Using the Transformer

BigBird

▸ Zaheer et al. (2020)

Learning goals

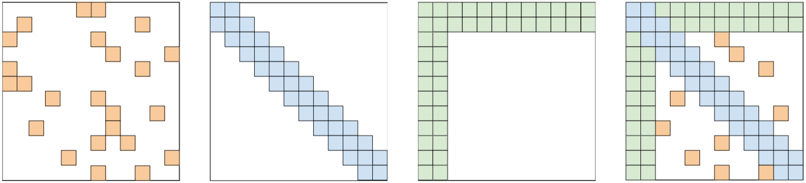
- Understand subtleties of Self-Attention
- BigBird architecture using patterns

ATTENTION IN THE ENCODER

In the Transformer:

- Independent, repeated application of the same process
- 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.

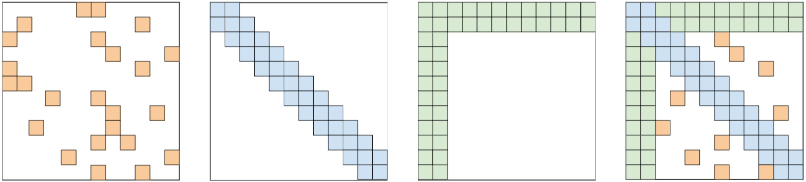
Source: Zaheer et al. (2020)

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)