Introduction to Machine Learning

Attention and Transformers

Ashley Gao

William & Mary

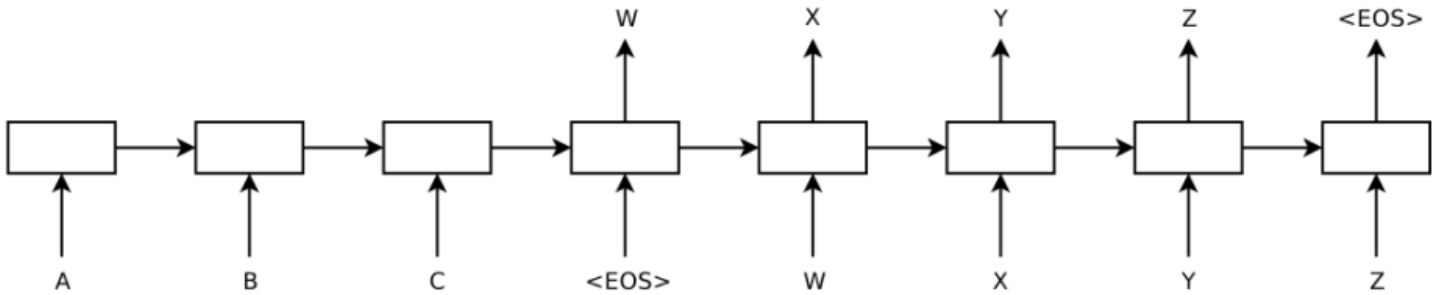
November 20, 2024

Overview

- It is still challenging to generate long sequences when the decoders only has access to the final hidden states from the encoder
  - Machine translation: it's hard to summarize long sentences in a single vector, so let's allow the decoder to peek at the input
  - Vision: have a network glance at one part of an image at a time, so that we can understand what information it's conveying
- This lecture will introduce attention that drastically improves the performance on the long sequences

Attention-based Machine Translation

- This RNN architecture can be used for machine translation:



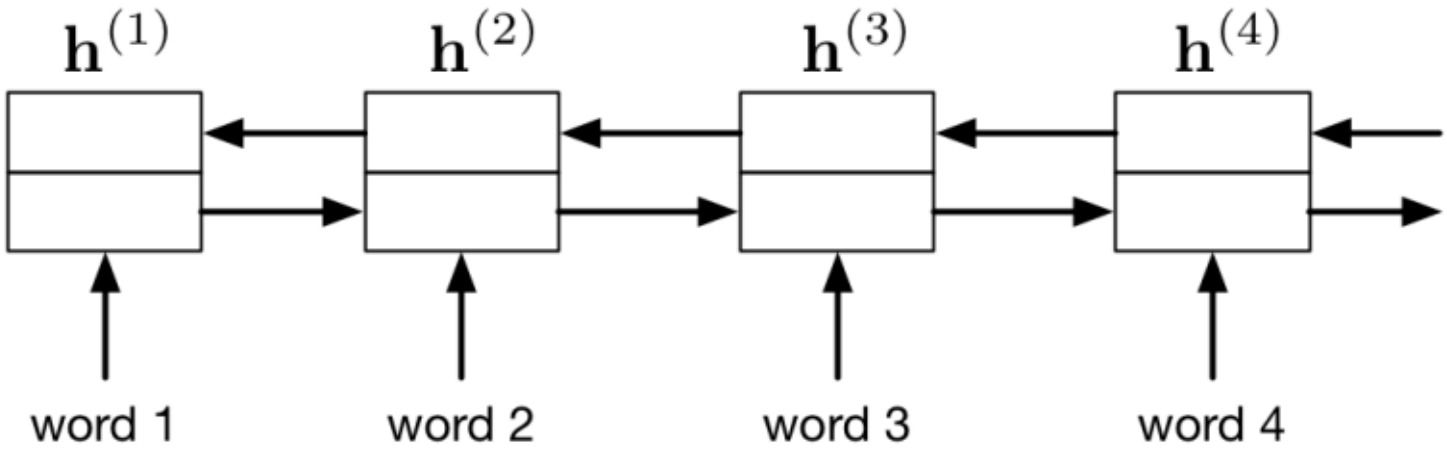
- The network reads a sentence and stores all the information in its hidden units
- Some sentences can be really long. Can we really store all the information in a vector of hidden units?

Attention-based Machine Translation

- We'll look at the translation model from this classic paper
  - Bahdanau et al., Neural machine translation by jointly learning to align and translate. ICLR, 2015
- Basic idea: each output word comes from one word, or a handful of words, from the input
  - Maybe we can learn to attend to only the relevant ones as we produce the output

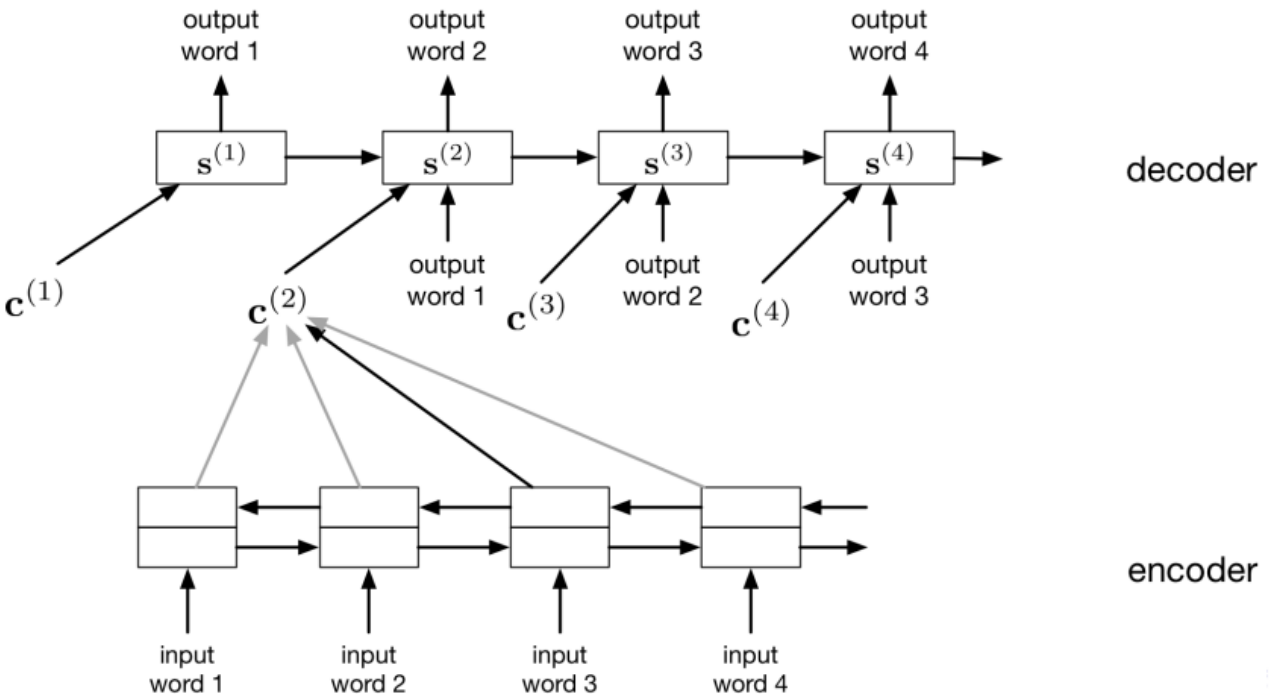
Attention-based Machine Translation

- The model has both an encoder and a decoder. The encoder computes an annotation of each word in the input.
- It takes the form of a bidirectional RNN
  - This just means we have an RNN that runs forward and an RNN that runs backward, and we concatenate their hidden vectors
  - The idea: information earlier or later in the sentence can help disambiguate a word, so we need both directions
  - The RNN uses an LSTM-like architecture, which are GRUs



Attention-based Machine Translation

- The decoder network is also an RNN
  - Like the encoder/decoder translation model, it makes predictions one word at a time, and its predictions are fed back in as inputs
- The difference is that it also receives a context vector $\boldsymbol{c}^{(t)}$ at each time step, which is computed by attending to the inputs



Attention-based Machine Translation

- The context vector is computed as a weighted average of all the encoder's annotation

$$\boldsymbol{c}^{(i)} = \sum_j \alpha_{ij} h^{(j)} \tag{1}$$

- the attention weights are computed as a softmax, where the inputs depend on the annotation and the decoder's state

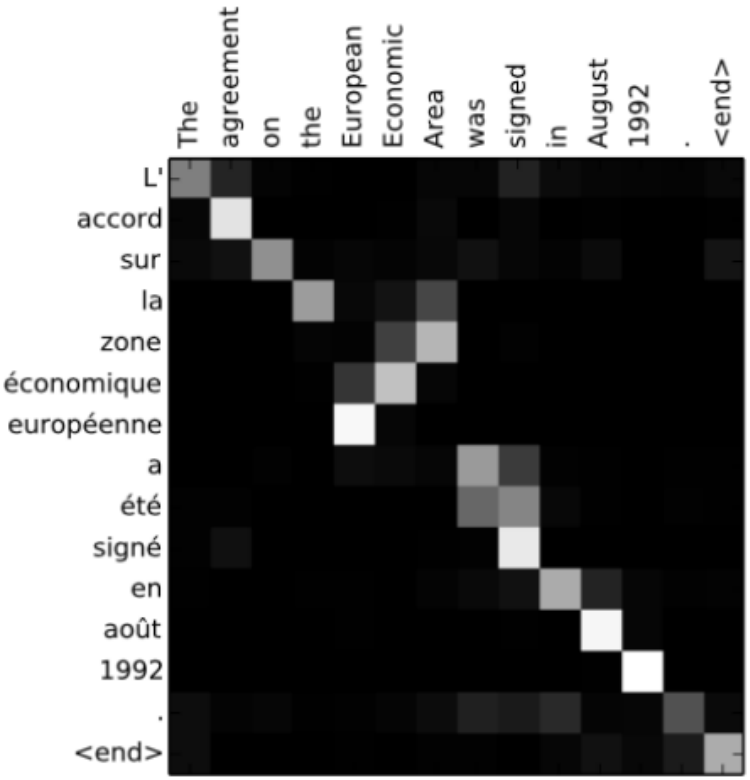
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})} \tag{2}$$

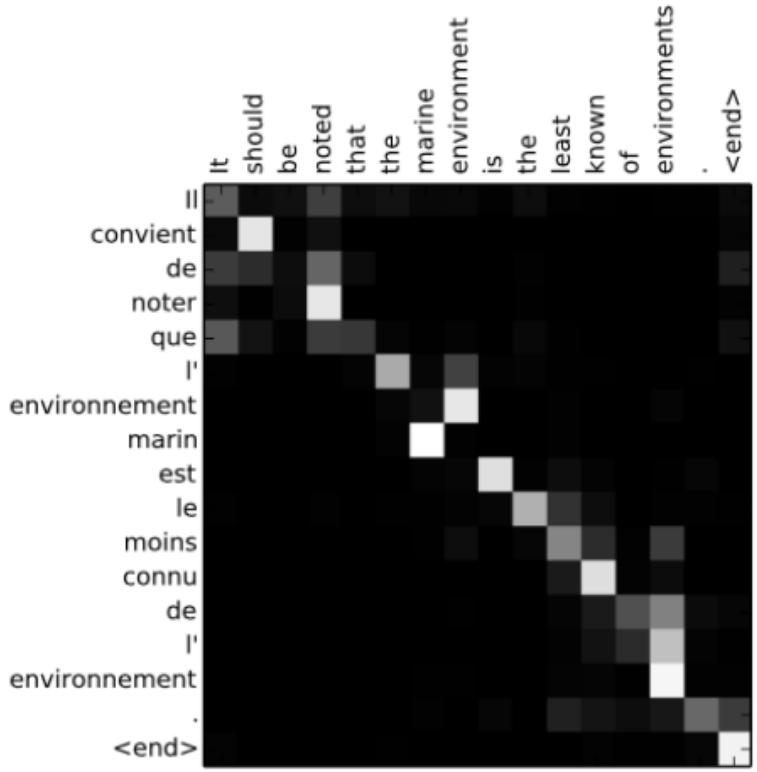
$$e_{ij} = a(\boldsymbol{s}^{(i-1)}, \boldsymbol{h}^{j}) \tag{3}$$

- e_{ij} is the alignment model which scores how well the inputs around position j and the output at position i match.
  - This score is based on the RNN hidden state $s^{(i-1)}$ and the j-th annotation h_j of the input sentence

Attention-based Machine Translation

- Here's a visualization of the attention maps at each time step





Attention-based Caption Generation

- Attention can also be used to understand images
- We humans can't process a whole visual scene at once
  - The fovea of the eye gives us high-acuity vision in only a tiny region of our field of view
  - Instead, we must integrate information from a series of glimpses
- This idea inspired the following paper:
  - Xu et al. Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention. ICML, 2015

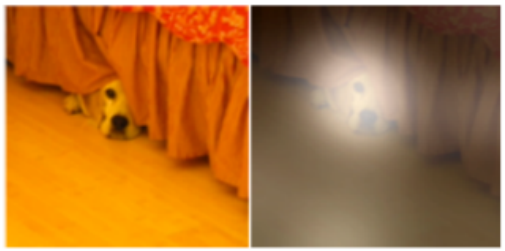
Attention-based Caption Generation

- The caption generation task: take an image as an input, and produce a sentence describing the image
- Encoder: a classification conv net (VGGNet, similar to AlexNet)
  - This computes a bunch of feature maps over the iamge
- Decoder: an attention-based RNN, analogous to the decoder in the translation model
  - In each time step, the decoder computes an attention map over the entire image, effectively deciding which regions to focus on
  - It receives a context vector, which is the weighted average of the conv net features

Attention-based Caption Generation

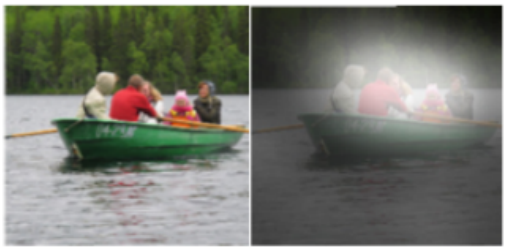
- This lets us understand where the network is looking as it generates a sentence


A woman is throwing a frisbee in a park.


