SparseBERT: Rethinking the Importance Analysis in Self-attention

Han Shi[1], Jiahui Gao[2], Xiaozhe Ren[3], Hang Xu[3], Xiaodan Liang[4], Zhenguo Li[3], James T. Kwok[1]

[1]Hong Kong University of Science and Technology

[2]The University of Hong Kong

[3]Huawei Noah's Ark Lab

[4]Sun Yat-sen University

Jun 12, 2021

Introduction

- Motivation: As the core component in Transformer-based architecture, understanding self-attention module is important.
- Prior Works:
  - Empirical: local and global attention are both important by attention visualization.
  - Theoretical: universal approximability of Transformer-based models.
- Contribution: We study the importance analysis in self-attention using differentiable search method. Furthermore, we propose a Differentiable Attention Mask (DAM) algorithm, which can be also applied in guidance of SparseBERT design further.

The self-attention layer output can be written as:

$$Attn(\boldsymbol{X}) = \boldsymbol{X} + \sum_{k=1}^{H} \sigma(\boldsymbol{X}\boldsymbol{W}_Q^k(\boldsymbol{X}\boldsymbol{W}_K^k)^\top)\boldsymbol{X}\boldsymbol{W}_V^k\boldsymbol{W}_O^{k\top}, \tag{1}$$

where H is the number of heads, σ is the softmax function, and $\boldsymbol{W}_Q^k, \boldsymbol{W}_K^k, \boldsymbol{W}_V^k, \boldsymbol{W}_O^k \in \mathbb{R}^{d \times d_h}$ (where $d_h = d/H$ is the dimension of a single-head output) are weight matrices for the query, key, value, and output, respectively of the kth head. In particular, the self-attention matrix

$$\boldsymbol{A}(\boldsymbol{X}) = \sigma(\boldsymbol{X}\boldsymbol{W}_Q(\boldsymbol{X}\boldsymbol{W}_K)^\top) \tag{2}$$

in (1) plays a key role in the self-attention layer [Park et al., 2019, Gong et al., 2019, Kovaleva et al., 2019].

Related Work

Self-attention: empirical understanding

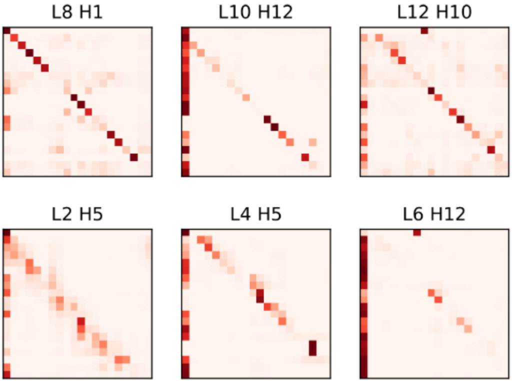


Figure: Self-attention matrix visualization [Gong et al., 2019].

Related Work

Self-attention: theoretical understanding

Let F_{CD} be the set of continuous functions $f : [0,1]^{n\times d} \mapsto \mathbb{R}^{n\times d}$. For any $p \geq 1$, the ℓ_p-distance between $f_1, f_2 \in F_{CD}$ is defined as $d_p(f_1, f_2) = (\int \|f_1(\boldsymbol{X}) - f_2(\boldsymbol{X})\|_p^p d\boldsymbol{X})^{1/p}$.

Theorem

Given $1 < p < \infty$, $\epsilon > 0$, for any $f \in F_{CD}$, there exists a transformer network, such that $d_p(f, g) < \epsilon$.

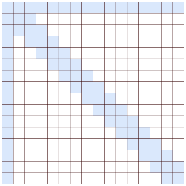
- Yun et al. [2019]: $g \in \mathcal{T}^{2,1,4}$ (vanilla transformer).
- Zaheer et al. [2020]: $g \in \mathcal{T}_D^{2,1,4}$ (containing star graph).
- Yun et al. [2020]: The sparsity patterns satisfy three assumptions.

$\rightarrow$ emphasize the importance of diagonal elements in the attention map.

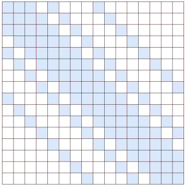
Related Work

Sparse Self-attention

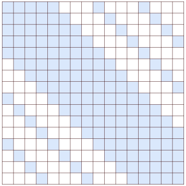
A number of sparse transformers have been recently proposed.



