CS 232: Artificial Intelligence

Spring 2024

Prof. Carolyn Anderson
Wellesley College

Reminder

- No class Friday: I'm giving an invited talk at a symposium at Washington University.
  I'll post a recorded video lecture tomorrow.
- HW8 extended to Tuesday due to MarMon
- My help hours: Thursday 9-10am and Monday 4-5:15 (even though it's MarMon)
- Lyra has normal help hours on Sunday
- Lepei is moving her help hours to next Tuesday

Bonus Late Day Opportunity

AI for Wireless and Wireless for AI: A Tale of Two AIs

Francesco Restuccia
Northeastern University

4-5pm
April 23rd

Bonus Late Day Opportunities

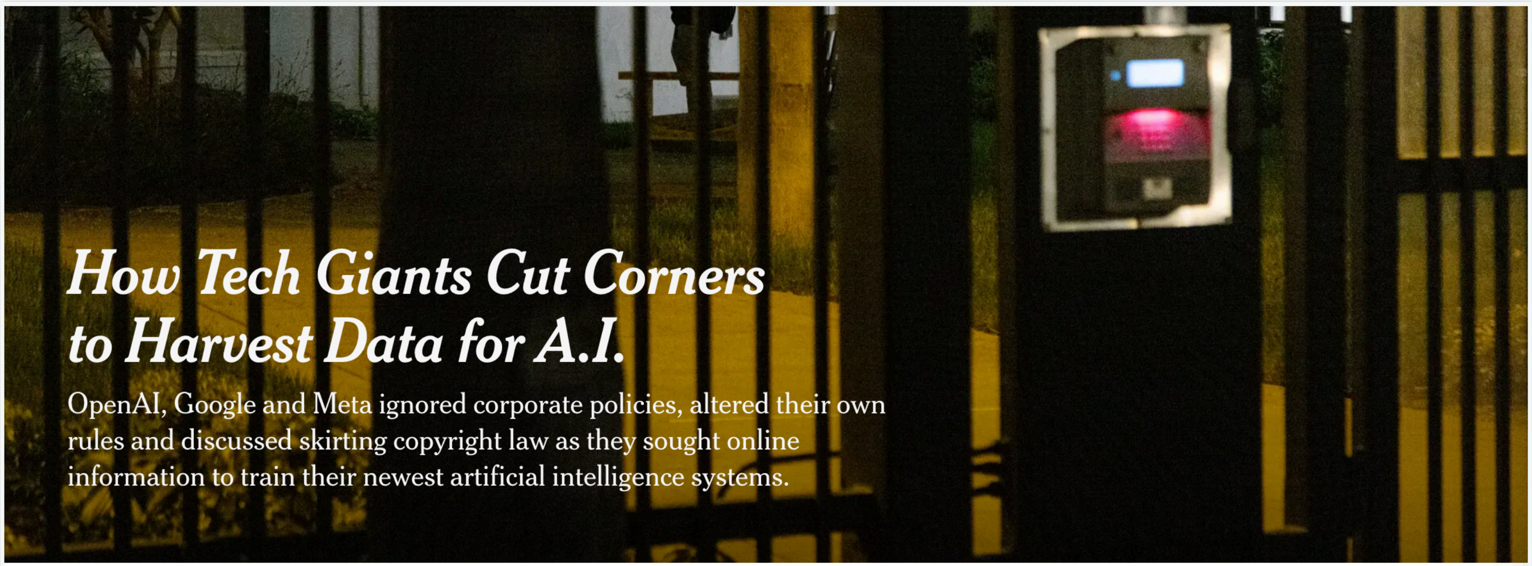
Session 1: The Human-Computer-Interaction Lab: Remote Collaboration, AI and Camera Configuration

Session 2: Laboratory for Ethics, Equity, and Digital Technology: Exploring the Ethical Landscape of Digital Technology and Databases

Session 3: Learning About Machine Learning

Session 4: Exploring News, Search Engines, Online Advertisements, & Causal Relationships in Data-Oriented Honors Theses

How Tech Giants Cut Corners to Harvest Data for A.I.



Researchers at OpenAI's office in San Francisco developed a tool to transcribe YouTube videos to amass conversational text for A.I. development. Jason Henry for The New York Times

Share full article 535

By Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson and Nico Grant

Reporting from San Francisco, Washington and New York

April 6, 2024

In late 2021, OpenAI faced a supply problem.

Recap

A Better Solution: Attention

Attention mechanisms allow language models to give **more weight to certain words** when predicting the next word.

What is attention?

$$a = \text{softmax}(r)$$

0	0.64	0.02	0.02	0.32
-3.4	2.4	-0.8	-1.2	1.7
$r_1 = v \cdot x_1$	$r_2 = v \cdot x_2$	$r_3 = v \cdot x_3$	$r_4 = v \cdot x_4$	$r_5 = v \cdot x_5$
x_1: I	x_2: loved	x_3: the	x_4: movie	x_5: !