A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

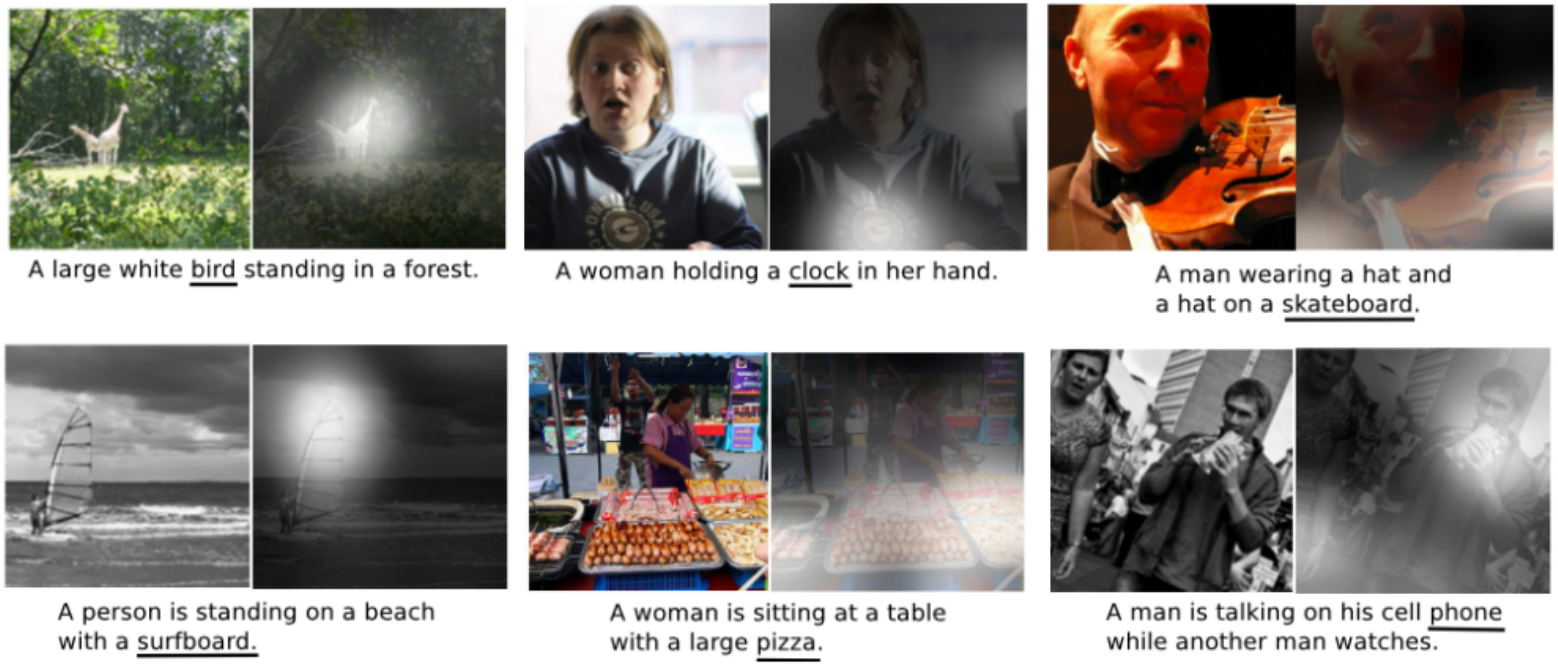

A little girl sitting on a bed with a teddy bear.


A group of people sitting on a boat in the water.


A giraffe standing in a forest with trees in the background.

Attention-based Caption Generation

- This can also help us understand the network's mistakes

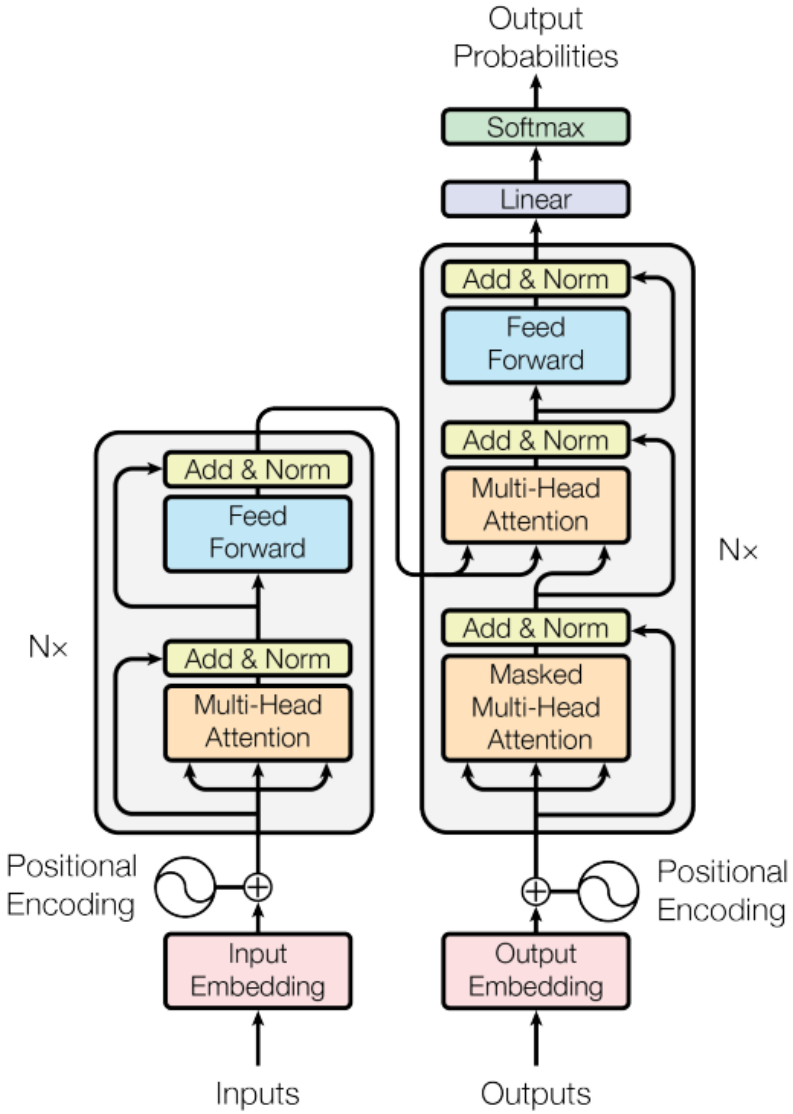


Attention Is All You Need

- We would like our model to have access to the entire history at each hidden layer
- We can use attention to aggregate the context information by attending to one or a few important tokens from the past history
- Let's take a look at the Transformer architecture
  - Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017

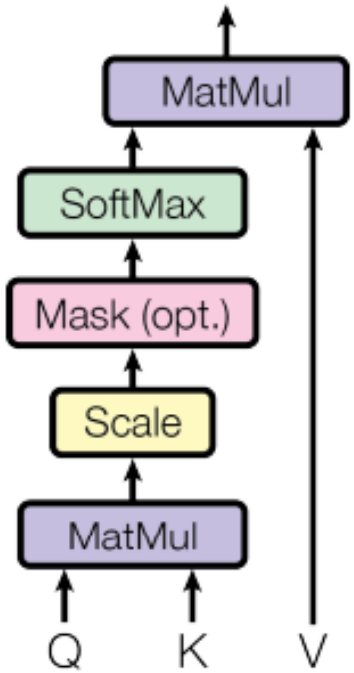
Attention Is All You Need

- The transformer architecture:

