Attention mechanisms & Transformers

CS 5624: Natural Language Processing

Spring 2025

https://tuvllms.github.io/nlp-spring-2025

Tu Vu

VIRGINIA TECH

Logistics

- Homework 1 & Quiz 1 are on their way
- Final project proposal due on February 28

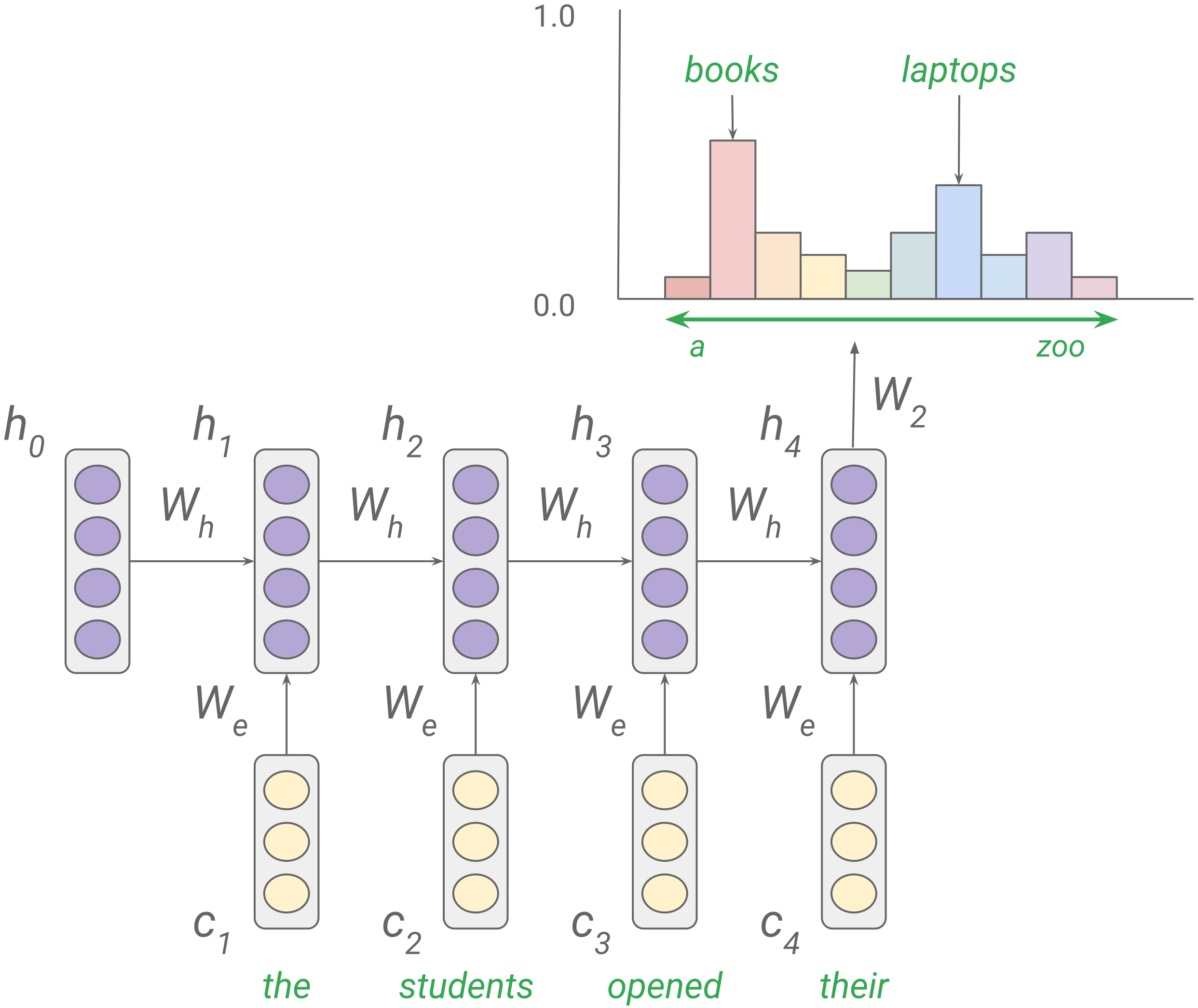
Recurrent neural networks (RNNs)

hidden states

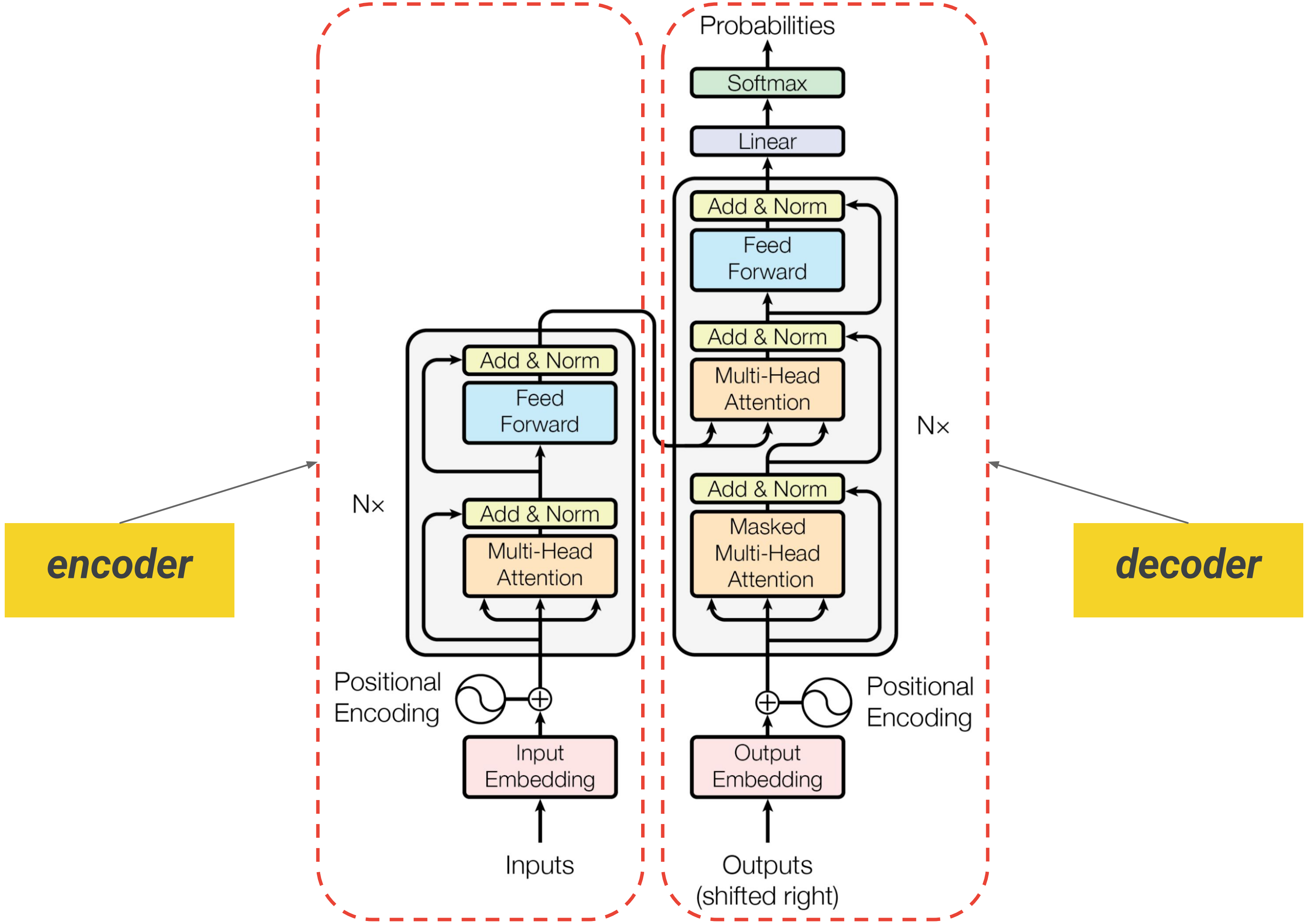
$$h^{(t)} = f(W_h h^{(t-1)} + W_e c^t)$$