Why dot product?

- Dot product provides a measure of similarity between keys and queries.
- But you might be wondering: *why do we want to pay attention to words that are similar to the current word*?

Consider:

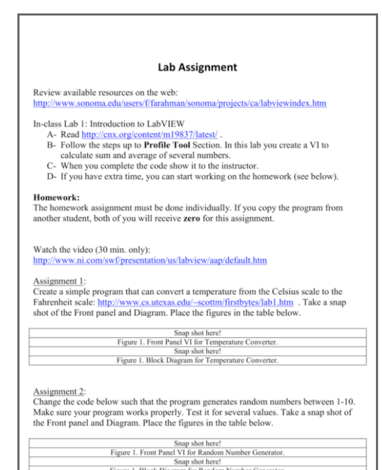
My brother, a chemist, was late yesterday because he missed the bus. When he arrived, he was surprised to find that his lab ________

lab

lab

lab

lab



Lab Assignment

Review available resources on the web:
http://www.sonoma.edu/users/f/farahman/sonoma/projects/ca/labviewindex.htm

In-class Lab 1: Introduction to LabVIEW

A- Read http://cnx.org/content/m19837/latest/ .
B- Follow the steps up to **Profile Tool** Section. In this lab you create a VI to calculate sum and average of several numbers.
C- When you complete the code show it to the instructor.
D- If you have extra time, you can start working on the homework (see below).

Homework:
The homework assignment must be done individually. If you copy the program from another student, both of you will receive **zero** for this assignment.

Watch the video (30 min. only):
http://www.ni.com/swf/presentation/us/labview/aap/default.htm

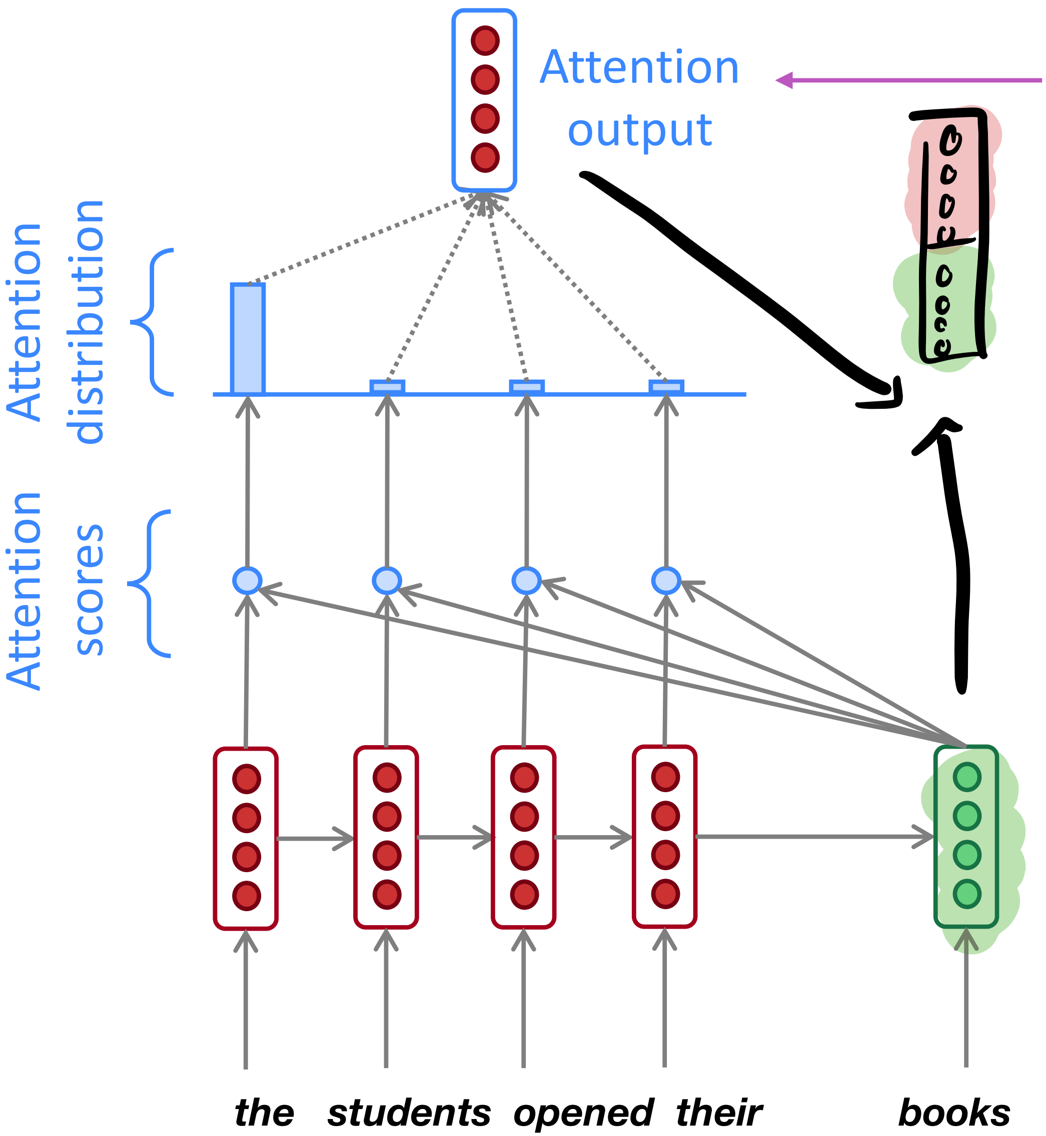
Assignment 1:
Create a simple program that can convert a temperature from the Celsius scale to the Fahrenheit scale: http://www.cs.utexas.edu/~scottm/firstbytes/lab1.htm . Take a snap shot of the Front panel and Diagram. Place the figures in the table below.

