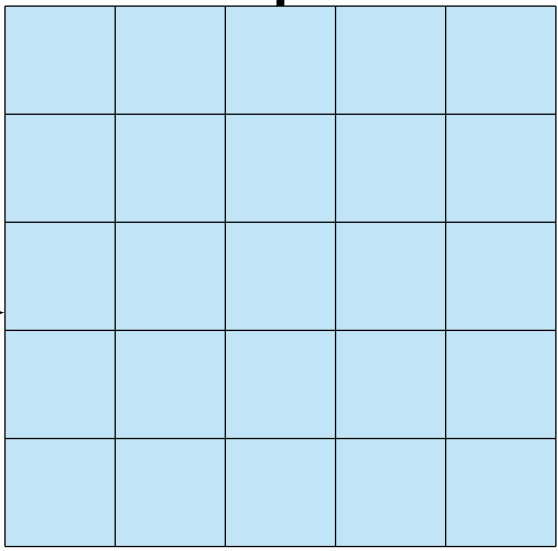


Dependency heads

Output probabilities

argmax



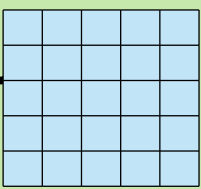
Self-attention weights

Tranformer Decoder

Previous target tokens

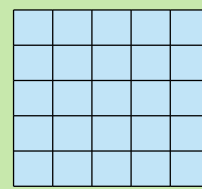
Transformer Encoder

Self-attention head #1

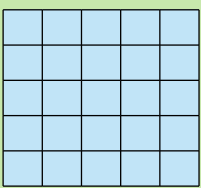


Layer 6

Self-attention head #8

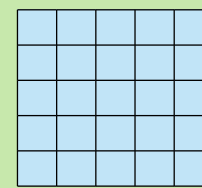


Self-attention head #1



Layer 1

Self-attention head #8



Source tokens