Attention Mechanism and Transformer Architecture

Instructor: Xiaoqian Wang

Review of RNN

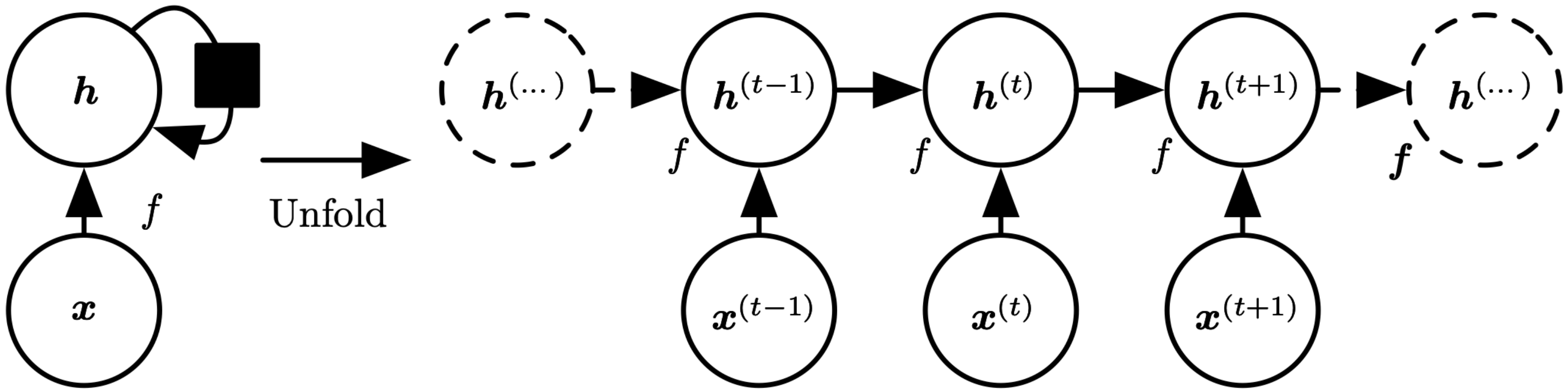


Figure 10.2

Figure from (Goodfellow 2016)

How to use RNN for sequence-to-sequence task?

- **Example**: Translation between French and English
- **Option 1:** one-to-one input/output RNN
  - Problem: Sequences could have different lengths.
  - Problem: The order of words is not the same in French and English.

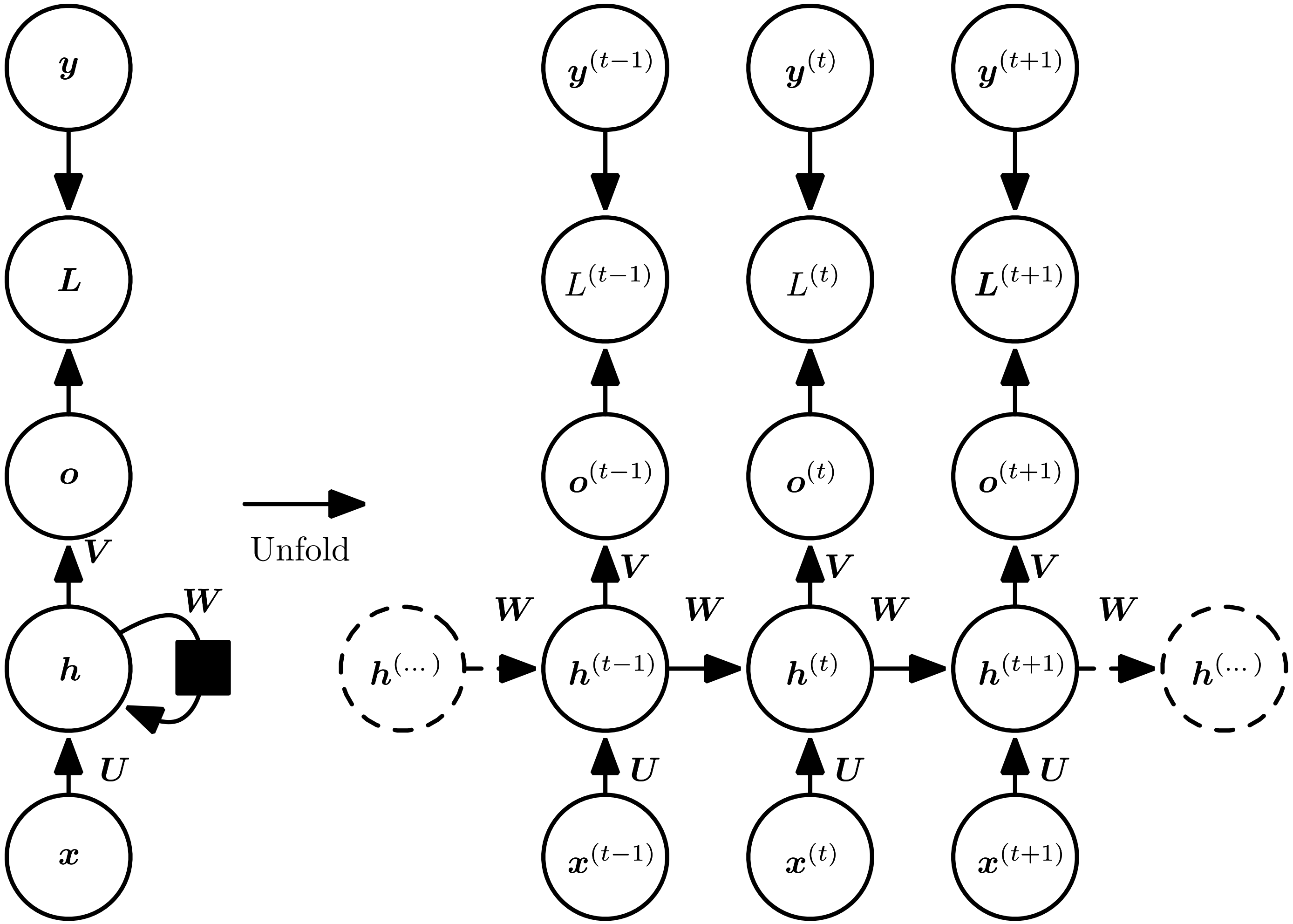


Figure from (Goodfellow 2016)

How to use RNN for sequence-to-sequence task?