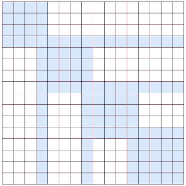
(a) Star.



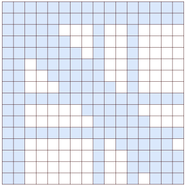
(b) LogSparse.



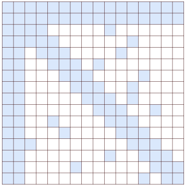
(c) Strided.



(d) Fixed.



(e) Longformer.



(f) BigBird.

Which Attention Positions are Important?

Continuous Relaxation

We associate an $\alpha_{i,j}$ with each position (i, j) in the self-attention matrix $\boldsymbol{A}(\boldsymbol{X})$, and define the attention probability as

$$P_{i,j} = \mathrm{sigmoid}(\alpha_{i,j}) \in [0, 1]. \tag{3}$$

For symmetry, we enforce $\alpha_{i,j} = \alpha_{j,i}$. Analogous to (1), the soft-masked self-attention is then

$$Attn(\boldsymbol{X}) = \boldsymbol{X} + \sum_{k=1}^{H} (\boldsymbol{P}^k \odot \boldsymbol{A}^k(\boldsymbol{X}))\boldsymbol{V}^k(\boldsymbol{X})\boldsymbol{W}_O^{k\top}, \tag{4}$$

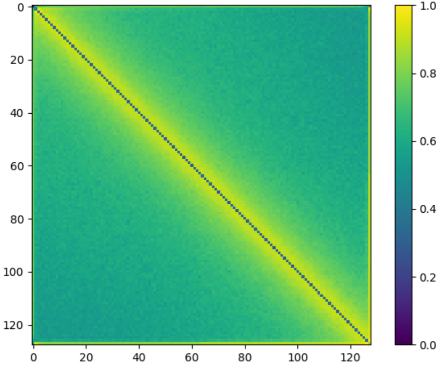
where $\odot$ is the element-wise product. Obviously, when $P_{i,j} = 1$ for all (i, j)'s, this reduces to Eq. (1).

However, the above multiplicative attention mask will result in unnormalized attention distributions. To solve this problem, we introduce the renormalization trick, which replaces the multiplicative attention mask with an additive mask before the softmax function as follows.

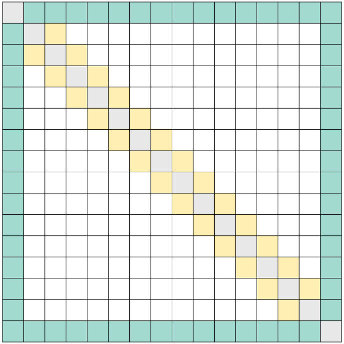
$$\hat{\boldsymbol{A}}(\boldsymbol{X}) = \sigma(\boldsymbol{X}\boldsymbol{W}_Q(\boldsymbol{X}\boldsymbol{W}_K)^\top + \boldsymbol{Q}), \tag{5}$$

$$Q_{i,j} = -c(1 - P_{i,j}), \tag{6}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is the addictive attention mask, c is a large constant such that $\hat{A}_{i,j} = 0$ if $P_{i,j} = 0$, and $\hat{A}_{i,j}$ reduces to the original attention score if $P_{i,j} = 1$.



(g) Attention distribution.



(h) Sketch.

Figure: Visualization of the attention distribution. In the figure on the right, the dark entries are for diag-attention, yellow for neighborhood attention and green for special attention.

Without diag-attention, the ith token output of the self-attention layer becomes:

$$Attn(\boldsymbol{X})_i = \boldsymbol{X}_i + \sum_{k=1}^{H} \sum_{j \neq i} A_{i,j}^{k}(\boldsymbol{X}) \boldsymbol{V}_j^{k}(\boldsymbol{X}) \boldsymbol{W}_O^{k\top}.$$

Let $\mathcal{T}^{H,d_h,d_{\text{ff}}}$ be a class of transformers without diag-attention stacks. The following Theorem shows that the self-attention mechanism without diag-attention is also a universal approximator:

Theorem

Given $1 < p < \infty$, $\epsilon > 0$ and $n > 2$, for any $f \in F_{CD}$, there exists a transformer network without diag-attention $g \in T^{2,1,4}$, such that $d_p(f, g) < \epsilon$.

