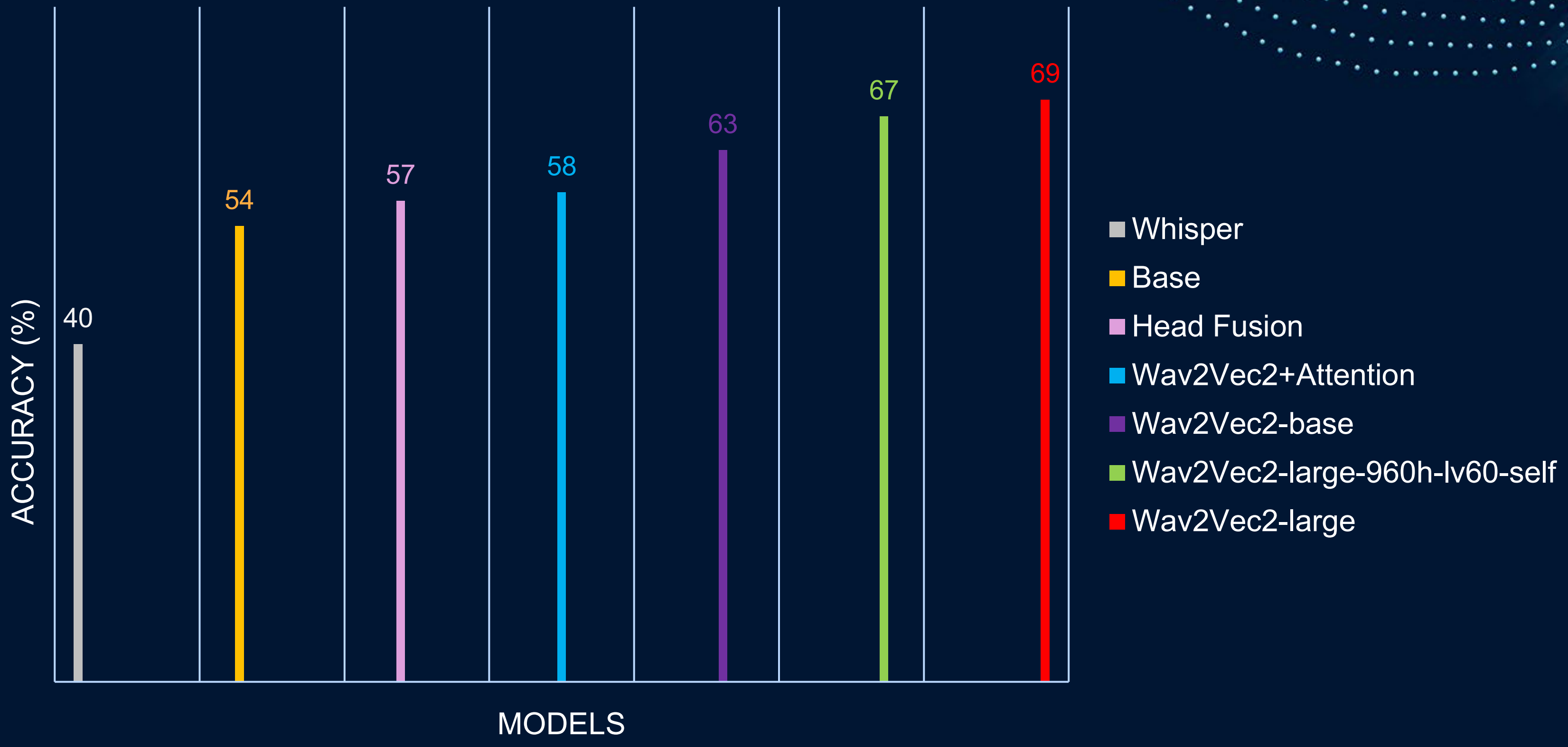
Emotion Recognition

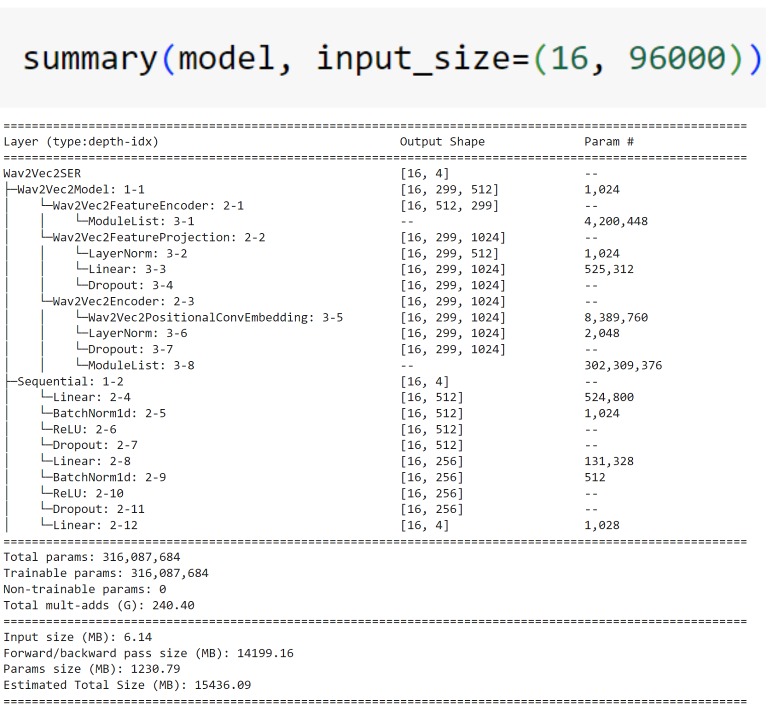
SMALL PROJECT

AAI3001_SP_1

ACCURACY OVER DIFFERENT MODELS



Comparison

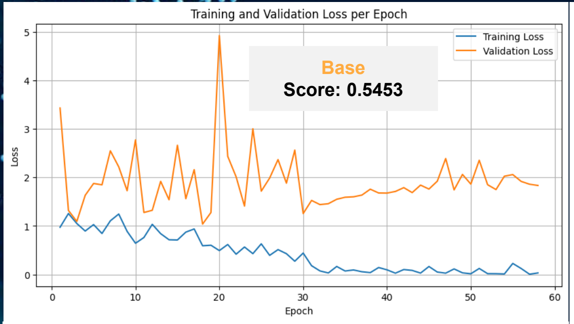


```
summary(model, input_size=(16, 96000))
```

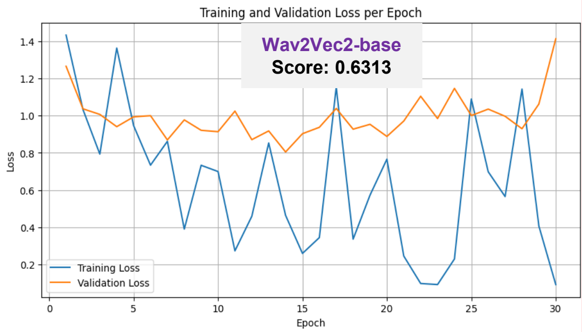
```
=========================================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
=========================================================================================================
Wav2Vec2SER                                             [16, 4]                   --
├─Wav2Vec2Model: 1-1                                    [16, 299, 512]            1,024
│    └─Wav2Vec2FeatureEncoder: 2-1                      [16, 512, 299]            --
│    │    └─ModuleList: 3-1                             --                        4,200,448
│    └─Wav2Vec2FeatureProjection: 2-2                   [16, 299, 1024]           --
│    │    └─LayerNorm: 3-2                              [16, 299, 512]            1,024
│    │    └─Linear: 3-3                                 [16, 299, 1024]           525,312
│    │    └─Dropout: 3-4                                [16, 299, 1024]           --
│    └─Wav2Vec2Encoder: 2-3                             [16, 299, 1024]           --
│    │    └─Wav2Vec2PositionalConvEmbedding: 3-5        [16, 299, 1024]           8,389,760
│    │    └─LayerNorm: 3-6                              [16, 299, 1024]           2,048
│    │    └─Dropout: 3-7                                [16, 299, 1024]           --
│    │    └─ModuleList: 3-8                             --                        302,309,376
├─Sequential: 1-2                                       [16, 4]                   --
│    └─Linear: 2-4                                      [16, 512]                 524,800
│    └─BatchNorm1d: 2-5                                 [16, 512]                 1,024
│    └─ReLU: 2-6                                        [16, 512]                 --
│    └─Dropout: 2-7                                     [16, 512]                 --
│    └─Linear: 2-8                                      [16, 256]                 131,328
│    └─BatchNorm1d: 2-9                                 [16, 256]                 512
│    └─ReLU: 2-10                                       [16, 256]                 --
│    └─Dropout: 2-11                                    [16, 256]                 --
│    └─Linear: 2-12                                     [16, 4]                   1,028
=========================================================================================================
Total params: 316,087,684
Trainable params: 316,087,684
Non-trainable params: 0
Total mult-adds (G): 240.40
=========================================================================================================
Input size (MB): 6.14
Forward/backward pass size (MB): 14199.16
Params size (MB): 1230.79
Estimated Total Size (MB): 15436.09
=========================================================================================================
```