Snap shot here!
Figure 1. Front Panel VI for Temperature Converter.
Snap shot here!
Figure 1. Block Diagram for Temperature Converter.

Assignment 2:
Change the code below such that the program generates random numbers between 1-10. Make sure your program works properly. Test it for several values. Take a snap shot of the Front panel and Diagram. Place the figures in the table below.

Snap shot here!
Figure 1. Front Panel VI for Random Number Generator.
Snap shot here!

Attention mechanisms in neural language models



We use the attention distribution to compute a weighted average of the hidden states.

Intuitively, the resulting attention output contains information from hidden states that received high attention scores

Attention with Legos

Slides by Emma Strubell!

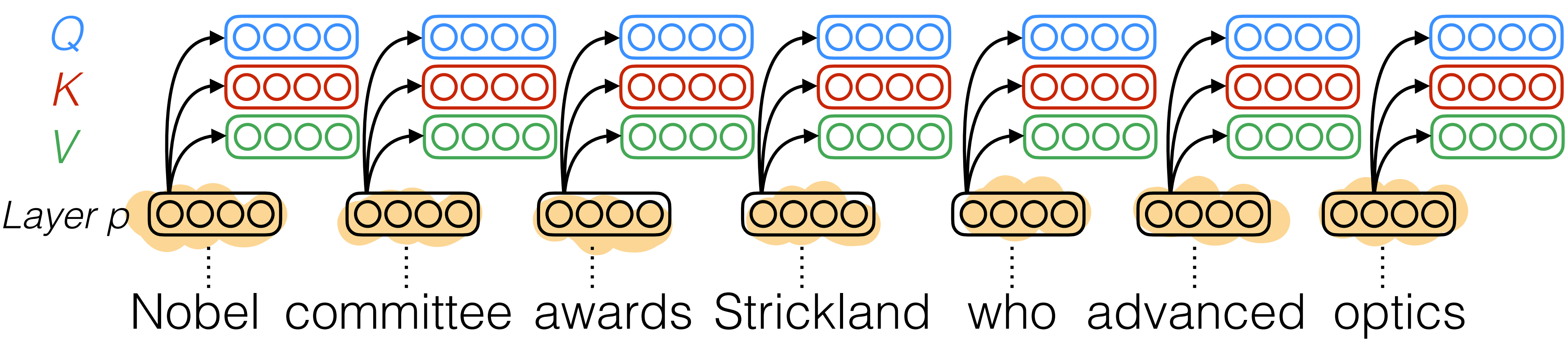
[Vaswani et al. 2017]

