RealFormer: Transformer Likes Residual Attention

Authors:
Ruining He, Anirudh Ravula, Bhargav Kanagal, Joshua Ainslie

Julia Fischer

Motivation

Motivation

- Transformer models are the backbone of NLP models like
  - BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)
  - GPT (**G**enerative **P**re-Trained **T**ransformer)
- self-supervised learning
- fine-tuning models for specific task

Transformer

Standard Transformer

- proposed by Vaswani et al. 2017
- consists of encoder and decoder
- 2 sub-layers inside each layer of a Transformer encoder/ decoder
  1. Multi-Head Attention module: compute output embeddings of a set of queries
  2. fully-connected Feed-Forward Network module with one hidden layer

Architecture

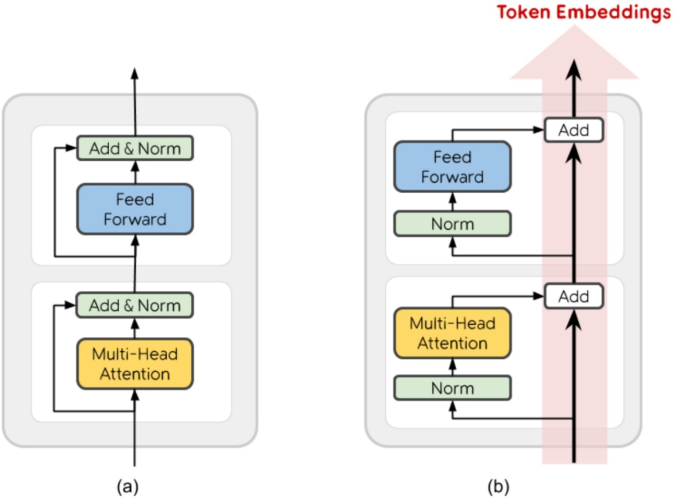


Figure 1: Comparison of (a) Post-LN layer and (b) Pre-LN layer in Transformer encoders; Image taken from He et al. 2020

RealFormer

Architecture

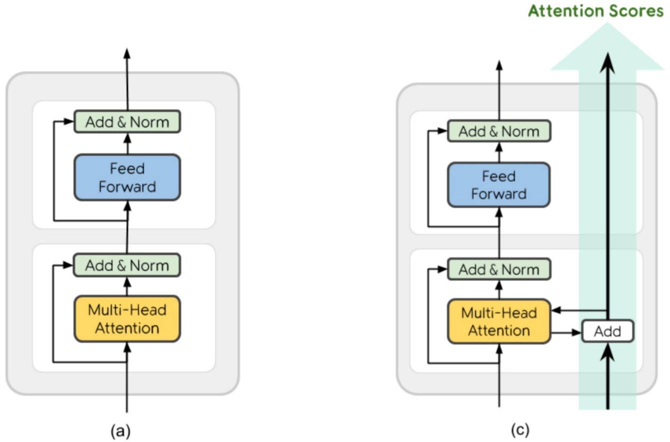


Figure 2: Comparison of (a) Post-LN layer and (c) RealFormer layer in Transformer encoders; Image taken from He et al. 2020

Residual **A**ttention **L**ayer Trans**former**

Advantages:

- implementation adds only a few lines to the code of the backbone
- no additional parameters
- straightforward application for Transformer variations

Disadvantages:

- might be sub-optimal for very deep networks

Experiments

BERT

BERT model

- proposed by Devlin et al. 2019
- setup based on official BERT repository
- compare 3 Transformer architectures on wide spectrum of sizes

Model	**#layers**	**h.s.**	**#heads**	**i.s.**	**#param**
BERT-Small	4	512	8	2048	30M
BERT-Base	12	768	12	3072	110M
BERT-Large	24	1024	16	4096	340M
BERT-xLarge	36	1536	24	6144	1B

Table 1: Model architecture for BERT. h.s.:hidden size, i.s.: intermediate size; the number of parameters is approximated; Adopted from He et al. 2020

Evaluation of pre-trained Models

Model	Post-LN	Pre-LN	RealFormer
BERT-Small	61.57%	61.67%	**61.70%**
BERT-Base	70.20%	69.74%	**70.42%**
BERT-Large	73.64%	73.21%	**73.94%**
BERT-xLarge	73.72%	73.53%	**74.76%**

Table 2: Masked Language Modeling (MLM) accuracy from the pre-trained models on the randomly held-out development set after pre-training 1M steps; Adopted from He et al. 2020