- **Example**: Translation between French and English
- **Option 2:** encoder-decoder structure
  - Hidden state **s** in the decoder

$$s_t = g(s_{t-1}, y_{t-1}, c)$$

  - Problem: difficult to encode a whole input sentence into a single fixed-length context vector **c**, especially for long sentences

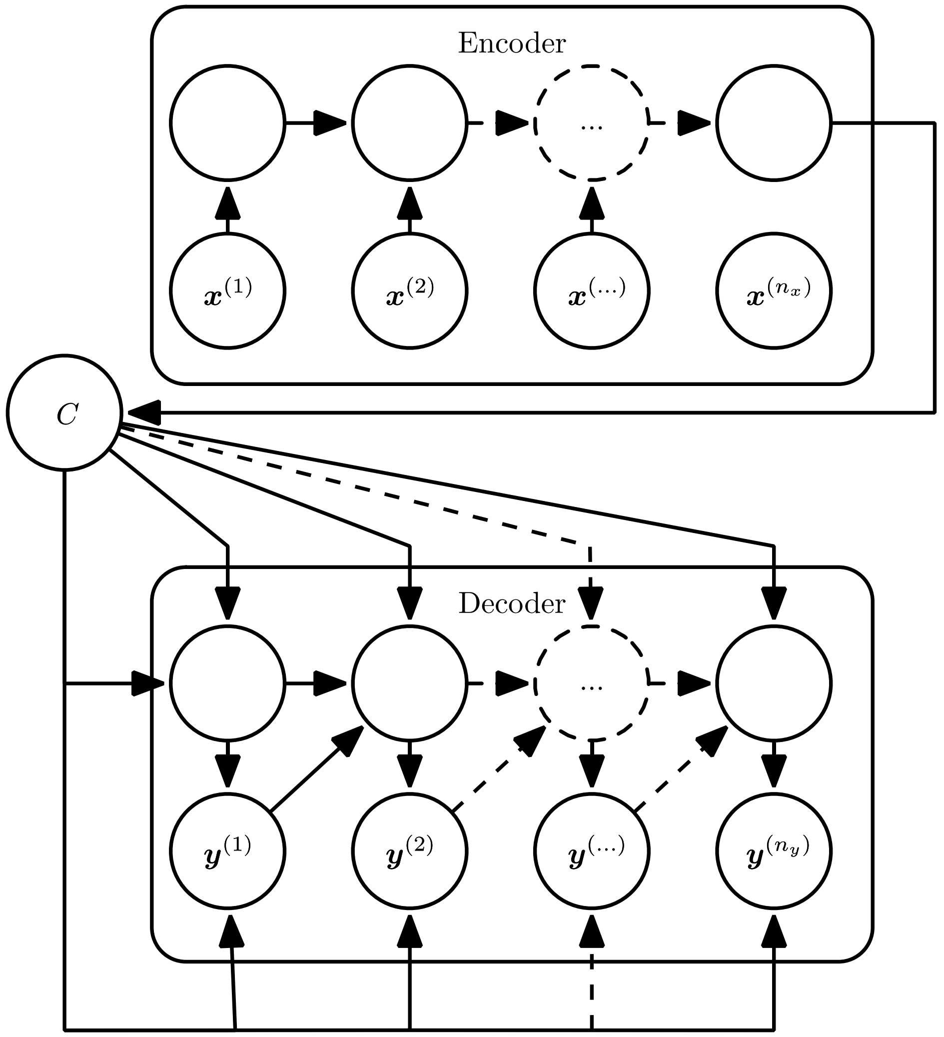


Figure from (Goodfellow 2016)

Attention mechanism in sequence-to-sequence task

- **Example**: Translation between French and English
- Attention mechanism to learn distinct context vectors:
    - Hidden state **s** in the decoder

