



# Analysis of Self-Attention Head Diversity for Conformer-based Automatic Speech Recognition

Kartik Audhkhasi, Yinghui Huang, Bhuvana Ramabhadran, Pedro J. Moreno

Google LLC, New York

{kaudhkhasi, huangyinghui, bhuv, pedro}@google.com

## Abstract

Attention layers are an integral part of modern end-to-end automatic speech recognition systems, for instance as part of the Transformer or Conformer architecture. Attention is typically multi-headed, where each head has an independent set of learned parameters and operates on the same input feature sequence. The output of multi-headed attention is a fusion of the outputs from the individual heads. We empirically analyze the diversity between representations produced by the different attention heads and demonstrate that the heads become highly correlated during the course of training. We investigate a few approaches to increasing attention head diversity, including using different attention mechanisms for each head and auxiliary training loss functions to promote head diversity. We show that introducing diversity-promoting auxiliary loss functions during training is a more effective approach, and obtain WER improvements of up to 6% relative on the LibriSpeech corpus. Finally, we draw a connection between the diversity of attention heads and the similarity of the gradients of head parameters.

**Index Terms:** automatic speech recognition, multi-headed attention, transformer

## 1. Introduction

Attention [1] has become an all-pervasive technology in modern neural network models used in several areas, such as natural language processing and machine translation [1–4], computer vision [5–9], and automatic speech recognition (ASR) [10–14]. A sequence-to-sequence learning problem entails training a model to take a data sequence as input and predict a data sequence as the output. A typical multi-layer neural network to solve such a problem predicts a sequence of output representations given a sequence of input representations at each layer. For each output time step, an *attention mechanism* in a layer assigns a probability density function over the input time sequence, where the probability of any input time step denotes the weight (attention) assigned to it while producing the output representation.

Self-attention [15] is the most popular form of attention, and uses the input sequence itself to derive the attention mechanism. We review dot product self-attention in the next section. Transformers [15] replaced recurrent and convolutional neural networks with self-attention and feed-forward layers, and offered significant speed-ups in training and inference time while providing quality improvements. It is common practice to use multiple attention mechanisms (*heads*) [16] in parallel in each layer and then combine their outputs. Each attention head uses an independent set of learned parameters. The intuition behind using multiple heads is that each head is free to attend to different regions of the input sequence and hence provides diverse representations. Vaswani et al. [15] show a text example where each attention head attends to a different input dependency for the output word (“making”). Some prior work has

investigated the redundancy of attention heads for NLP tasks. Voita et al. [17] show that in a Transformer model, all except a few key attention heads can be pruned without a significant drop in performance on a English-Russian machine translation task. Kovaleva et al. [18] make a similar conclusion about BERT attention heads on the GLUE tasks.

A few works in the NLP (especially machine translation) literature have attempted to explicitly promote attention heads to be diverse during model training. Lin et al. [16] proposed adding a L2 penalty to the training loss that promotes the attention probabilities from different heads to be orthogonal. They evaluated the models on 3 different NLP tasks – author profiling, sentiment classification and textual entailment. Li et al. [19] proposed similar loss terms and showed improvements on English-German and Chinese-English machine translation tasks. Chien et al. [20] presented a disentangled masked attention that represents each attention head’s probability distribution by a latent topic model. They additionally minimized an upper-bound on the mutual information between query vectors from pairs of attention heads to encourage them to be diverse. Correia et al. [21] proposed a transformer model with sparse attention heads and adaptive sparsity. Their machine translation experiments showed that the resulting attention heads have better diversity than the baseline, and also obtain performance improvements. An et al. [22] gave a Bayesian view of multi-headed attention and used particle filtering for model training.

To the best of our knowledge, prior work has not analyzed and explicitly improved the diversity of attention heads for the state-of-the-art Conformer [23] (convolution-augmented Transformer) model on automatic speech recognition (ASR). One related work is by Lohrenz et al. [24], who propose smoothing the attention probability distribution for all heads by interpolating with a uniform distribution, and obtain WER improvements on LibriSpeech. We make the following contributions:

1. We present an in-depth empirical analysis of the diversity of Conformer attention heads on the public LibriSpeech corpora and show that multiple attention heads indeed become significantly correlated during training.
2. We experiment with a few approaches to increasing attention head diversity such as using different attention mechanisms for each head and introducing diversity-promoting loss functions during training. We show that the latter gives up to 6% relative improvement in WER.
3. We analyze the connection between attention head diversity and similarity between gradients of head parameters, and conclude that more diverse heads have less correlated gradients.