		Models				
		Base	Wav2Vec2	Wav2Vec2-base	Wav2Vec2-960h-lv60-self	Wav2Vec2-large
Hyperparameters	Learning Rate	1e-3	5e-5	1e-4	2e-4	0.0001
	Momentum	-	-	-	-	-
	Optimizer	Adam	AdamW	AdamW	AdamW	AdamW
	Batch Size	128	16	32	16	16
	Epoch	100 (20)	100	30	30	200 (83)
Regularizations	Weight Decay	1e-5	0.01	0.01	5e-5	0.0001
	Dropout	0.3	0.3	0.3	0.3	0.6
	Augmentation	-	Time Stretch Pitch Shift Add Noise Time Mask	Add Noise	Add Noise	Add Noise

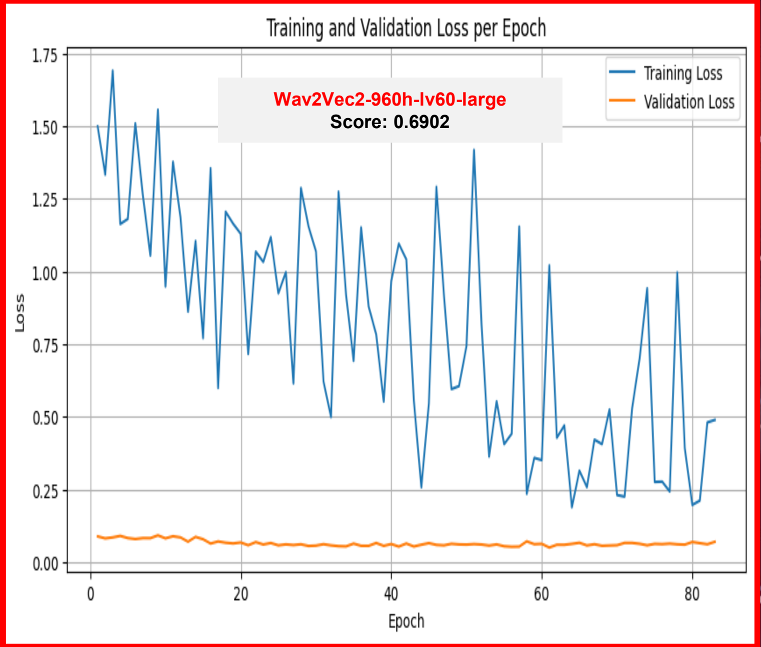
Analysis


Training and Validation Loss per Epoch

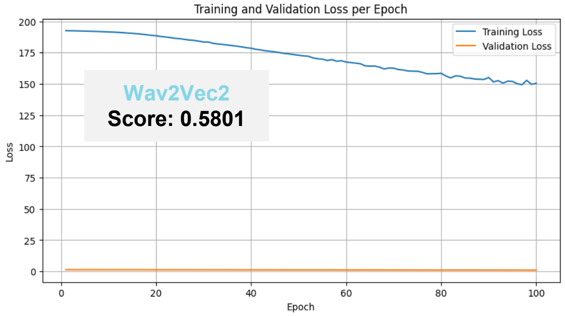
Base
Score: 0.5453


Training and Validation Loss per Epoch

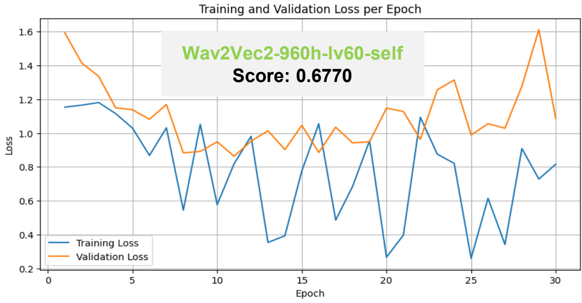
Wav2Vec2-base
Score: 0.6313


Training and Validation Loss per Epoch

Wav2Vec2-960h-lv60-large
Score: 0.6902

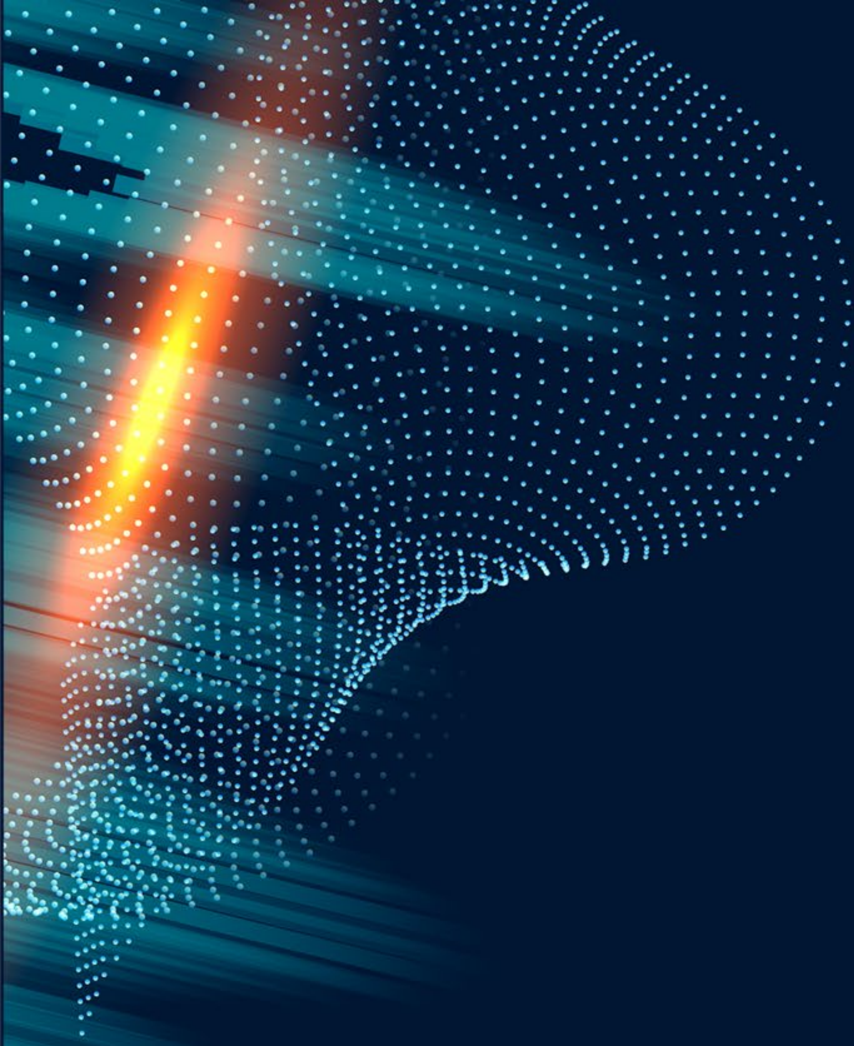

Training and Validation Loss per Epoch

Wav2Vec2
Score: 0.5801


Training and Validation Loss per Epoch

Wav2Vec2-960h-lv60-self
Score: 0.6770

- Wav2Vec2 Pre-Trained models perform better
- High instability: Due to high VRAM requirements, batch size must be reduced, which can be improved with better hardware



Thank You

Credits: Brian Tham & Hong Ziyang

References

1. M. Xu, F. Zhang and W. Zhang, "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset," in IEEE Access, vol. 9, pp. 74539-74549, 2021, doi: 10.1109/ACCESS.2021.3067460. keywords: {Speech recognition;Emotion recognition;Data models;Deep learning;Text recognition;Magnetic heads;Training data;Speech emotion recognition;convolutional neural network;attention mechanism;noise reduction} https://ieeexplore.ieee.org/abstract/document/9381872

2. M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246. keywords: {Feature extraction;Convolution;Recurrent neural networks;Speech recognition;Emotion recognition;Solid modeling;Task analysis;Attention mechanism;convolutional recurrent neural networks (CRNN);speech emotion recognition (SER)}, https://ieeexplore.ieee.org/abstract/document/9381872

3. Meta: https://huggingface.co/collections/facebook/wav2vec-20-651e865258e3dee2586c89f5

4. OpenAI: https://huggingface.co/collections/openai/whisper-release
6501bba2cf999715fd953013