$$s_t = g(s_{t-1}, y_{t-1}, \boxed{c_t})$$

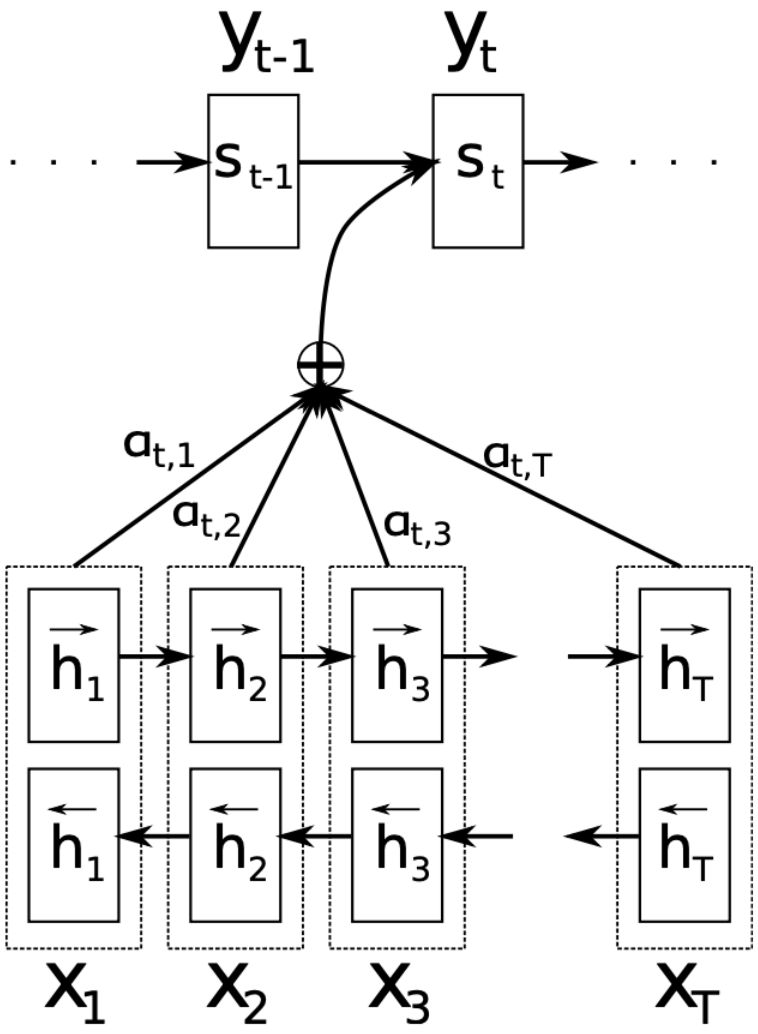


Figure from (Bahdanau et al. 2015)

Attention mechanism in sequence-to-sequence task

- **Example**: Translation between French and English
- Attention mechanism to learn distinct context vectors:
  - Hidden state **s** in the decoder

$$s_t = g(s_{t-1}, y_{t-1}, \boxed{c_t})$$

where $$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

$$e_{ij} = a(s_{i-1}, h_j)$$

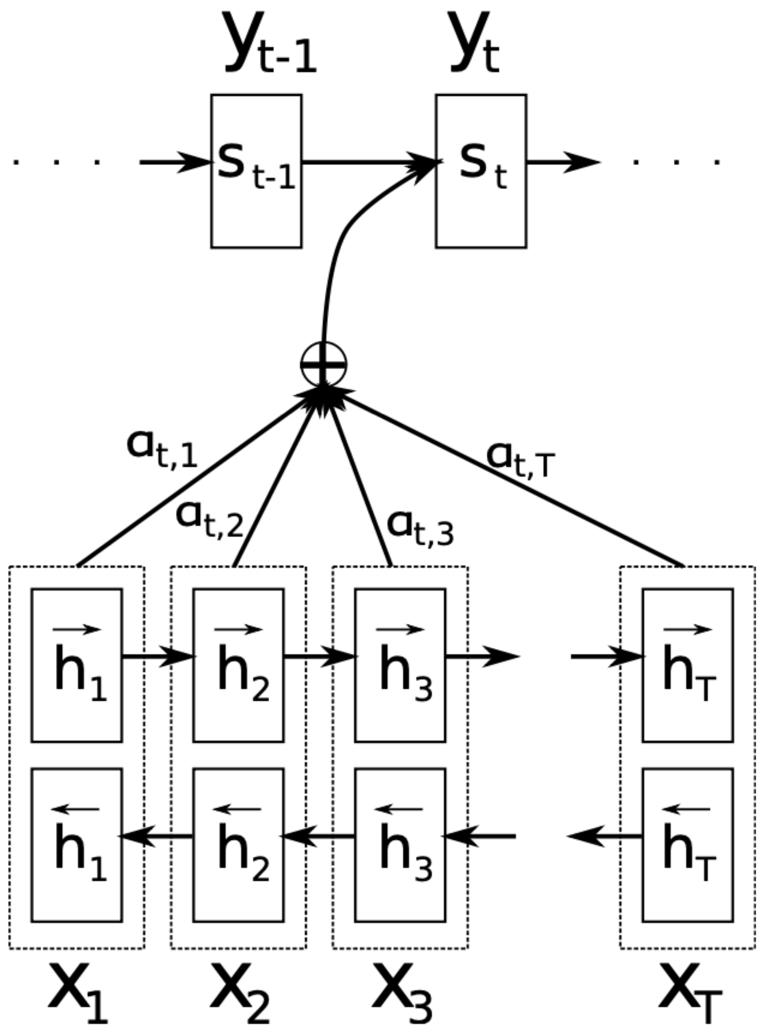


Figure from (Bahdanau et al. 2015)

Attention mechanism in sequence-to-sequence task

- Alignment model scores how well the inputs around position j and the output at position i match:

$$e_{ij} = a(s_{i-1}, h_j)$$

- The attention weights are normalized via softmax such that $\sum_j \alpha_{ij} = 1$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

- In other words, the attention mechanism determines which encoder outputs the network should "focus" on