The next section gives an overview of multi-headed self-attention. Section 3 presents various diversity losses we used for measuring and promoting attention head diversity during

model training. We present our experiments in Section 4 and conclude the paper in Section 5.

## 2. Multi-Headed Self-Attention

We first describe the standard dot product multi-headed self-attention to set the notation and for completeness. Let  $\mathbf{X}$  be the  $T \times D$  matrix of  $D$ -dimensional input vectors over  $T$  time steps. Let  $N$  denote the number of attention heads. The  $n^{\text{th}}$  head computes the following three matrices:

$$\begin{aligned} \mathbf{Q}_n &= \mathbf{X}\mathbf{W}_n^{\text{query}} && \text{(Query)} \\ \mathbf{K}_n &= \mathbf{X}\mathbf{W}_n^{\text{key}} && \text{(Key)} \\ \mathbf{V}_n &= \mathbf{X}\mathbf{W}_n^{\text{value}} && \text{(Value)} \end{aligned} \quad (1)$$

where  $\mathbf{W}_n^{\text{query}}$ ,  $\mathbf{W}_n^{\text{key}}$  and  $\mathbf{W}_n^{\text{value}}$  are  $D \times H$  weight matrices that transform the input matrix  $\mathbf{X}$  into three  $T \times H$  matrices –  $\mathbf{Q}_n$  (query),  $\mathbf{K}_n$  (key), and  $\mathbf{V}_n$  (value). We then compute dot product between the query and key matrices followed by row-wise softmax to obtain the  $T \times T$  attention probability matrix

$$\mathbf{A}_n = \text{softmax}(\mathbf{Q}_n \mathbf{K}_n^T / H). \quad (2)$$

The value matrix is then multiplied by the attention probability matrix to obtain the  $T \times H$  output *context* matrix from the  $n^{\text{th}}$  head:

$$\mathbf{Y}_n = \mathbf{A}_n \mathbf{V}_n. \quad (3)$$

The  $t^{\text{th}}$  row of  $\mathbf{Y}_n$  is a convex combination of all rows of  $\mathbf{V}_n$  using attention weights given by the  $t^{\text{th}}$  row of  $\mathbf{A}$ . Next, the output matrices from the  $N$  attention heads are concatenated along rows and down-projected to produce the final  $T \times D$  output of the attention layer:

$$\mathbf{Y} = [\mathbf{Y}_1 \ \mathbf{Y}_2 \ \dots \ \mathbf{Y}_N] \mathbf{W}^{\text{out}} \quad (4)$$

where  $\mathbf{W}^{\text{out}}$  is the  $NH \times D$  projection matrix. The next section discusses various losses we used for measuring and promoting self-attention head diversity.

## 3. Measuring and Promoting Attention Head Diversity

For the sake of illustration, consider the computation of diversity between the  $T \times H$  output representations (*context vectors*)  $\mathbf{Y}_m$  and  $\mathbf{Y}_n$  produced by the  $m^{\text{th}}$  and  $n^{\text{th}}$  attention heads. We compute the correlation coefficient between two heads  $m$  and  $n$  as

$$d^Y(m, n) = \frac{1}{T} \text{sum}(\tilde{\mathbf{Y}}_m \odot \tilde{\mathbf{Y}}_n) \quad (5)$$

where  $\tilde{\mathbf{Y}}_m$  and  $\tilde{\mathbf{Y}}_n$  denote the matrices of row unit vectors,  $\odot$  denotes the element-wise product of two matrices and  $\text{sum}$  is the sum of all entries. Given an  $N \times N$  matrix of the above correlation coefficients, we would like to compute a scalar loss that captures the average diversity of the  $N$  heads. Each  $d^Y(m, n) \in [-1, 1]$  and correlation coefficients of both 1 and  $-1$  would correspond to minimal diversity. In other words, perfectly diverse heads produce perfectly decorrelated/orthogonal representations. Hence, similar to [16], we define the following diversity loss:

$$\mathcal{L}^{\text{diversity}} = \frac{1}{N^2} \sum_{m=1}^N \sum_{n=1}^N (d^Y(m, n) - I(m, n))^2 \quad (6)$$

where  $I(m, n)$  denotes the  $(m, n)^{\text{th}}$  entry of the identity matrix. This loss is proportional to the Frobenius norm of the difference between the  $N \times N$  diversity loss and identity matrices.

