Using the Transformer

Efficient Transformers

Learning goals

- Understand the efficiency problems and shortcomings of transformer-based models
- Learn about some strategies to alleviate them

THE $\mathcal{O}(N^2)$ PROBLEM

Quadratic time & memory complexity of Self-Attention

- *Inductive bias of Transformer models:*
  Connect all tokens in a sequence to each other
- **Pro:** Can (theoretically) learn contexts of arbitrary length
- **Con:** Bad scalability limiting (feasible) context size

Resulting Problems:

- Several tasks require models to consume longer sequences
- *Efficiency:* Are there more efficient modifications which achieve similar or even better performance?

EFFICIENT TRANSFORMERS

Broad overview on so-called "X-formers" ▸ Tay et al. (2020)

- Efficient & fast Transformer-based models
  $\rightarrow$ Reduce complexity from $\mathcal{O}(n^2)$ to (up to) $\mathcal{O}(n)$
- Claim on-par (or even) superior performance
- Different techniques used:
  - Fixed/Factorized/Random Patterns
  - Learnable Patterns (extension of the above)
  - Low-Rank approximations or Kernels
  - Recurrence (see e.g. ▸ Transformer-XL (Dai et al., 2019))
  - Memory modules

Side note:

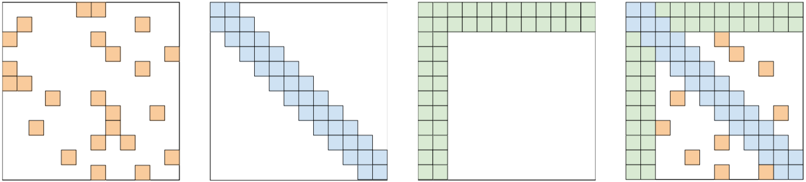
- Most Benchmark data sets not explicitly designed for evaluating long-range abilities of the models.
- Recently proposed: ▸ *Longe Range Arena: A benchmark for efficient transformers* (Tay et al., 2020)

INTRODUCING PATTERNS

Reasoning:

- Making every token attend to every other token might be unnecessary
- Introduce sparsity in the commonly dense attention matrix

Example:



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

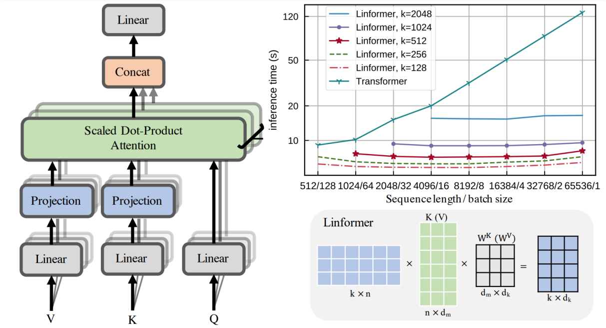
Figure 1: Building blocks of the attention mechanism used in BIGBIRD. White color indicates absence of attention. (a) random attention with $r = 2$, (b) sliding window attention with $w = 3$ (c) global attention with $g = 2$. (d) the combined BIGBIRD model.

Source: Zaheer et al. (2020)

LINEAR SELF-ATTENTION

Reasoning ▸ Wang et al. (2020)

- Most information in the Self-Attention mechanism can be recovered from the first few, largest singular values
- Introduce additional *k*-dimensional projection before self-attention



Source: Wang et al. (2020)

DEBERTA

Disentangled Attention ▸ He et al. (2020)

- Each token represented by two vectors for content ($\mathbf{H}_i$) and relative position ($\mathbf{P}_{i|j}$)
- Calculation of the Attention Score:

$$\begin{aligned} A_{i,j} &= \{\boldsymbol{H_i}, \boldsymbol{P_{i|j}}\} \times \{\boldsymbol{H_j}, \boldsymbol{P_{j|i}}\}^\intercal \\ &= \boldsymbol{H_i}\boldsymbol{H_j}^\intercal + \boldsymbol{H_i}\boldsymbol{P_{j|i}}^\intercal + \boldsymbol{P_{i|j}}\boldsymbol{H_j}^\intercal + \boldsymbol{P_{i|j}}\boldsymbol{P_{j|i}}^\intercal \end{aligned}$$

- with content-to-content, content-to-position, position-to-content and position-to-position attention

DISENTANGLED ATTENTION

Standard (Single-head) Self-Attention:

$$\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{W_q}, \boldsymbol{K} = \boldsymbol{H}\boldsymbol{W_k}, \boldsymbol{V} = \boldsymbol{H}\boldsymbol{W_v}, \boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^\intercal}{\sqrt{d}}$$

$$\boldsymbol{H_o} = \text{softmax}(\boldsymbol{A})\boldsymbol{V}$$

Disentangled Attention*:

$$\boldsymbol{Q_c} = \boldsymbol{H}\boldsymbol{W_{q,c}}, \boldsymbol{K_c} = \boldsymbol{H}\boldsymbol{W_{k,c}}, \boldsymbol{V_c} = \boldsymbol{H}\boldsymbol{W_{v,c}}, \boldsymbol{Q_r} = \boldsymbol{P}\boldsymbol{W_{q,r}}, \boldsymbol{K_r} = \boldsymbol{P}\boldsymbol{W_{k,r}}$$

$$\tilde{A}_{i,j} = \underbrace{\boldsymbol{Q_i^c}\boldsymbol{K_j^c}^\intercal}_{\text{(a) content-to-content}} + \underbrace{\boldsymbol{Q_i^c}\boldsymbol{K_{\delta(i,j)}^r}^\intercal}_{\text{(b) content-to-position}} + \underbrace{\boldsymbol{K_j^c}\boldsymbol{Q_{\delta(j,i)}^r}^\intercal}_{\text{(c) position-to-content}}$$

$$\boldsymbol{H_o} = \text{softmax}\left(\frac{\tilde{\boldsymbol{A}}}{\sqrt{3d}}\right)\boldsymbol{V_c}$$

*Position-to-position part is removed since it, according to the authors, does not provide much additional information as *relative* position emebeddings are used.