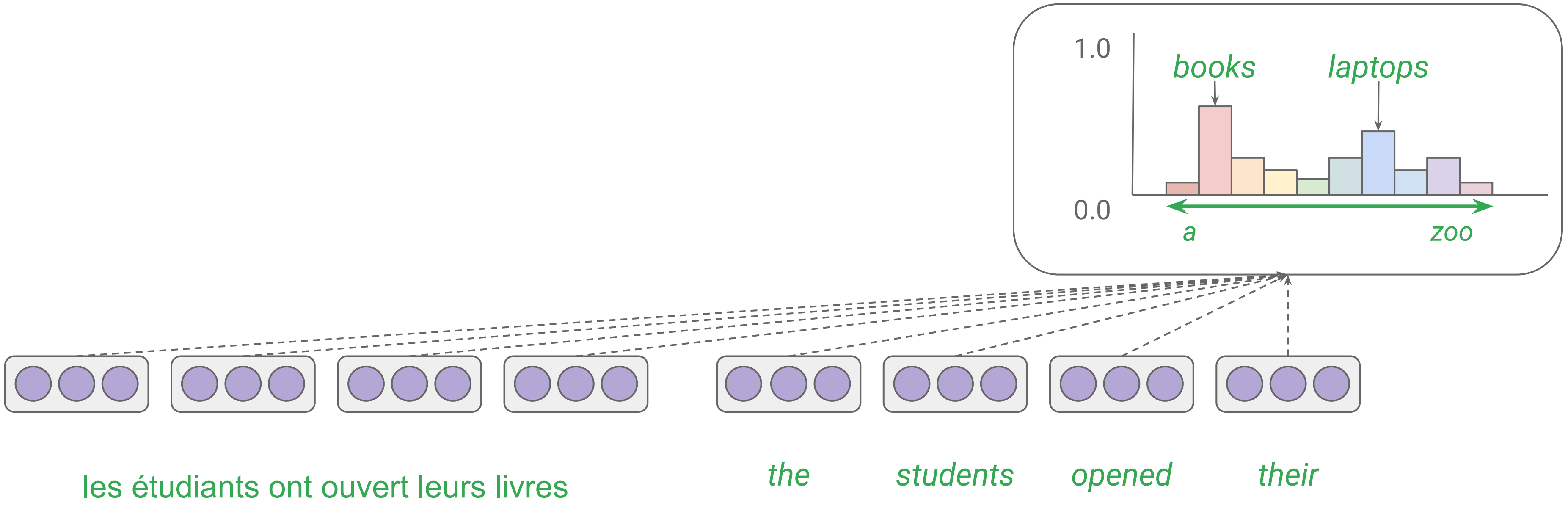
output distribution

$$\hat{y} = softmax(W_2 h^{(n-1)})$$



Encoder-decoder architecture


encoder
decoder
Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
Positional Encoding
Output Embedding
Outputs (shifted right)

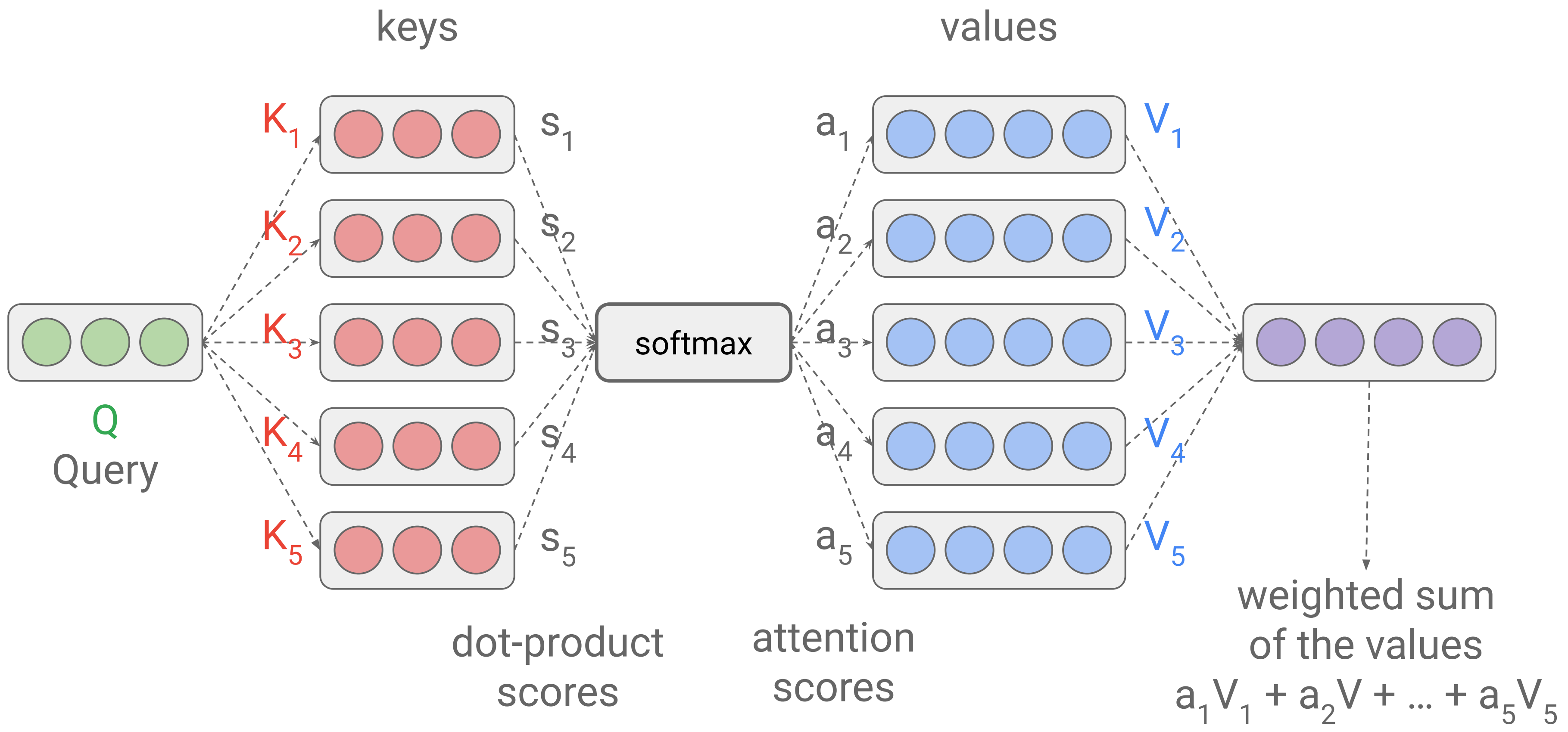

1.0
books
laptops
0.0
a
zoo
les étudiants ont ouvert leurs livres
the
students
opened
their

