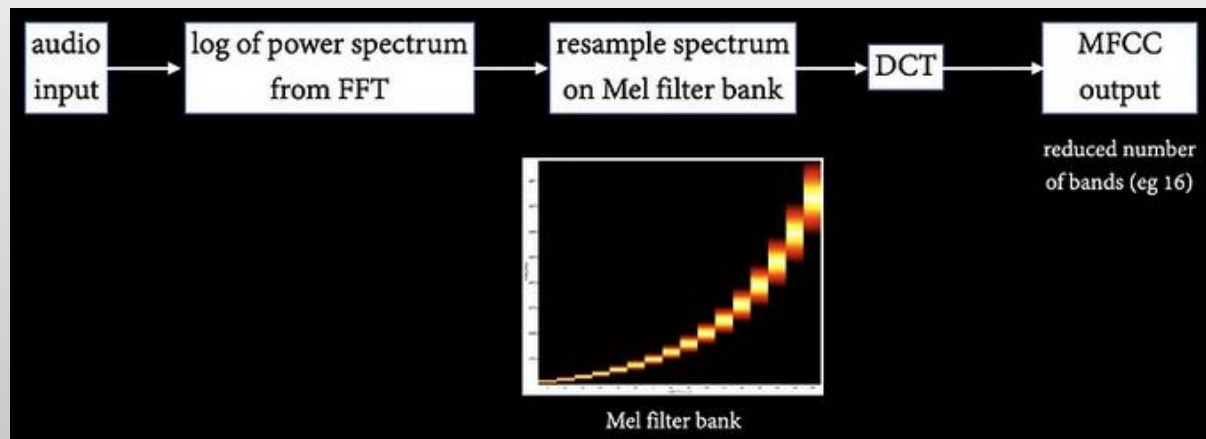
The screenshot shows the GitHub repository page for 'MachineListeningProject' by user 'onahte'. The repository is private and has 17 commits. The file list includes: __pycache__, papers, CONFIG.py, MachineListening_FinalProject.ipynb, Model.py, Modelx.py, PreProcess.py, Readme.md, SRData.py, main.py, renameSpecs.py, resizeSpecs.py, and updateDataSetList.py. The README section is visible, titled 'CS698 Machine Listening' and 'Speaker Recognition VIT', authored by Ethan Oh. The right sidebar shows repository statistics: CS698 Final Project, 0 stars, 1 watching, 0 forks, and no releases or packages published. It also lists languages (Jupyter Notebook 99.6%, Python 0.4%) and suggested workflows like 'Publish Python Package' and 'SLSA Generic generator'.



