Also claim quadratic dependency to linear

- Universal approximator of sequence functions
- Turing complete, preserving properties of quadratic, full attention model

Benefits of O(1) global tokens (such as CLS) that attend to entire sequence as part of the sparse attention mechanism

- Can handle sequences of length 8x what was previously possible

Improves performance on NLP

- Just because attention isn't complete, accuracy is still higher because more context?

Pretrain- fine tune paradigm

Attention lets each token to be processed in parallel which is better than LSTM or RNN which needs to be sequential
Sequence length is very limited due to quadratic requirement

- Tasks like QA and document classification need longer sequences
- Also for extracting contextual representations of genomic sequences like DNA
    - Promoter- region and chromatin profile prediction

It wasn't clear from the design if self attention was as effective as RNNS

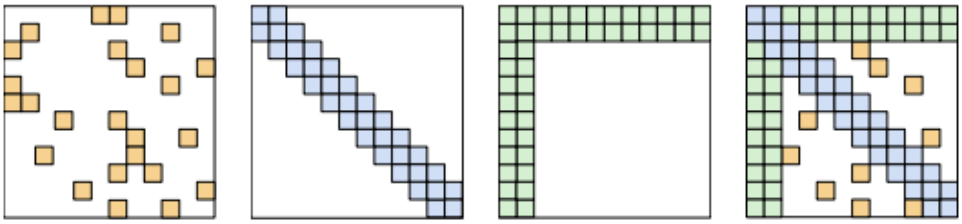
- Self attention does not obey sequence order as it is permutation equivariant
- Resolved as someone showed transformers are expressive enough to capture all continuous sequence to sequence functions with a compact domain
- Someone else shows they are Turing complete
- Can we achieve the empirical benefits of a fully quadratic self-attention scheme using fewer inner-products? Do these sparse attention mechanisms preserve the expressivity and flexibility of the original network?

Sparse attention mechanism

- Take inspiration from graph sparsification methods
- Understand where proof for expressiveness breaks down when full- attention is relaxed to form proposed attention pattern
- Mechanism
    - Global tokens attending on all parts of sequence
    - All tokens attending to a set of local neighbor tokens
    - All tokens attending to a set of random tokens

1. Embraces length of limitation of transformer and develops methods around it such as sliding window
    a. Use some other mechanism to select a smaller subset of relevant contexts to feed in the transformer and optionally iterate, call transformer block multiple times with different contexts each time
    b. Require significant engineering effort and are hard to train

2. Methods that do not require full attention, reducing memory and computation requirements (this paper)



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

What is a “sparse shift operator”
“Turing completeness”
Limitations