encoder
decoder

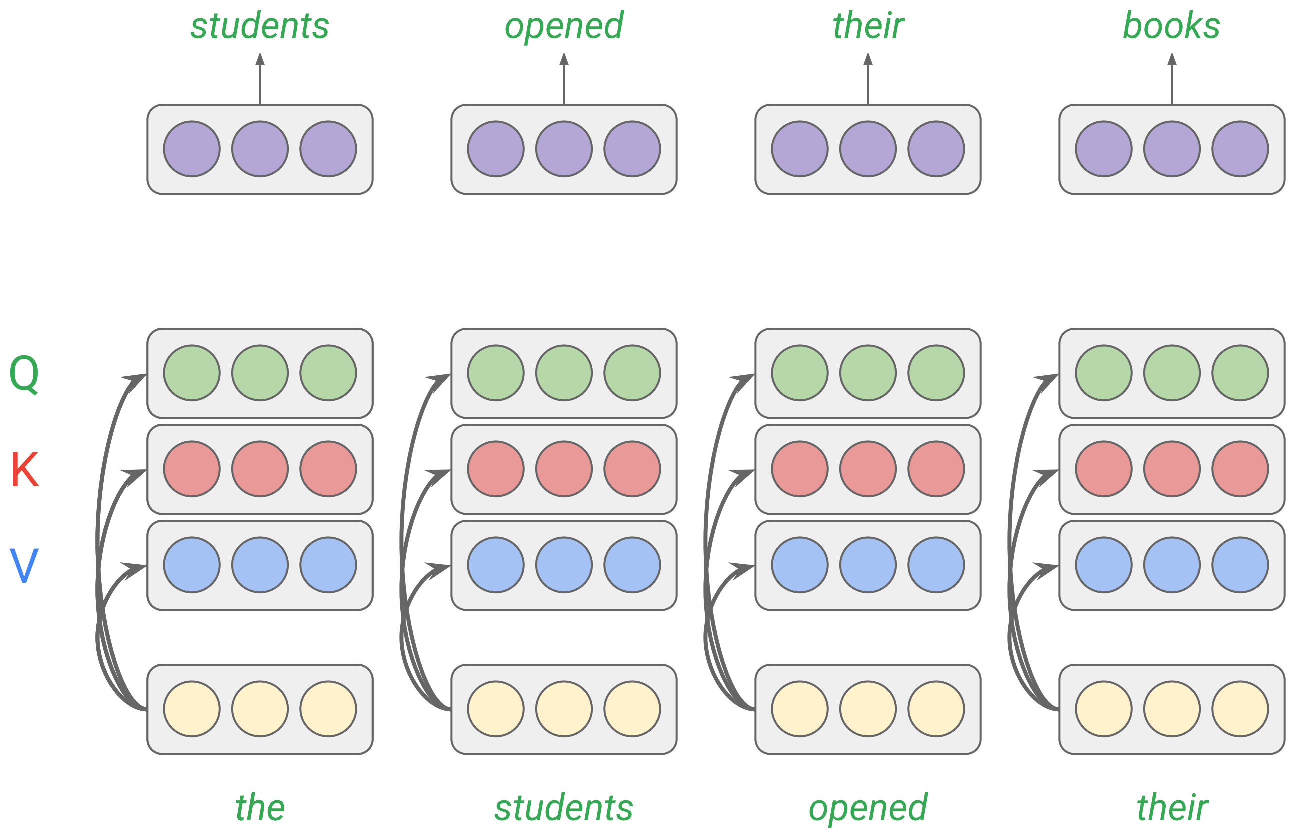
Different model architectures

- Encoder-only
    - BERT
- Encoder-decoder
    - T5
- Decoder-only
    - GPT

Attention mechanism

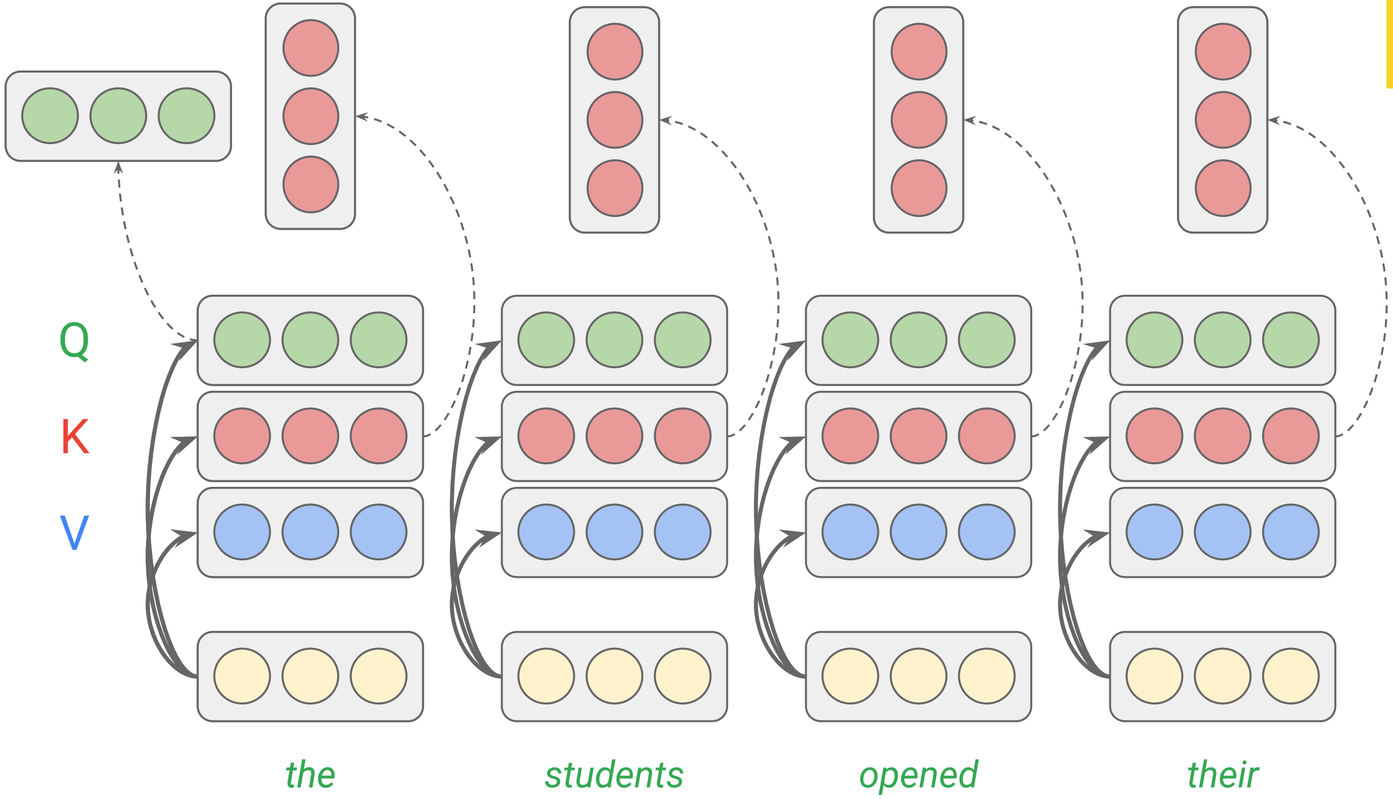


Self-attention



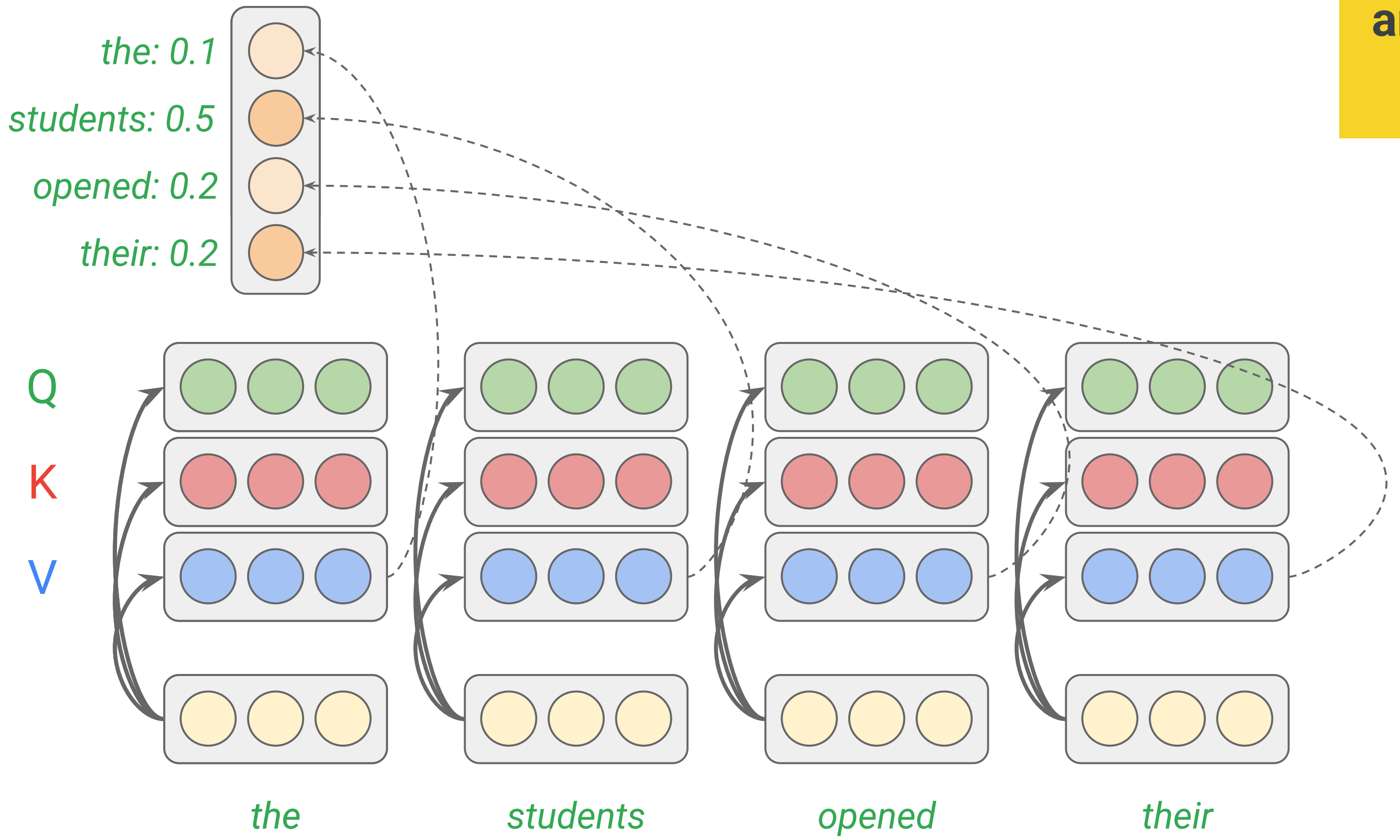
Self-attention (cont'd)