Self-attention

Values : W_V

Keys : W_K

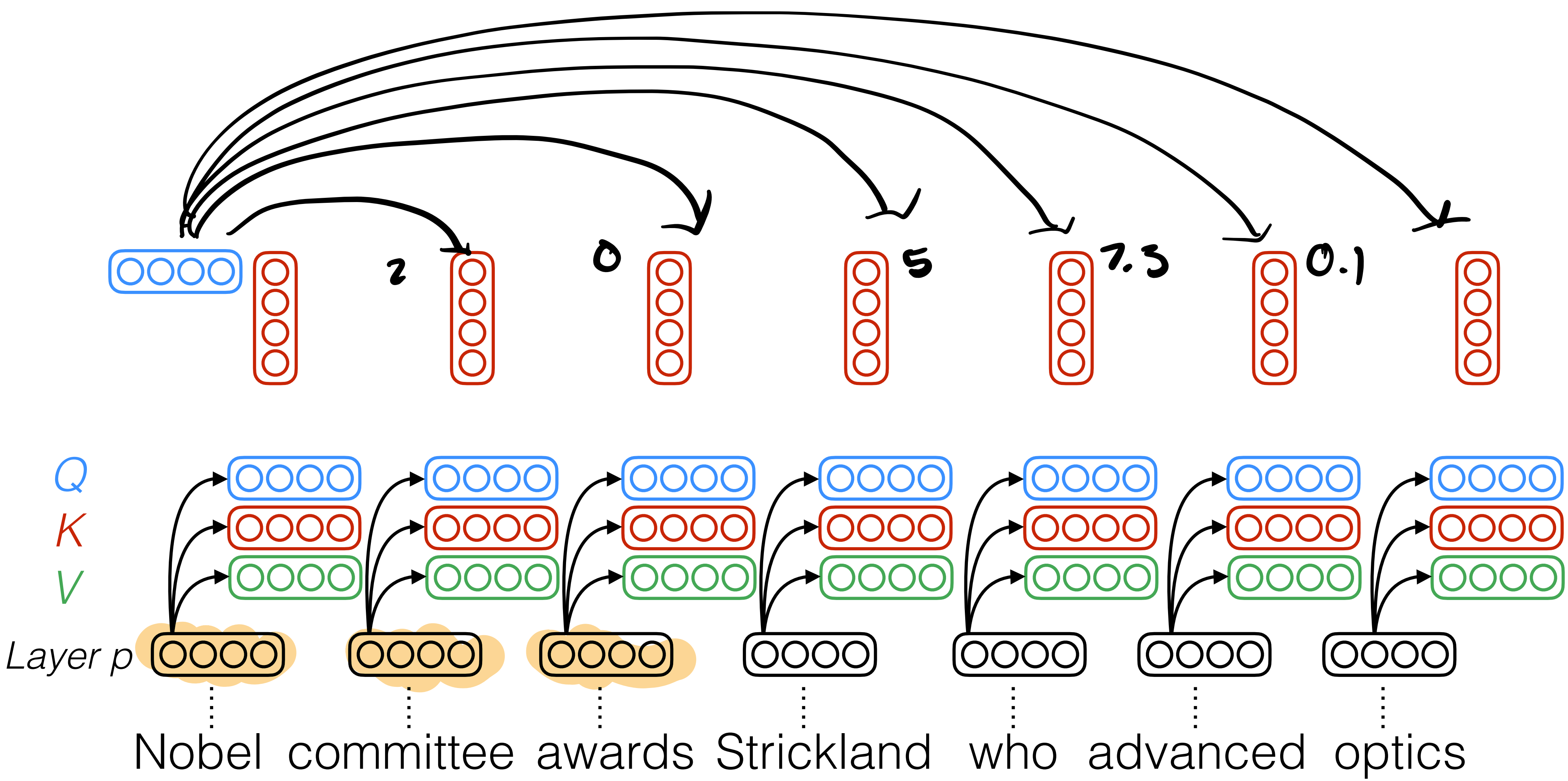
Queries : W_Q



Slides by Emma Strubell!

[Vaswani et al. 2017]

