Similarity and Content-based Phonetic Self Attention for Speech Recognition

Kyuhong Shim and Wonyong Sung

Department of Electrical and Computer Engineering, Seoul National University, Korea

Abstract

Transformer-based speech recognition models have achieved great success due to the self-attention (SA) mechanism that utilizes every frame in the feature extraction process. Especially, SA heads in lower layers capture various phonetic characteristics by the query-key dot product, which is designed to compute the pairwise relationship between frames. In this paper, we propose a variant of SA to extract more representative phonetic features. The proposed phonetic self-attention (phSA) is composed of two different types of phonetic attention: one is similarity-based and the other is content-based. In short, similarity-based attention captures the correlation between frames while content-based attention only considers each frame without being affected by other frames. We identify which parts of the original dot product equation are related to two different attention patterns and improve each part with simple modifications. Our experiments on phoneme classification and speech recognition show that replacing SA with phSA for lower layers improves the recognition performance without increasing the latency and the parameter size.

Index Terms: speech recognition, self attention, transformer, phoneme classification, phonetic attention

1. Introduction

End-to-end automatic speech recognition (ASR) has made great progress in line with the advances in deep neural networks (DNNs). Among various architectures, Transformer [1] models have shown state-of-the-art performance [2, 3, 4, 5, 6] in ASR. Most Transformer-based ASR models stack the same layer multiple times without considering the difference between layer positions, although the behaviors are very different [7, 8, 9]. If we can identify the role of each layer, we can improve the model architecture by exploiting domain-specific knowledge, resulting in a more heterogeneous composition of layers. However, because end-to-end DNNs performs as a black box, it is difficult to design and apply specific modifications for relevant layers.

Recently, a study suggested that the role of self-attention (SA) in Transformer-based ASR models can be distinguished into two types, phonetic and linguistic localization [10]. Two roles contribute to speech recognition in a row: the ASR system first extracts phonetically meaningful features by reducing the pronunciation variations and then combines such information into textual features to produce natural output sentences. These two-stage processes, which correspond to phonetic and linguistic localization, seem to be natural because ASR is a many-to-one problem in that multiple speeches can be transcribed as the same text. The study discovered that phonetic localization mainly appears in lower layers while linguistic localization happens in upper layers [10], and their attention patterns are also very different. The findings imply that we can identify layers of a certain role, and we may boost the performance by improving such layers to perform their role better.

Among the two types of roles mentioned above, we focus on improving phonetic localization based on a deeper understanding of the behavior. Here, we call SA heads that perform phonetic localization a phonetic (attention) head. From the observation of the attention weights produced by phonetic heads, we can separate two distinct types of attention patterns. The first type is similarity-based phonetic attention that gives a larger attention weight value on similarly pronounced frames. For example, frames corresponding to phoneme class ‘S’ often show large attention weight for frames corresponding to ‘S’, ‘Z’, ‘SH’, and vice versa [10]. The second type is content-based phonetic attention that attends to certain phonemes regardless of the query. In other words, a certain attention head may be highly optimized for detecting a specific phoneme class. We suggest that each phonetic head can be more specialized from the decomposition of similarity-based and content-based attention mechanisms.

In this paper, we propose phonetic self-attention (phSA), a variant of SA that extracts similarity and content-based phonetic features in phonetic localization. We modify the query-key dot product term inside the SA mechanism to capture similarity and content separately. In particular, we improve the dot product by (1) decomposing the two terms to remove shared parameters and (2) inserting trainable non-linearity functions. We evaluate the proposed phSA using phoneme classification and speech recognition and achieve considerably improved recognition performance on both tasks. In addition, we empirically show that similarity-based and content-based phonetic attention produce relatively concentrated and distributed attention probabilities, respectively.

2. Motivation

2.1. Dot Product in Self Attention

Self-attention (SA) is the key component of Transformer that computes the relationship between every pair of frames. For a sequence of speech frame features \( X = \{x_1, x_2, ..., x_T\} \) as an input, SA first projects features into three components, namely query \((Q)\), key \((K)\), and value \((V)\). SA utilizes multiple attention heads with different parameters to capture diverse relationships in each layer. Without loss of generality, we explain the behavior of a single attention head below. \( Q, K, V \) are linear projections of input as follows:

\[
Q, K, V = XW_{(Q, K, V)} + b_{(Q, K, V)} \tag{1}
\]

where \( X \in \mathbb{R}^{T \times d_h} \), \( W \in \mathbb{R}^{d_h \times d_h} \) and \( b \in \mathbb{R}^{1 \times d_h} \) are input, weight, and bias, respectively. \( d_h \) is the dimension of each attention head. The attention map \( A \) is then calculated as:

\[
A = \text{Softmax} \left( \frac{Q^T K}{\sqrt{d_h}} \right) \in \mathbb{R}^{T \times T}. \tag{2}
\]

Each element of the attention map represents how much one frame focuses on the other one, which is, in practice, imple-
Figure 1: Visualization of PAR from selected attention heads in the baseline model. Two rows show representative examples of similarity-based and content-based phonetic attention, respectively. Brighter points indicate higher attention weight.