Pre-training curves

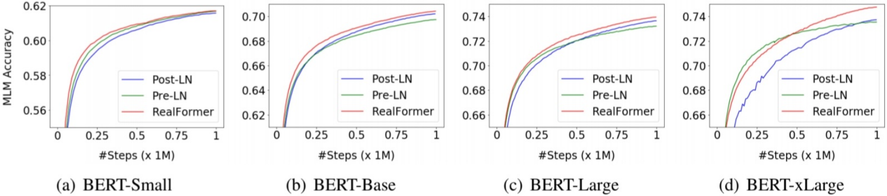


Figure 3: Development set MLM accuracy; Image taken from He et al. 2020

GLUE

Task	Post-LN	Pre-LN	RealFormer
MNLI-m	85.96 ± 0.11	85.03 ± 0.12	**86.28** ± 0.14
MNLI-nm	85.98 ± 0.14	85.05 ± 0.19	**86.34** ± 0.30
QQP	91.29 ± 0.10	91.29 ± 0.16	**91.34** ± 0.03
QQP (F1)	**88.34** ± 0.15	88.33 ± 0.26	88.28 ± 0.08
QNLI	92.26 ± 0.15	**92.35** ± 0.26	91.89 ± 0.17
SST-2	92.89 ± 0.17	93.81 ± 0.13	**94.04** ± 0.24
CoLA (MC)	58.85 ± 1.31	58.04 ± 1.50	**59.83** ± 1.06
STS-B (PC)	90.08 ± 0.27	90.06 ± 0.33	**90.11** ± 0.56
STS-B (SC)	89.77 ± 0.26	89.62 ± 0.28	**89.88** ± 0.54
MRPC	**87.50** ± 0.67	86.76 ± 5.64	87.01 ± 0.91
MRPC (F1)	**91.16** ± 0.45	90.69 ± 3.16	90.91 ± 0.65
RTE	71.12 ± 2.52	68.59 ± 1.52	**73.65** ± 0.90
Overall	84.01	83.47	**84.53**

Figure 4: GLUE development set results of fine-tuning BERT-Large models. All numbers are scaled by 100. Numbers in smaller font are standard deviations; Taken from He et al. 2020

SQuAD

SQuAD	Public	Post-LN	Pre-LN	RealFormer
v1.1 (F1)	90.9	91.68 ± 0.12	91.06 ± 0.09	**91.93** ± 0.12
v1.1 (EM)	84.1	85.15 ± 0.13	83.98 ± 0.24	**85.58** ± 0.15
v2.0 (F1)	81.9	82.51 ± 0.12	80.30 ± 0.12	**82.93** ± 0.05
v2.0 (EM)	78.7	79.57 ± 0.12	77.35 ± 0.16	**79.95** ± 0.08

Table 3: SQuAD development set results of fine-tuning BERT-Large models. All numbers are scaled by 100. Numbers in smaller font are standard deviations. Public: Post-LN results from Devlin et al. 2019; Adopted from He et al. 2020

How well does RealFormer perform with half the pre-training budget?

Task	Post-LN (500K)	Post-LN (1M)	RealFormer (500K)
GLUE	83.84	84.01	84.34
v1.1 (F1)	91.49 $\pm$ 0.18	91.68 $\pm$ 0.12	91.56 $\pm$ 0.09
v1.1 (EM)	84.87 $\pm$ 0.24	85.15 $\pm$ 0.13	85.06 $\pm$ 0.12
v2.0 (F1)	81.44 $\pm$ 0.50	82.51 $\pm$ 0.12	82.52 $\pm$ 0.55
v2.0 (EM)	78.64 $\pm$ 0.48	79.57 $\pm$ 0.12	79.54 $\pm$ 0.54
Overall	83.97	84.37	**84.51**

Table 4: Downstream development set results of finetuning BERT-Large with Post-LN and RealFormer pretrained with different number of steps. All numbers are scaled by 100. Numbers in smaller font are standard deviations; Adopted from He et al. 2020

Does a larger learning rate help?

