Supplementary Materials — Jump Self-attention: Capturing High-order Statistics in Transformers

Haoyi Zhou
BDBC
Beihang University
Beijing, China 100191
haoyi@buaa.edu.cn

Siyang Xiao
BDBC
Beihang University
Beijing, China 100191
xiaosy@act.buaa.edu.cn

Shanghang Zhang
School of Computer Science
Peking University
Beijing, China 100871
shanghang@pku.edu.cn

Jieqi Peng
BDBC
Beihang University
Beijing, China 100191
pengjq@act.buaa.edu.cn

Shuai Zhang
BDBC
Beihang University
Beijing, China 100191
zhangs@act.buaa.edu.cn

Jianxin Li*
BDBC†
Beihang University
Beijing, China 100191
lijx@buaa.edu.cn

Appendix

A Related Work

The reading comprehension [8], question and answering [4], sentiment analysis [1], image captioning [7] and other tasks encourage linguistic-meaningful solution [17] with higher-level understanding of the language other than superficial correspondences between inputs and outputs. The giant model [3] shows the superior performance on various down-stream tasks through the large-scale pre-training. Recently, CLIP [11, 10], CogView[6] and DALL·E [14] learn a multi-modal embedding space to estimate the semantic similarity between texts and images, and a handful of prompt techniques [9, 15] explores the remapping of semantic representation. They reveal the possibility of capturing semantic relations based on the self-attention mechanism. Others apply the knowledge graph in Transformers [18, 2], but it is resource-hungry and needs human-aid crafting. The most related work is Triplet Attention [19], it utilizes the diversity to enhance the attention feature map and focus on the attention's dissimilarity. However, we aim to measure high-order statistics in attention's similarity to enhance the dot-product self-attention.

B Mathematical notions

For better understanding, we collected some definition of the matrices in Table 1. Besides, the *miracle heads* refer to these most effective heads in capturing the jump self-attention, which is hardly acquired in the standard Transformer architecture. If the magnitude of jump connections ρ is fixed, finding the statistics are combination problem. But the optimization becomes intractable.

*Jainxin Li is the corresponding author.

†BDBC is the abbreviation for the Beijing Advanced Innovation Center for Big Data and Brain Computing.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: The mathematical notations as preliminary.

Notation	Meaning
$\mathbf{S}$	The matrix of self-attention score.
$\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$	The matrix of projected inputs.
$\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$	The transformation matrix.
$\mathbf{A}$	The JAT-defined adjacency matrix.
$\mathbf{U}$	The temporary matrix help to define $\mathbf{A}$.

C The model architecture

We present the overall architecture of proposed Jump Self-attention model in Fig.(8), especially on the heads splits and connection flows. To simplify the comparison, we use the same feature map between the canonical self-attention and jump one. They may be different from individual projects in practice. The red rectangle refers to the enhanced "jump" connections from JAT.

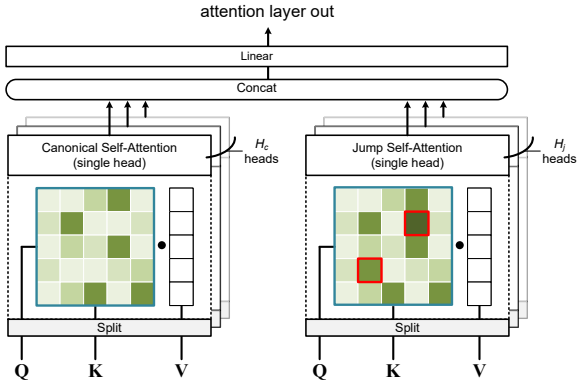


Figure 1: The overview of the proposed Jump Self-attention. We show the pipeline of capturing high-order statistics. Our proposed JAT adopts the GCN operator to enhance the weak "jump" connections in dot-product attention's feature map. The attention layer output is based on the concatenation of two groups of heads.

D Experiments settings of SQuAD

We also conduct JAT's experiments on the performance of the question-answering task. The Stanford Question Answering Dataset (SQuAD v1.1/v2.0) [12, 13] is a collection of 150k crowd-sourced question/answer pairs. The task is to predict the answer text span in the passage given a question and a passage from Wikipedia containing the answer. **Settings:** We maintain the same fine-tuning strategy as in the GLUE experiment. **Metric:** Exact match (EM), which is the number of exactly correct answers, and F1 scores, which captures the precision and recall that words chosen as part of the answer are part of the answer. **Platform:** Intel Xeon 3.2GHz + The Nvidia V100 GPU (32 GB) X 4.

E Experiments details on the Case Study

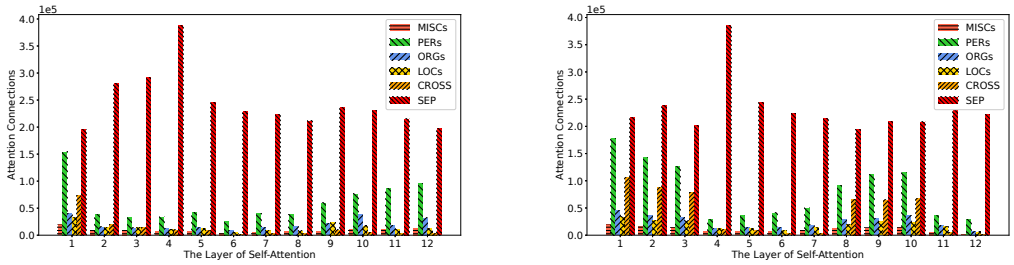
E.1 Setup

CoNLL-2003: We conduct a case study experiment on Named-Entity Recognition (NER) task using the standard CoNLL-2003 dataset, which concentrates on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. **Settings:** We use a batch size of 32 and fine-tune a pre-trained BERT$_{base}$ model and a BERT-JAT model whose layers {1,2,3,8,9,10} have 6 JAT heads for 5 epochs. The other settings

follow past works' recommendation. **Metric:** F1 scores and accuracy are used. **Platform:** Intel Xeon 3.2GHz + The Nvidia V100 GPU (32 GB) X 4.