The relative positional encoding (RPE) has been widely used for Transformer models for ASR \[2, 3, 13, 14\]. For exam-
Table 1: Phoneme classification accuracy (%) of different dot product variants evaluated on LibriSpeech dataset. M2 is the dot product of the original self-attention, and M5 is the dot product of the proposed phonetic self-attention.

| Model | Dot-product | dev-clean | dev-other | test-clean | test-other |
|-------|-------------|-----------|-----------|------------|------------|
| M1    | $(XW_Q)(XW_K)^T$ | 81.92     | 73.42     | 81.86      | 73.63      |
| M2    | $(XW_Q)(XW_K)^T + (XW_Kb^T)^T$ | 81.84     | 73.37     | 81.79      | 73.55      |
| M3    | $(XW_Q)(XW_K)^T + (Xe)^T$ | 81.93     | 73.26     | 81.82      | 73.52      |
| M4    | $(XW_Q)(XW_K)^T + (\phi(XW_C)c^T)^T$ | 82.40     | 73.89     | 82.25      | 74.20      |
| M5    | $\psi_e((XW_Q)(XW_K)^T) + \psi_e(\phi(XW_C)c^T)^T$ | 82.66     | 74.20     | 82.53      | 74.48      |

4.2. Phoneme Classification

To evaluate the phonetic feature extraction performance, we train the models for phoneme classification. Table 1 compares the vanilla SA (M2), phSA (M5), and other variants. M2 is the original dot-product, and M1 is the same version without bias parameter that only focus on similarity-based relationships. M3 is identical to the M2 but differs in the implementation that the parameter $W_K$ is not shared. M1, M2, and M3 show almost similar accuracy with less than 0.1% difference. In contrast, M4 shows a noticeable gain in phoneme classification accuracy compared to M2 and M3. The proposed phSA (M5) achieves the highest accuracy among the dot-product variants. The results verify that our architectural modifications, M2 → M4 (Sec. 3.1) and M4 → M5 (Sec. 3.2), each contributes to better phonetic feature extraction.

4.3. Speech Recognition

Table 2 shows the end-to-end speech recognition performance with the proposed phSA. Compared to the baseline, replacing the vanilla SA to phSA reduces the word error rate (WER) on every data subset, especially for the challenging LibriSpeech dev-other and test-other datasets. We empirically show that adopting phSA only for lower layers (under 8-th layer) achieves the best performance. This observation is aligned with previous analysis [10] that the lower layers more focus on phonetic information than upper layers. For example, replacing phSA for lower 6 layers decreases the WER from 8.23% to 7.77% (5.6% relative reduction) and 8.21% to 7.93% (3.4% relative reduction) for dev-other and test-other datasets, respectively. In contrast, utilizing phSA for 12 layers shows worse performance than the baseline, and using phSA for the entire (16) layers suffers from significant performance degradation.
Table 3: Effect of similarity-based (S) and content-based (C) attention by removing each component. Entropy (mean ± std) of phSA attention maps and word error rate (%) are reported.

|    | Entropy | dev-other | test-other |
|----|---------|-----------|------------|
| ✓  | 1.91 ± 0.12 | 7.77 | 7.93 |
| ✓  | 2.02 ± 0.14 | 8.16 (+0.39) | 8.54 (+0.61) |
| ✓  | 2.39 ± 0.04 | 9.20 (+1.43) | 9.35 (+1.42) |

4.4. Discussion

4.4.1. Speed and Parameter Size

Although we add several new computation steps for phSA, we observe that the training and inference time does not change much. The main reason is that the removal of RPE can compensate for the additional cost of phSA, in both latency and parameter size. For example, the wall-clock training time of the phSA (2nd row in Table 2) is about 5% faster than the baseline (1st row) with almost the same number of parameters.

4.4.2. Comparison of Similarity and Content

To understand the relative importance between similarity-based and content-based attention, we evaluate the recognition performance without each component. Table 3 presents the word error rate of the model using only similarity-related or content-related computation. Specifically, we fix the parameters of the converged model with 6 phSA layers and discard either term of the phSA dot product (Eq. (8)). Removing similarity-based attention (bottom row of Table 3) degrades the performance more than removing content-based one (top row of Table 3), which implies that the phonetic features extracted from similarity are more important than content-based attention; however, both are indispensable for speech recognition.