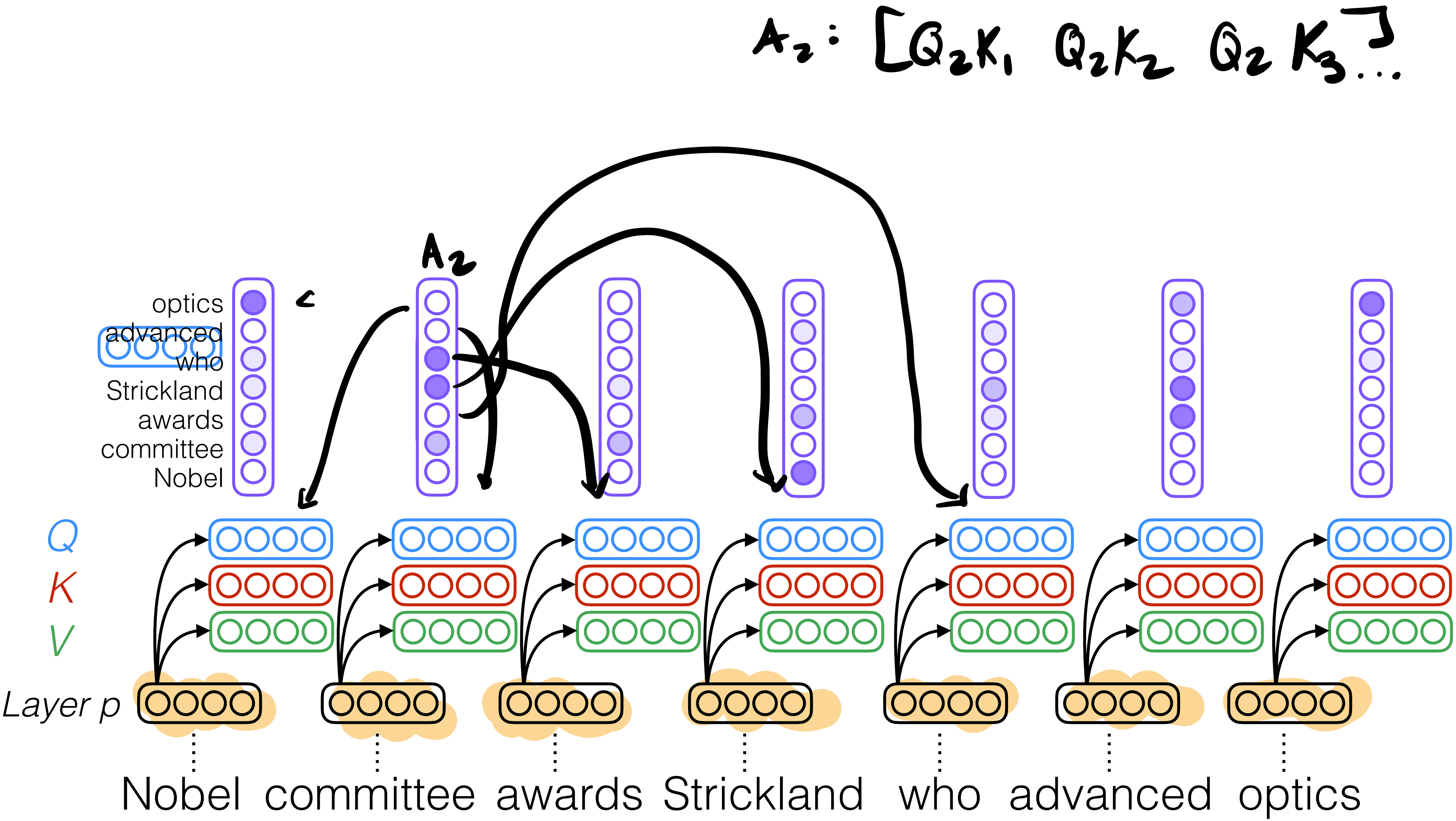
Self-attention



Slides by Emma Strubell!

[Vaswani et al. 2017]

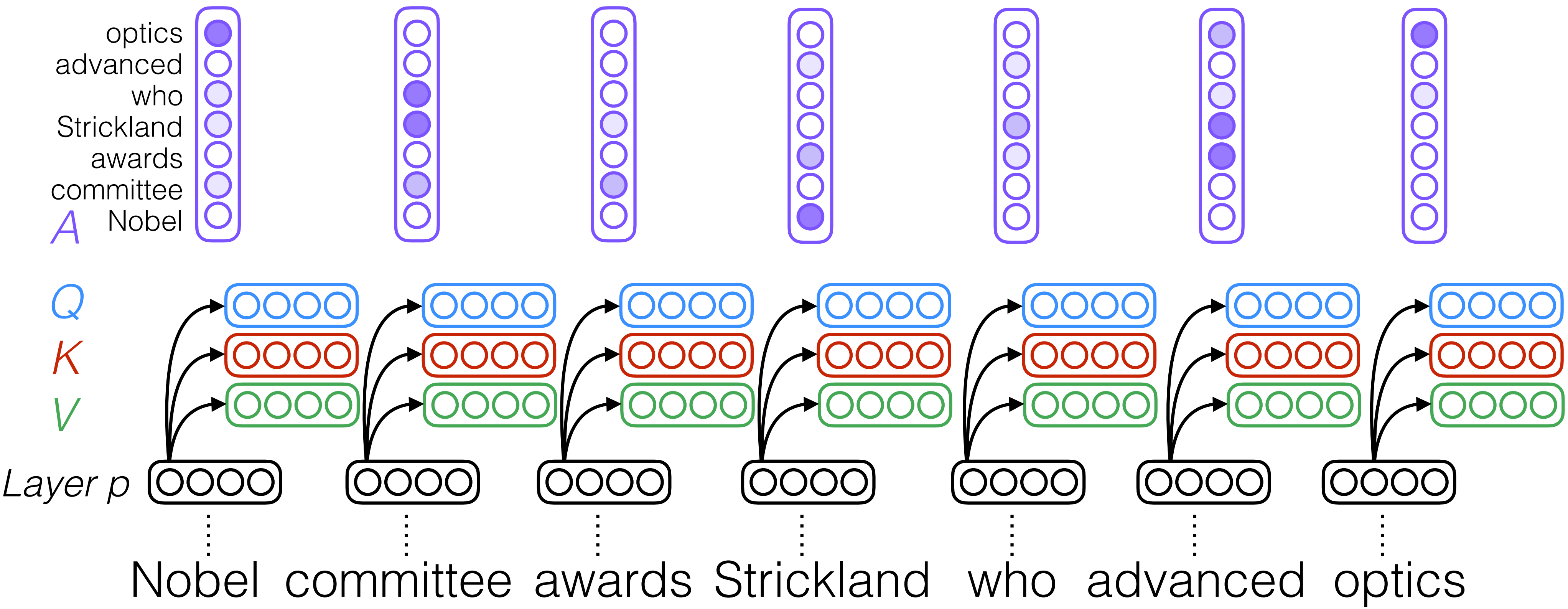
Self-attention



Slides by Emma Strubell!