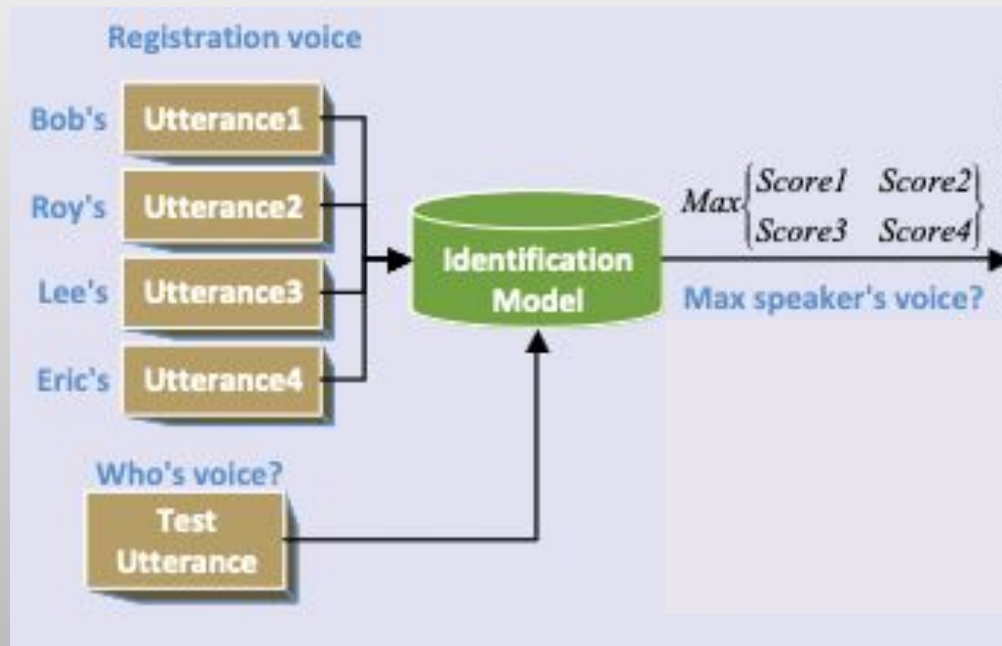
Speaker Recognition

- 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]; VG-GNet [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

The collage features several video thumbnails and titles:

- Conformer-1: a new large scale/robust speech recognition model** by AssemblyAI. The thumbnail shows a woman speaking. The title indicates it's a new large scale/robust speech recognition model. The video has 298K views and is 8 months old.
- Conformer-2: A state-of-the-art speech recognition model** by AssemblyAI. The thumbnail shows a woman speaking. The title indicates it's a state-of-the-art speech recognition model. The video has 3.5K views and is 4 months old.
- Conformer Parrot: a Faster and Stronger End-to-end SpeechConversion and Recognition Model for...** by INTERSPEECH2021. The thumbnail shows a man speaking. The title indicates it's a faster and stronger end-to-end speech conversion and recognition model. The video has 253 views and is 1 year old.
- Deep Sparse Conformer for Speech Recognition** by Xianchao Wu. The thumbnail shows a 3D visualization of a neural network. The title indicates it's a deep sparse conformer for speech recognition. The video has 146 views and is 11 months old.
- Reworked Conformer #23** by Dan K2 #23. The thumbnail shows a man speaking. The title indicates it's a reworked conformer model. The video has 770 views and is 1 year old.

An abstract digital landscape featuring pixelated, jagged mountains. The left side shows a mountain range in shades of teal and cyan, while the right side features a larger, more prominent mountain range in vibrant yellow and orange. The foreground is dark, with scattered vertical bars of various colors (pink, yellow, green) and a glowing, pixelated path leading towards the mountains. The overall aesthetic is reminiscent of early computer graphics or digital art.