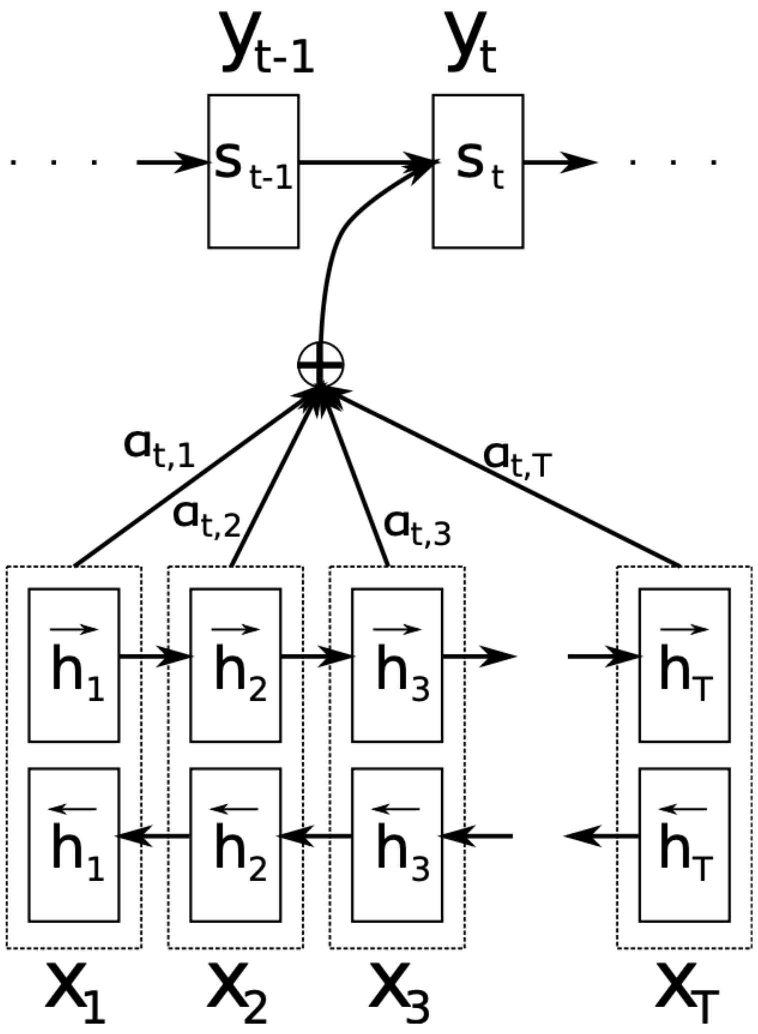
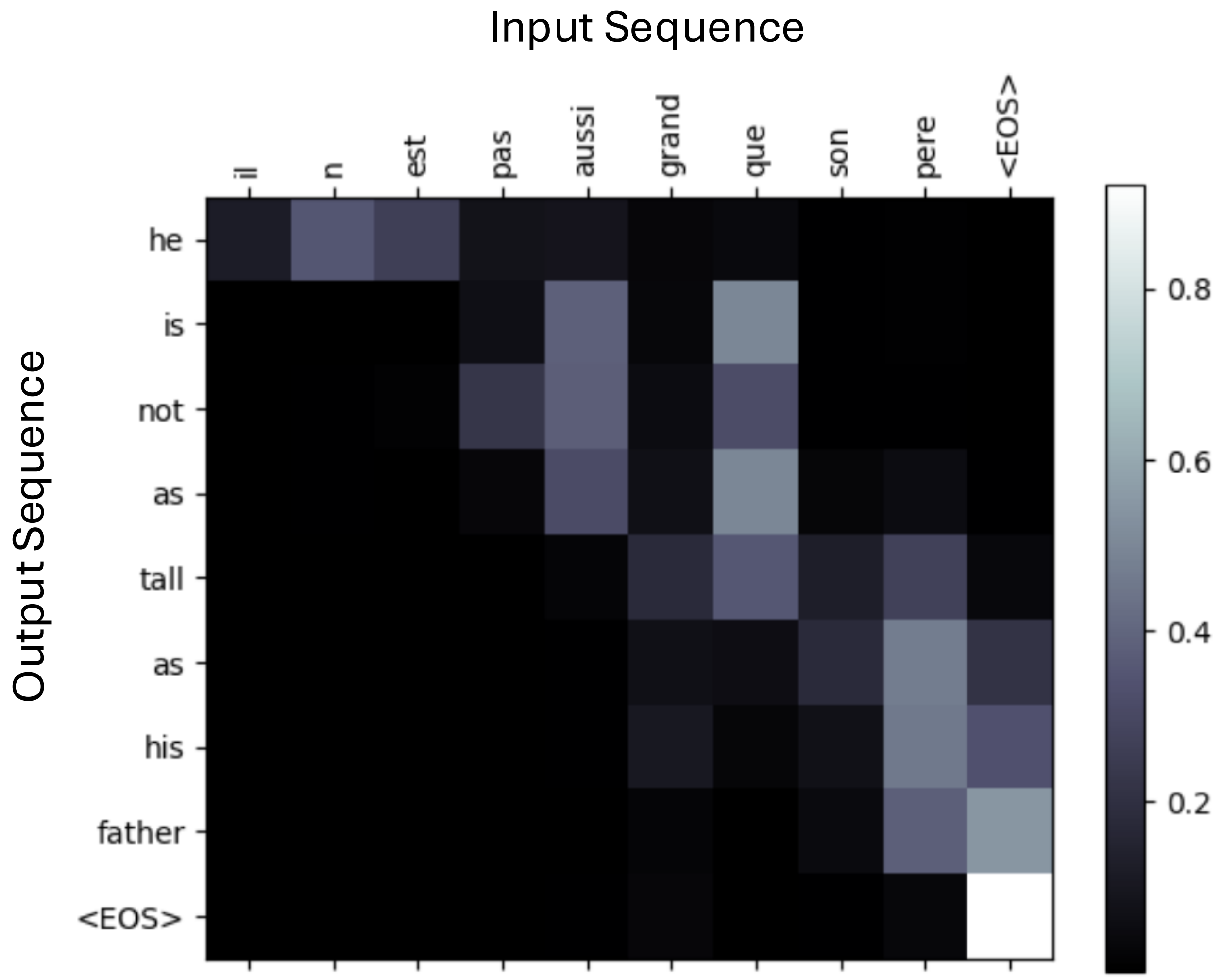


Figure from (Bahdanau et al. 2015)