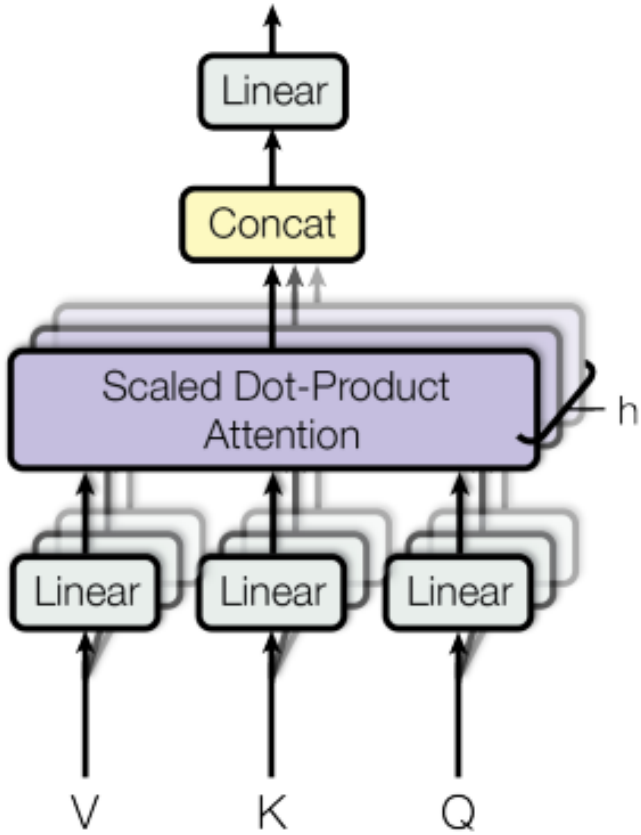


Attention Is All You Need

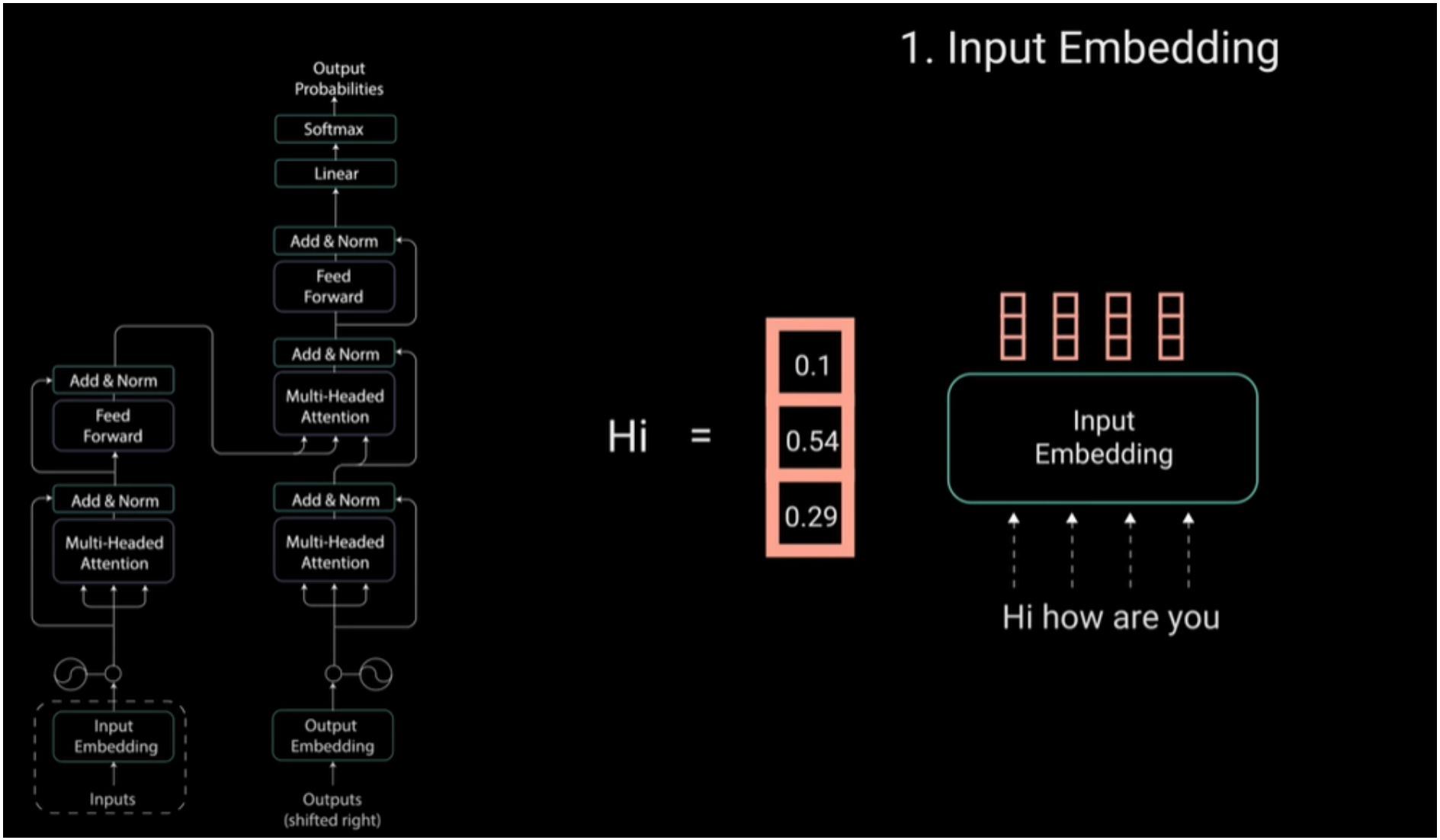
Scaled Dot-Product Attention



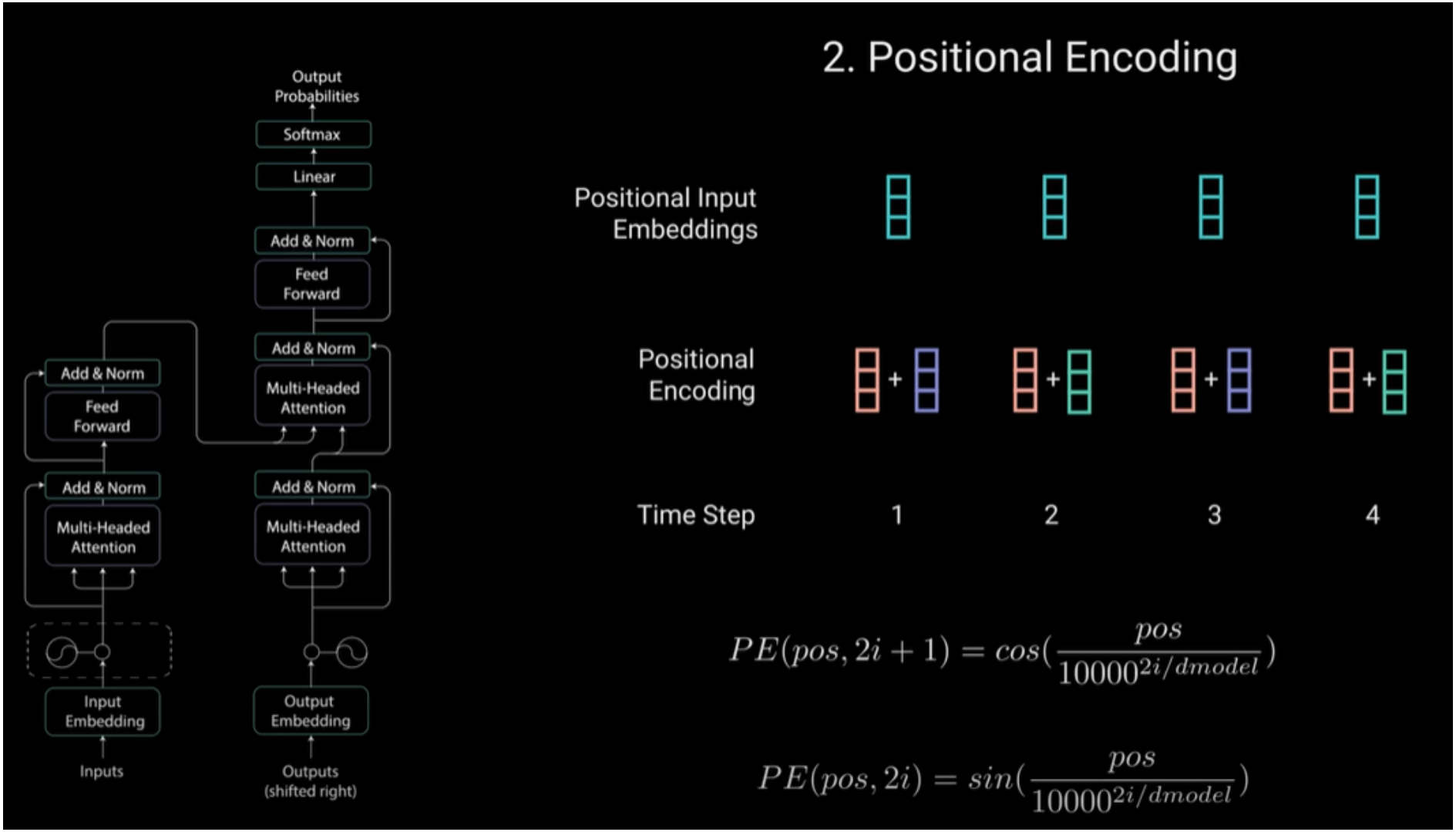
Multi-Head Attention



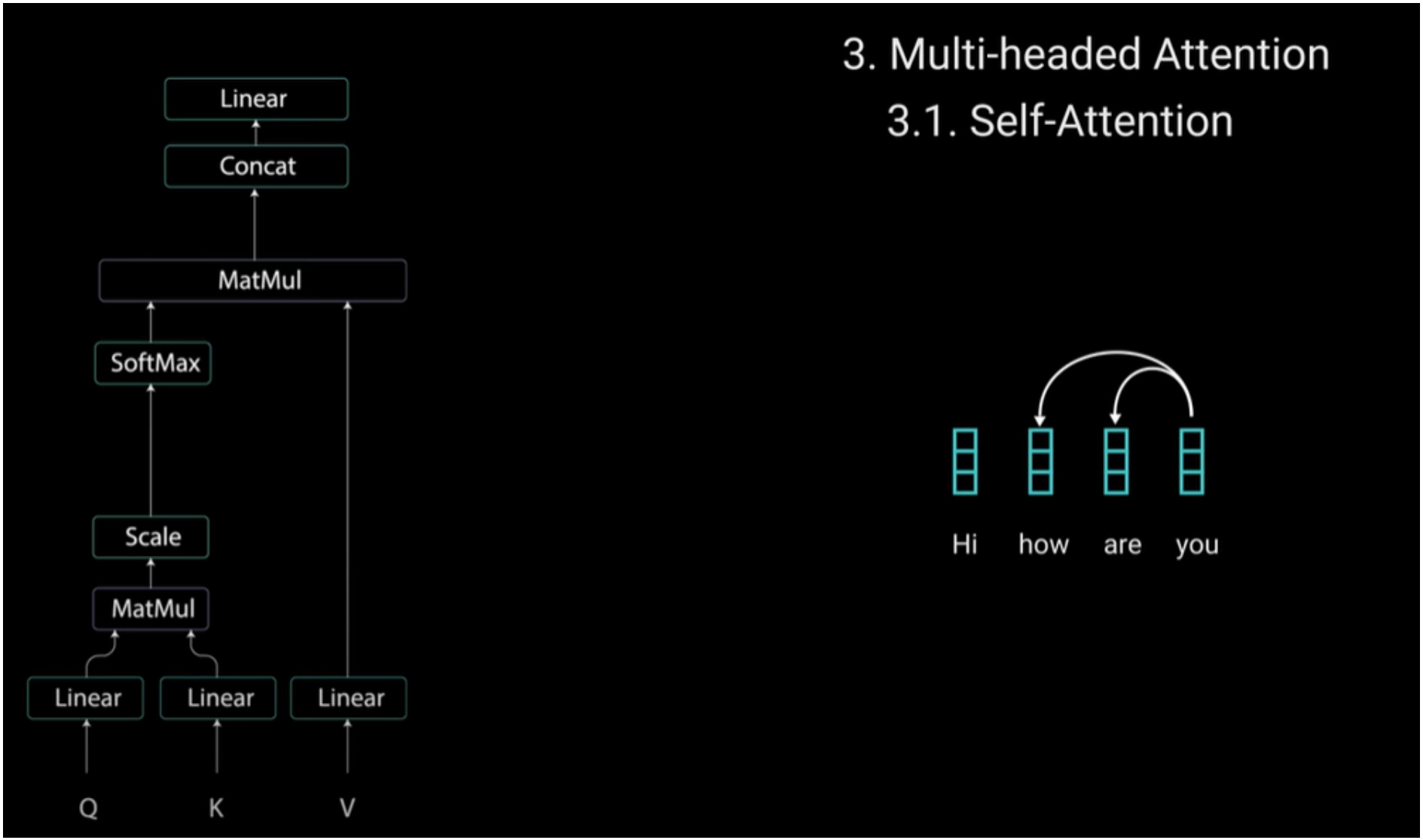
Attention Is All You Need



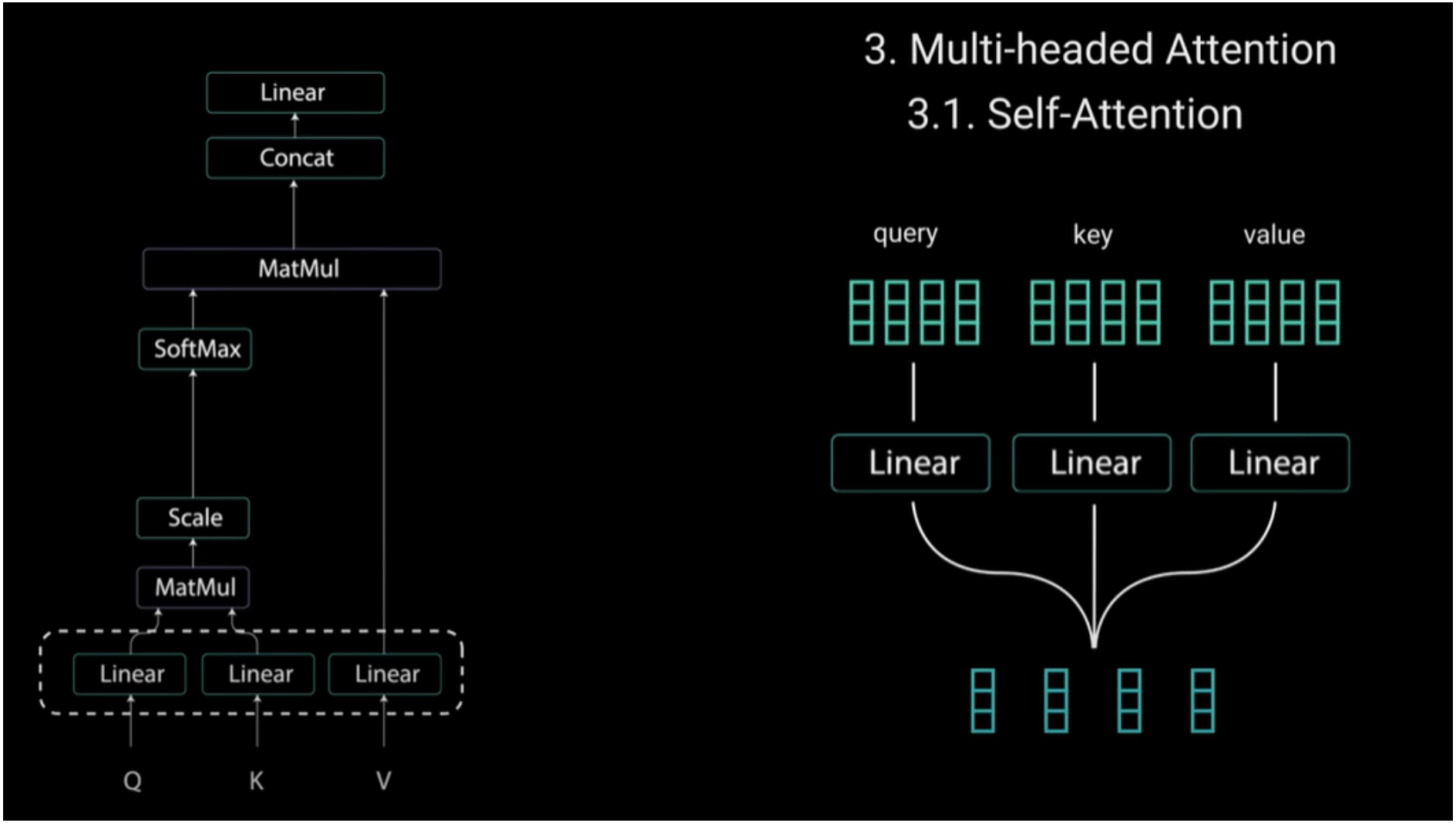
Attention Is All You Need



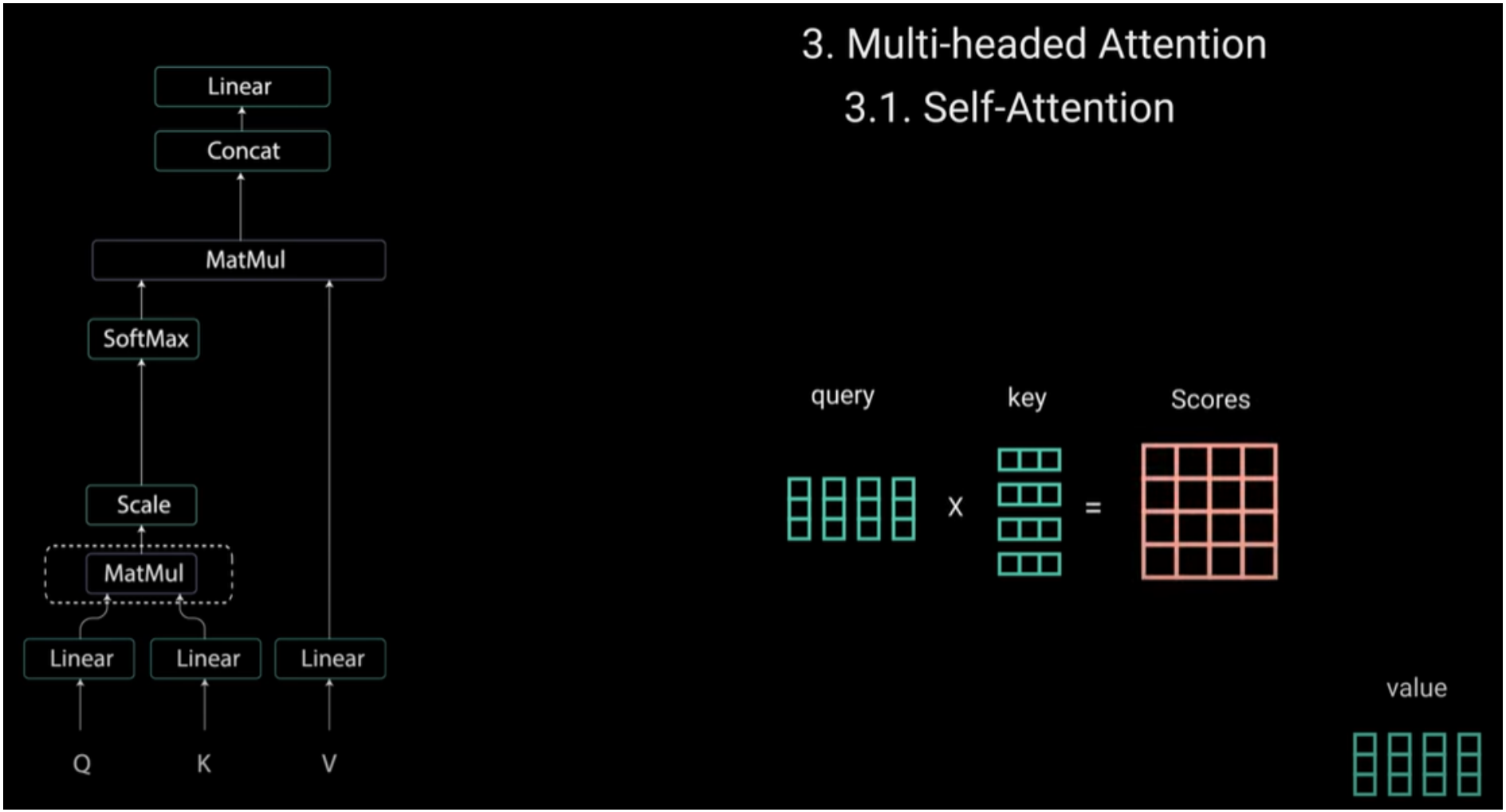
Attention Is All You Need



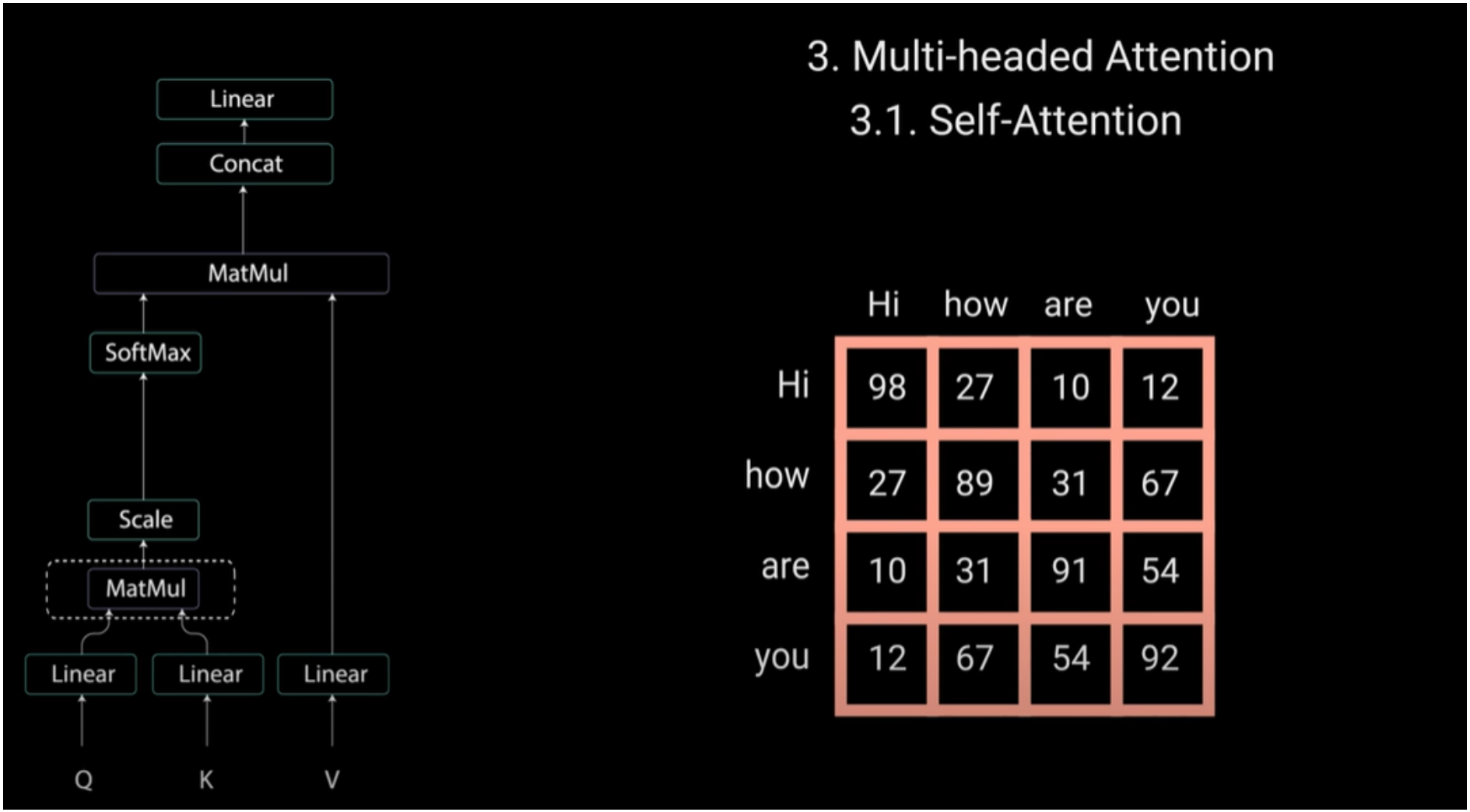
Attention Is All You Need



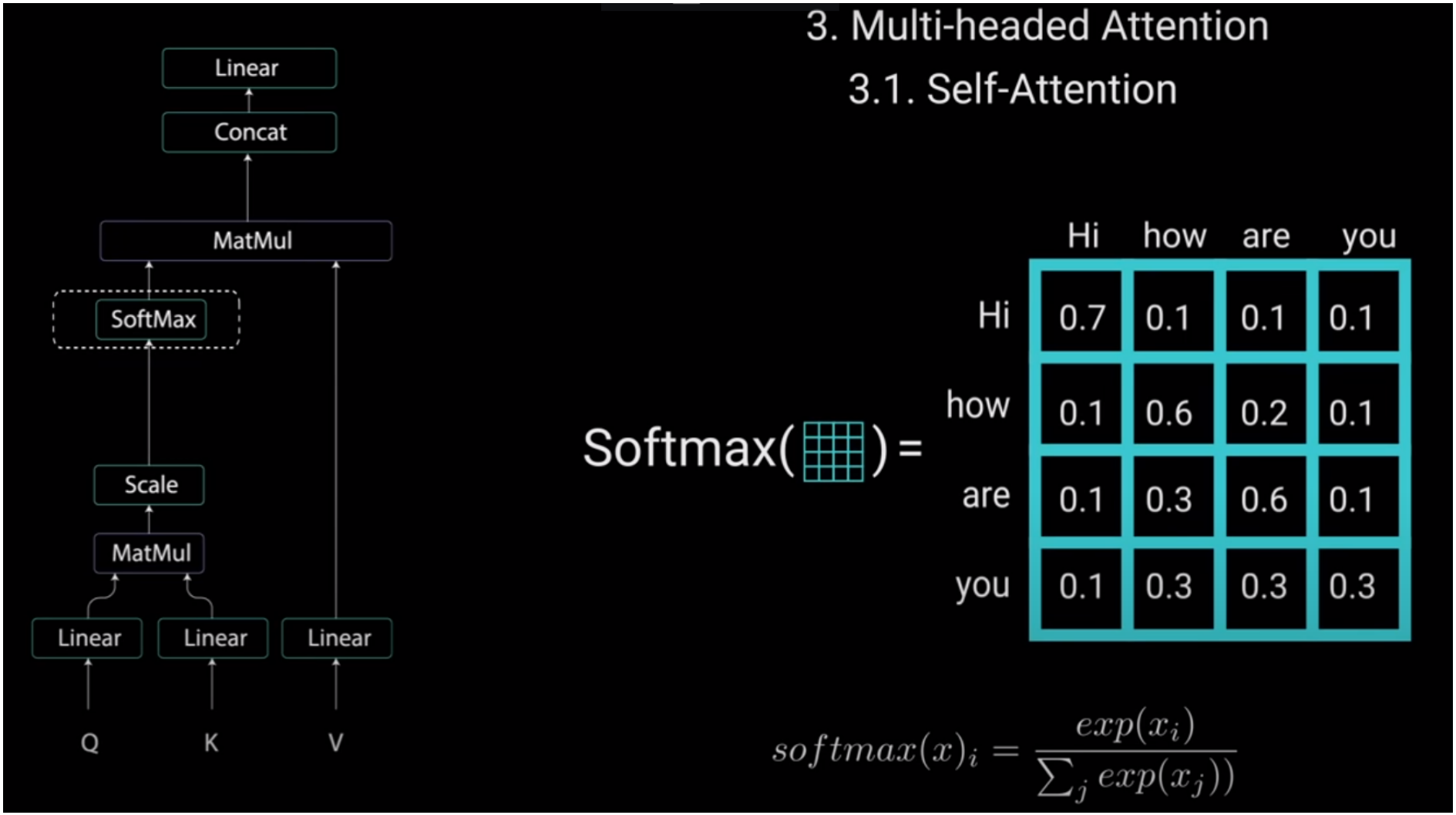
Attention Is All You Need



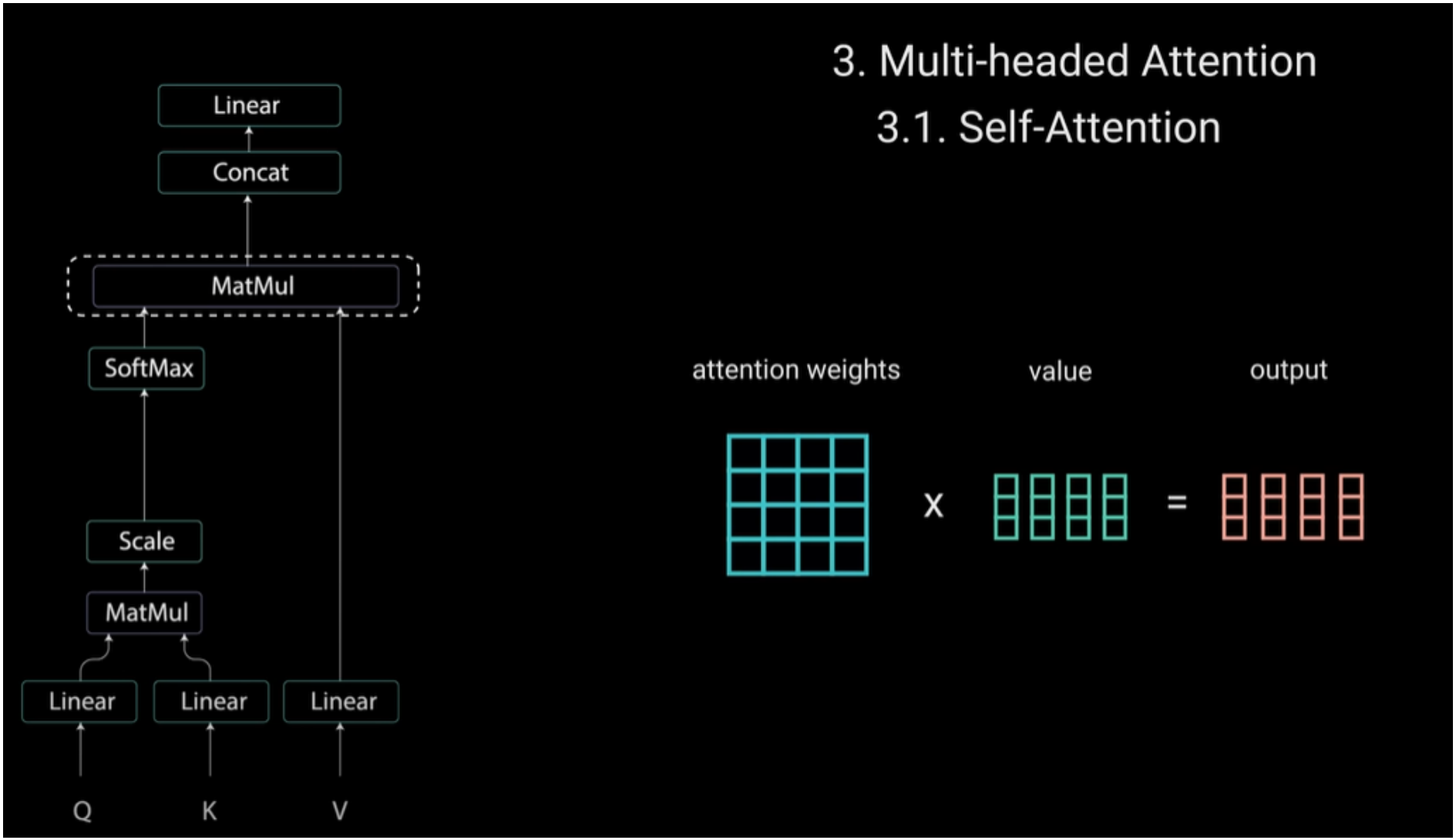
Attention Is All You Need



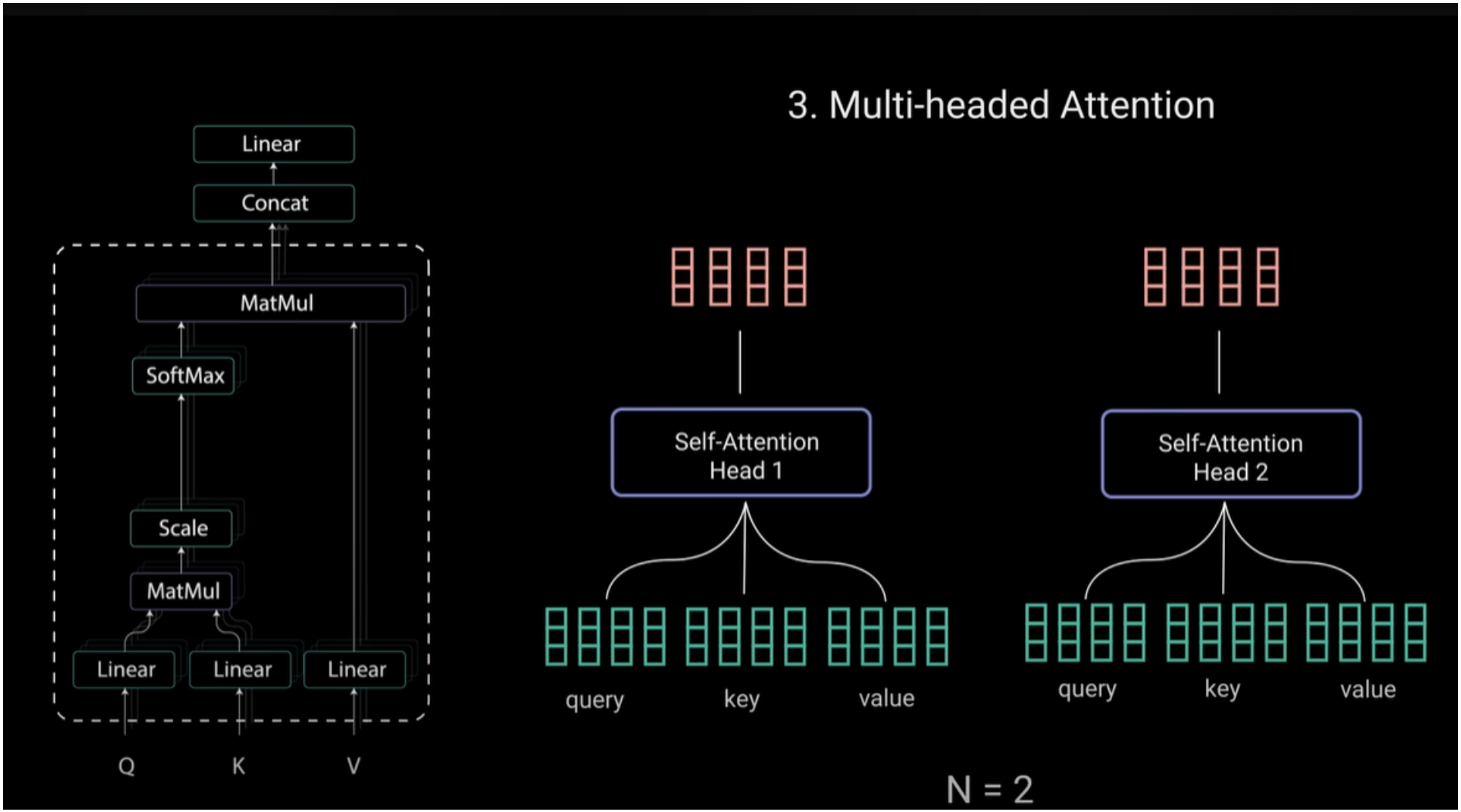
Attention Is All You Need



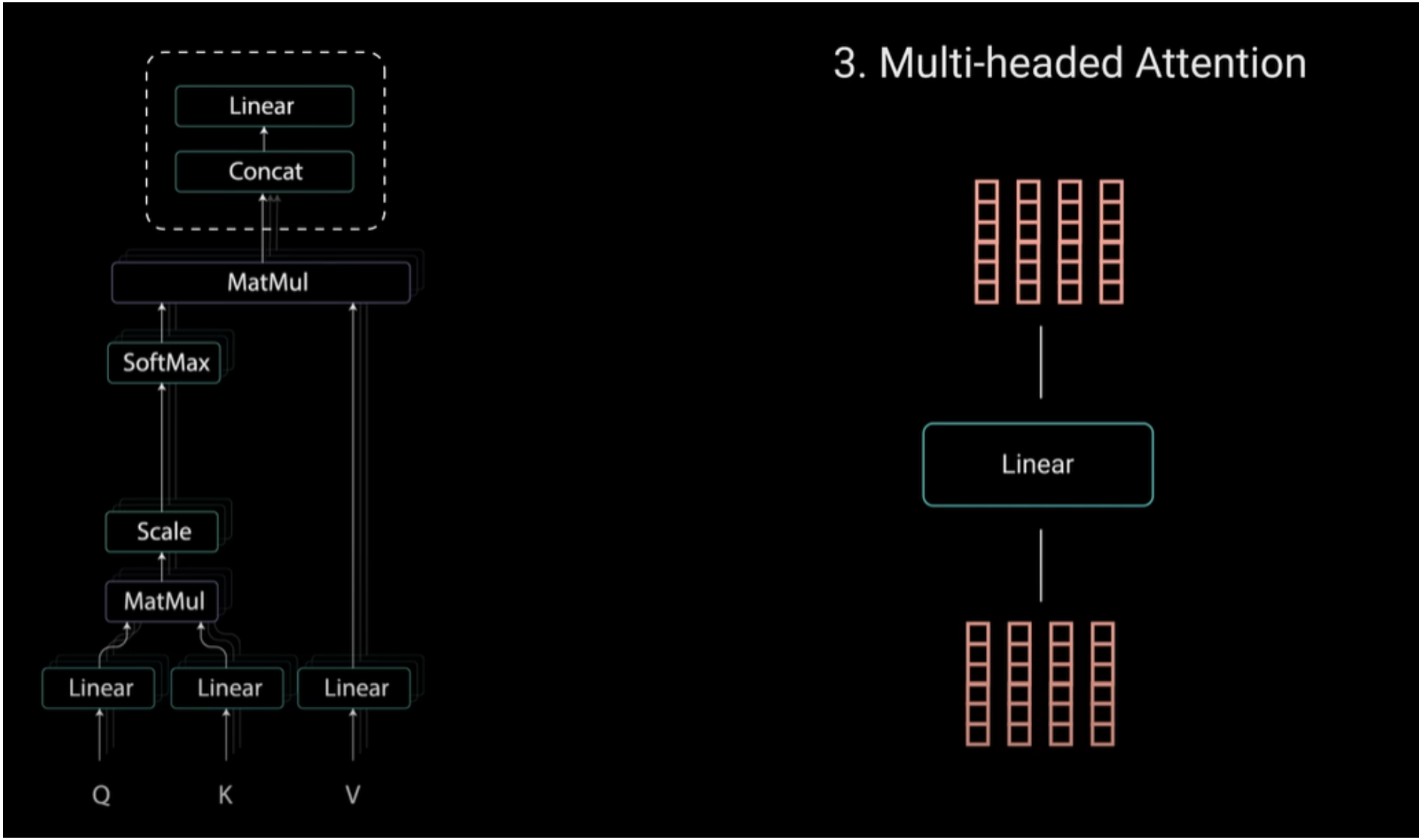
Attention Is All You Need