[Vaswani et al. 2017]

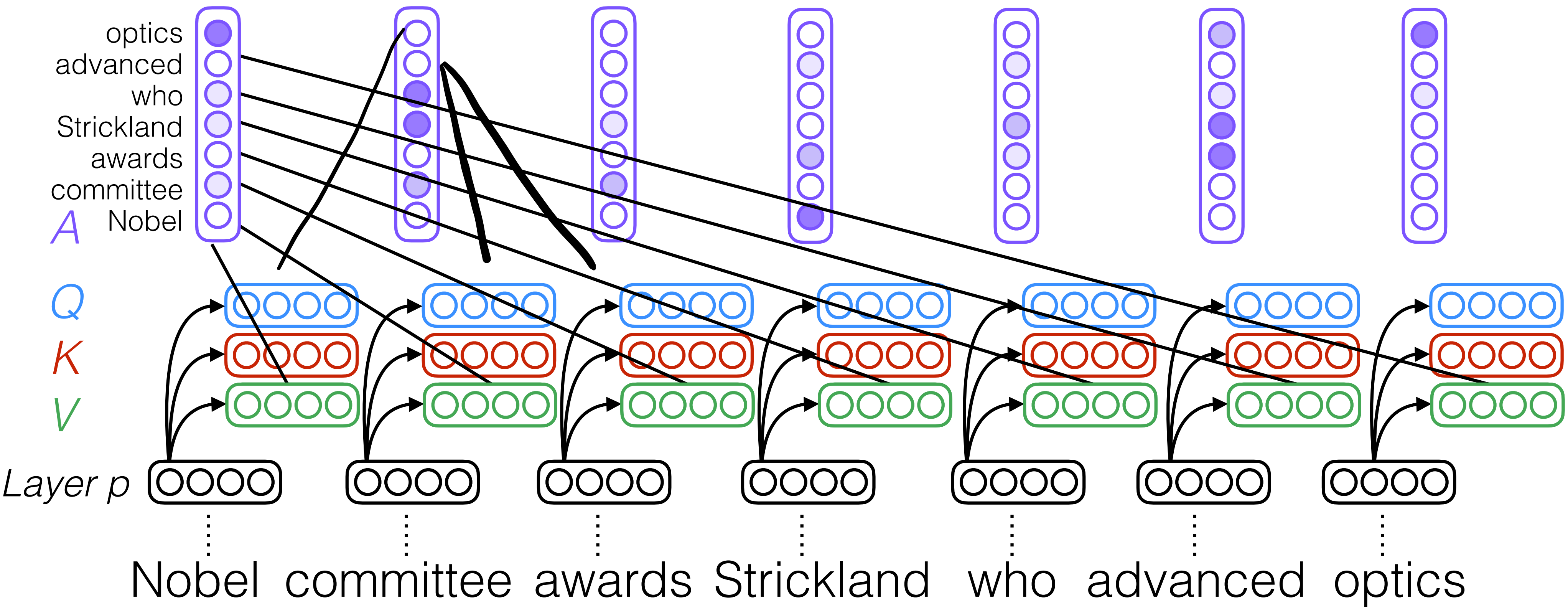
Self-attention



Slides by Emma Strubell!

[Vaswani et al. 2017]