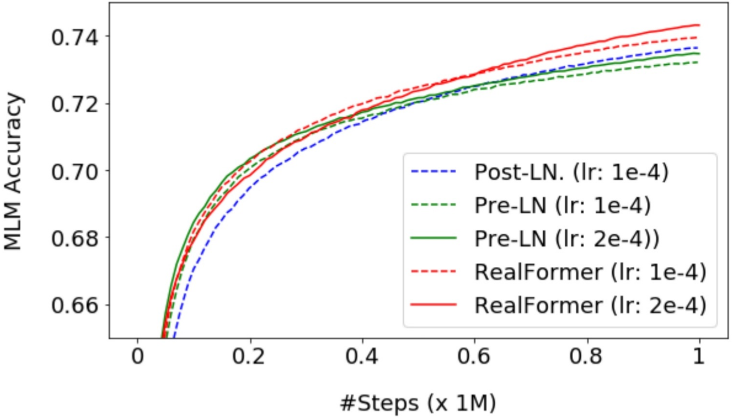


Figure 5: Development set MLM accuracy of BERTLarge with different learning rates; Image taken from He et al. 2020

Is attention sparser in RealFormer?

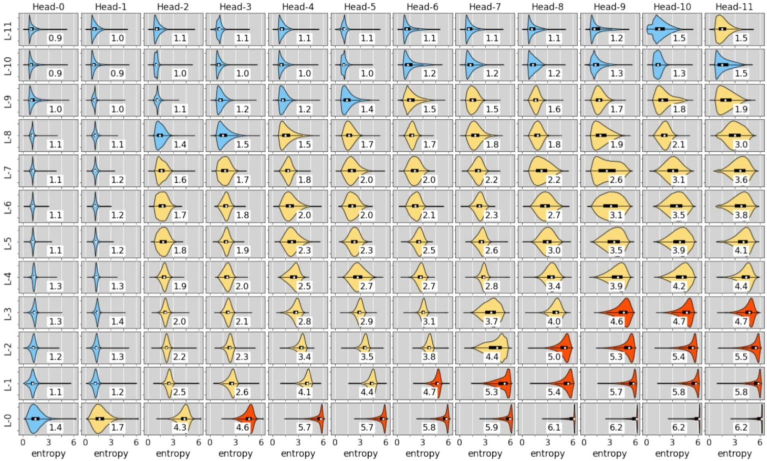
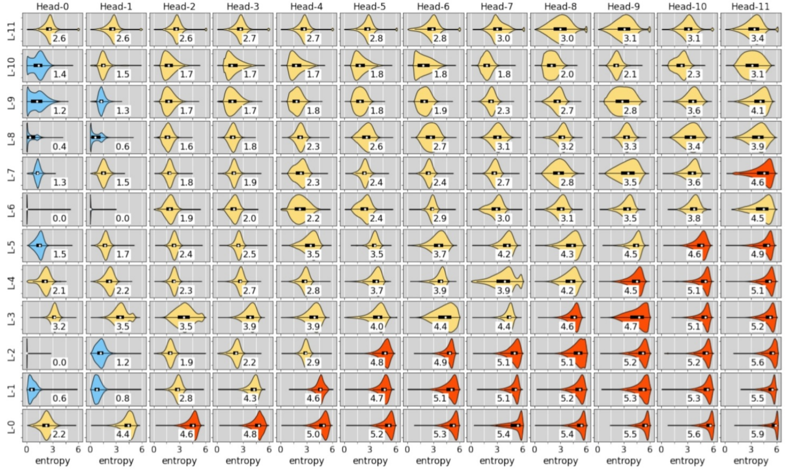


Figure 6: Distribution of entropies of the attention probabilities using the pre-trained BERT-Base with **RealFormer**. RED (median > 4.5), YELLOW ($1.5 \leq$ median ≤ 4.5), BLUE (median < 1.5), i.e., colder colors mean sparser attention; Image taken from He et al. 2020

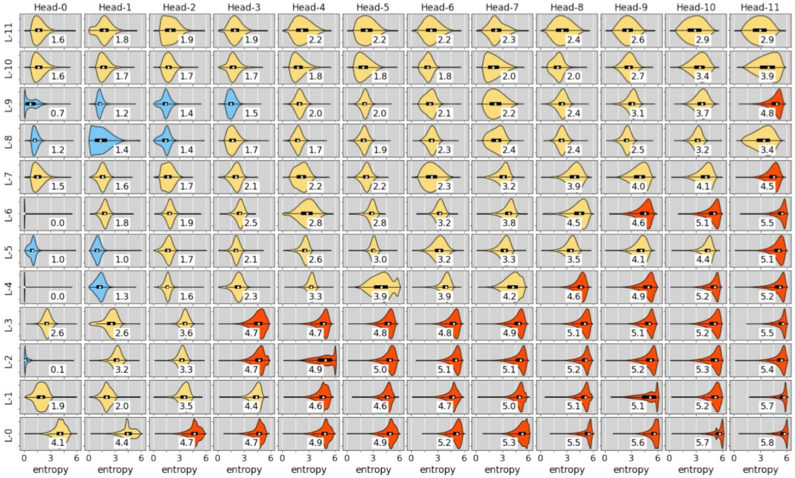
Is attention sparser in RealFormer?