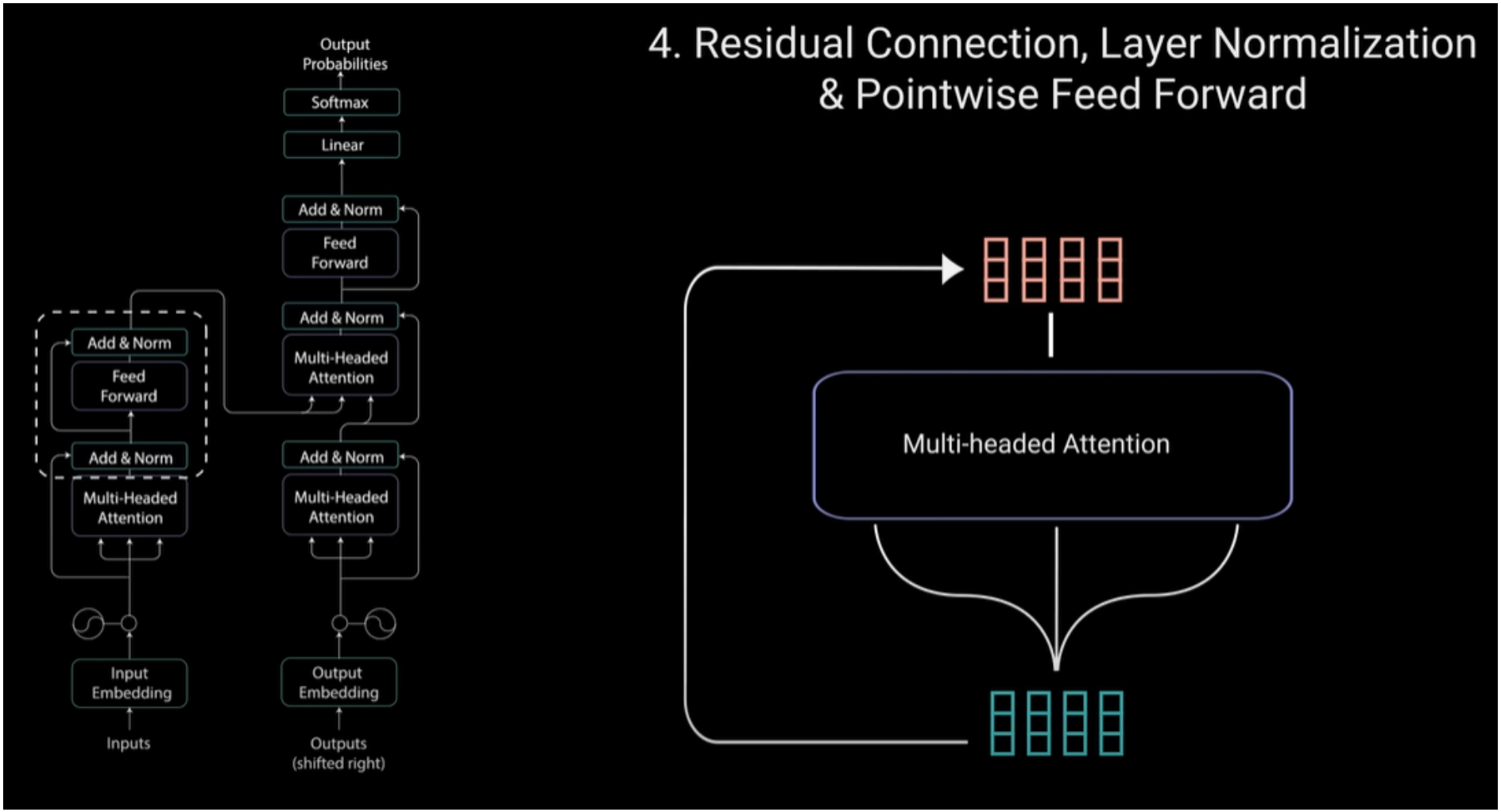
Attention Is All You Need



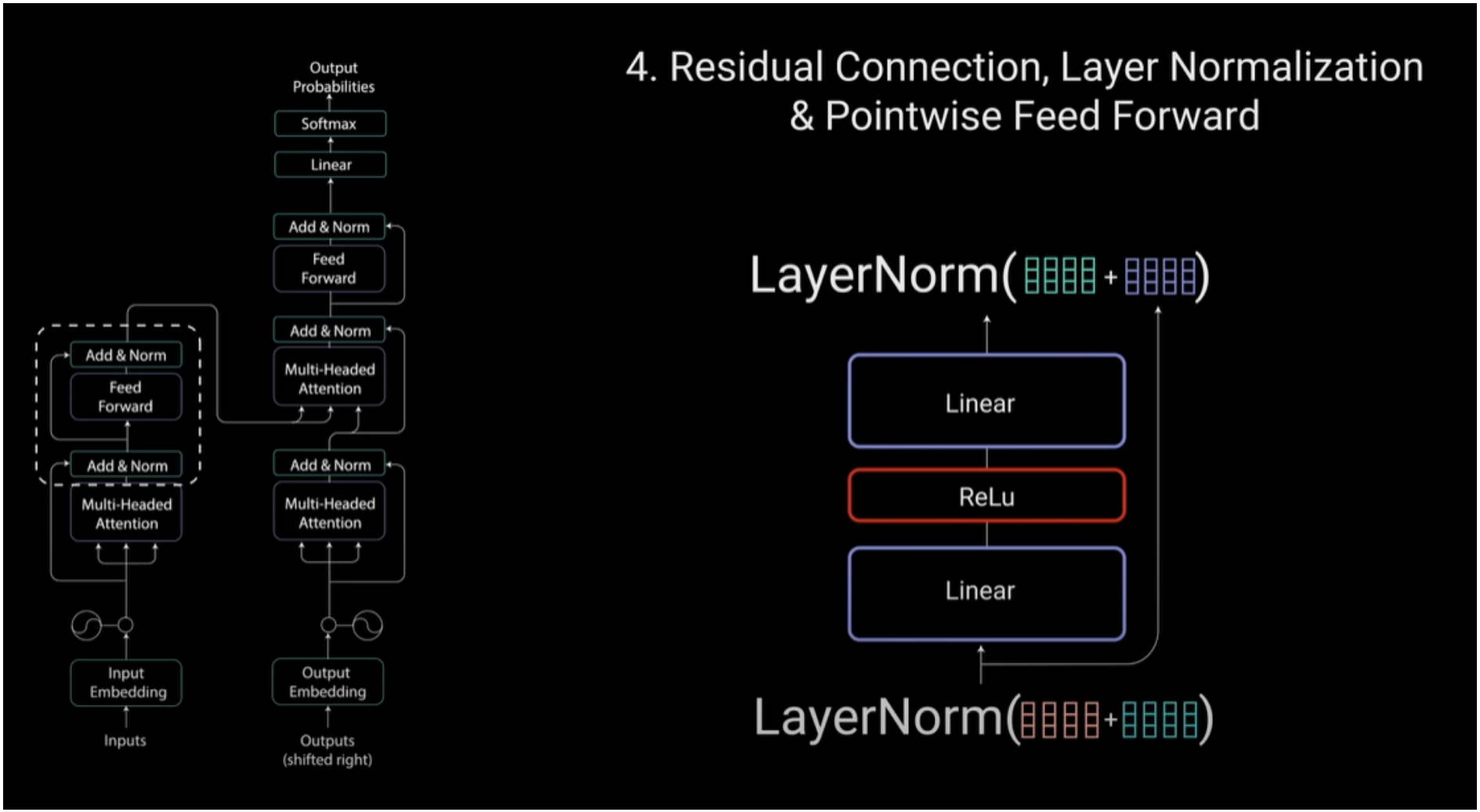
Attention Is All You Need



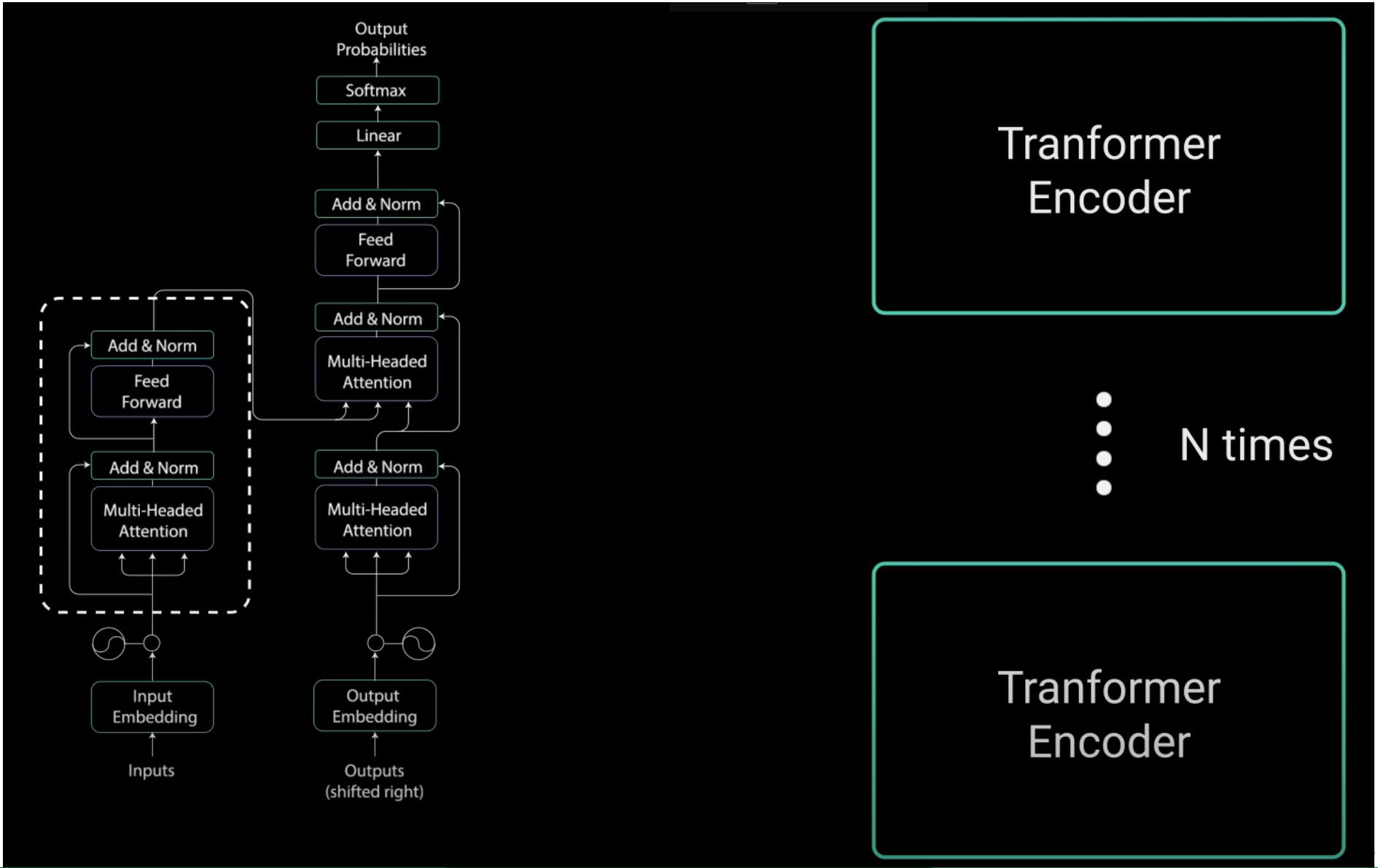
Attention Is All You Need



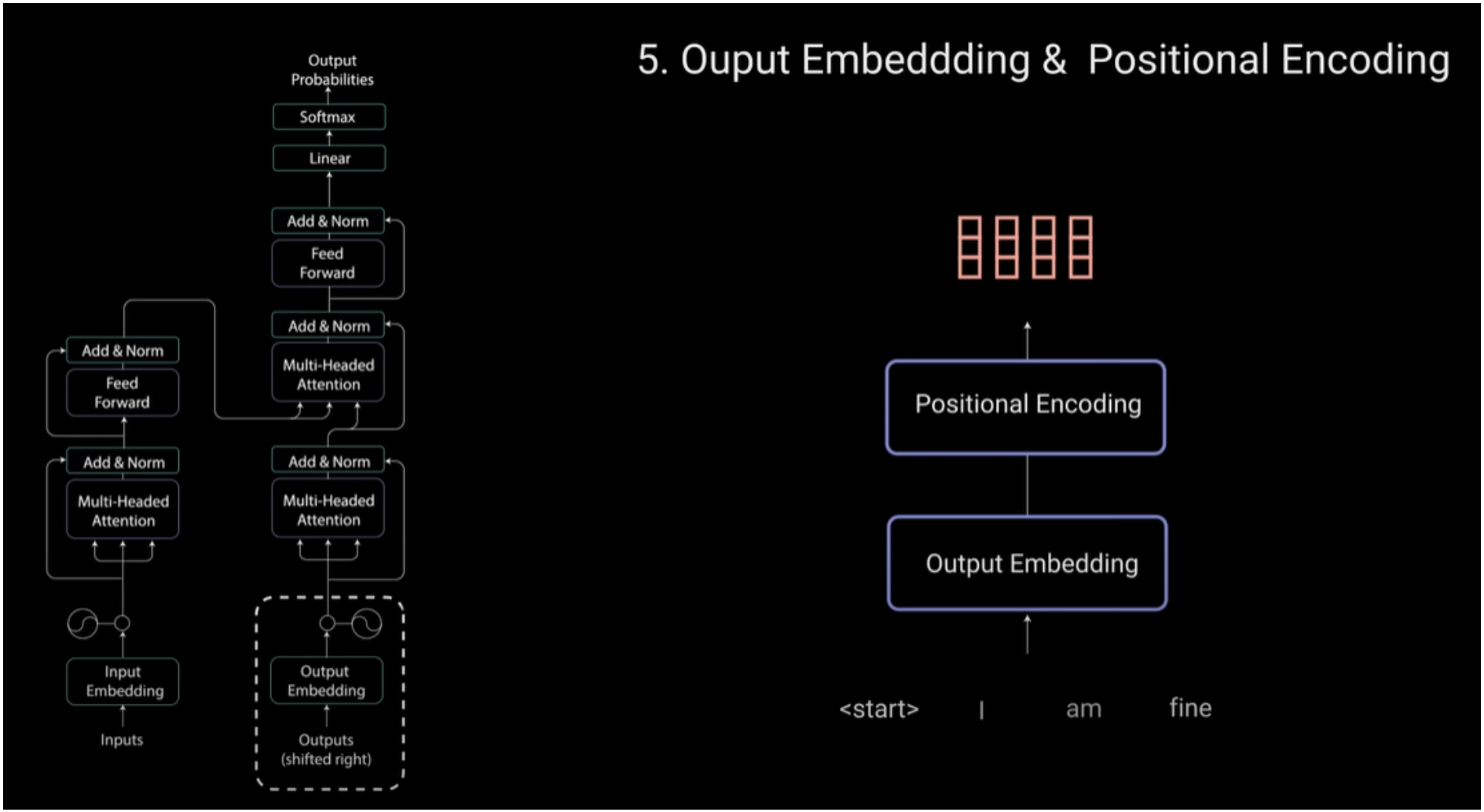
Attention Is All You Need



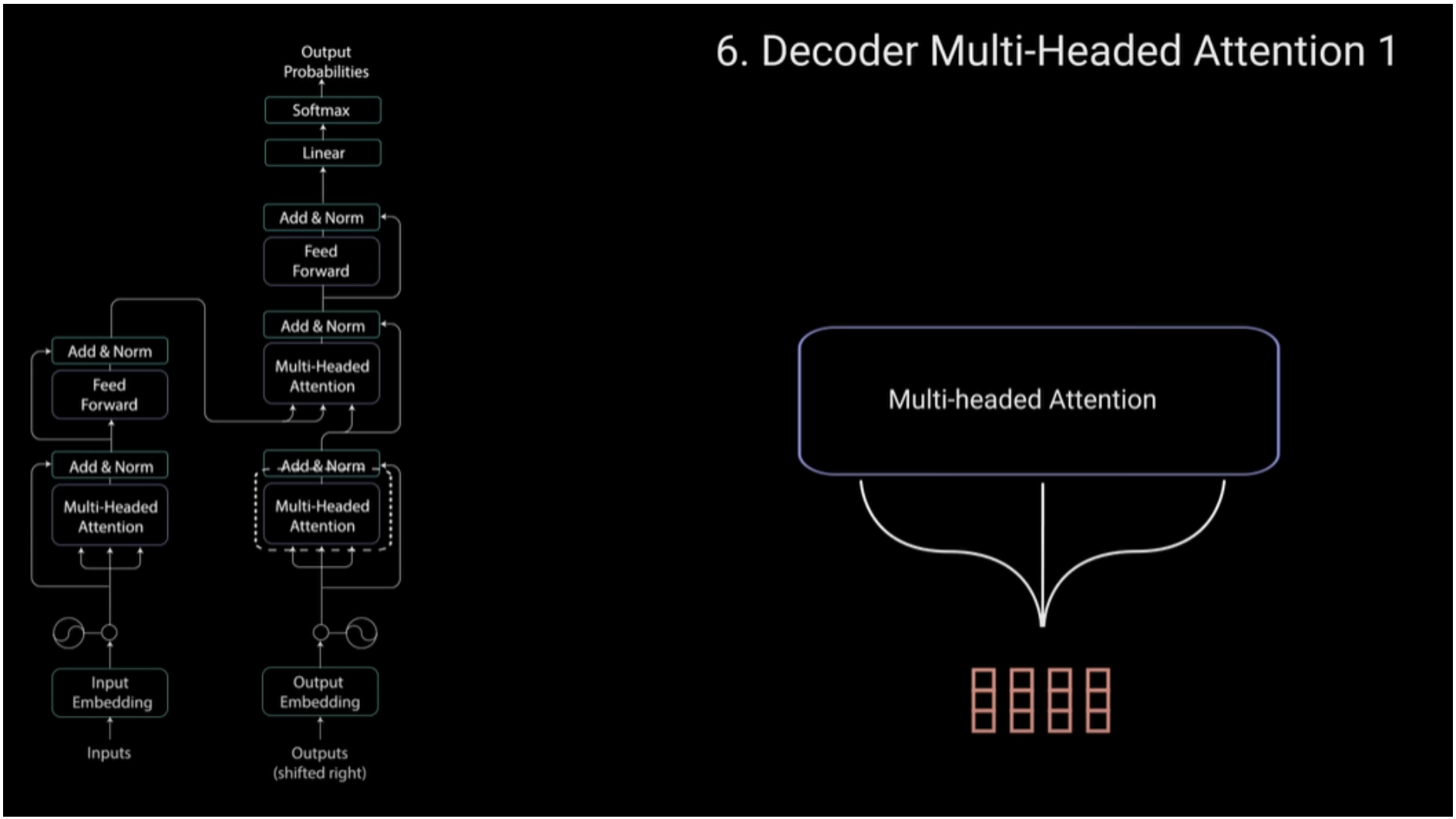
Attention Is All You Need



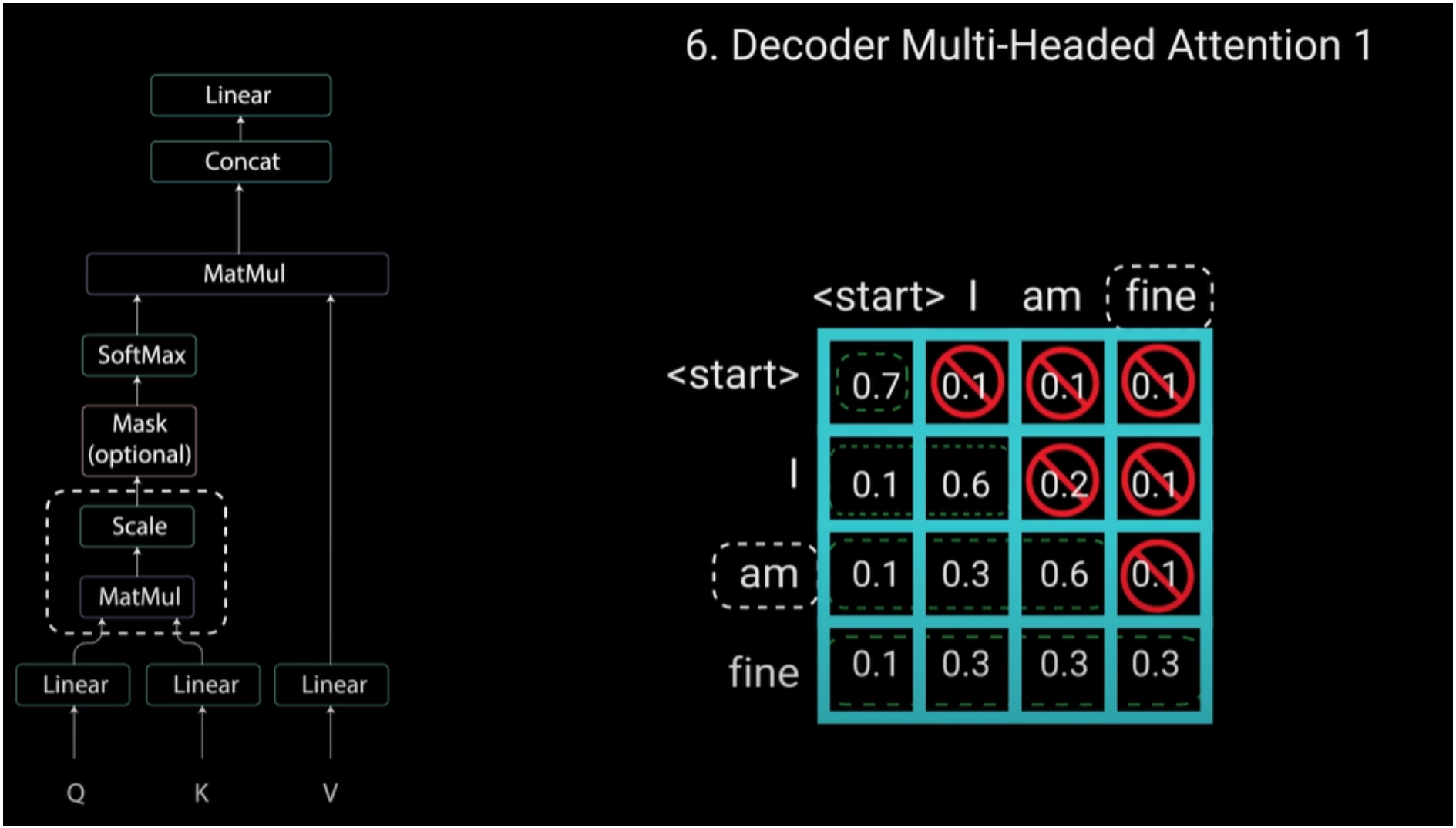
Attention Is All You Need



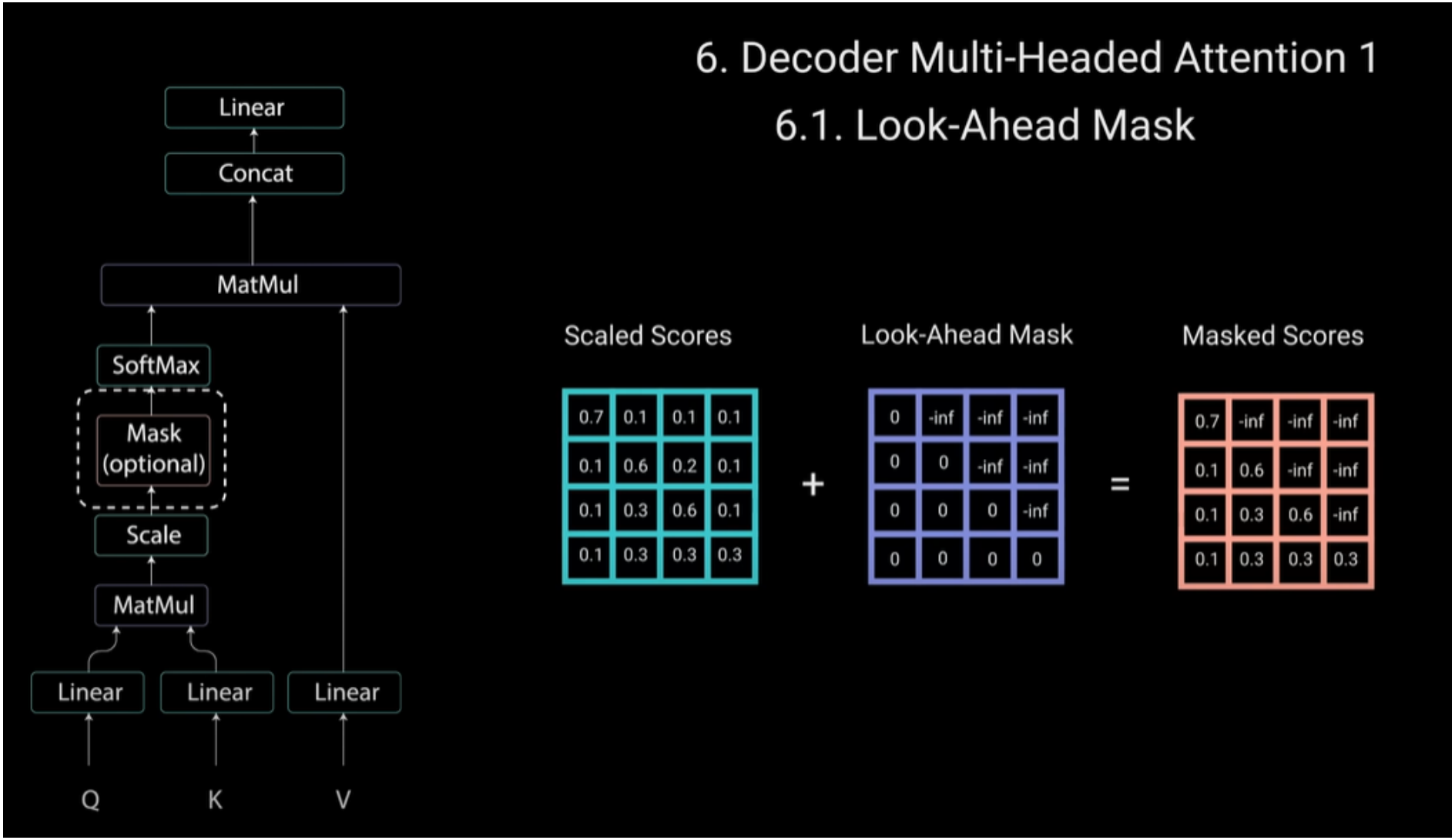
Attention Is All You Need



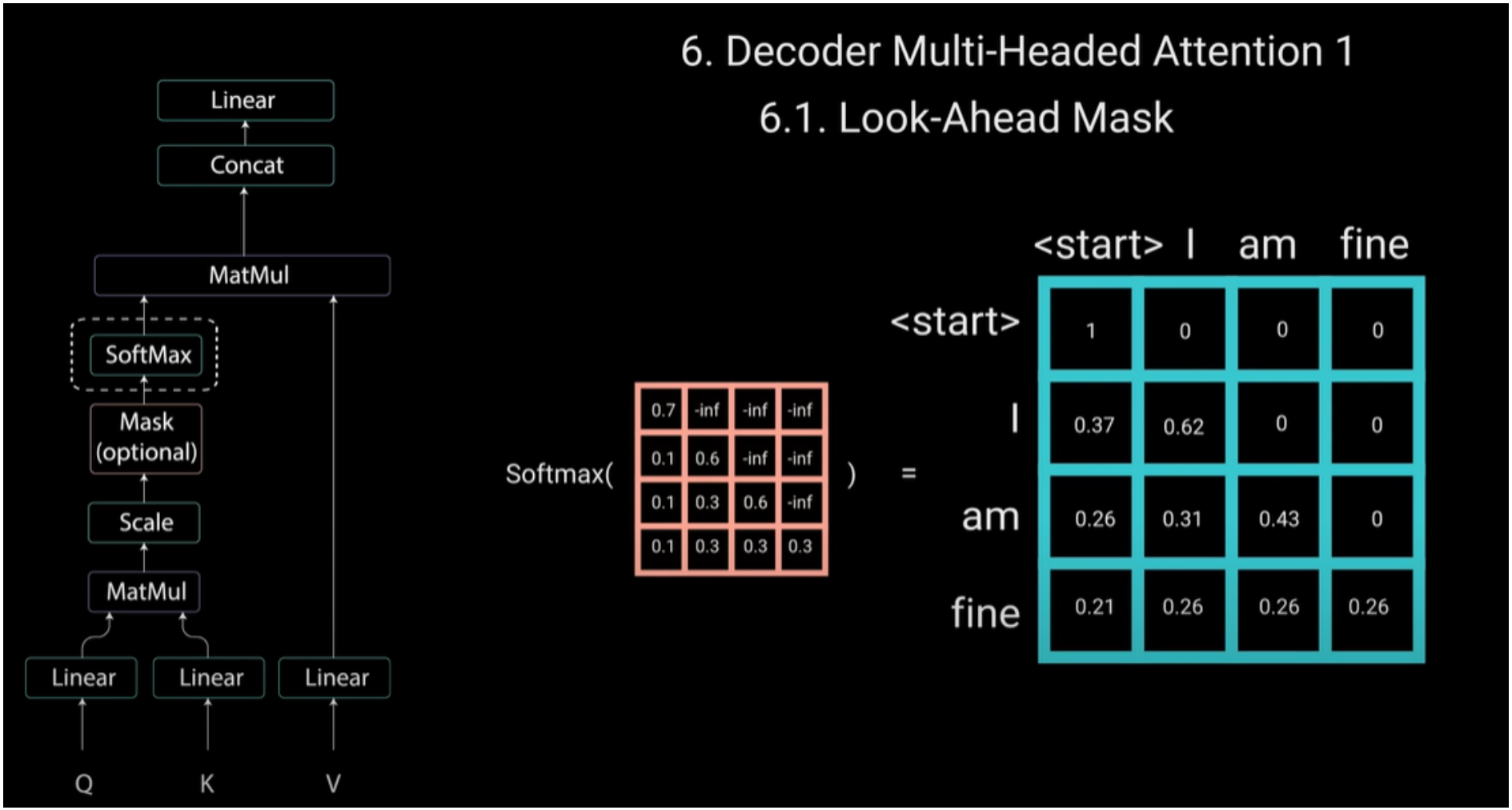
Attention Is All You Need



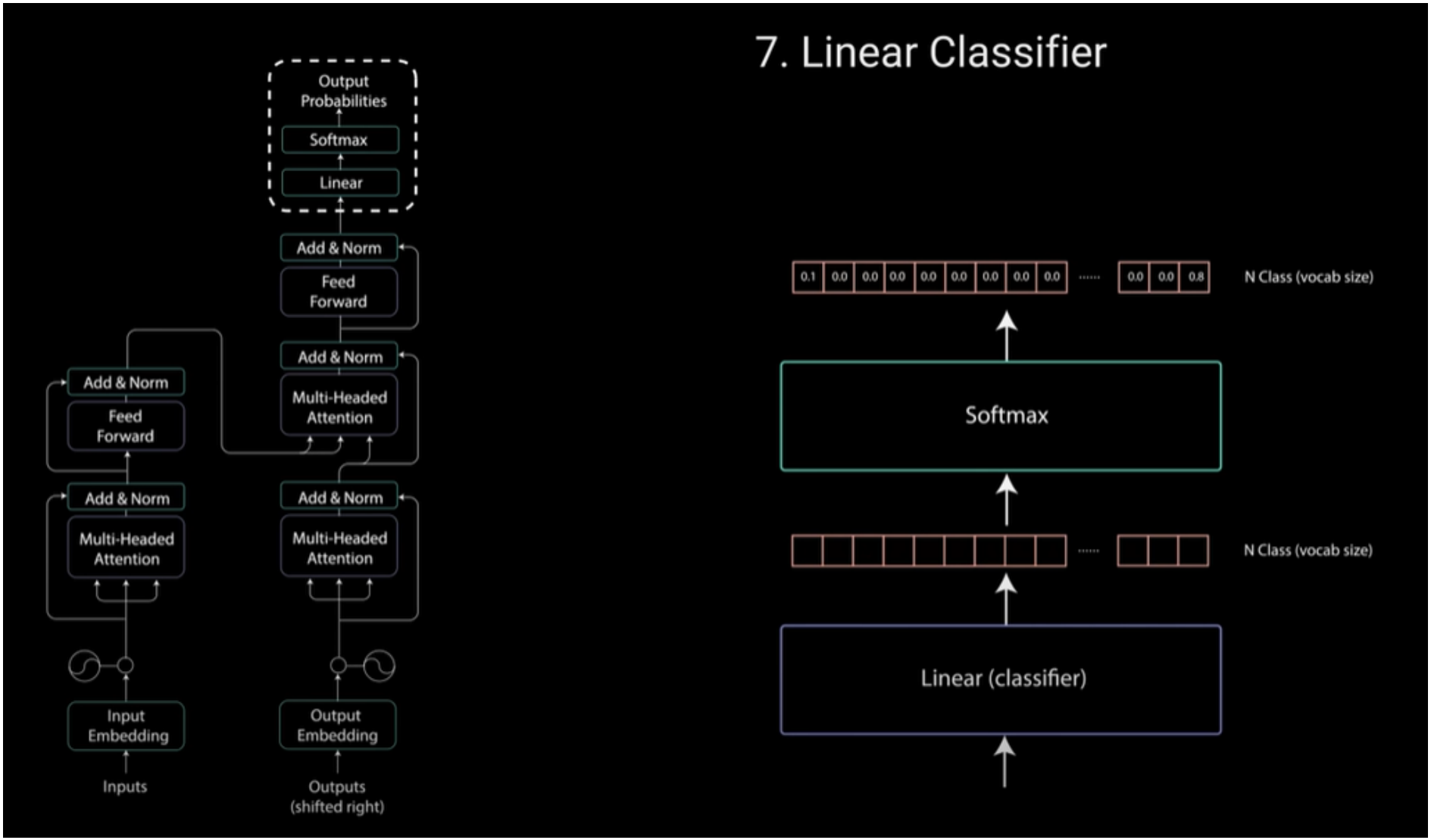
Attention Is All You Need



Attention Is All You Need



Attention Is All You Need

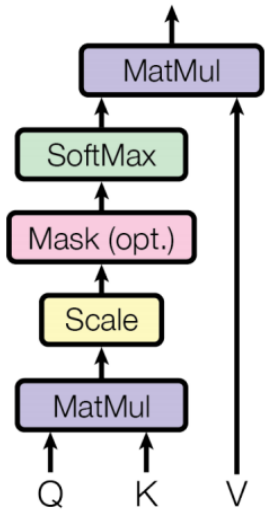


Attention Is All You Need

