Understanding Self-Attention of Self-Supervised Audio Transformers

Shu-wen Yang\textsuperscript{1}, Andy T. Liu\textsuperscript{12}, Hung-yi Lee\textsuperscript{12}

\textsuperscript{1}College of Electrical Engineering and Computer Science, National Taiwan University
\textsuperscript{2}Graduate Institute of Communication Engineering, National Taiwan University

\{r08944041, f07942089, hungylee\}@ntu.edu.tw

Abstract

Self-supervised Audio Transformers (SAT) enable great success in many downstream speech applications like ASR, but how they work has not been widely explored yet. In this work, we present multiple strategies for the analysis of attention mechanisms in SAT. We categorize attentions into explainable categories, where we discover each category possesses its own unique functionality. We provide a visualization tool for understanding multi-head self-attention, importance ranking strategies for identifying critical attention, and attention refinement techniques to improve model performance.

Index Terms: Self-supervised Learning, Self-attention, Transformer Encoders, Interpretablity

1. Introduction

Adapting the idea of self-supervised learning\cite{1, 2, 3, 4} to continuous speech has received much attention in recent work\cite{5, 6, 7, 8, 9, 10}, where Transformer Encoders with multi-head self-attention\cite{11} are pre-trained on a large amount of audio data in a self-servised scheme. Once pre-trained, they are used to improve various downstream supervised tasks, including phone classification, speaker recognition, SLU, and ASR. Despite the great success of these Self-supervised Audio Transformers (SAT)\cite{1}, their internal attention are often neglected and not explored, as we have little understanding of how they work, or the knowledge they acquire from a large amount of unlabeled data. Understanding how SAT models draw conclusions is crucial for both their improvement and application. In the area of natural language processing (NLP), explaining and interpreting pre-trained black-box models like BERT have been a well-explored topic\cite{12, 13, 14, 15, 16, 17}. However, the analysis of models that are pre-trained on speech has not seen such widespread exploration, and remains an important and challenging endeavor for the speech community.

In this work, we propose to analyze the multi-head self-attention mechanism of SAT through the following methods: visualization, categorization, functionality study, and importance ranking. We found that the self-attentions of SAT models tend to converge into three categories: global attentions, vertical attentions, and diagonal attentions. Diagonal attentions either highly attend to self neighbor or are highly correlated with phoneme boundaries; vertical attentions often concentrate on specific phonemes. As for noisy global attentions, we provide a visualization tool to draw insights about their implicit operations. Through our quantized ranking analysis, we conclude that diagonal attentions outrank the most in terms of importance, followed by vertical attentions. Last but not least, we introduce attention refinement methods which allow us to improve learned representations by moderately removing global attentions or constraining attention span, resulting in a faster inference time and higher performance.

2. Self-Supervised Audio Transformers

The main idea of NLP BERT pre-training\cite{1, 2, 3, 4} is to corrupt the input word tokens by randomly masking or permuting them with a probability policy, layers of Transformer Encoder\cite{11} are trained together with a classifier that estimates the masked words at the output. Primarily inspired by this idea, previous works\cite{5, 7, 6, 8, 9, 10} proposed self-supervised learning for audio with Transformer Encoders. In this work, we refer to these types of models as Self-Supervised Audio Transformers, SAT. Unlike BERT where the inputs are discrete text tokens, the inputs of SATs are acoustic features (e.g., MFCC, FBANK, Mel-Spectrogram), which are continuously long and locally smooth. Also, different from BERT where the model is trained by estimating discrete tokens, SATs change the classifier to a feed-forward network and minimize reconstruction error between real frames and predicted frames.