E.2 Results

We preprocess the texts before we feed them to the model, including adding 'CLS' and 'SEP' special tokens whose named-entity labels are set to -100, and split the words into subwords. We select the major connections in attention feature map where their scores are greater than (mean + std). We studied the named-entities of the pair-wise tokens corresponding to those major connections and use 'MISCs','PERs','ORGs','LOCs' to represent the pair-wise attention connections of tokens with the same named-entities, which means that both tokens are miscellaneous, persons, locations or organizations. We use 'SEP' to represent connections of two tokens where one of them is a special token 'CLS' or 'SEP', and we use 'CROSS' to represent connections of tokens with different named-entities.



(a) BERT$_{base}$. (b) BERT-JAT$_{l\{1,2,3,8,9,10\}}$.

Figure 2: The pair-wise attention connections of BERT$_{base}$ and BERT-JAT on dataset CoNLL2003, our model BERT-JAT reduce the connections to the special tokens 'CLS' or 'SEP' in lower layers.

We visualize the results containing the 'SEP' connection in Fig.(2). From Fig.(2a), we can find that the number of 'SEP' connection is more than other connections, which means that there are many tokens connected to special tokens 'CLS' or 'SEP' in self-attention, and these connections have redundant and over-expressed information. This observation is also consistent with previous research [16, 5]. From Fig.(2b), we can find that JAT reduce the number of 'SEP' connection at the layers {1,2,3}, noted that the number of JAT heads in these layers is 6, only half of the total attention heads, which shows that JAT can reduce redundant information in attention and enhance the model's expression ability.

We further remove the 'SEP' connection and visualize the results in Fig.(7). From Fig.(7a) we can find that there are not many 'CROSS' connections in the attention of BERT$_{base}$, and most of the connections are between tokens that belong to a person's entity. From Fig.(7b), we can observe that JAT significantly increase the number of connections between tokens with different named-entities at the layers {1,2,3,8,9,10} having JAT heads, which enhances the model's ability to discover high-order information and makes the model's representation more diverse.

E.3 Discussion

We summarized the analysis of high-order statistics in Sec 3.1. The Fig.(2) is a concrete example. And the NER case study in Sec 4.5 can be considered as an empirical evaluation of capacity limitation. The 'CROSS' connections represent the high-order connections, which are rare in BERT$_{\text{base}}$ while are more in BERT-JAT.

F The performance of training from scratch

Without the fine-tuning framework, where Roberta is a representative SOTA, we want to separately investigate the JAT's performance when training from scratch. We have performed experiments at different data scales on CoLA. Table 2 shows that the model receives constant gains after applying

the JAT. Although the performance gain decreases a little, we acquire an even higher performance than JAT in the fine-tuning manner ($66.5 > 65.4$). This shows the potential of building models with the JAT architecture.

Table 2: Training from the scratch at different CoLA data scale.

Model	20%	40%	60%	80%	100%
RoBERTa	51.0	55.0	59.4	60.1	63.7
RoBERTa-JAT	52.5	56.7	60.6	60.9	66.5
Δ Performance	+1.5	+1.7	+1.2	+0.8	+0.8

G Limitations / Broder Impact

Our method can enable the Transformer model to capture high-order connections by applying the JAT techniques. It can potentially benefit many communities in society. For example, the large-scale pre-train model (GPT-3, T5, etc.) can use our method to enhance the model performance without more examples. It can save energy consumption during collecting examples and training models for complex tasks. As another example, the popular CLIP model can use our method to enhance its performance on various tasks, especially in a dangerous environment. Our method requires GPUs and high-speed networks to run. For underdeveloped regions such as rural areas, our method may not be feasible to use. Our method develops on the input data, which may have domain shift and population bias. Such bias may render the JAT to produce suboptimal results. If the method fails, the Transformer may generate incorrect outputs. In mission-critical applications, the usage of this method should be guided by experienced ML experts and domain experts.

Appendix References

[1] Ismail El Bazi and Nabil Laachfoubi. Arabic named entity recognition using deep learning approach. *International Journal of Electrical and Computer Engineering*, 9(3):2025–2032, 2019.

[2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779, 2019.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NIPS*, 2020.

[4] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *EMNLP*, pages 2174–2184, 2018.

[5] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. pages 276–286, 2019.

[6] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. pages 19822–19835, 2021.

[7] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6):118:1–118:36, 2019.

[8] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*, pages 785–794, 2017.

[9] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, 2021.

[10] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. pages 16399–16408, 2022.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021.

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392, 2016.

[13] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *ACL*, pages 784–789, 2018.

[14] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, volume 139, pages 8821–8831, 2021.

[15] Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. Constrained language models yield few-shot semantic parsers. In *EMNLP*, pages 7699–7715, 2021.

[16] Jesse Vig. A multiscale visualization of attention in the transformer model. In *ACL*, pages 37–42, 2019.

[17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

[18] Liang Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: BERT for knowledge graph completion. *arXiv:1909.03193*, 2019.

[19] Haoyi Zhou, Jianxin Li, Jieqi Peng, Shuai Zhang, and Shanghang Zhang. Triplet attention: Rethinking the similarity in transformers. In *KDD*, pages 2378–2388, 2021.