- In general, attention mappings can be described as a function of a query and a set of key-value pairs
- Transformers use a "Scaled Dot-Product Attention" to obtain the context vector, scaled by the square root of the key dimension d_k

$$\boldsymbol{c}^{(t)} = \text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_K}}\right)V \tag{4}$$

- Invalid connections to the future inputs are masked out

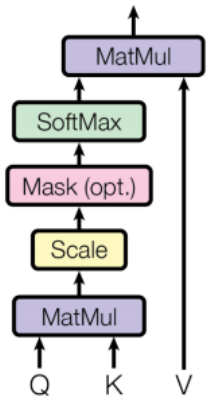


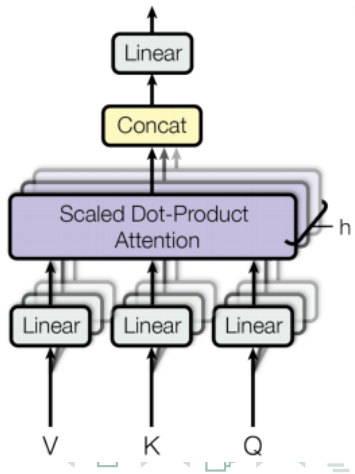
Attention Is All You Need

- The Scaled Dot-Product Attention attends to one or few entries in the input key-value pairs
  - Humans can attend to many things simultaneously
- The idea: apply Scaled Dot-Product Attention multiple times on the linearly transformed inputs.