Representation Name	Diversity loss
Context ( $\mathbf{Y}$ )	$d^Y(m, n)$
Attention Probability ( $\mathbf{A}$ )	$d^A(m, n)$
Query ( $\mathbf{Q}$ )	$d^Q(m, n)$
Key ( $\mathbf{K}$ )	$d^K(m, n)$
Value ( $\mathbf{V}$ )	$d^V(m, n)$

Table 1: Various attention diversity losses used in this paper.

Multi-headed attention provides several representations for computing the diversity loss as defined above. Table 1 lists the diversity losses we used. In order to promote diversity during training, we add the diversity loss as an auxiliary loss function:

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{RNNT}} + \lambda \mathcal{L}^{\text{diversity}} \quad (7)$$

where  $\lambda$  is a scalar hyper-parameter and  $\mathcal{L}^{\text{RNNT}}$  denotes the Recurrent Neural Network Transducer (RNNT) loss, defined as the negative log-likelihood of the output label sequence given the input acoustic feature sequence. We also considered using the negative Kullback-Liebler (KL) divergence as the diversity loss  $d^A(m, n)$  between the attention probability matrices. However, that would lead to unstable training because we would like to maximize KL divergence and it is unbounded from above.

## 4. Experiments

### 4.1. Dataset and Model Architecture

We use the well-benchmarked, publicly-available Librispeech [25] corpus for all our experiments. Librispeech comprises of read 960 hours of speech from audiobooks from over 2000 speakers. We did not include any additional audio or text data in our experiments.

The Conformer-RNNT model follows an encoder-decoder architecture detailed in [23] and consists of a 17-layer Conformer acoustic encoder with self-attention in all layers, a 2-layer unidirectional LSTM decoder, and a joint network. The full network is trained end-to-end using the RNN-T loss. The Conformer acoustic encoder uses multi-headed attention over the whole utterance with 8 attention heads. We use a model dimension of 512 for the encoder, which results in a total of 118M parameters. The input acoustic features for our LibriSpeech experiments are the same as used in [23]. The target vocabulary for all models are wordpieces with a vocabulary of 1024. All models are randomly-initialized and trained with an effective batch size of 4096 in Lingvo [26] on Tensor Processing Unit slices. Our ablation experiments did not use relative positional embedding in the attention layer. We later show results with one setting on a model with relative positional embedding.

Table 2 shows several baseline models with different attention context sizes. We next present the analysis of attention head diversity for the full context baseline system.

### 4.2. Analysis of Attention Head Diversity

We computed various attention head diversity losses presented in (6) and Table 1 using the baseline model for the Librispeech dev and test sets at the convergence of training. Note that these scores are summed over all 17 layers of the Conformer acoustic encoder.

Attention	dev	dev-other	test	test-other
Full context	2.0	5.1	2.2	5.2
L=256, R=256	2.1	5.0	2.2	5.2
L=128, R=128	2.1	5.3	2.2	5.3
L=64, R=64	2.1	5.7	2.4	5.7
L=32, R=32	2.2	6.0	2.4	6.0
L=16, R=16	2.3	6.7	2.4	6.5

Table 2: Baseline WERs for different left (L) and right (R) context sizes measured in terms of time steps.

Diversity loss	dev	dev-other	test	test-other
$d^A(m, n)$	6.37	6.02	6.31	6.10
$d^Q(m, n)$	0.53	0.59	0.54	0.55
$d^K(m, n)$	0.57	0.61	0.61	0.58
$d^V(m, n)$	0.13	0.14	0.13	0.14
$d^Y(m, n)$	0.19	0.20	0.20	0.20

Table 3: Attention diversity losses summed over all layers of the Conformer acoustic encoder for the baseline full-context Librispeech model.