all **computations are parallelized**



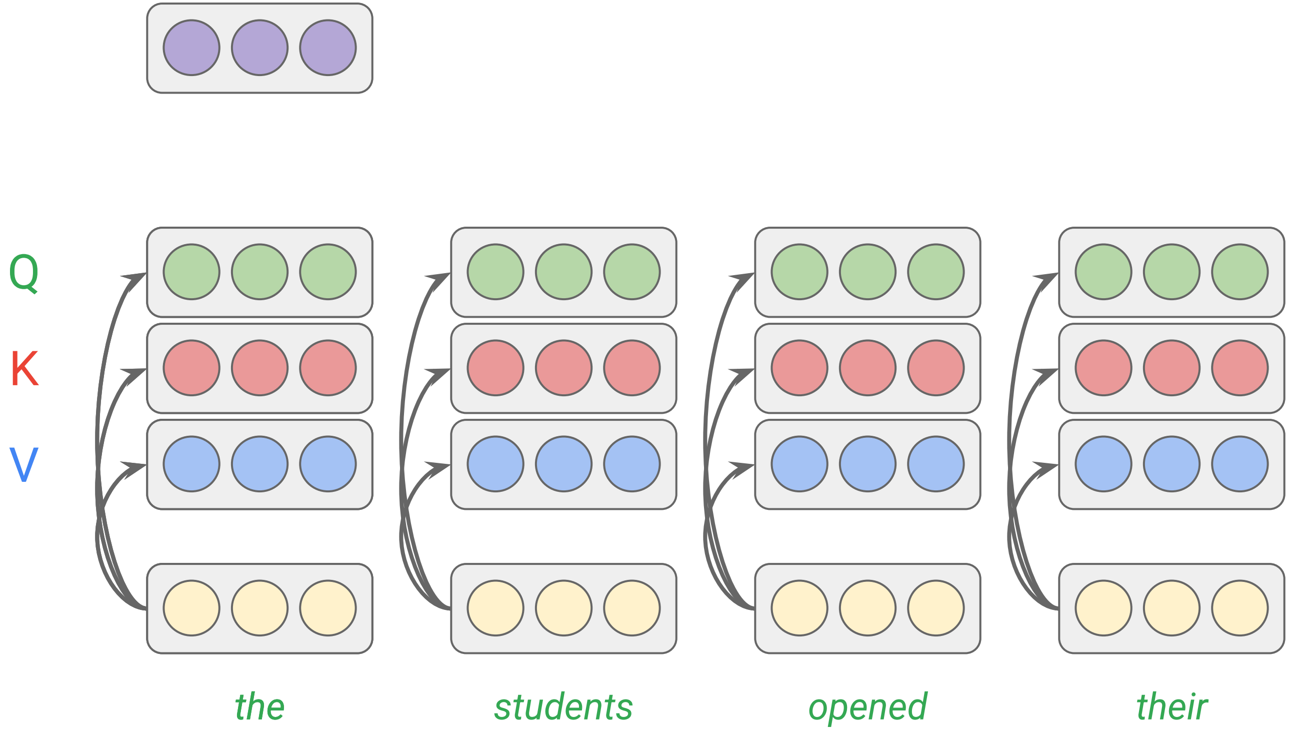
Self-attention (cont’d)

all computations are parallelized



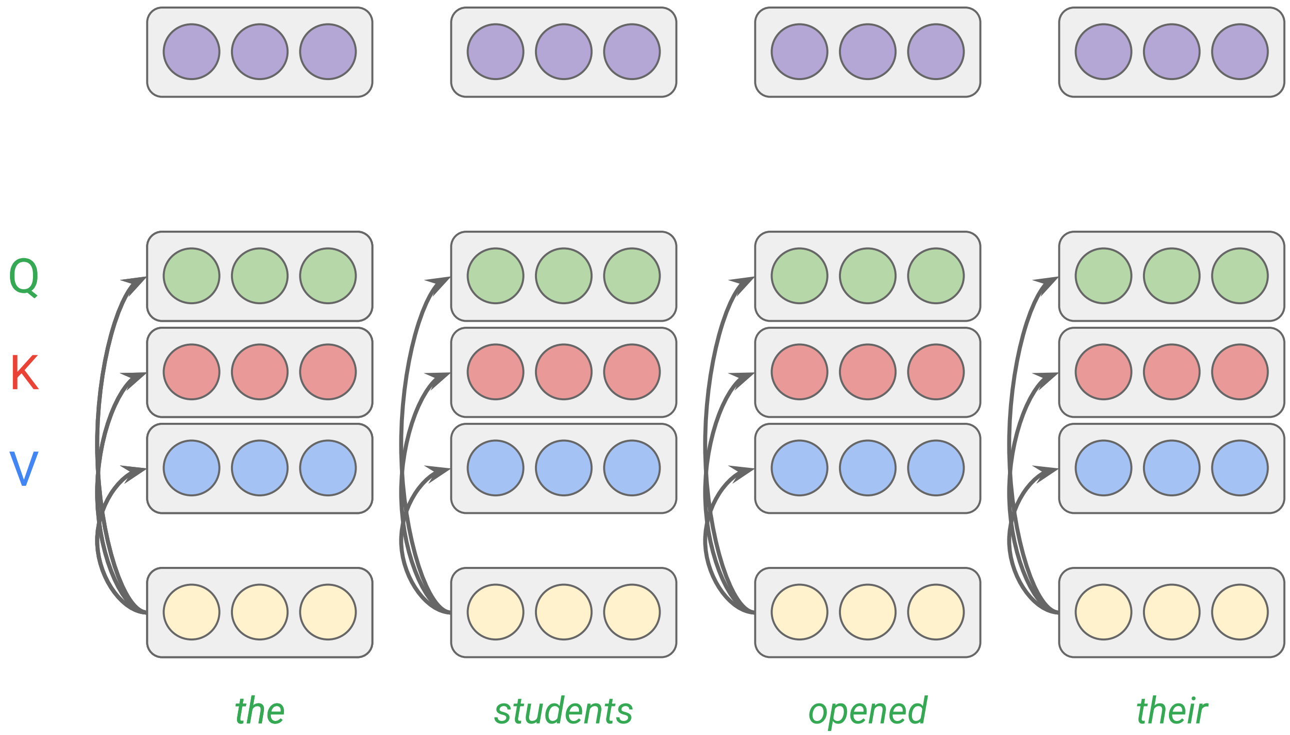
Self-attention (cont'd)

all **computations are parallelized**

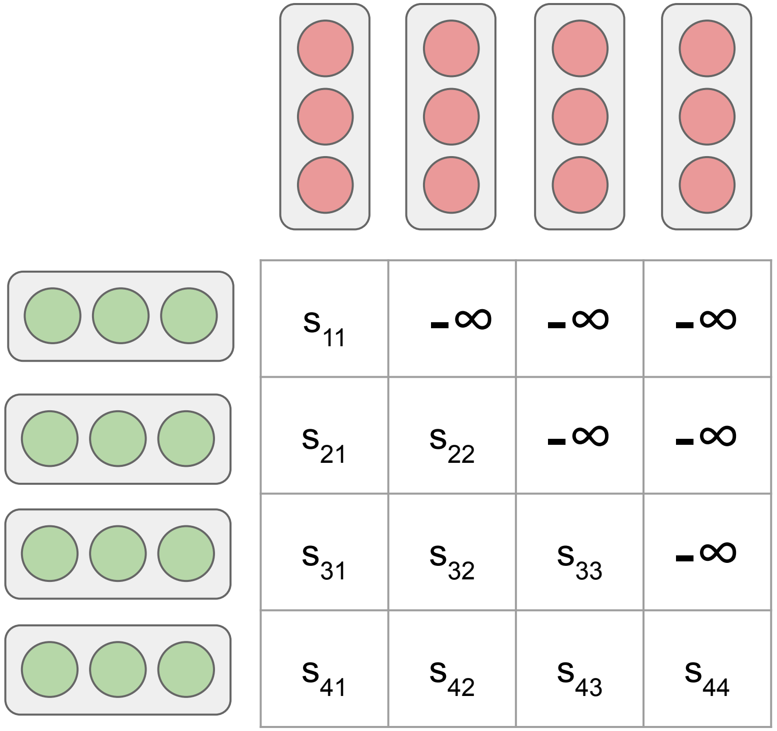


Self-attention (cont'd)