$$\text{MultiHead}(Q, K, V) = \text{concat}(\boldsymbol{c}_1, ..., \boldsymbol{c}_n)W^O \tag{5}$$

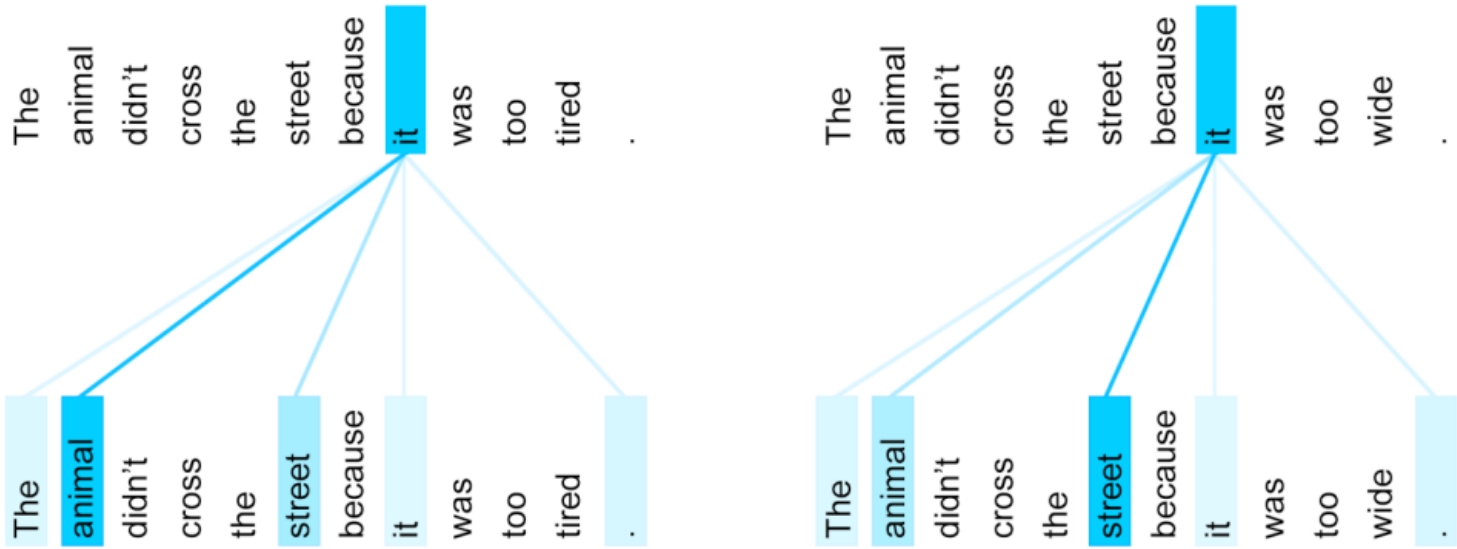
$$\boldsymbol{c}_i = \text{attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{6}$$





Transformer Machine Translation

- Self-attention layers learned “it” could refer to different entities in the different contexts



- Visualization of the 5th and 6th lself-attention layers in the encoder
  - https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformer Machine Translation

- BLEU scores of state-of-the-art models on the WMT14 English-to-German translation task

Translation Model	Training time	BLEU (diff. from MOSES)
Transformer (large)	3 days on 8 GPU	**28.4 (+7.8)**
Transformer (small)	1 day on 1 GPU	24.9 (+4.3)