We conclude from Table 3 that the attention probabilities from the various heads show the highest diversity loss and hence the highest correlation. This implies that the different attention heads are often focusing on the same frames of the input sequence. The query and key diversity losses are the next highest, though significantly lower than the attention probability diversity loss. The value and context vector diversity losses are the lowest. We next discuss a simple method to introduce attention head diversity – by using a mixture of different attention mechanisms.

### 4.3. Mixture of Different Attention Mechanisms

Intuitively, one would expect that using different attention mechanisms for subsets of heads would automatically generate enough diversity, and would result in a better WER than the baseline. In order to check this hypothesis, we trained our models with the following mixtures of different attention mechanisms:

1. **Different left and right contexts** - We mixed some attention mechanisms with different left (L) and right (R) context widths from Table 2.
2. **Multi-headed softmax and FAVOR attention** - We mixed full context multi-headed softmax attention and FAVOR attention, which is used in Performers. FAVOR attention [27] approximates softmax attention by computing query-key similarity via dot product in a random kernel space. Since FAVOR attention uses a very different way to compute dot products versus standard softmax attention, we hoped that combining the two would naturally lead to some diversity.

Table 4 shows the WERs for a few relevant baselines from Table 2 with the same form of attention mechanism used for all 8 heads. It also shows the WERs for a few attention mixtures. We observe that none of the attention mechanism mixtures are able to improve the WER over the baseline models. This indicates that simply combining different attention mechanisms is not enough to ensure diversity of the attention heads.

Attention	dev	dev-other	test	test-other
<b>:Number of Heads</b>				
<i>Full context</i>				
Full-ctx Softmax: 8	2.0	5.1	2.2	5.2
Full-ctx Softmax: 4	2.0	5.1	2.2	5.1
FAVOR: 4				
<i>Limited context</i>				
(L=256, R=256): 8	2.1	5.0	2.2	5.2
(L=256, R=256): 4	2.1	5.3	2.3	5.4
(L=256, R=0): 4				
(L=128, R=128): 8	2.1	5.3	2.2	5.3
(L=128, R=128): 2	2.1	<b>5.2</b>	2.3	5.4
(L=64, R=64): 2				
(L=32, R=32): 2				
(L=16, R=16): 2				

Table 4: WERs for mixtures of different attention mechanisms. The rows with all 8 attention heads using the same attention mechanism are the baselines.

### 4.4. Diversity-promoting Auxiliary Training Losses

Next, we explored the impact of explicitly introducing diversity promoting auxiliary loss functions during model training. We focused on the full context softmax Conformer model in this section. We experimented with each diversity loss in Table 1 as an auxiliary loss function during RNNT training as shown in (7). We tuned the weight  $\lambda$  of this auxiliary loss on the dev and dev-other test sets. Table 5 shows the Librispeech WERs. We observe that including diversity losses seems to consistently improve the WER over the baseline model.

Diversity score	dev	dev-other	test	test-other
Baseline	2.0	5.1	2.2	5.2
$d^A(m, n)$	2.0	<b>4.8</b>	<b>2.1</b>	<b>5.1</b>
$d^Y(m, n)$	2.0	<b>4.8</b>	2.2	<b>5.1</b>
$d^Q(m, n)$	2.0	<b>4.8</b>	<b>2.1</b>	<b>5.0</b>
$d^K(m, n)$	2.1	<b>4.9</b>	<b>2.1</b>	<b>5.1</b>
$d^V(m, n)$	2.1	<b>4.9</b>	<b>2.1</b>	<b>5.0</b>

Table 5: WERs using different diversity scores as auxiliary loss functions during training. All models use the full context.

Next, we analyzed the impact of including the diversity auxiliary loss function during training on the diversity losses themselves. Table 6 shows the results. When compared to Table 3, we observe that all the diversity losses are significantly reduced, indicating that the different attention heads are indeed diverse. Attention probabilities are the most directly-interpretable representations computed in an attention layer. The reduction in its diversity loss from approximately 6.0 (Table 3) to 0.4 (Table 6) indicates that the different attention heads are now focusing on different subsets of input frames.

