Paper Skimming Session:
ETC: Encoding Long and Structured Inputs in Transformers

Nicolay Rusnachenko

rusnicolay@gmail.com
https://nicolay-r.github.io/

Newcastle University
England

The Alan Turing Institute

Long-Ranged Input for Transformers

Main limitation for input $X \in \mathbb{R}^N$:

- $O(N^2)$ original self-attention[1] computation complexity;

How to address this problem:

1. Sparse version of self-attention: Reformer, Longformer[2]
2. #1 with Global Attention
3. + **Structurization**[3] – limit attention within sentences, paragraphs, etc. via *masking*

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150* (2020).
[3] Joshua Ainslie et al. "ETC: Encoding long and structured inputs in transformers". In: *arXiv preprint arXiv:2004.08483* (2020).

Relative Position Encoding

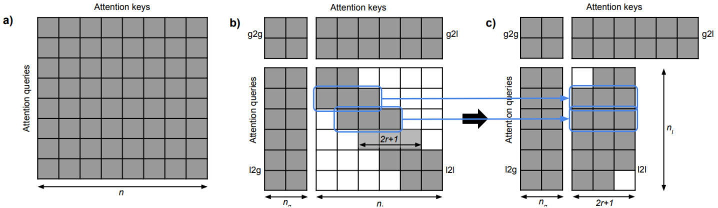
BERT[4] exploits absolute position encoding $X \in \mathbb{R}^N$.
ETC proposes **relative**:

- Now position is label $l_{i,j}$ of **connection** of $x_i \in X$ with other X
- Distance clipping: k – limit window
  - l_k outside after i,
  - l_{-k} outside radius k before i.
- **Result** in α_l^K – learnable vectors of relative positions

[4] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

Global + Local Attention

- n_l – main input components: **now windowed** (sparsed)
- n_g – global input components ($n_g << n_l$)



$$z_i^g = \sum_{j=1}^{n_g} \boxed{\alpha_{ij}^{g2g}} x_j^g W^V$$

position

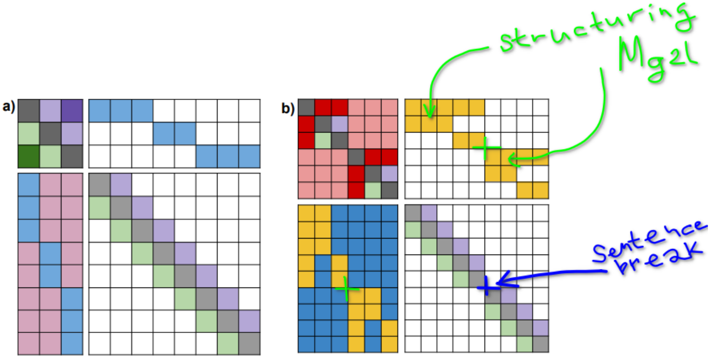
$$\alpha_{ij}^{g2g} = \frac{\exp(e_{ij}^{g2g})}{\sum_{\ell=1}^{n} \exp(e_{i\ell}^{g2g})}$$

Masking

$$e_{ij}^{g2g} = \frac{x_i^g W^Q (x_j^g W^K + a_{ij}^K)^T}{\sqrt{d_z}} - \boxed{(1 - M_{ij}^{g2g})C}$$

Structuring via Masking

- Using masking: $M_{l2l}, M_{l2g}, M_{g2l}, M_{l2l}$ (edges between tokens)
- colors – different connection types: *part-of*, *is-a*, etc.
  - **blue** – l2g connection with global tokens.
- Structuring: segments (sentences), using `[SENT_SEP]` special token
- **Masking find its application in pre-training.**



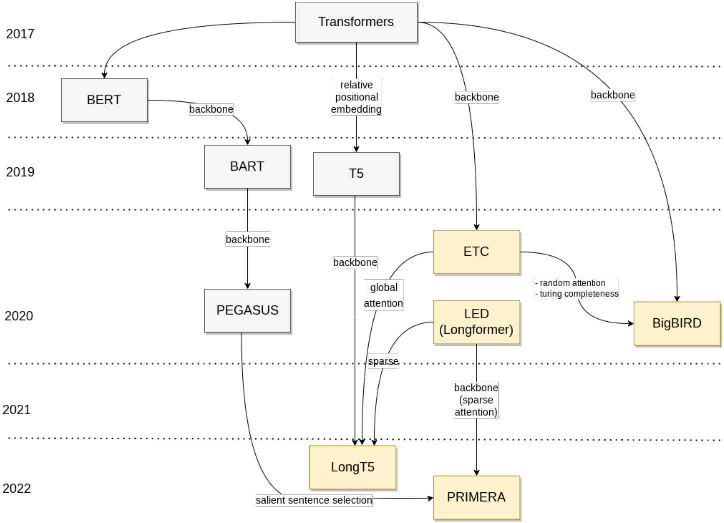
Results (NQ[5])

- Significant improvement when ETC $4K$ input (110M)[1] vs. $\text{BERT}_{\text{base}}$ (109M).
- Next improvement: double radius $\approx$ usage $8K$ input. (169M)
- Next improvement: Switch to ETC large + **Weights lifting from RoBERTa[liu2019roberta]. (558M)**

1 shared, no CPC, no hard g2l

[5] Tom Kwiatkowski et al. "Natural Questions: A Benchmark for Question Answering Research". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 452–466. doi: 10.1162/tacl_a_00276. url: https://aclanthology.org/Q19-1026.

Affection on Future Models for Text Summarization



Conclusion

Main Contributions as as follows:

- Sparsed attention as in BigBIRD, Longformer
- Structuring during pretraining stage
- Studies address transformer encoding part $\rightarrow$ weights lifting from BERT/RoBERTa due to a minor modifications towards attention complexity computation reduction