# 10.7. The Transformer Architecture

Lecture based on "Dive into Deep Learning" **http://D2L.AI** (Zhang et al., 2020)

Prof. Dr. Christoph Lippert

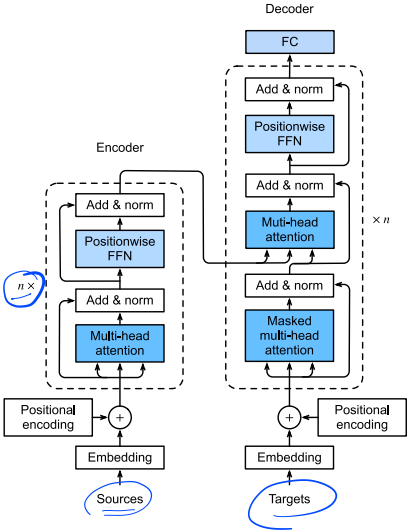
Digital Health & Machine Learning

# Motivation

- Self-attention allows for both parallel computation and has the shortest maximum path length.
- Earlier self-attention models still relied on RNNs for input representations,
- the transformer model is solely based on attention mechanisms without any convolutional or recurrent layer.
- Though originally proposed for sequence to sequence learning on text data, transformers have been pervasive in modern deep learning

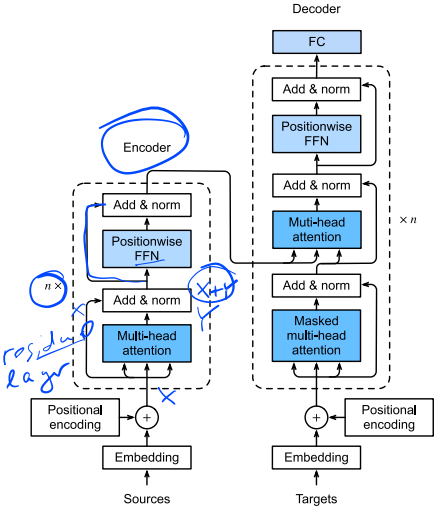
  applications to language, vision, speech, and reinforcement learning.

# Model




- The transformer is composed of an encoder and a decoder.
- The input (source) and output (target) sequence embeddings are added with positional encoding
- the encoder and the decoder that stack modules based on self-attention.

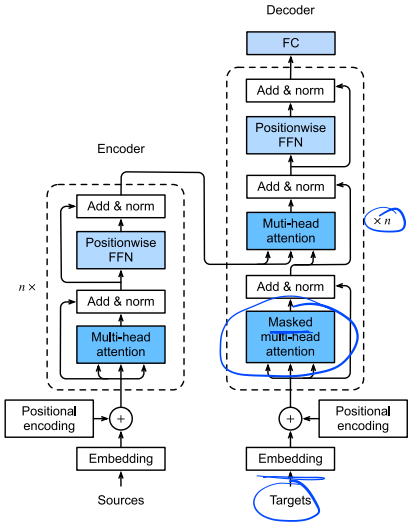
# Encoder




- For any input $\mathbf{x} \in \mathbb{R}^d$ at any position of the sequence $\text{sublayer}(\mathbf{x}) \in \mathbb{R}^d$
- This way the residual connection is feasible. $\mathbf{x} + \text{sublayer}(\mathbf{x}) \in \mathbb{R}^d$
- The Encoder outputs a $d$-dimensional vector representation for each position of the input sequence.

- The encoder is a stack of multiple identical layers, where each layer has two sublayers
- The first is a multi-head self-attention pooling and the second is a positionwise feed-forward network.
- Inspired by ResNet, a residual connection is employed around both sublayers
- This addition from the residual connection is immediately followed by layer normalization.

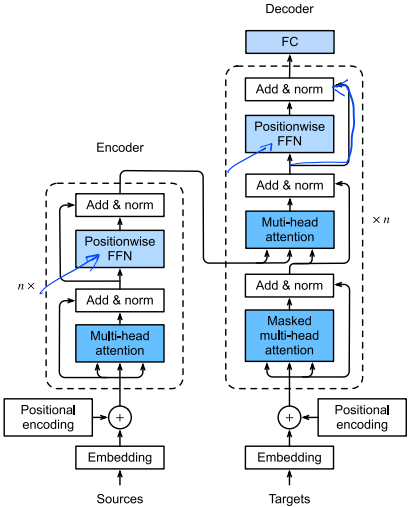
# Decoder




- In the encoder-decoder attention, queries are from the outputs of the previous decoder layer, and the keys and values are from the transformer encoder outputs.
- each position in the decoder is allowed to only attend to all positions in the decoder up to that position using *masked* attention,
- This *masked* attention preserves the auto-regressive property, ensuring that the prediction only depends on those output tokens that have been generated.

- The decoder is also a stack of multiple identical layers with residual connections and layer normalizations
- The decoder inserts an encoder-decoder attention sublayer, between the other sublayers.
- In the encoder-decoder attention, queries are from the outputs of the previous decoder layer, and the keys and values are from the transformer encoder outputs.
- In the decoder self-attention, queries, keys, and values are all from the outputs of the previous decoder layer.

# Positionwise Feed-Forward




The *positionwise* feed-forward networks transform the representation at all the sequence positions using the same MLP.

- input tensor of shape (batch size, number of time steps or sequence length in tokens, number of hidden units or feature dimension)
- output tensor of shape (batch size, number of time steps, num_outputs).

# Summmary

- The transformer is an instance of the encoder-decoder architecture, though either the encoder or the decoder can be used individually in practice.
- In the transformer, multi-head self-attention is used for representing the input sequence and the output sequence, though the decoder has to preserve the auto-regressive property via a masked version.
- Both the residual connections and the layer normalization in the transformer are important for training a very deep model.
- The positionwise feed-forward network in the transformer model transforms the representation at all the sequence positions using the same MLP.