Theorem

Given $1 < p < \infty$, $\epsilon > 0$ and $n > 2$, for any $f \in F_{CD}$, there exists a transformer network without diag-attention $g \in T^{2,1,4}$, such that $d_p(f, g) < \epsilon$.

The proof outline (following Yun et al. [2019]):

- **Step 1**: Approximate F_{CD} with the set of piecewise-constant functions $\bar{F}_{CD}$.
- **Step 2**: Approximate $\bar{F}_{CD}$ with the modified transformer blocks $\bar{\mathcal{T}}^{H,d_h,d_{\mathrm{ff}}}$, which replace the softmax operator and ReLU with the hardmax operator and a piece-wise linear functions.
- **Step 3**: Approximate the modified transformer blocks $\bar{g} \in \bar{\mathcal{T}}^{2,1,1}$ with standard transformer blocks $g \in \mathcal{T}^{2,1,4}$.

- GLUE: The General Language Understanding Evaluation benchmark is a collection of diverse natural language understanding tasks.

Table: Performance (in %) of the various BERT-base variants on the GLUE data set.

	MNLI (m/mm)	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	Average
Development Set									
BERT-base (ours)	85.4/85.8	88.2	91.5	92.9	62.1	88.8	90.4	69.0	83.8
BERT-base (randomly dropped)	84.6/85.2	87.7	91.1	92.7	62.0	88.9	89.3	68.9	83.4
BERT-base (no diag-attention)	85.6/85.9	88.2	92.0	93.8	63.1	89.2	91.2	67.9	**83.9**
Test Set									
BERT-base [Devlin et al., 2019]	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT-base (ours)	84.8/84.1	71.3	90.9	93.4	52.3	85.3	88.3	66.9	79.7
BERT-base (randomly dropped)	84.5/83.5	70.3	91.1	93.4	52.0	85.8	87.4	66.7	79.4
BERT-base (no diag-attention)	85.5/84.9	71.3	91.1	93.4	53.3	86.3	88.9	67.9	**80.3**

Without diag-attention

Empirical Verification

- SWAG: The Situations With Adversarial Generations dataset contains $113k$ sentence-pair completion examples that evaluate grounded commonsense inference.
- SQuAD: The Stanford Question Answering Dataset is a collection of crowd-sourced question/answer pairs.

Table: Performance (in %) of the various BERT-base variants on the SWAG and SQuAD development sets.

	SWAG	SQuAD v1.1		SQuAD v2.0	
	acc	EM	F1	EM	F1
BERT-base [Devlin et al., 2019]	81.6	80.8	88.5	-	-
BERT-base (ours)	82.5	79.7	87.1	72.9	75.5
BERT-base (randomly dropped)	81.6	79.7	87.0	71.5	74.2
BERT-base (no diag-attention)	83.5	80.3	87.9	73.2	75.9

Instead of using the sigmoid function to output the attention probability as in (3), we use the Gumbel-sigmoid:

$$M_{i,j} = \text{Gumbel-sigmoid}(\alpha_{i,j}) = \text{sigmoid}((\alpha_{i,j} + G_1 - G_2)/\tau),$$

where G_1, G_2 are independent Gumbel noises generated from the uniform distribution U as:

$$G_k = -\log(-\log(U_k)), U_k \sim U(0, 1),$$

and τ is a temperature hyperparameter. To balance mask sparsity with performance, we add the sum absolute values of the attention mask to the loss, as:

$$\mathcal{L} = l(\text{BERT}(\boldsymbol{X}, \boldsymbol{A}(\boldsymbol{X}) \odot \boldsymbol{M}(\alpha); \boldsymbol{w})) + \lambda \|\boldsymbol{M}(\alpha)\|_1, \tag{7}$$

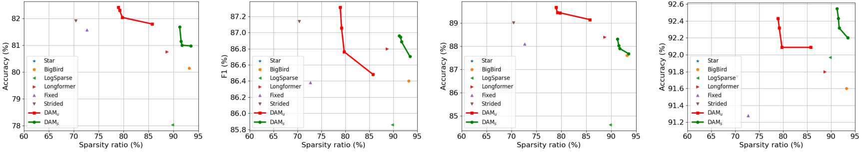
where $l(\text{BERT}(\boldsymbol{X}, \boldsymbol{A}(\boldsymbol{X}); \boldsymbol{w}))$ is the pre-training loss, and λ is a trade-off hyperparameter.