Among all the variants of SATs, we address our focus on SATs that take continuous acoustic frames as input\cite{5, 7, 6}, rather than ones that operate on quantized discrete vectors of speech\cite{8, 9, 10}. In our analysis, we particularly consider the framework described in the Mockingjay\cite{5} literature, mainly due to its designs to address the continuously long and locally smooth problem of speech. In Mockingjay, two techniques of downsampling and consecutive masking are introduced to resolve these issues. Downsampling is applied on input features to adapt SATs to long sequences. To reduce the length of frames by a factor of $R_{factor}$, consecutive frames of $R_{factor}$ amount are reshaped and stacked into one frame\cite{5, 18, 19}. On the other hand, consecutive masking is applied during pre-training to avoid the model from exploiting the local smoothness of acoustic frames. Instead of masking a single frame, consecutive frames of $C_{num}$ are masked to zero. To study the attentions of SAT models, we use the prevailing framework of the LARGE model described in Mockingjay\cite{5}, which consists of 12 layers of Transformer Encoders. We train three models on the LibriSpeech\cite{20} train-clean-360 subset with identical settings as in\cite{5}, except for $C_{num} \in \{3, 6, 9\}$, and we name them as M3, M6, M9, respectively.

3. Notations

We start by defining notation for self-attention mechanism and SAT representations. Given a length $T$ sequence of vectors $\mathbf{x}=x_1, ..., x_T \in \mathbb{R}^d$, we denote $A_{u}^{h} \in \mathbb{R}^{T \times T}$ as attention weights for all query-key pairs of a head $h$ when propagating an utterance $u$. Hence, $A_{u}^{h}[q, k] \in \mathbb{R}$ is the attention weight of $x_q$ attending to $x_k$. We use $q$ for timestamp of query, $k$ for times-
A head $h$ can be classified into three categories, we define three metrics to quantify the attention maps of a head: (1) **global** attention maps with clear diagonal; (2) **vertical** attention maps with clear vertical lines; and (3) **diagonal** attention maps with block diagonal like Fig 2(a). There are attention maps of clear block diagonal like Fig 2(a). The borders of blocks might be phoneme boundaries, as illustrated in Fig 2(b). It seems diagonal attention knows phoneme boundaries. We conduct phoneme segmentation to examine the correlation. We mainly follow the algorithm proposed in [21], which first calculates a similarity matrix from a sequence of features, containing all pairwise distances between features, and then extract boundary points from the similarity matrix. For segmentation with an attention map, its rows are considered as the feature sequence for computing the similarity matrix. Examples of similarity matrices are shown in Fig 2(c)(d) when segmentation features are the attention maps and MFCCs, respectively. We slightly modify the boundary-point-extraction algorithm in [21] for segmentation. Little performance difference is found between MFCC and $x^0$ (Mel-scale spectrogram).

Fig 2(e) verifies the correlation between block diagonal attentions and phoneme boundaries, that attentions clearly surpass MFCC under the same setting. As for R-value, under the same setting, we hand-pick a visually block diagonal head for each of M3, M6, and M9. We choose MFCC as baseline feature since it is the most prevailing feature for segmentation features are the attentions map and MFCCs, respectively. The reason is that when $C_{num} = 9$ masks are used, the performance will be poorer.

**5. Phoneme Segmentation**

There are attention maps of clear block diagonal like Fig 2(a). The borders of blocks might be phoneme boundaries, as illustrated in Fig 2(b). It seems diagonal attention knows phoneme intervals. We conduct phoneme segmentation to examine the correlation. We mainly follow the algorithm proposed in [21], which first calculates a similarity matrix from a sequence of features, containing all pairwise distances between features, and then extract boundary points from the similarity matrix. For segmentation with an attention map, its rows are considered as the feature sequence for computing the similarity matrix. Examples of similarity matrices are shown in Fig 2(c)(d) when segmentation features are the attention maps and MFCCs, respectively. We slightly modify the boundary-point-extraction algorithm in [21] for segmentation. Little performance difference is found between MFCC and $x^0$ (Mel-scale spectrogram). Fig 2(e) verifies the correlation between block diagonal attentions and phoneme boundaries, that attentions clearly surpass MFCC under the same setting. As for R-value, under the same setting, we hand-pick a visually block diagonal head for each of M3, M6, and M9. We choose MFCC as baseline feature since it is the most prevailing feature for segmentation. Little performance difference is found between MFCC and $x^0$ (Mel-scale spectrogram).
neighbor structure. Figures similar to Fig 2(d) are shown (e) attends to sil (f) not attends to sil (g) attends to ch, jh, s, sh (h) not attends to s, sh