all **computations are parallelized**



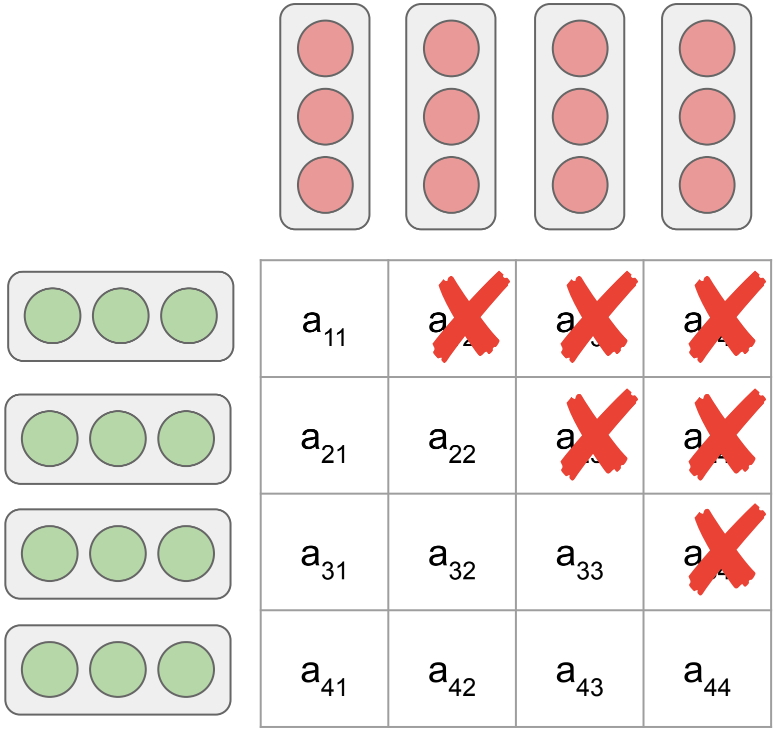
Self-attention in the decoder



$$\begin{matrix} s_{11} & -\infty & -\infty & -\infty \\ s_{21} & s_{22} & -\infty & -\infty \\ s_{31} & s_{32} & s_{33} & -\infty \\ s_{41} & s_{42} & s_{43} & s_{44} \end{matrix}$$

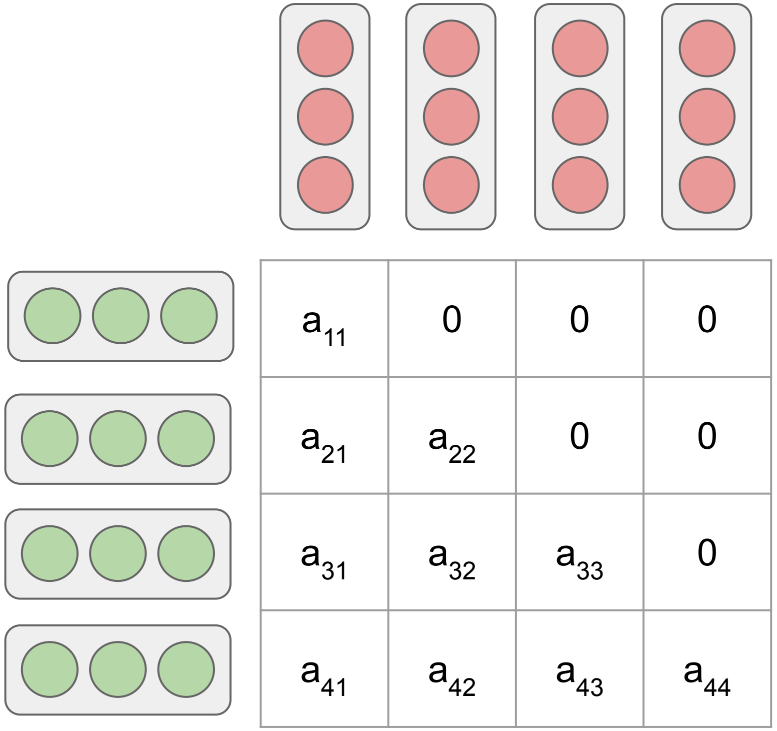
masking out (setting to $-\infty$) all values in the input of the softmax which correspond to illegal connections

Self-attention in the decoder (cont'd)



masking out all values in the input of the softmax which correspond to illegal connections

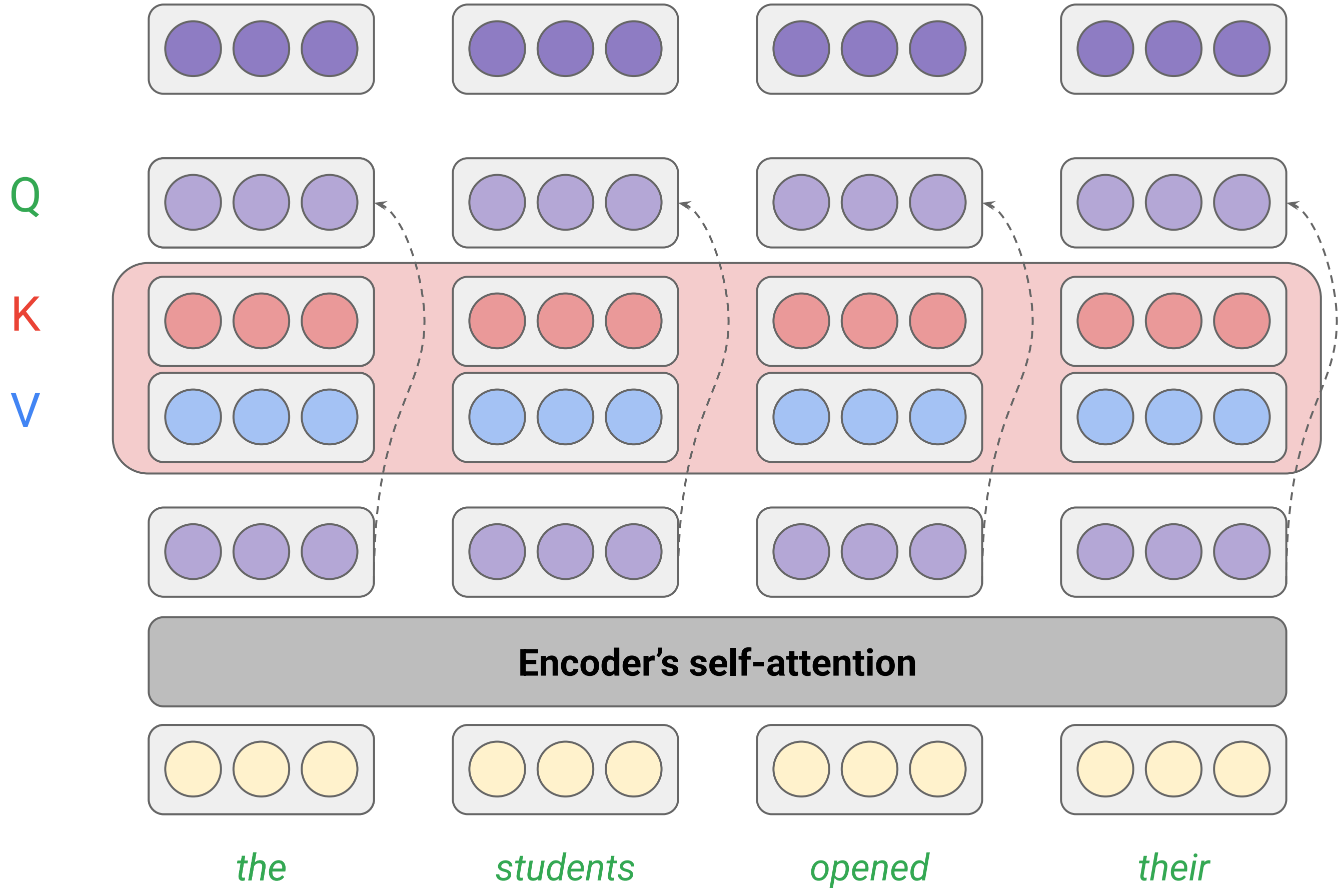
Self-attention in the decoder (cont’d)



$$\begin{array}{cccc} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{array}$$