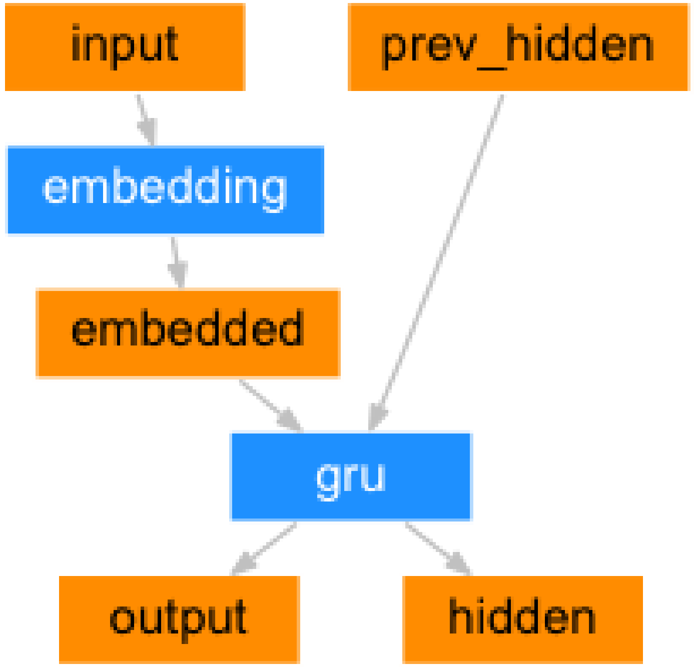
Visualization of attention weights as interpretation

- Visualization of the attention weights α shows what the model pays attention to when making the prediction.



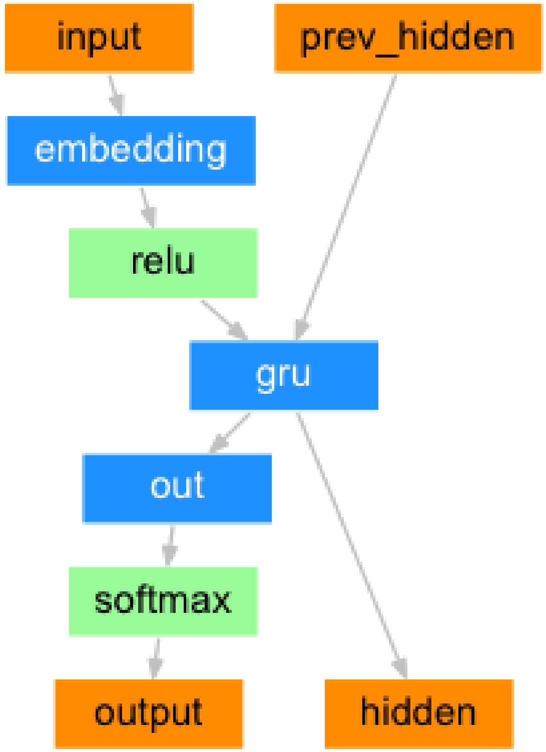
Demo of sequence to sequence task

- **Example**: Translation between French and English
  - Tutorial from https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
  - **Encoder** is simple RNN