To further understand the impact of training with the attention probability diversity loss, we computed the  $N \times N$  cosine similarity matrix of the  $N$  attention heads for each Conformer layer on a Librispeech utterance. Figure 1 shows the  $8 \times 8$  cosine similarity matrix for the even-numbered layers of the Conformer model using both the baseline model and the model trained with attention probability diversity loss. We observe that the baseline model contains highly correlated and redundant attention heads. In particular the final layer (layer

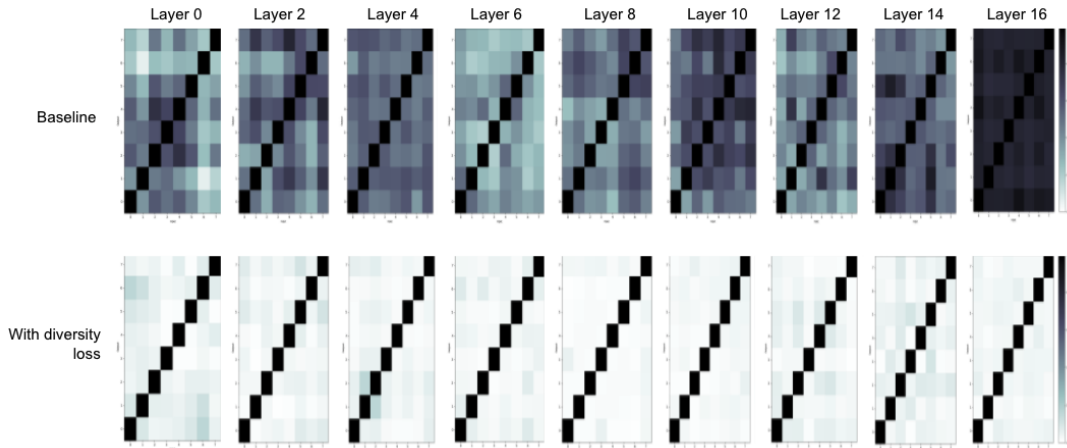


Figure 1: This figure shows the cosine similarity between the 8 attention heads for the baseline model and the model trained with the attention probability diversity loss. We show images only for the even-numbered Conformer layers. Darker colors imply higher cosine similarity (closer to 1).

Diversity loss	dev	dev-other	test	test-other
$d^A(m, n)$	0.41	0.37	0.45	0.39
$d^Q(m, n)$	0.02	0.02	0.02	0.02
$d^K(m, n)$	0.00	0.00	0.00	0.00
$d^V(m, n)$	0.01	0.01	0.01	0.01
$d^Y(m, n)$	0.04	0.05	0.05	0.05

Table 6: Attention diversity scores summed over all layers of the Conformer encoder for the models trained with diversity loss.

16) shows an extremely high correlation between all pairs of attention heads. The attention heads for each layer become significantly more decorrelated upon training with the diversity loss, and this also results in a WER improvement.

Finally, we trained a different and stronger Conformer model with relative position encodings and evaluated the impact of using the attention context diversity loss. Table 7 shows the WERs of the two models. We observe that including the diversity auxiliary loss function obtains a WER improvement even over a stronger baseline model on the harder dev-other and test-other sets.

Diversity score	dev	dev-other	test	test-other
Baseline	1.9	4.3	2.1	4.6
$d^Y(m, n)$	1.9	<b>4.2</b>	2.1	<b>4.4</b>

Table 7: WERs using a baseline Conformer model with relative position encoding and a model using context diversity loss as auxiliary loss functions during training.

#### 4.5. Analysis of gradient similarity

We next asked the question – Do more diverse attention heads also receive more diverse gradients during backpropagation for a given batch of data? We focused on the query weight matrix for the purposes of illustration. We considered three models with different choices of the diversity loss weight  $\lambda$  – baseline ( $\lambda = 0$ ),  $\lambda = 0.001$ , and  $\lambda = 1.0$ . Higher values of  $\lambda$  force the attention heads to be more diverse during training.

