Workshop Week 1: Introduction to Posthoc Interpretability Methods

Payanshi Jain: 14230143

June 9, 2023

Lab 1: Probing Language Models

```
16
('Australian', 2) :neutral
('actor\\/director', 2) :neutral
('John', 2) :neutral
('Polson', 2) :neutral
('and', 2) :neutral
('award-winning', 4) :positive
('English', 2) :neutral
('cinematographer', 2) :neutral
('Giles', 2) :neutral
('Nuttgens', 2) :neutral
('make', 2) :neutral
('a', 2) :neutral
('terrific', 4) :positive
('effort', 2) :neutral
('at', 2) :neutral
('disguising', 2) :neutral
('the', 2) :neutral
('obvious', 1) :negative
('with', 2) :neutral
('energy', 3) :positive
('and', 2) :neutral
('innovation', 3) :positive
('.', 2) :neutral
```

Figure 1: *ToSubmit 1:* Assigning negative for < 2, neutral for $= 2$, and positive for > 2. In sentence 16, I observed positive and negative sentiments, Green, Yellow, and Red lines represent positive, natural, and negative sentiments respectively.

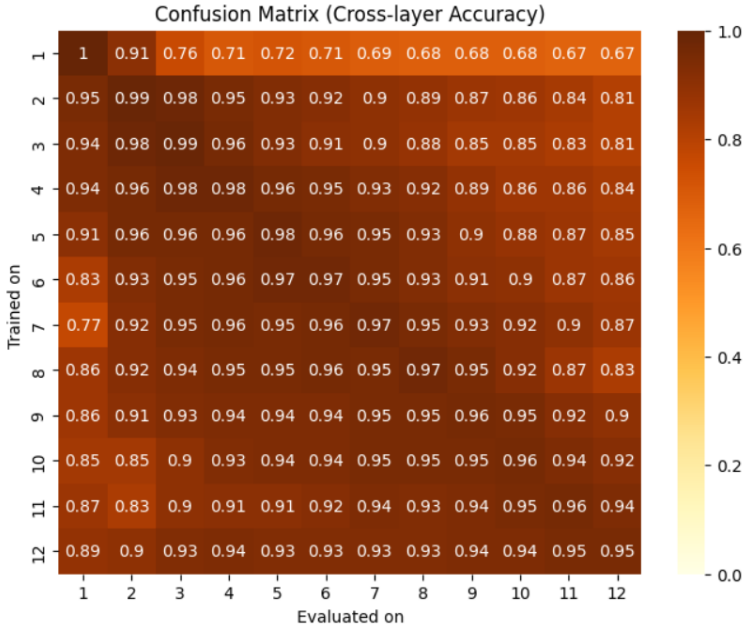
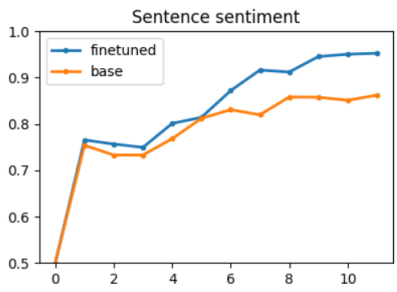
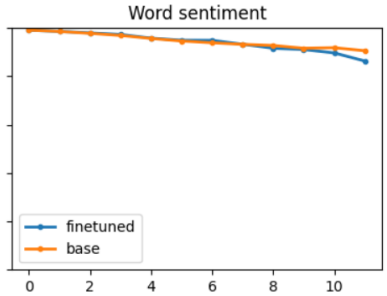


Figure 2: *ToSubmit 3:* The diagonal values, where the layer being trained matches the layer being evaluated, show higher accuracy, suggesting that the features learned at a particular layer are better suited for that same layer. However, the accuracy decreases when features learned at one layer are evaluated on another layer, indicating that these features may not generalize well across layers.