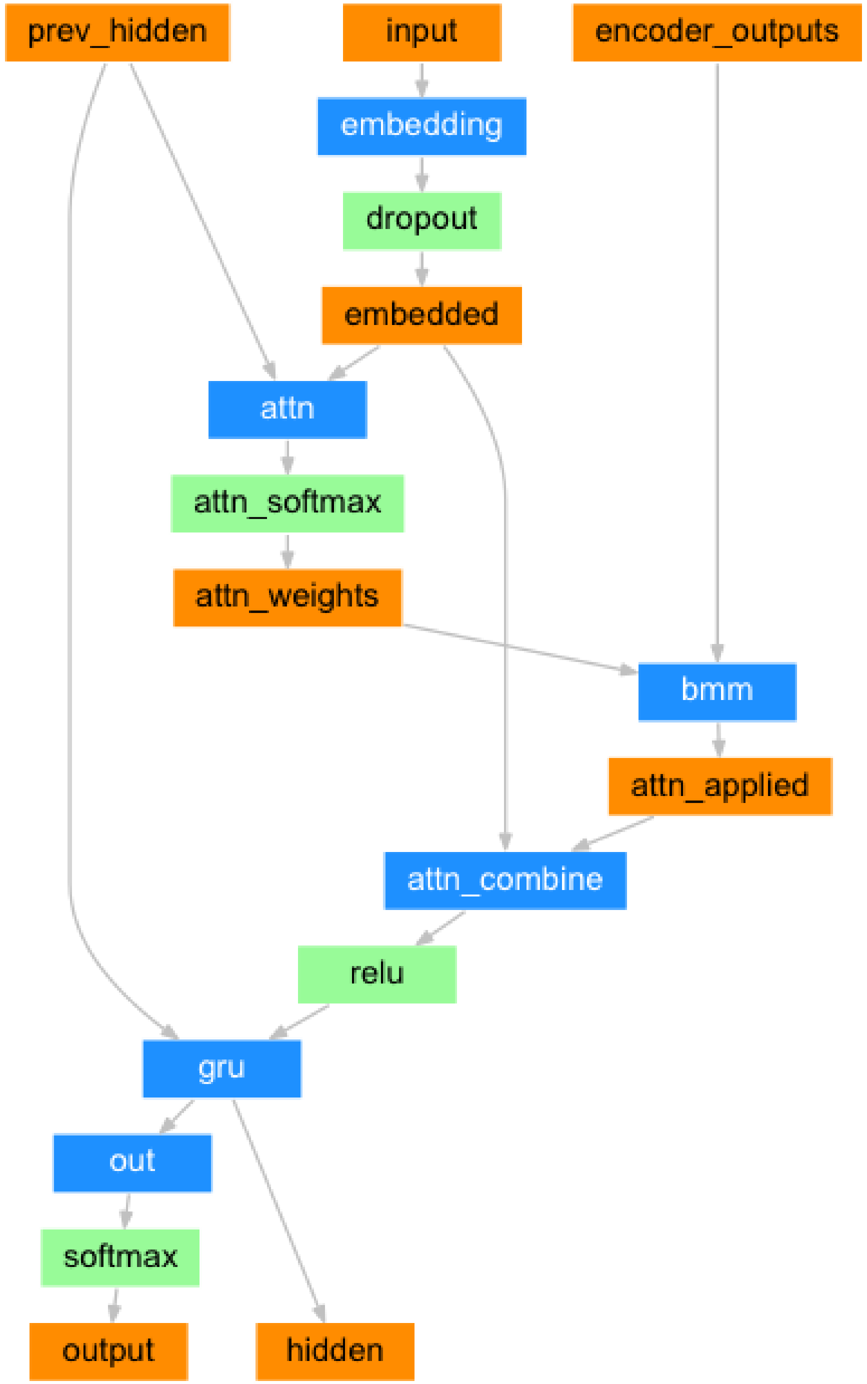
Demo of sequence to sequence task

- **Example**: Translation between French and English
  - Tutorial from https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
  - **RNNdecoder** includes a simple RNN



Demo of sequence to sequence task

- **Example**: Translation between French and English
  - Tutorial from https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
  - **AttnDecoderRNN** includes the attention mechanism



Our seq-2-seq attention is a special case of this more general attention mechanism

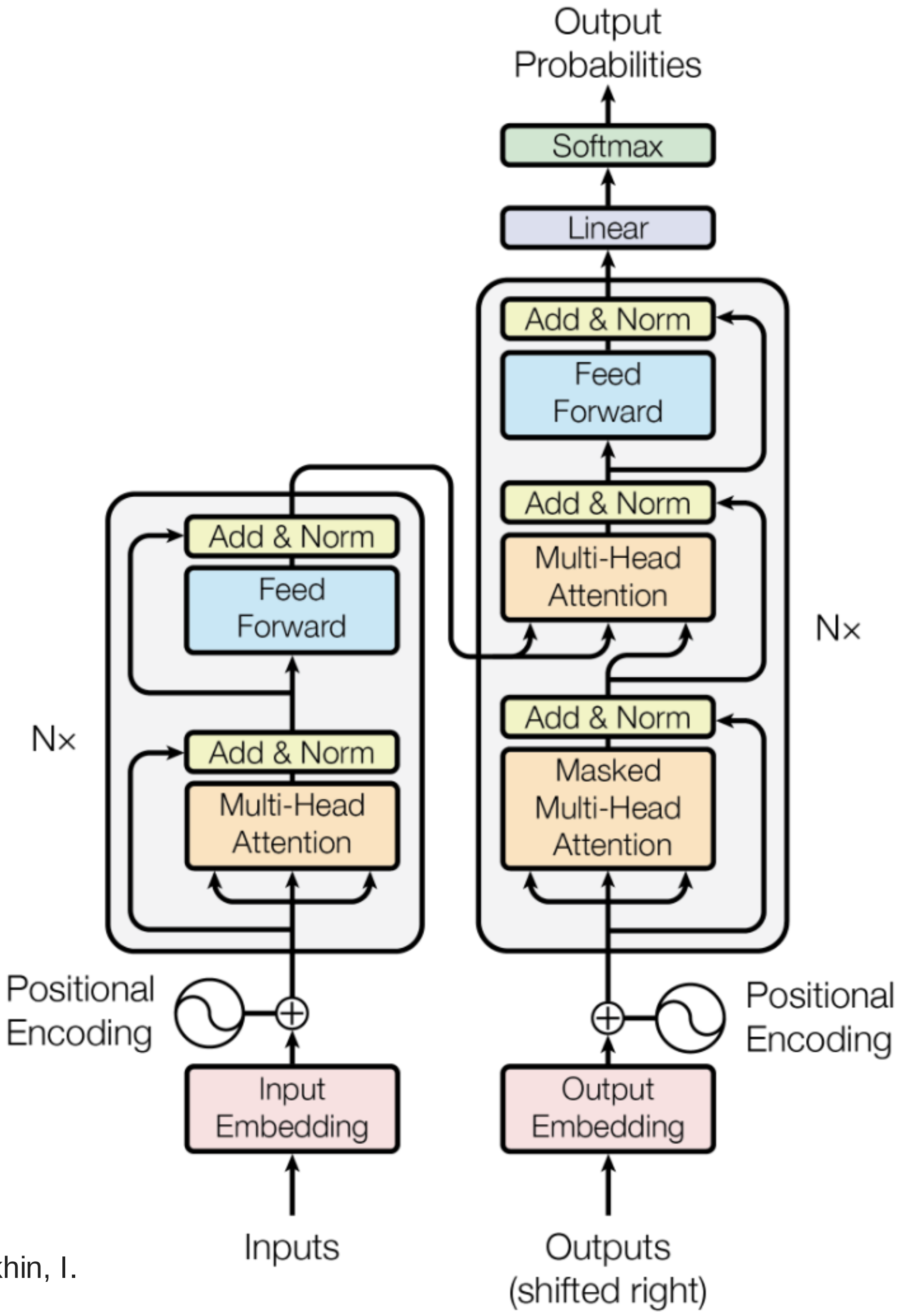
- The output of attention is a weighted average of the values

$$A(q, K, V) = \sum_i \alpha_i v_i = \sum_i \left(\frac{\exp\big(e(q, k_i)\big)}{\sum_j \exp\Big(e\big(q, k_j\big)\Big)} \right) v_i$$

  - q is the **query** input, K is the key matrix, V is the value matrix
  - α_i is the **attention weight** for the i-th value
  - $e(q, k_i)$ is the **attention score** (pre-softmax) based on the i-th key

Transformer

- No RNN structure, only attention, parallel computing
- Long-range interaction
- Positional encoding
- Multi-head attention