Given a pair  $(m, n)$  of attention heads, we computed the

gradient of the query weight matrices  $\mathbf{W}_m^{\text{query}}$  and  $\mathbf{W}_n^{\text{query}}$  using a batch of training data. In similar fashion to the diversity loss computation in (5) and (6), we first computed the cosine similarity between the gradient tensors, and then computed the mean of squares of the  $N \times N$  cosine similarity matrix minus the identity matrix over all possible pairs of attention heads and layers. Table 8 shows the query gradient similarity loss for the baseline

$\lambda$	dev	dev-other	test	test-other
0	0.021	0.026	0.021	0.026
0.001	0.016	0.017	0.016	0.018
1.0	0.015	0.017	0.015	0.017

Table 8: Gradient similarity loss between query weight matrices for different choices of query diversity loss weight ( $\lambda$ ) used during model training.

( $\lambda = 0$ ) and two models trained with diversity loss. We observe that the model trained with the most weight ( $\lambda = 1.0$ ) to the diversity loss also saw the most diverse gradients received by the individual heads, as indicated by the lowest value of the gradient diversity loss. Hence, the weight matrices of more diverse attention heads update very differently given the same batch of data, when compared to less diverse attention heads.

## 5. Conclusion

We analysed attention head diversity for one of the state-of-the-art models used for ASR – Conformer, on the public Librispeech data set. We evaluated diversity losses computed over all intermediate and final representations in each self-attention layer of the model. The paper showed that the various attention heads do tend to become highly correlated during the course of model training. In particular, the attention probabilities show the highest correlation across attention heads. We then show that adding auxiliary training loss functions that promote head diversity do make the heads less redundant and improve the model’s WER. Finally, we evaluated the connection between attention head diversity and gradient similarity of the head weight matrices. We showed that more diverse attention heads also received more diverse gradients.

## 6. References

- [1] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [3] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *IEEE Computational intelligence magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [4] A. Gatt and E. Kraehmer, “Survey of the state of the art in natural language generation: Core tasks, applications and evaluation,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.
- [5] V. Mnih, N. Heess, A. Graves *et al.*, “Recurrent models of visual attention,” *Advances in neural information processing systems*, vol. 27, 2014.
- [6] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
- [7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
- [8] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, “Draw: A recurrent neural network for image generation,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 1462–1471.
- [9] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.
- [10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” *Advances in neural information processing systems*, vol. 28, 2015.
- [11] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.
- [12] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, “Self-attentional acoustic models,” *arXiv preprint arXiv:1803.09519*, 2018.
- [13] 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 rnn-t loss,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.
- [14] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, “Transformer-transducer: End-to-end speech recognition with self-attention,” *arXiv preprint arXiv:1910.12977*, 2019.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [16] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” *arXiv preprint arXiv:1703.03130*, 2017.
- [17] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,” *arXiv preprint arXiv:1905.09418*, 2019.
- [18] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, “Revealing the dark secrets of BERT,” *arXiv preprint arXiv:1908.08593*, 2019.
- [19] J. Li, Z. Tu, B. Yang, M. R. Lyu, and T. Zhang, “Multi-head attention with disagreement regularization,” *arXiv preprint arXiv:1810.10183*, 2018.
- [20] J.-T. Chien and Y.-H. Huang, “Disentangled mask attention in transformer,” *openreview.net*, 2021.
- [21] G. M. Correia, V. Niculae, and A. F. Martins, “Adaptively sparse transformers,” *arXiv preprint arXiv:1909.00015*, 2019.
- [22] B. An, J. Lyu, Z. Wang, C. Li, C. Hu, F. Tan, R. Zhang, Y. Hu, and C. Chen, “Repulsive attention: Rethinking multi-head attention as Bayesian inference,” *arXiv preprint arXiv:2009.09364*, 2020.
- [23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [24] T. Lohrenz, P. Schwarz, Z. Li, and T. Fingscheidt, “Relaxed attention: A simple method to boost performance of end-to-end automatic speech recognition,” *arXiv preprint arXiv:2107.01275*, 2021.
- [25] 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)*. IEEE, 2015, pp. 5206–5210.
- [26] J. Shen, P. Nguyen, Y. Wu, Z. Chen, M. X. Chen, Y. Jia, A. Kannan, T. Sainath, Y. Cao, C.-C. Chiu *et al.*, “Lingvo: a modular and scalable framework for sequence-to-sequence modeling,” *arXiv preprint arXiv:1902.08295*, 2019.
- [27] K. Choromanski, V. Likhoshesterov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser *et al.*, “Rethinking attention with performers,” *arXiv preprint arXiv:2009.14794*, 2020.