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

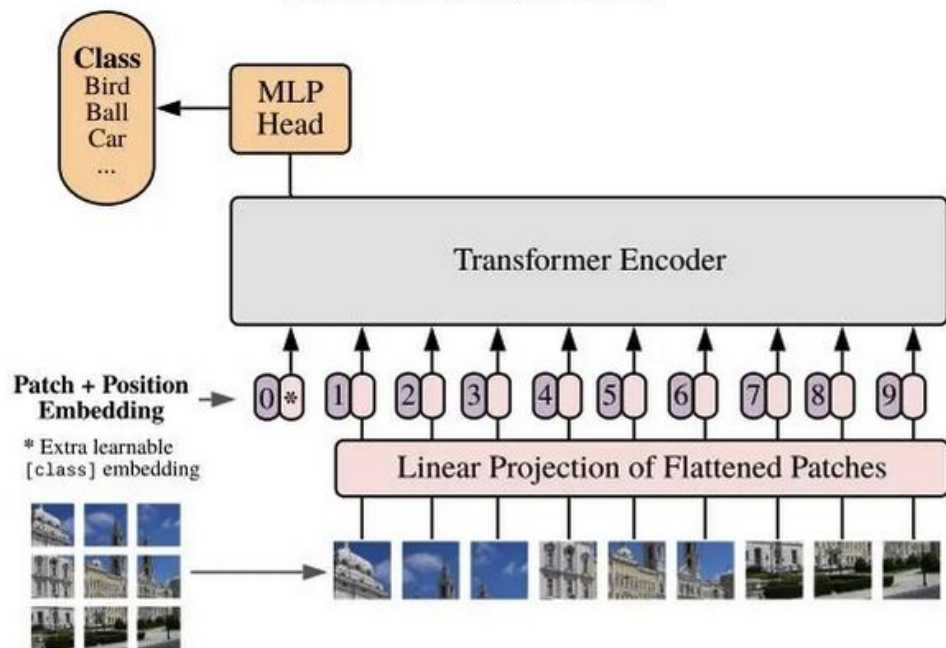
Google Research, Brain Team

{adosovitskiy, neilhoulbsby}@google.com

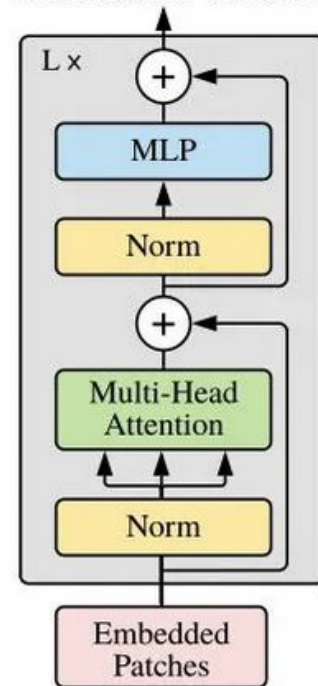
- Transformer reimagined for images
- Competitive with SOTA CNNs
- Patching

VisionTransformer

Vision Transformer (ViT)



Transformer Encoder



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 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 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

An abstract, colorful visualization of data, possibly a 3D bar chart or a complex waveform, rendered in vibrant yellow, orange, red, and blue against a dark background. The visualization appears to be composed of many small, vertical bars or segments, creating a textured, almost crystalline appearance. The colors transition from dark blue at the top to bright yellow and orange at the bottom, suggesting a gradient or a specific data range.

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

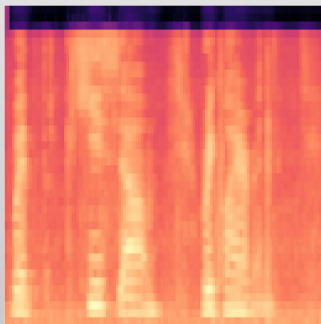


Dataset

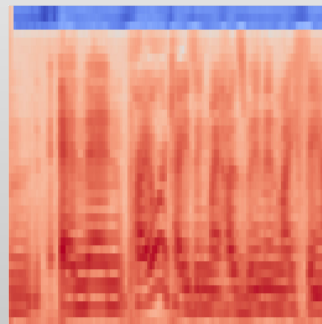
- VoxCeleb1
 - 113_985 clips
 - YouTube audio
 - 932 classes

Dataset

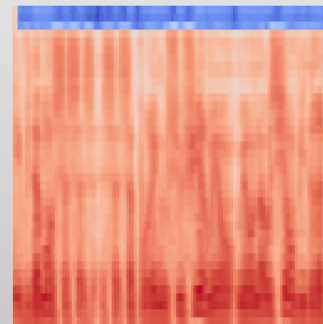
Mel Spectrogram



0



639



924



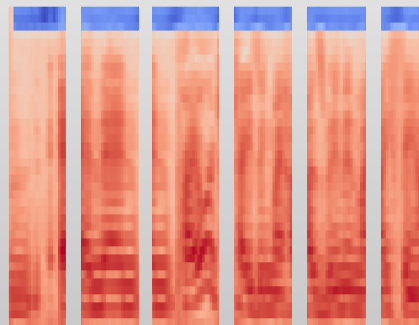
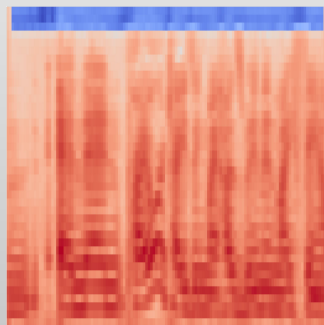
PUTTING

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:

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

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



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>