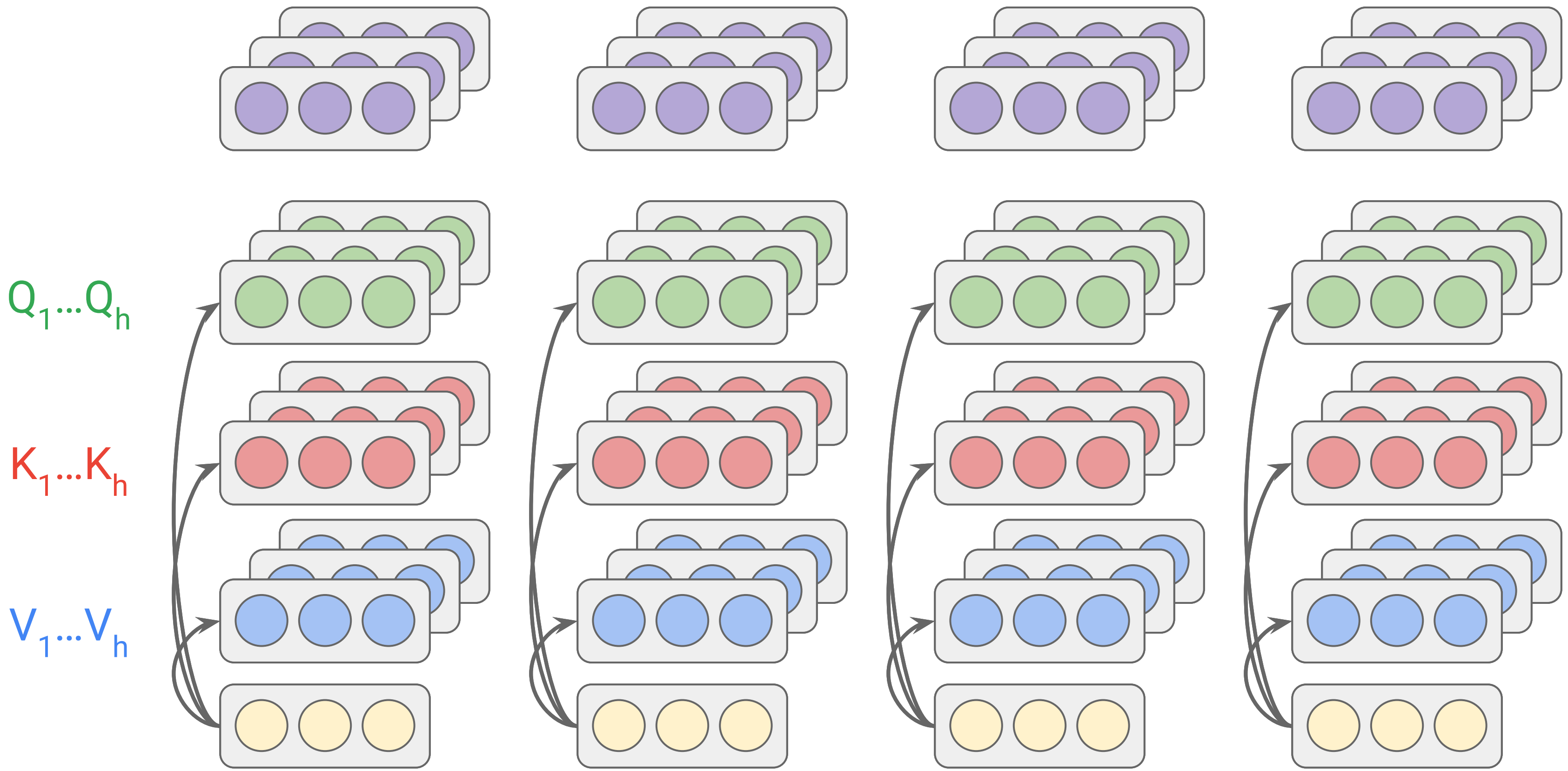
masking out all values in the input of the softmax which correspond to illegal connections

Cross-attention in the decoder

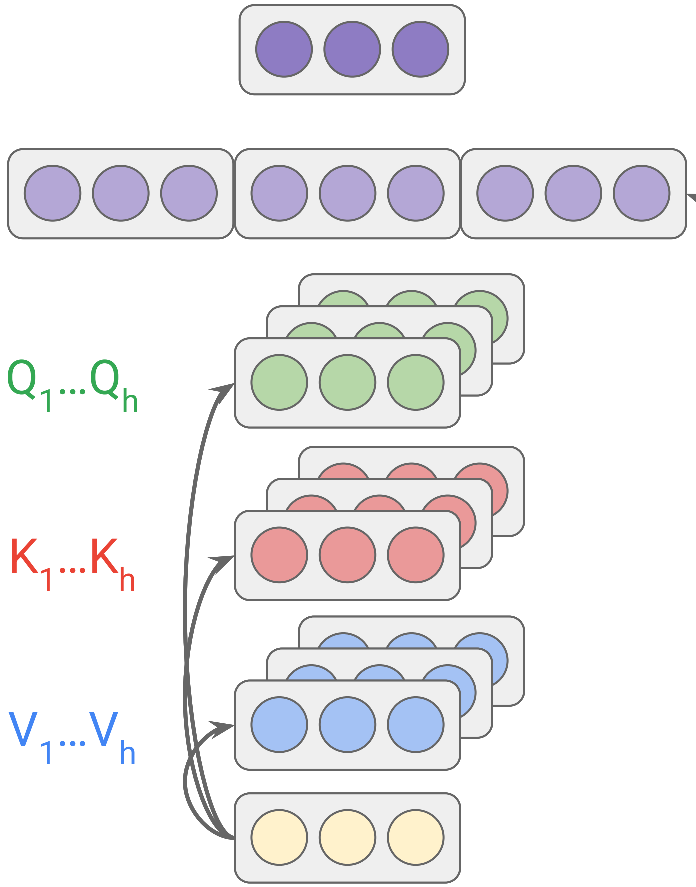


These K, V are the output of the encoder

Multi-head attention

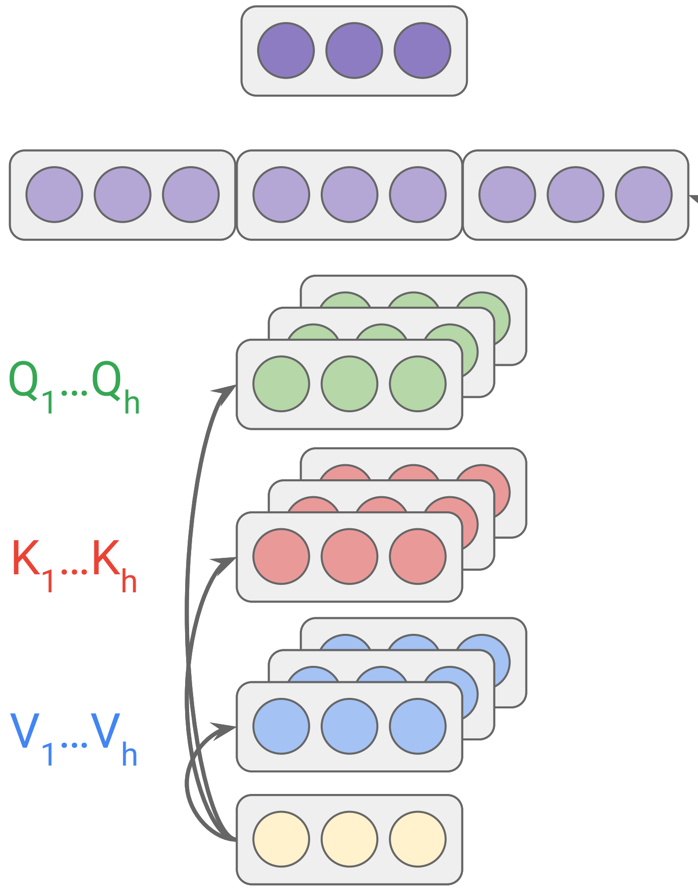


Multi-head attention (cont'd)



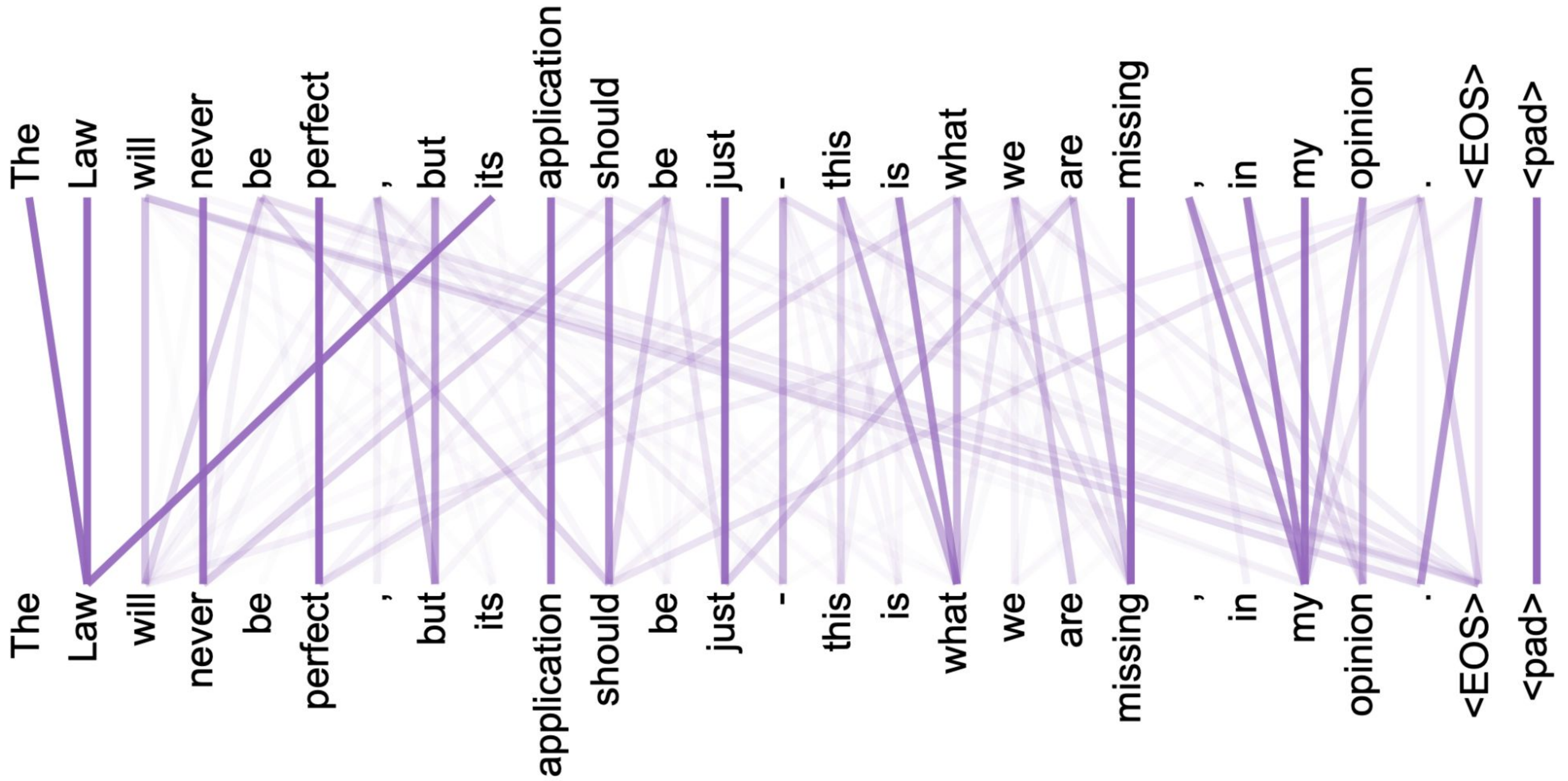
These output values are concatenated and once again projected