(a) Post-LN

Figure 7: Distribution of entropies of the attention probabilities using the pre-trained BERT-Base with **Post-LN**; Image taken from He et al. 2020

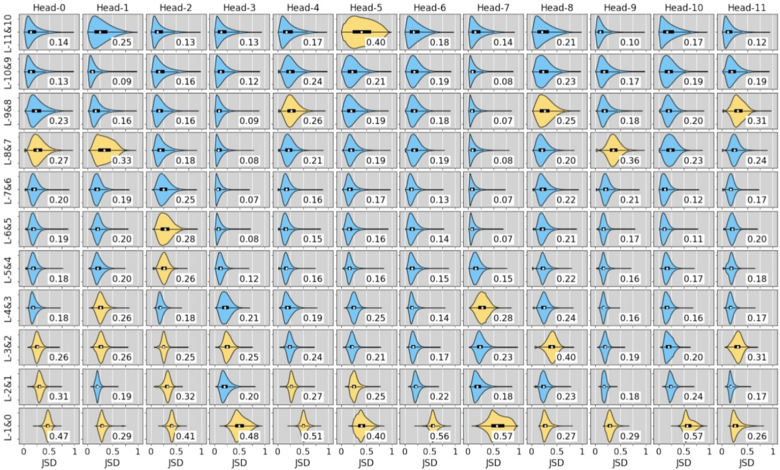
Is attention sparser in RealFormer?



(b) Pre-LN

Figure 8: Distribution of entropies of the attention probabilities using the pre-trained BERT-Base with **Pre-LN**; Image taken from He et al. 2020

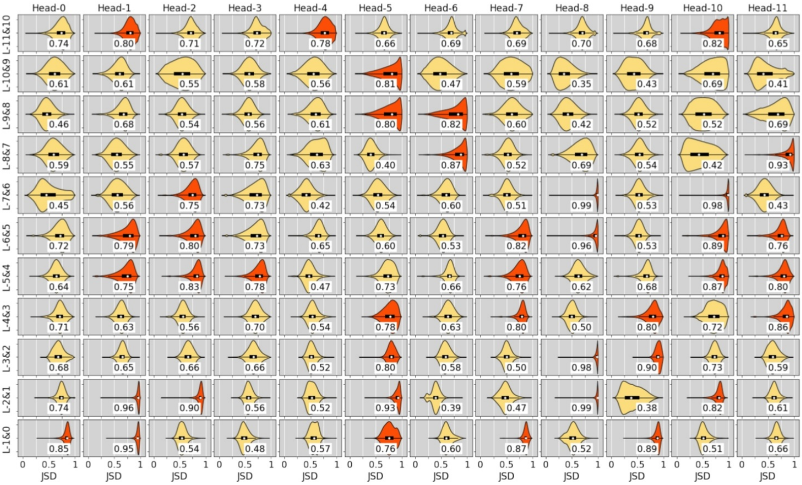
Do attention heads in layer L resemble those in layer $L - 1$?



(a) RealFormer

Figure 9: Distribution of JSD of attention probabilities in (vertically) adjacent attention heads using the pre-trained BERT-Base with **RealFormer** Transformer. Colder color means more "similar" attention heads across adjacent layers; Image taken from He et al. 2020

Do attention heads in layer L resemble those in layer $L-1$?



(b) Post-LN

Figure 10: Distribution of JSD of attention probabilities in (vertically) adjacent attention heads using the pre-trained BERT-Base with **Post-LN** Transformer; Image taken from He et al. 2020

Is residual attention really necessary?