(a) At Sentence-level, the accuracy rapidly increases as the number of layers increases, meaning deeper layers are better in capturing sentence-level sentiments.



(b) At word-level, accuracy slightly decreases as the number of layers increases, meaning word-level sentiments get difficult to capture as the number of layers increase.

Figure 3: *ToSubmit 2:* The probing accuracies over layers plots for the sentence-level (CLS) and the word-level probing results on fine-tuned and base model.

Lab 2: Probing Audio Models

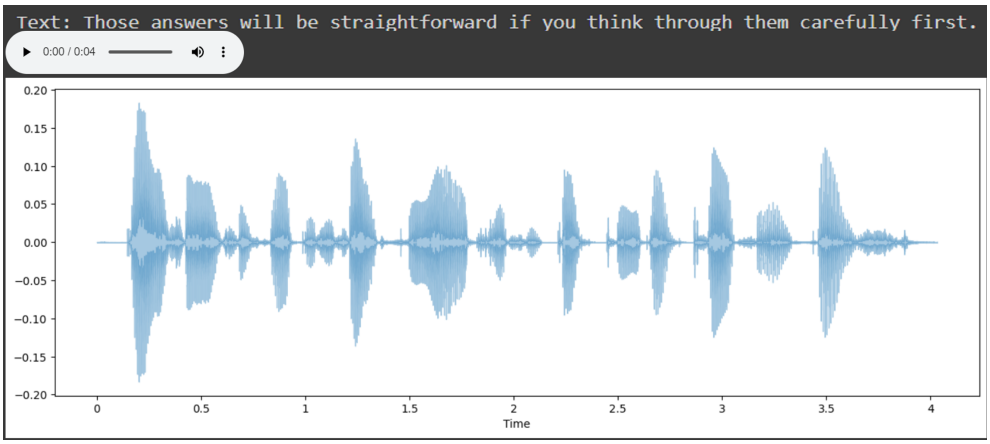


Figure 4: *ToSubmit 1:* The input text and the Waveform signal

```
Input shape: torch.Size([1, 64512])
'THOSE ANSWERS WILL BE STRAIGHTFORWARD IF YOU THINK THROUGH THEM CAREFULLY FIRST'
```

Figure 5: *ToSubmit 1:*The transcription that Wav2Vec2 generated for the above waveform

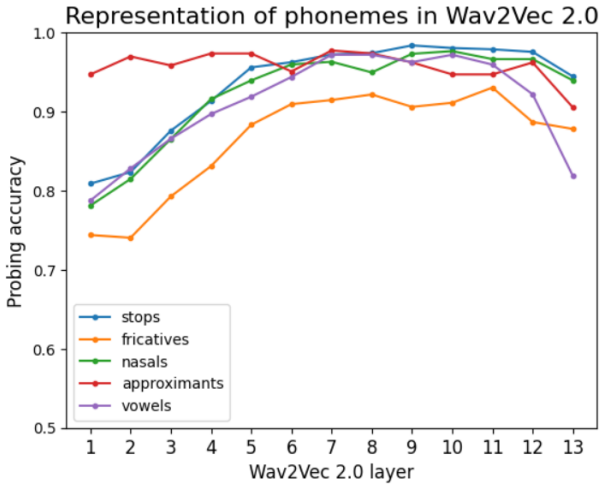


Figure 6: *ToSubmit 2:*The analysis of probing accuracies over layers reveals distinct patterns for different phoneme categories. Specifically, the accuracies for stops, fricatives, nasals, and vowels initially show an increasing trend as the number of layers increases, reaching a plateau, and then slightly declining. On the other hand, the accuracies for approximants start at a higher level and remain relatively stable, with a slight decline observed towards the last layer. It suggests that the model's representations become increasingly specialized and informative as the layers progress, but may reach a point of diminishing returns. This information highlights the model's progression in capturing phoneme representations which potentially can improve Acoustic modelling for the traditional ASR pipeline.