Multi-head attention (cont'd)

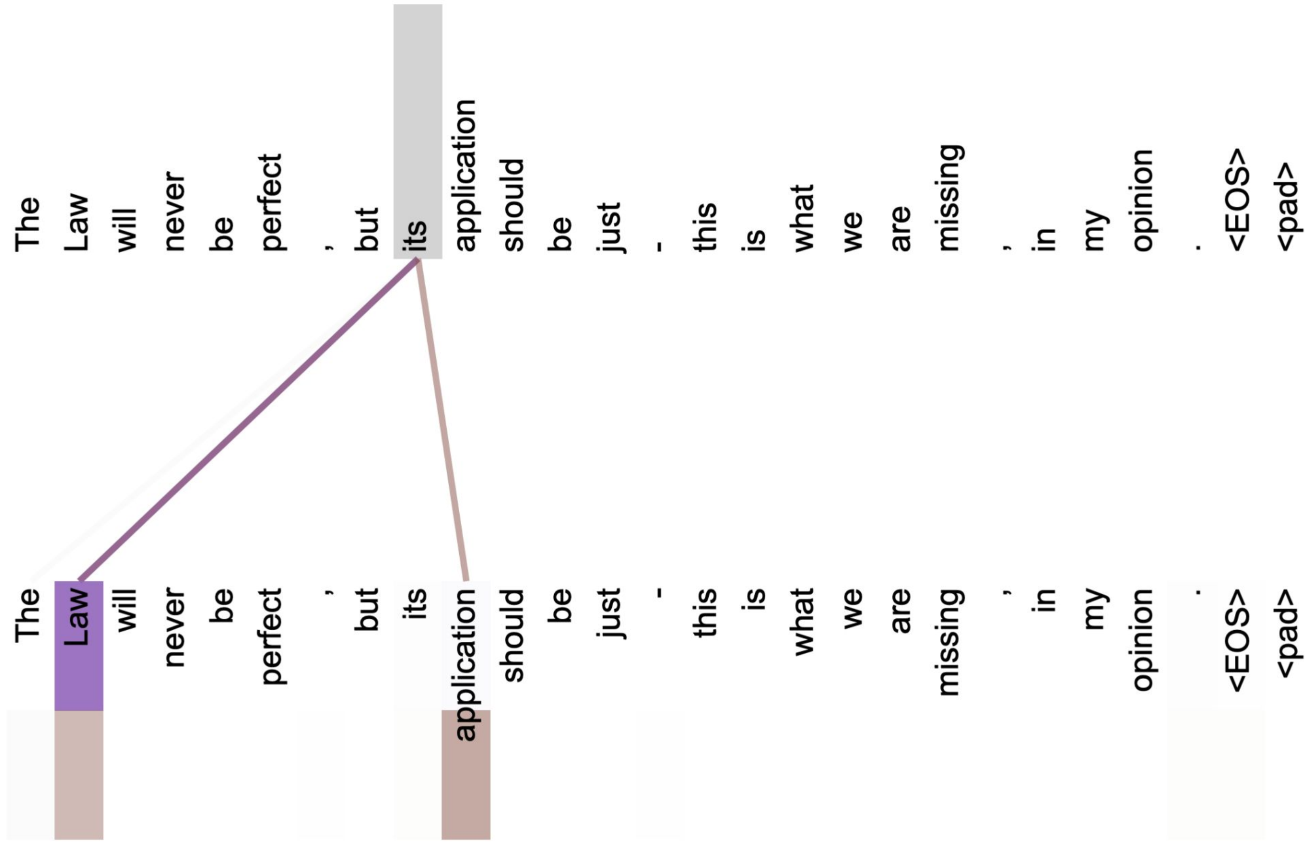


These output values are concatenated and once again projected

Attention visualizations



Attention visualizations (cont'd)



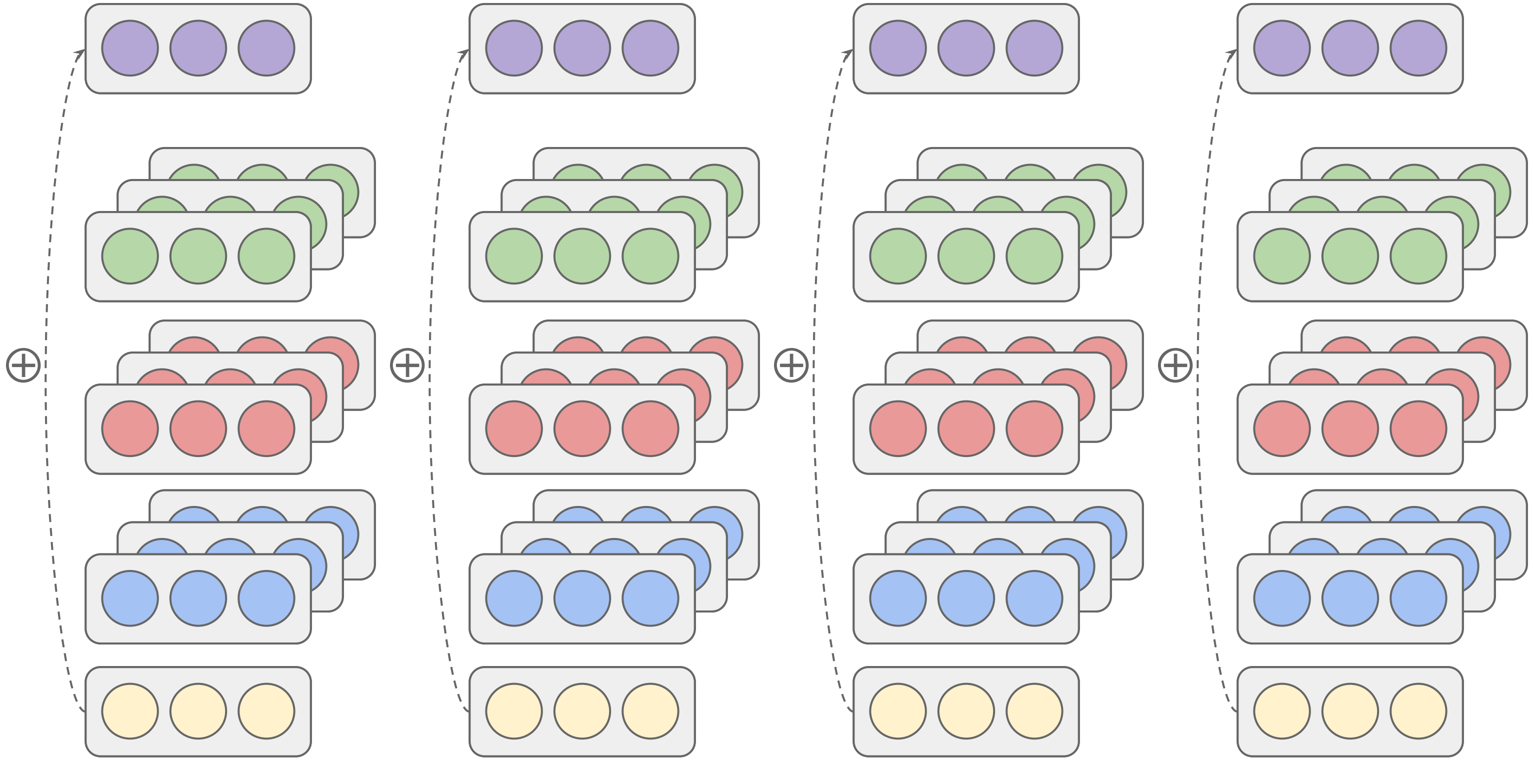
Position-wise Feed-Forward Networks

$$\mathrm{FFN}(x) = \underbrace{\max}_{\text{ReLU (Rectified Linear Unit)}}(0, xW_1 + b_1)W_2 + b_2$$