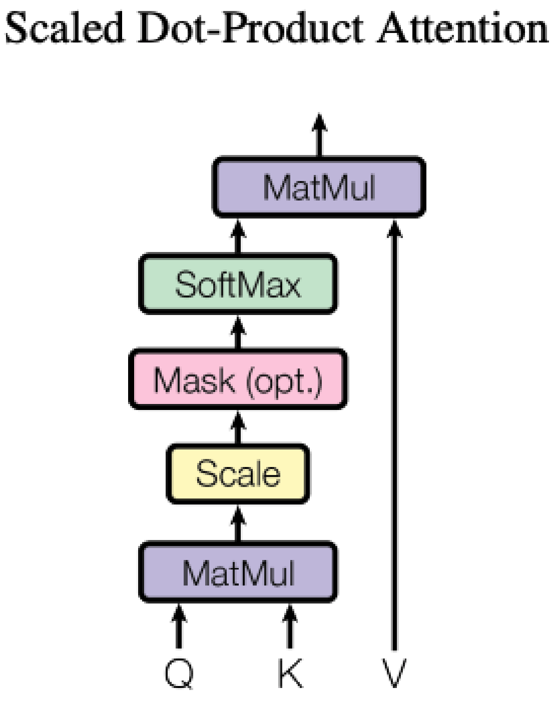
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Scaled Dot-Product Attention

- $X = [x_1, x_2, \dots, x_L]^T \in \mathbb{R}^{n \times d_{model}}$, n is sequence length
- Self attention to compute Q, K and V
  - $Q = XW_Q \in \mathbb{R}^{n \times d_k}$
  - $K = XW_K \in \mathbb{R}^{n \times d_k}$
  - $V = XW_V \in \mathbb{R}^{n \times d_v}$
- Dot-product attention with a scaling factor $\frac{1}{\sqrt{d_k}}$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Additive attention vs. dot-product attention



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

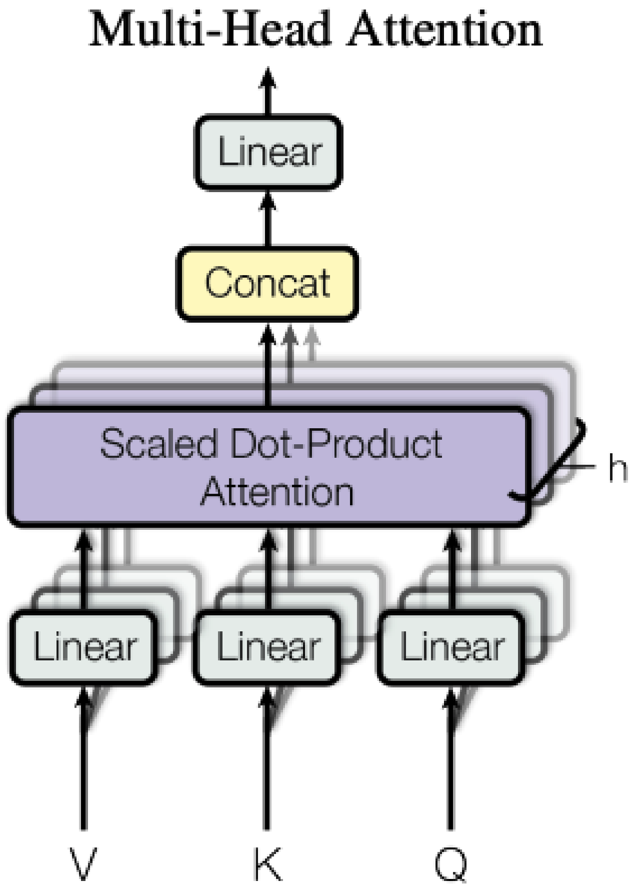
Multi-head attention

- Use h attention modules and concatenate their outputs

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

- Allowing the model to jointly attend to information from different representation subspaces at different positions