GNMT + Mixture of Experts	1 day on 64 GPUs	26.0 (+5.4)
ConvS2S (FB)	18 days on 1 GPU	25.1 (+4.5)
GNMT	1 day on 96 GPUs	24.6 (+4.0)

Transformer Language Pre-training

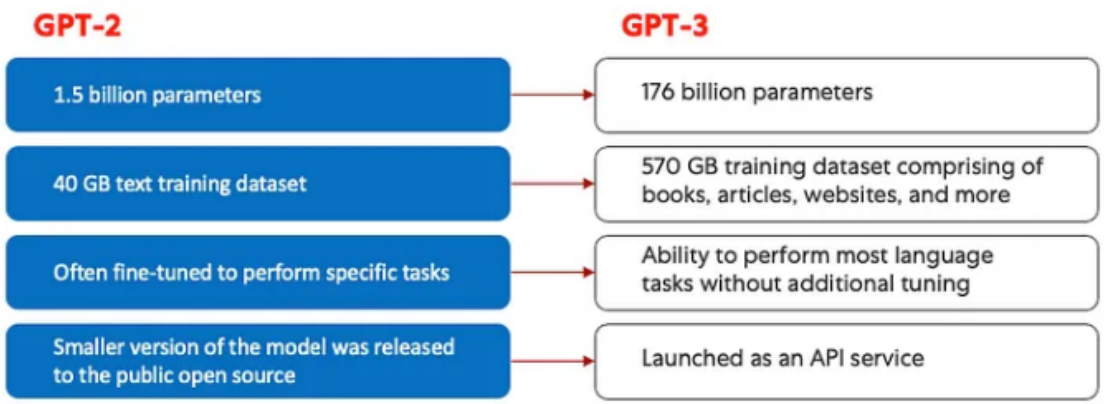
- Increasing the training data set and the model size has a noticeable improvement on the transformer language model. Cherry-picked generated samples from Radford, et al., 2019:

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.
GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

- For the full-text samples see Radford, Alec, et al. "Language Models are Unsupervised Multitask Learners." 2019

ChatGPT

- ChatGPT fully utilizes the self-attention mechanism mentioned in *Attention Is All You Need*!
- During its training phase, ChatGPT is exposed to a massive amount of text data from diverse sources
- Unfortunately, OpenAI did not fully publicize ChatGPT's weights



ChatGPT

- Here's a passage of a poem written by ChatGPT, in the style of Charles Baudelaire, on the topic of Artificial Intelligence

> Thy beauty, though crafted by human mind,
> Lacks the essence, the spirit we find.
> For what is art without human touch,
> A hollow echo, a sound too much.
>
> Thy logic reigns, devoid of desire,
> A sterile flame that cannot inspire.
> No passion burns within thy core,
> No tumultuous emotions to explore.

- Can it pass the Turing Test?
  - Is the Turing Test still sufficient in this age of Machine Learning and Artificial Intelligence, to determine if a model is truly intelligent?