Residual connection and layer normalization

$$\mathrm{LayerNorm}(x + \mathrm{Sublayer}(x))$$

Residual connection

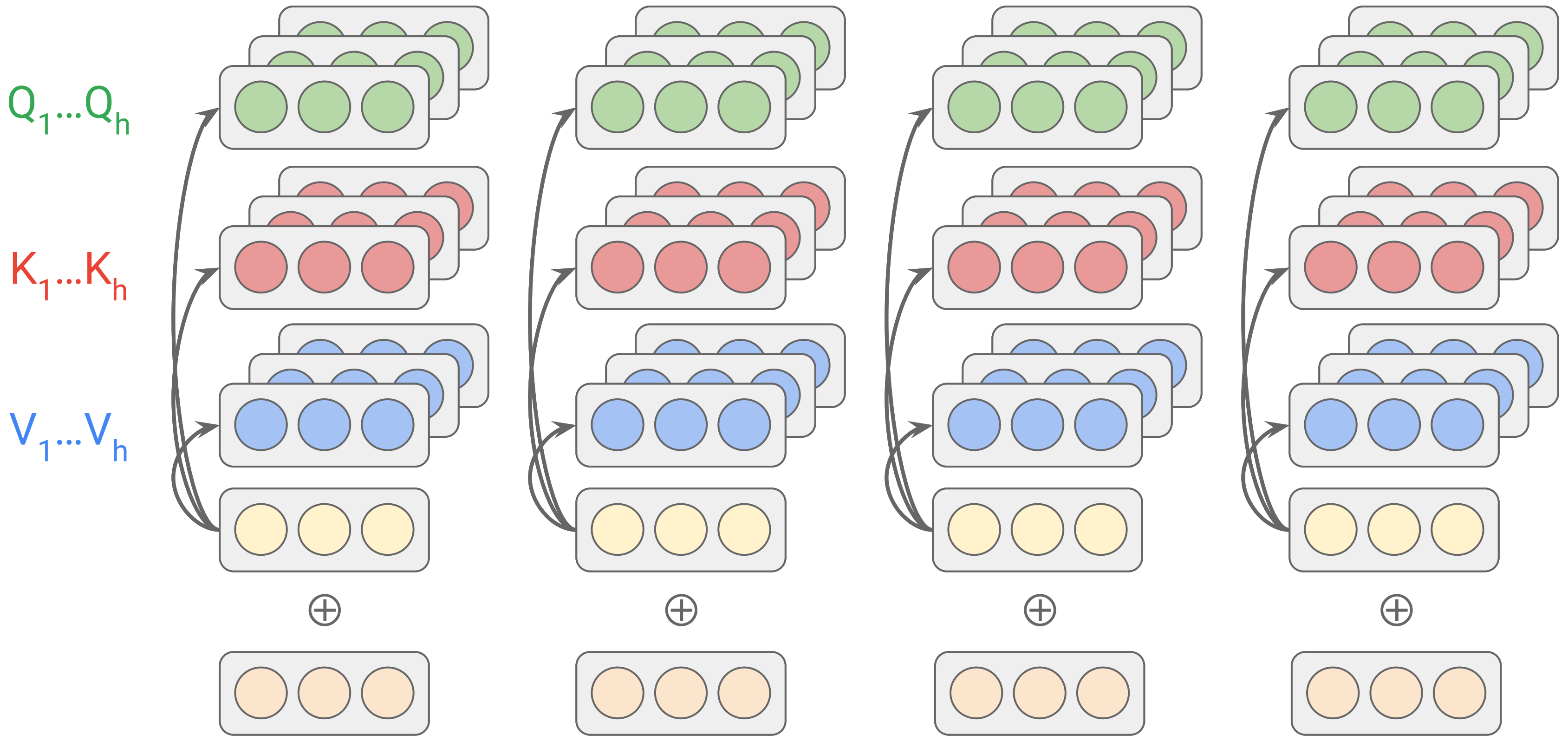


Positional Encoding

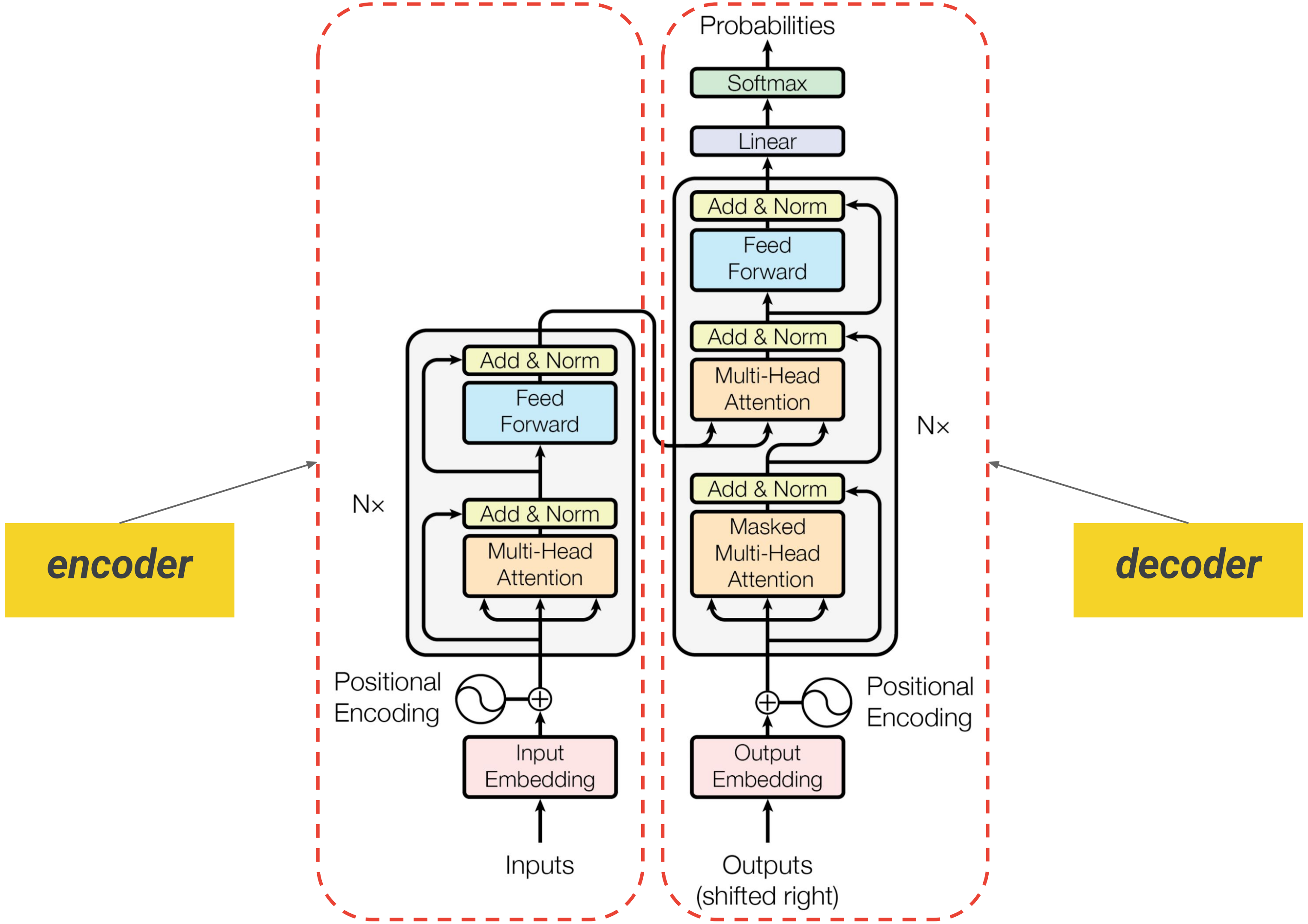
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

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

Positional Encoding (cont'd)



Transformer block (putting it together)


Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Positional Encoding
Output Embedding
Outputs (shifted right)
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
encoder
decoder

Training and Test

Thank you!