- Reduced dimension of each head



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

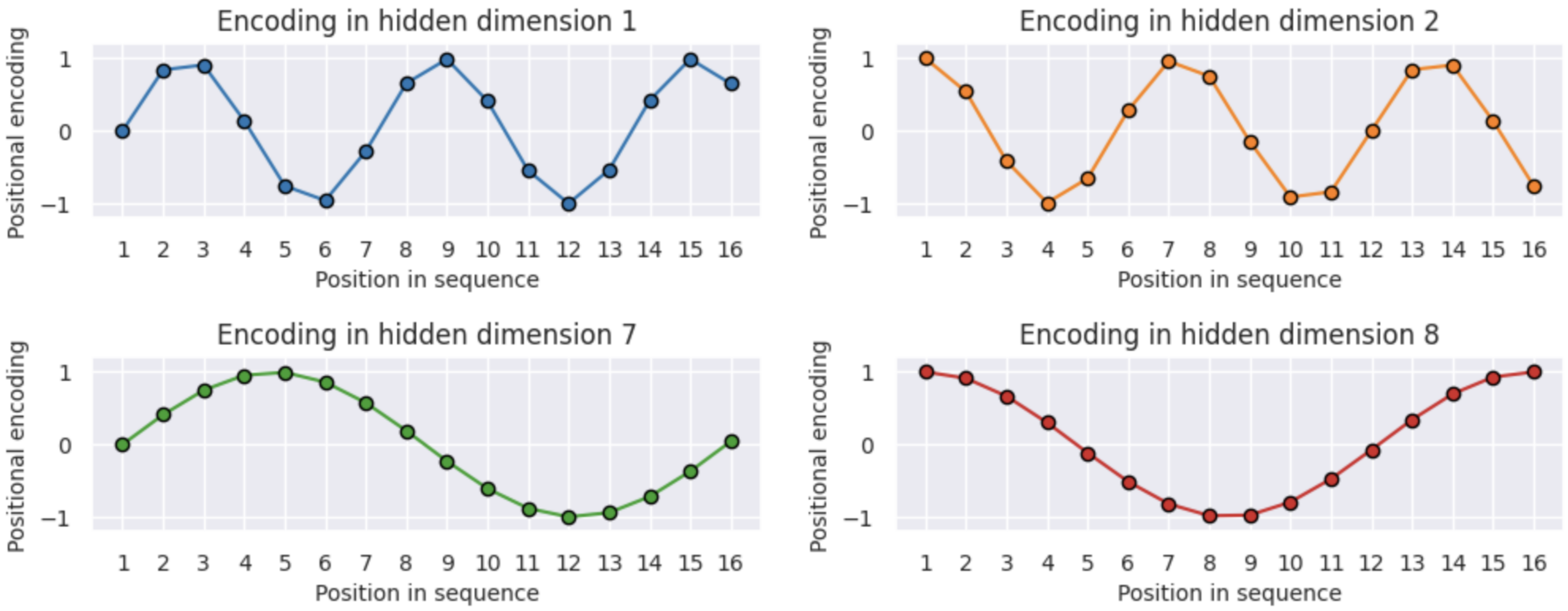
Positional Encoding

- Positional encoding has the same dimension d_{model} as the word embedding x
- Alternating between sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

where pos is the position and i is the dimension



Positional Encoding

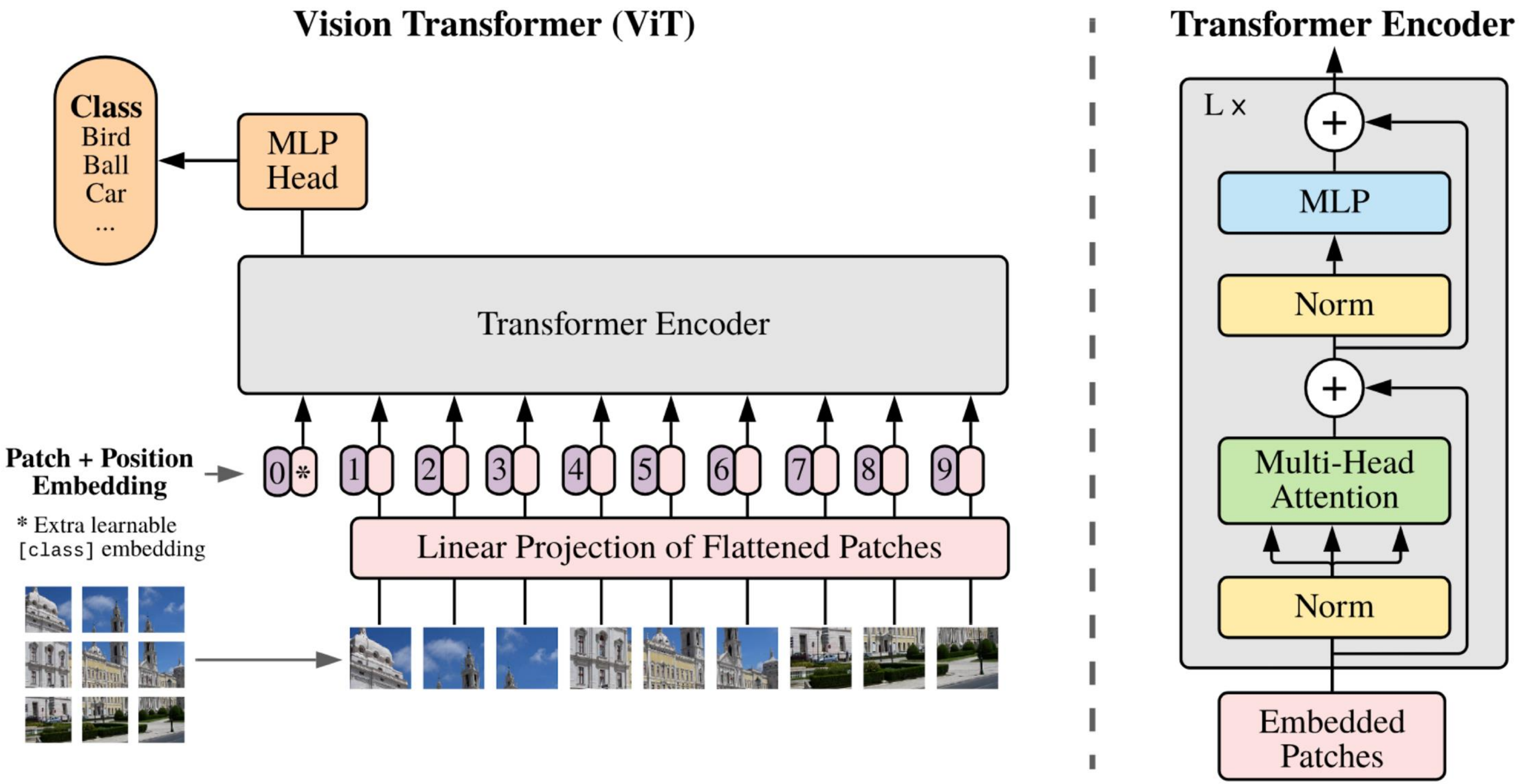
- The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$.

- This enables relative and absolute positioning information to be encoded
  - Enables relative positioning because for any k, PE_{pos+k} can be represented as a linear function of PE_{pos}

- There are other choices of positional encodings, either learned and fixed. Refer to Gehring, Jonas, et al. "Convolutional sequence to sequence learning." International conference on machine learning. PMLR, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Transformers are applicable to various contexts



Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In International Conference on Learning Representations. 2020.