Self-Attention in Transformer

Three uses of multi-head attention in Transformer

- Encoder-Decoder Attention:
  - Q: previous decoder layers; K = V: outputs of encoder

  $\Rightarrow$ Decoder positions attend to all positions of the input.
- Encoder Self-Attention:
  - Q = K = V: outputs of the previous layer of the encoder

  $\Rightarrow$ Encoder positions attend to all positions of previous layer.
- Decoder Self-Attention:
  - Q = K = V: outputs of the previous decoder layer.
  - Masking used to prevent depending on future outputs.

  $\Rightarrow$ Decoder attends to all its previous outputs.