Algorithm 1 Differentiable Attention Mask (DAM).

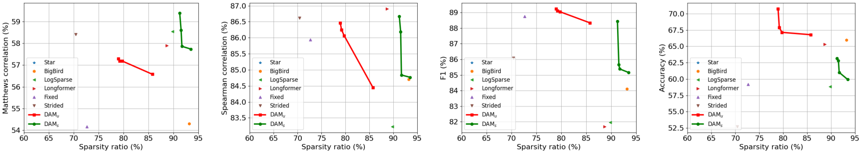
1: initialize model parameter $\boldsymbol{w}$ and attention mask parameter $\boldsymbol{\alpha}$.
2: **repeat**
3: generate mask $M_{i,j} \leftarrow$ gumbel-sigmoid$(\alpha_{i,j})$;
4: obtain the loss with attention mask $\mathcal{L}$;
5: update parameter w and α simultaneously;
6: **until** convergence.
7: **return** attention mask $\boldsymbol{M}$.

Structured Varient

- the first and last row/column of the attention mask to be active.
- the generated mask has $M_{i,j} = M_{i+k,j+k}$ for integer k



(a) MNLI. (b) QQP. (c) QNLI. (d) SST-2.

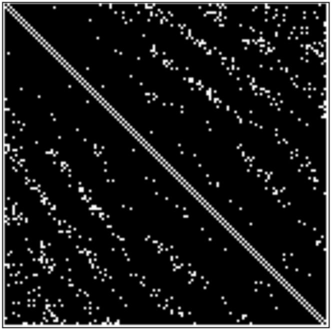


(e) CoLA. (f) STS-B. (g) MRPC. (h) RTE.

Figure: Performance of the BERT-base for different attention masks on the GLUE development set. MNLI shows the average performance on the MNLI-m and MNLI-mm sections.

SparseBERT

Mask Visualization



(a) unstructured.

(b) structured.

Figure: Visualization of the attention masks generated by DAM. Here, white means with-attention and dark means no-attention.

Table: Ablation study on the importance of diag-attention in different attention masks. Here, "w/" means using diag-attention and "w/o" means without using diag-attention. As can be seen, dropping diag-attention increases sparsity ratio without harming the performance.

	Strided		Fixed		Longformer		LogSparse		BigBird		Star		DAM$_s$($\lambda = 10^{-4}$)		DAM$_s$($\lambda = 10^{-1}$)	
	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o
Sparsity (%)	70.4	71.2	72.7	73.4	88.7	89.5	89.8	90.6	93.2	93.9	96.1	96.9	90.4	91.2	92.7	93.5
GLUE (%)	79.5	80.2	79.7	79.6	80.1	80.1	77.9	77.8	79.4	79.5	78.9	78.6	80.5	80.9	79.3	79.6

References I

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

L. Gong, D. He, Z. Li, T. Qin, L. Wang, and T. Liu. Efficient Training of BERT by Progressively Stacking. In *International Conference on Machine Learning*, 2019.

O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the Dark Secrets of BERT. In *Empirical Methods in Natural Language Processing*, 2019.

C. Park, I. Na, Y. Jo, S. Shin, J. Yoo, B. Kwon, J. Zhao, H. Noh, Y. Lee, and J. Choo. SANVis: Visual Analytics for Understanding Self-Attention Networks. In *IEEE Visualization Conference*, 2019.

C. Yun, S. Bhojanapalli, A. Rawat, S. Reddi, and S. Kumar. Are Transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.

C. Yun, Y. Chang, S. Bhojanapalli, A. Rawat, S. Reddi, and S. Kumar. $O(n)$ Connections are Expressive Enough: Universal Approximability of Sparse Transformers. In *Neural Information Processing Systems*, 2020.

M. Zaheer, G. Guruganesh, K. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al. Big Bird: Transformers for Longer Sequences. In *Neural Information Processing Systems*, 2020.

Thank You!