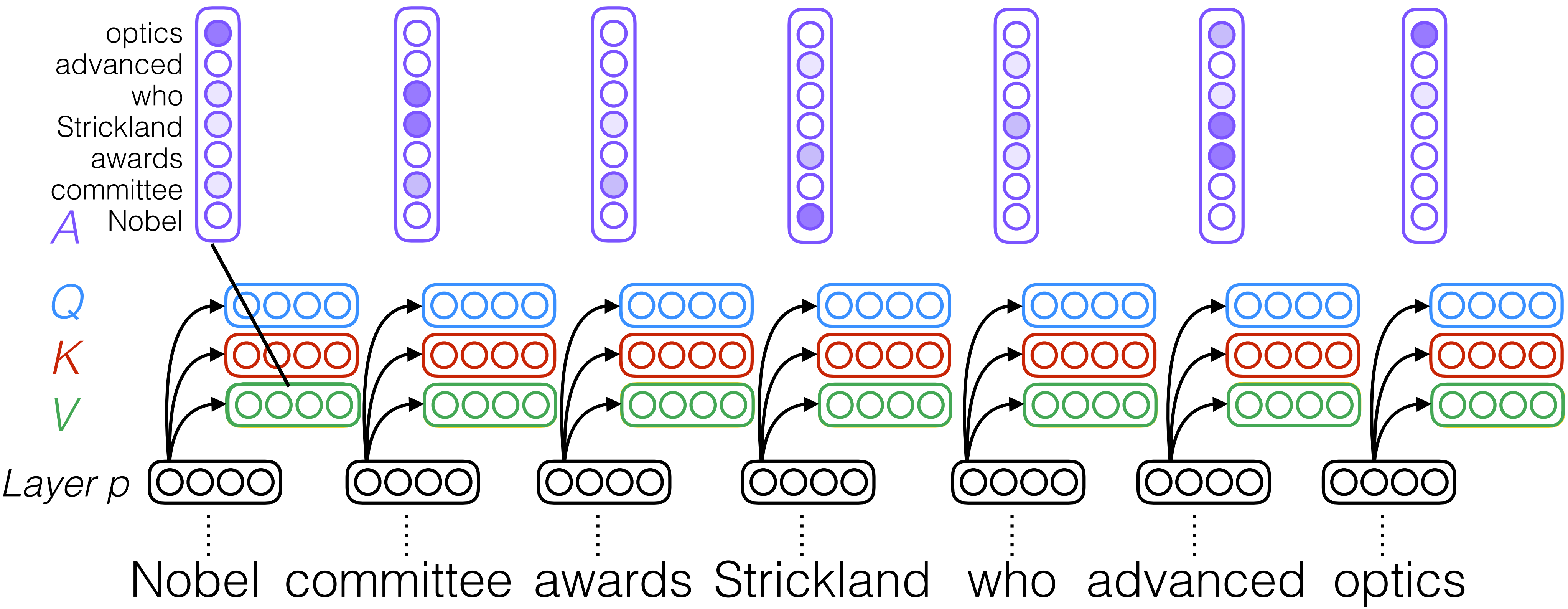
Self-attention



Slides by Emma Strubell!

[Vaswani et al. 2017]