Dropout	Post-LN	Pre-LN	RealFormer
0%	71.16%	69.80%	**71.30%**
10%	73.64%	73.21%	**73.94%**
20%	73.21%	72.97%	**73.66%**

Table 5: Development set MLM accuracy of BERT-Large with different dropout rates; Adopted from He et al. 2020

ADMIN

Adaptive Model Initialization

- proposed by Liu et al. 2020
- state-of-the-art Neural Machine Translation model
- ADMIN adopts Post-LN as backbone
- compare with RealFormer with running mean
- use 2 NMT benchmarks: WMT'14 En-De and WMT'14 En-Fr
- training setup from Liu et al. 2020 given in the official ADMIN repository

Result

Model	En-De			En-Fr	
	6L-6L	12L-12L	18L-18L	6L-6L	60L-12L
Post-LN	27.80	failed	failed	41.29	failed
Pre-LN	27.27	28.26	28.38	40.74	43.10
ADMIN	27.90	28.58	29.03	41.47	43.80
ADMIN*	28.06	28.85	29.11	41.65	43.72
Ours	**28.17**	**29.06**	**29.35**	**41.92**	**43.97**

Table 6: Test set BLEU scores on two WMT'14 benchmarks using different sizes of models. xL-yL: #Encoder layers-#Decoder layers. First three rows are from Liu et al. 2020. Ours is switching the backbone of ADMIN from Post-LN to RealFormer.
*: Our run of ADMIN using the same setups as RealFormer; Adopted from He et al. 2020

ETC

Extended Transformer Construction

- recent sparse attention mechanism to handle long context
- proposed by Ainslie et al. 2020 and Zaheer et al. 2020
- state-of-the-art results on 4 NL benchmarks

	Instances		Instance length	
Datasets	Training	Dev	Median	Max
NQ	307373	7830	4004	156551
HotpotQA	90447	7405	1227	3560
WikiHop	43738	5129	1541	20337
OpenKP	133724	6610	761	89183

Table 7: Statistics of the datasets adopted from Ainslie et al. 2020. Length in word piece tokens

- experiments based on GitHub ETC repository
- use ETC-Large model (24 layers, 1024 hidden size, 16 heads)

Result

Task	Metric	ETC-Large	Ours
WikiHop	Accuracy	78.92 ± 0.14	**79.21** ± 0.38
HotpotQA	Ans. F1	80.38 ± 0.13	**80.86** ± 0.16
	Sup. F1	89.07 ± 0.06	**89.21** ± 0.12
	Joint F1	73.12 ±0.19	**73.57** ± 0.19
Natural Questions	Long Ans. F1	77.70 ± 0.15	**77.93** ± 0.31
	Short Ans. F1	58.54 ± 0.41	**59.10** ± 0.81
	Average F1	68.07 ± 0.17	**68.51** ± 0.56
OpenKP	F1@3	44.06 ± 0.08	**44.27** ± 0.08

Table 8: Performance on the development set. All numbers are scaled by 100. Numbers in smaller font are standard deviations; Adopted from He et al. 2020

WikiHop leaderboard

WikiHop

#	Model / Reference	Affiliation	Date	Accuracy[%]
1	RealFormer-large (single)	[anonymized]	January 2021	84.4
2	ETC-large (single)	[anonymized]	May 2020	82.3
3	Longformer (single)	AI2	March 2020	81.9
4	Path-based GCN (ensemble)	Zhejiang University (ZJU)	September 2019	78.3
5	Chen et al. (2019)	UT Austin	September 2019	76.5
6	QIT (ensemble)	[anonymized]	March 2023	76.5
7	ChainEx (single)	[anonymized]	May 2019	74.9
8	JDReader (ensemble)	JD AI Research	March 2019	74.3

Figure 11: WikiHop leaderboard

Conclusion

Take Home Message

- RealFormer: simple, generic and cost-effective technique
- Proven improvement in tasks like:
  - ⋆ Masked Language Modeling
  - ⋆ Neural Machine Translation
  - ⋆ Long document modeling
- Results in sparser attention:
  - ⋆ Within individual heads
  - ⋆ Across heads in adjacent layers

References I

Ainslie, Joshua et al. (2020). “ETC: Encoding long and structured inputs in transformers”. In: *arXiv preprint arXiv:2004.08483.*

Devlin, Jacob et al. (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

He, Ruining et al. (2020). “Realformer: Transformer likes residual attention”. In: *arXiv preprint arXiv:2012.11747.*

References II

Liu, Liyuan et al. (Nov. 2020). "Understanding the Difficulty of Training Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 5747–5763.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *arXiv preprint arXiv:1706.03762*.

Zaheer, Manzil et al. (2020). "Big bird: Transformers for longer sequences". In: *Advances in neural information processing systems* 33, pp. 17283–17297.

End

Thank you for listening!
Do you have any questions?