6. Phonoeme Relation Map

To study the functionality of global and vertical heads, we propose to align attentions to phonoeme relations to see whether some heads focus on looking for specific phonoeme relations in the utterances. For a sequence of input features \( \mathbf{x} \), there exists frame-wise phonoeme labels \( y \in Y \), where \( Y \) is a predefined phone set. We consider \( x_{t+h} \) attending to \( x_t \) as when observing phonoeme \( y_t \) the head would look for phonoeme \( y_t \). We quantify a phonoeme relation \( Y_m \rightarrow Y_n \) inside a head \( h \) by summing up all attention weights \( A_h^{k} \) whose phonoeme relation \( y_m \rightarrow y_n \) equals \( Y_m \rightarrow Y_n \), over the entire speech corpus. More specifically, we plot a phonoeme relation map (PRM) \( P_h \in \mathbb{R}^{Y \times |Y|} \) by the following equations:

\[
P_h[m, n] = \mathbb{E}_{x \sim U} \left[ \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} I_{y_m = Y_m, y_n = Y_n, A_h^{k} = [y_m, y_n]} \right] \tag{4}
\]

\[
P_h[m, n] = \frac{P_h[m, n] - P_U[m, n]}{P_U[m, n]} \tag{5}
\]

where \( 1 \leq m, n \leq |Y| \) and \( P_U \in \mathbb{R}^{Y \times |Y|} \) and \( P_U \) is the distribution of all possible phonoeme relations in speech corpus \( U \), normalizing the effect of dominating relations like \( \text{sil} \rightarrow \text{sil} \) which appears in all utterances. As a result, positive values in \( P_h \) represent preference for specific phonoeme relations; negative values represent the opposite.

PRMs are plotted using TIMIT [22] with 39 phonemes, and results of several heads are shown in Fig 3. Since diagonal heads are interpretable themselves, we focus on vertical and global heads. There are several operations: attending to silence, identity, specific phonemes, and not attending to these (not operations). We observe tendency of vertical heads either focus or neglect specific phonemes for all queries, and we bridge their connections. For later discussion, we use focus and neglect to refer to these types of behaviors.

While a PRM characterizes all phonoeme relations of a head, we further define concentration \( C_h \in \mathbb{R}^{|Y|} \) of a head \( h \), where each \( C_h[n] \in \mathbb{R} \) quantifies the amount of focus (when positive) or neglect (when negative) of a head on specific phonoeme \( Y_n \), over all queries:

\[
C_h[n] = \frac{1}{|Y|} \sum_{m=1}^{|Y|} P_h[m, n] \tag{6}
\]

Fig 4 verifies the connection between verticality and concentration. We report the maximum focus or neglect for each head. Fig 4(a) points out that heads with high verticality do focus on specific phonemes; Fig 4(b) points out even a slight increase of the verticality \( V \) of a head has correlation to concentration, for both focus and neglect. Some low-verticality heads with extreme neglect at the bottom-left of Fig 4(b) are diagonal heads, which always attend to their neighbors dynamically and show extreme neglect for all phonemes.

7. Importance Ranking

To evaluate the importance of different attention patterns, we conducted two pruning-based probing experiments. We ablate partial functionality of self-attention directly at inference time in two aspects: (1) ablates an entire head; (2) ablates the visible span for all heads. If an attention pattern is essential, ablating it should exhibit immediate loss in terms of the quality of final evaluations. We examine representation quality by three probing tasks: spectrogram reconstruction, phonoeme classification, and speaker recognition. For the first task, we examine the richness in terms of spectrogram details of refined representations. We reuse the reconstruction head during pre-training and measure L1 loss compared to the original. For the latter two tasks, we examine the usefulness of refined representations on downstream tasks. For phonoeme and speaker classifications, we train the downstream models using LibriSpeech [20] train-clean-100 subset and fixed 50k steps. In frame-level setting, we use single-hidden MLP; in utterance-level setting, we use mean-pooling followed by a linear transform. Phonoeme classification is conducted under frame-level setting; speaker recognition is under frame-level and utterance-level. Following [5], phonoeme labels are obtained by the Montreal Force Aligner [28], and all evaluations are done on the LibriSpeech test-clean subset.

