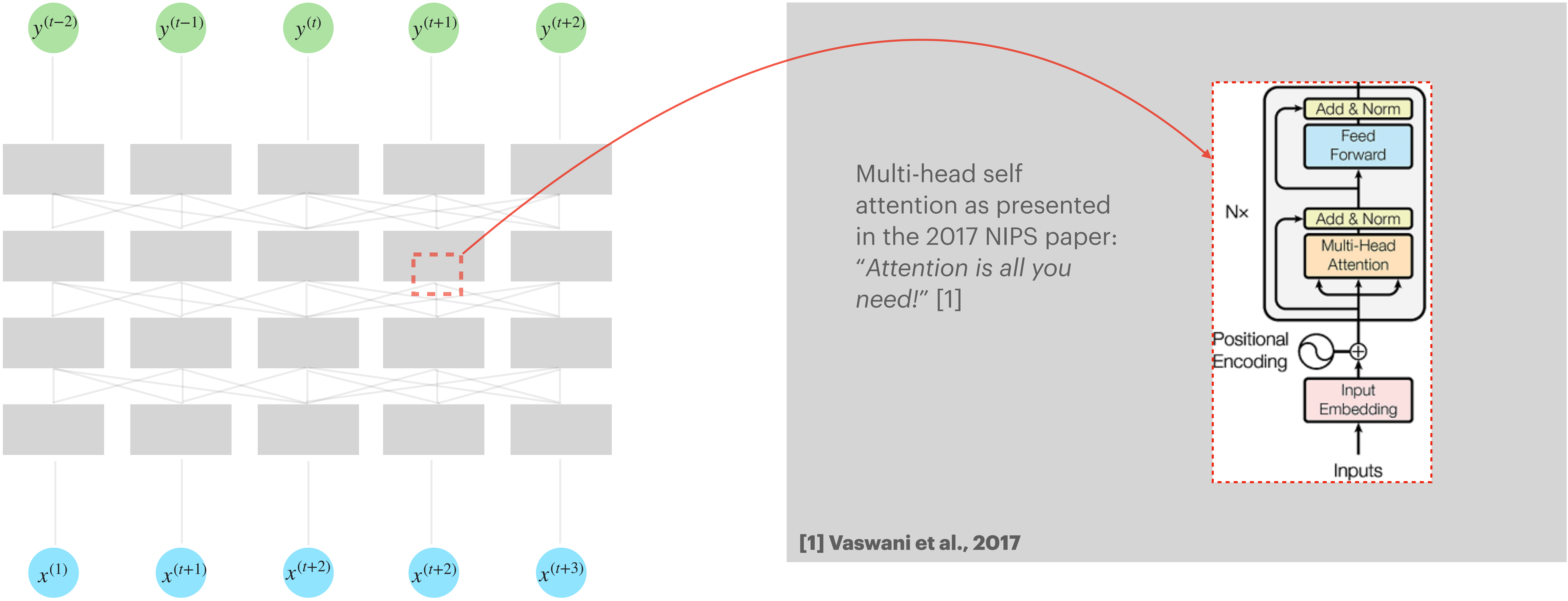
# Transformer networks

- Transformer networks describe text sequences of a fully connected graph
  - The embedding at each sequence position, at each layer, is a function of the embeddings at all sequence positions from the previous layer.
  - The computation that produces the embedding at each sequence position at each layer is referred to as multi-head self attention (more on this in lecture 09).
  - This architecture is advantageous from both in terms of computational and optimization.



Multi-head self attention as presented in the 2017 NIPS paper: *“Attention is all you need!”* [1]

**[1] Vaswani et al., 2017**

# LM evaluation using perplexity

$$
\begin{aligned}
PPL(\mathbf{X}^{(i)}) &= \exp\left\{ -\frac{1}{T}\sum_{t=1}^{T} \log P\left( \mathbf{x}^{(t)_i} \,|\, \mathbf{x}^{(t-n)_i} \ldots \mathbf{x}^{(t-1)_i}; \boldsymbol{\theta} \right) \right\} \\
&= \exp\left\{ H(P_D, P_{\boldsymbol{\theta}}) \right\}
\end{aligned}
$$