In addition, we calculate the average per-head entropy of attention probability for two settings and observe the meaningful difference. Similarity-based attention probabilities are more concentrated (lower entropy) and content-based attention probabilities are more distributed (higher entropy). In other words, the similarity-based term emphasizes the difference while the content-based term enhances the uniformness. We believe that the proposed phSA encourages two terms to be specialized for different attention patterns.

4.4.3. PReLU Negative Slope

The range of PReLU negative slopes is very different for similarity-based and content-based terms after training. Figure 2 shows the negative slopes of the phSA after training. Four dots in each layer indicate PReLU parameters in four attention heads. x-axis indicates the layer index. Note that the range of y-axis is very different, (0 ~ 14) for (a) and (0 ~ 0.8) for (b).

4.5. Related Work

Architectural modifications for Transformer-based ASR models have been of great interest. Many works focus on reducing the heavy computational cost caused by SA [20, 21, 22, 23]. For example, Efficient Conformer [20] proposed grouped SA and downsampling block to shorten the length of the sequence to be processed. Our work is very distinct from previous works in two points. First, phSA is designed to enhance the quality of intermediate feature representation, therefore improving the recognition performance. Second, only a lower part of the model is changed to phSA so that the model utilizes two different types of self-attention mechanisms together.

Pretraining-based approaches have been proven effective in improving the ability to capture useful phonetic information for various downstream tasks. For example, Wav2Vec2.0 [24], XLSR [25], TERA [26], and ACPC [27] presented various self-supervised speech pretraining methods and showed that phonologically meaningful features can be captured while learning the general characteristics of speech. However, these models use identical Transformer architecture for every layer without considering the different behaviors of each. Explicit pretraining objectives have also been introduced for learning the useful phonetic features during pretraining. For example, UniSpeech [28] and BERTphone [29] exploited CTC loss using phoneme sequence as label. The drawback of the aforementioned studies is that they require an additional pretraining stage before fine-tuning the model for ASR.

6. Conclusion

In this paper, we proposed a variant of self-attention (SA), named phonetic self-attention (phSA), to improve the ASR performance. Especially, we investigated the phonetic behavior of attention heads and distinguished two different attention patterns, similarity-based and content-based attention. The proposed phSA emphasized the two behaviors by applying simple and effective modifications to the original dot-product in SA. In addition, the effect of each behavior is controlled by additional trainable parameters. From the phoneme classification experiments, we showed that phSA is more suitable than the vanilla SA for phonetic feature extraction. By replacing SA in lower layers with phSA, we improved the speech recognition performance on the end-to-end Transformer-based ASR model.

7. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korea government (MSIT) (No. 2021R1A2C1013513). This work was also supported in part by Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd.
8. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[2] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, “Transformer transducer: A streamable speech recognition model with transformer encoders and mni-t loss,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7829–7833.

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in Proc. Interspeech 2020, 2020, pp. 5036–5040.

[4] E. G. Ng, C.-C. Chiu, Y. Zhang, and W. Chan, “Pushing the limits of non-autoregressive speech recognition,” arXiv preprint arXiv:2104.03416, 2021.

[5] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” arXiv preprint arXiv:2010.10504, 2020.

[6] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, “Speechstew: Simply mix all available speech recognition data to train one large neural network,” arXiv preprint arXiv:2104.02133, 2021.

[7] S. Yang, A. T. Liu, and H. yi Lee, “Understanding self-attention of self-supervised audio transformers,” in Proc. Interspeech 2020, 2020, pp. 3785–3789.

[8] S. Zhang, E. Lowezmi, P. Bell, and S. Renals, “On the usefulness of self-attention for automatic speech recognition with transformers,” in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 89–96.

[9] -----, “Stochastic attention head removal: A simple and effective method for improving transformer based asr models,” arXiv preprint arXiv:2011.04004, 2020.

[10] K. Shim, J. Choi, and W. Sung, “Understanding the role of self-attention for efficient speech recognition,” in International Conference on Learning Representations, 2022.

[11] S. Elfwing, E. Uchibe, and K. Doya, “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning,” Neural Networks, vol. 107, pp. 3–11, 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[13] N.-Q. Pham, T.-L. Ha, T.-N. Nguyen, T.-S. Nguyen, E. Salesky, S. Stuker, J. Niehues, and A. Waibel, “Relative Positional Encoding for Speech Recognition and Direct Translation,” in Proc. Interspeech 2020, 2020, pp. 31–35.

[14] T. Likhomanenko, Q. Xu, G. Synnaeve, R. Collobert, and A. Ro gozhnikov, “Cape: Encoding relative positions with continuous augmented positional embeddings,” Advances in Neural Information Processing Systems, vol. 34, 2021.

[15] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhut dinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2978–2988.

[16] J. Park, C. Kim, and W. Sung, “Convolution-based attention model with positional encoding for streaming speech recognition on embedded devices,” in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 30–37.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2022, pp. 5206–5210.