7.1. Head-based Pruning

For each head \( h \), we first compute values of \( G(h), V(h), D(h) \), and then cumulatively prune heads from high to low for each metric by setting \( A_h^0 = 0 \), resulting in three curves as shown in Fig 5(a)(b)(c). We can thus rank the importance of the three categories by observing which pruning results in a larger performance drop. We find ranking results are consistent for different \( C_{team} \), so we only show the result of M3. There are several...
findings: (1) Diagonal heads are the most important. Performances on all three tasks drop significantly with only 24 heads pruned. (2) Vertical heads rank second. While pruning them does not hurt reconstruction or phoneme classification much, it drops faster than global heads in speaker recognition. This suggests that vertical attentions have more relation to speaker identity. (3) Global heads have the least importance that pruning them has the least effect on all tasks. (4) Both global and vertical heads are harmful to the phonetic structure. Fig 5(b) shows that pruning them even improve the phoneme classification accuracy. For vertical heads, we speculate that the vertical heads might focus on distant phonemes when forming a new representation independently (disrespectfully) of query phoneme, which might corrupt the local phonetic structure. (5) In Fig 5(b), when we prune according to diagonality, phoneme accuracy drops dramatically for the first 24 heads pruned, while it surprisingly increases as we prune more heads. This is because when pruning more than 24 diagonal heads, we start to prune the heads that are more vertical or global than diagonal, supporting the previous finding that vertical and global attentions are harmful for phonemic information.

We further show the result of ranking the importance of heads based on their maximum attention weights, denoted as weight in Fig 5(a)(b)(c), which has been shown to be a strong baseline in the previous work [29]. Fig 5(c) shows pruning based on globalness has less influence than weight. Fig 6(a) visualizes the difference between two ranking strategies. Although they agree on which heads are essential, they slightly diverge on which are not. Their decision boundaries in terms of the first 24 heads to be pruned are shown by red and blue lines in Fig 6(a), which is the direct cause of their performance differences. Globalness prunes blue dots while leaving red dots unpruned; and vise versa for weight (while they all prune green dots). Since globalness performs better than weight after pruning, this suggests that heads of red dots are more important than blue dots. We select the head with the highest ranking difference from both red and blue dots, and plot their PRMs in Fig 6(b) and (c), respectively. We find that while Fig 6(b) shows strong neglect, (c) does not possess observable operation. In fact, heads of red dots are mostly with clear neglect\(^*\). We argue that this is the main reason why globalness performs better after pruning, that heads with neglect are essential to speaker identity, and globalness defined by entropy is able to recognize neglect and score them higher. On the other hand, weight is confused by attentions with large weights but without meaningful operation, suggesting that weight do not always reflect the importance of heads. When we plot the PRMs for dots located at upper left (the red box in Fig 6(a)), they typically show neglect. These observations are consistent for all M3, M6, M9. We speculate that these heads might learn to neglect less useful frames, like sil in Fig 6(b), and focus more on other frames with more speaker information [30]. Based on the above observations, we choose globalness as our refinement metric. Fig 5(d)(e)(f) show pruning results for M3, M6, M9. The importance of global heads become less for larger \(C_{max}\), and we keep observing performance boost for phoneme classification. Despite all three models drop for speaker recognition, the drop is mitigated dramatically in utterance-level setting (a more common scenario), suggesting that global heads are not necessary when speaker classification is performed on utterance level. In conclusion, we can prune SAT heads for more than 50% without sacrificing performance.

### 7.2. Span-based Pruning

Since most of the heads have attention span over a long range (no matter what category it belongs to), we further conduct attention-span pruning to examine if global information is genuinely not helpful for extracting phonetic information. We limit the visible span of all heads by length \(r\), either to the left or right. That is, we set \(A^h_{r}(q, k) = 0\) for any \(|q - k| > r\). Results are presented in Fig 5(g)(h)(i).

### 8. Conclusion

In this paper, we present multiple strategies for analyzing the self-attention mechanism in SATs, including phoneme segmentation, phoneme relation map, and two aspects of pruning. We find several attention functionality and operations. We identify critical attentions and show our visualization tool useful for understanding pruning behavior. Finally, we conclude that we can refine representations and speed up inference time for a given SAT in two aspects: removing global heads or constraining attention span. More materials are provided in our demo page².
9. References