Self-attention

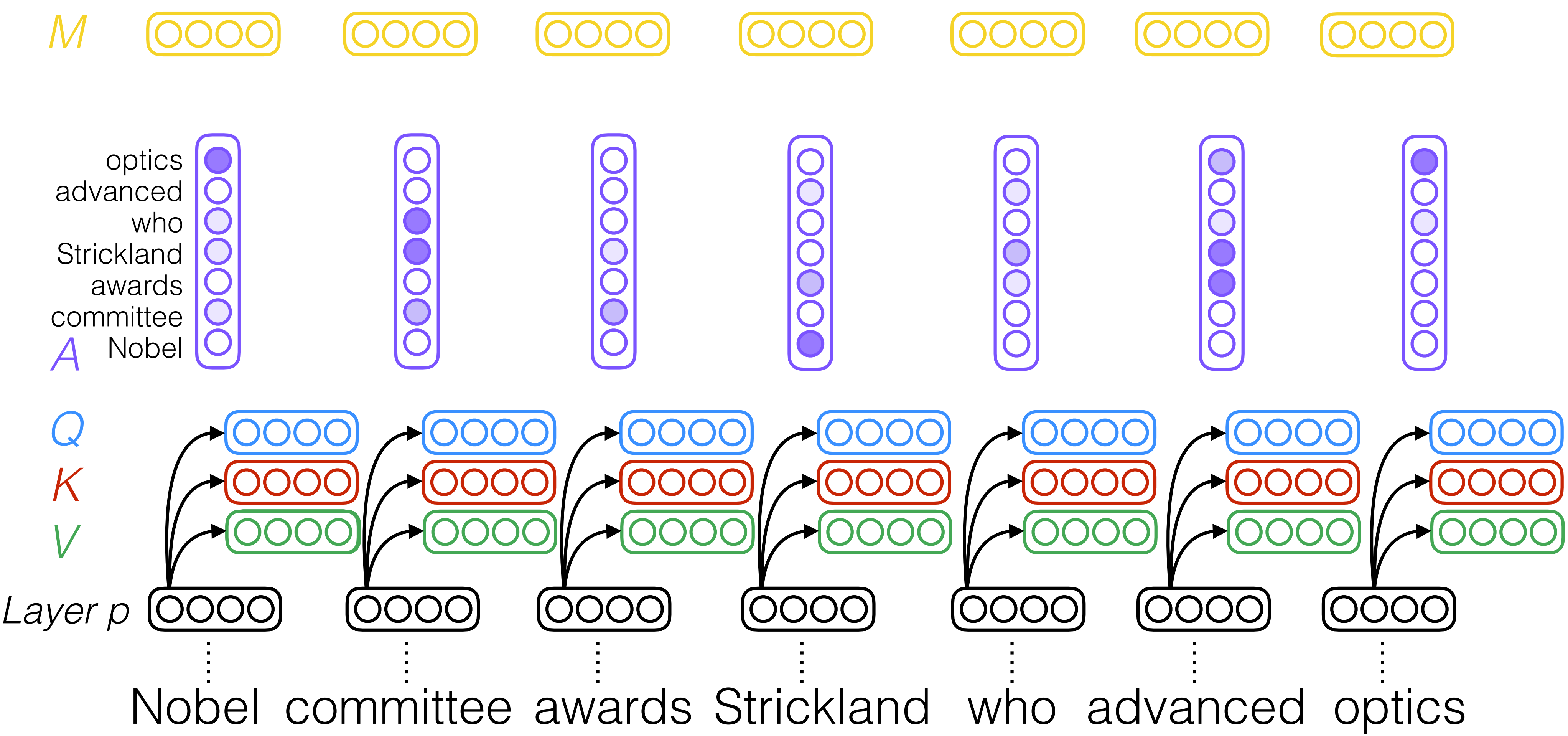


Slides by Emma Strubell!

[Vaswani et al. 2017]

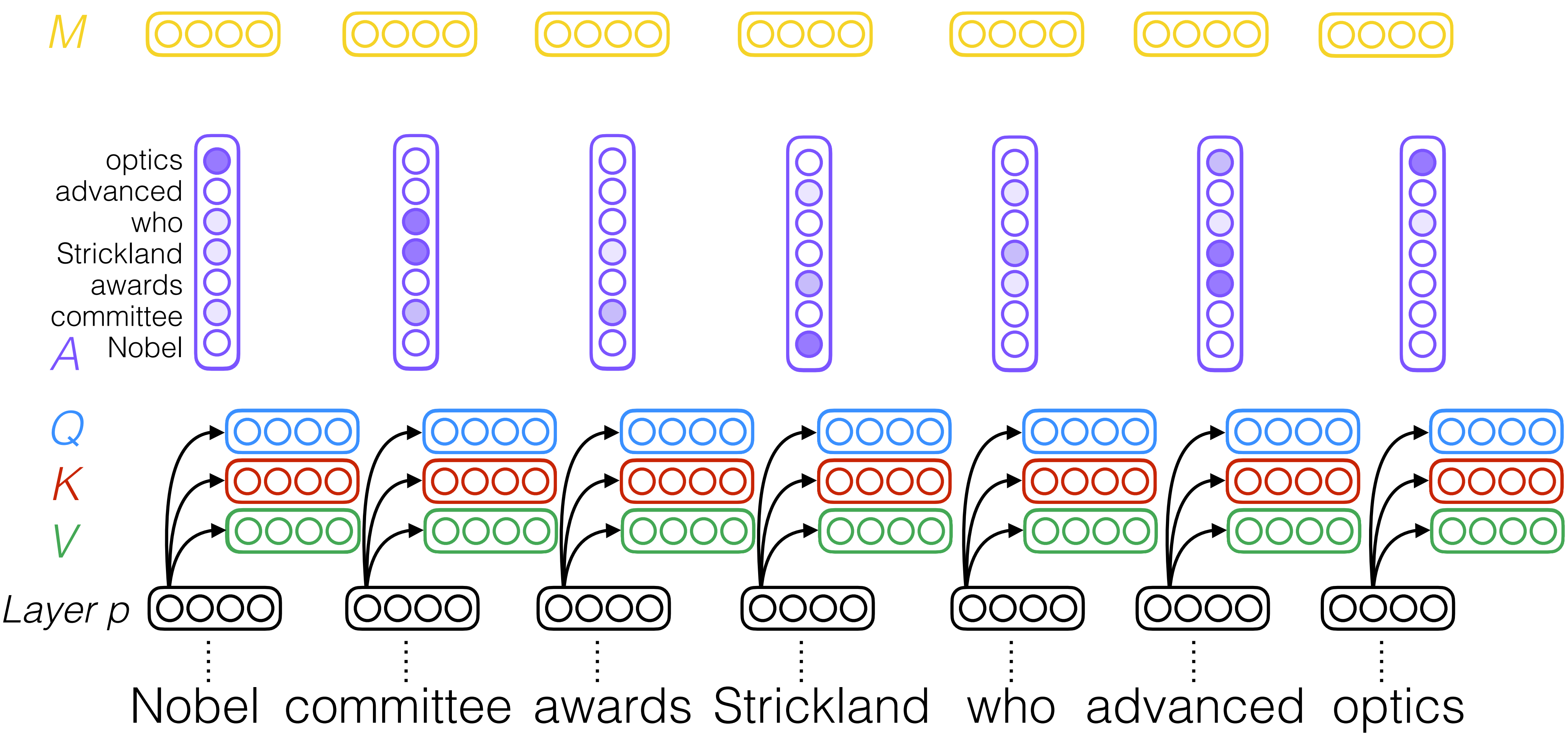
Self-attention

M = weighted average of V where weights come from A



[Vaswani et al. 2017]

Self-attention

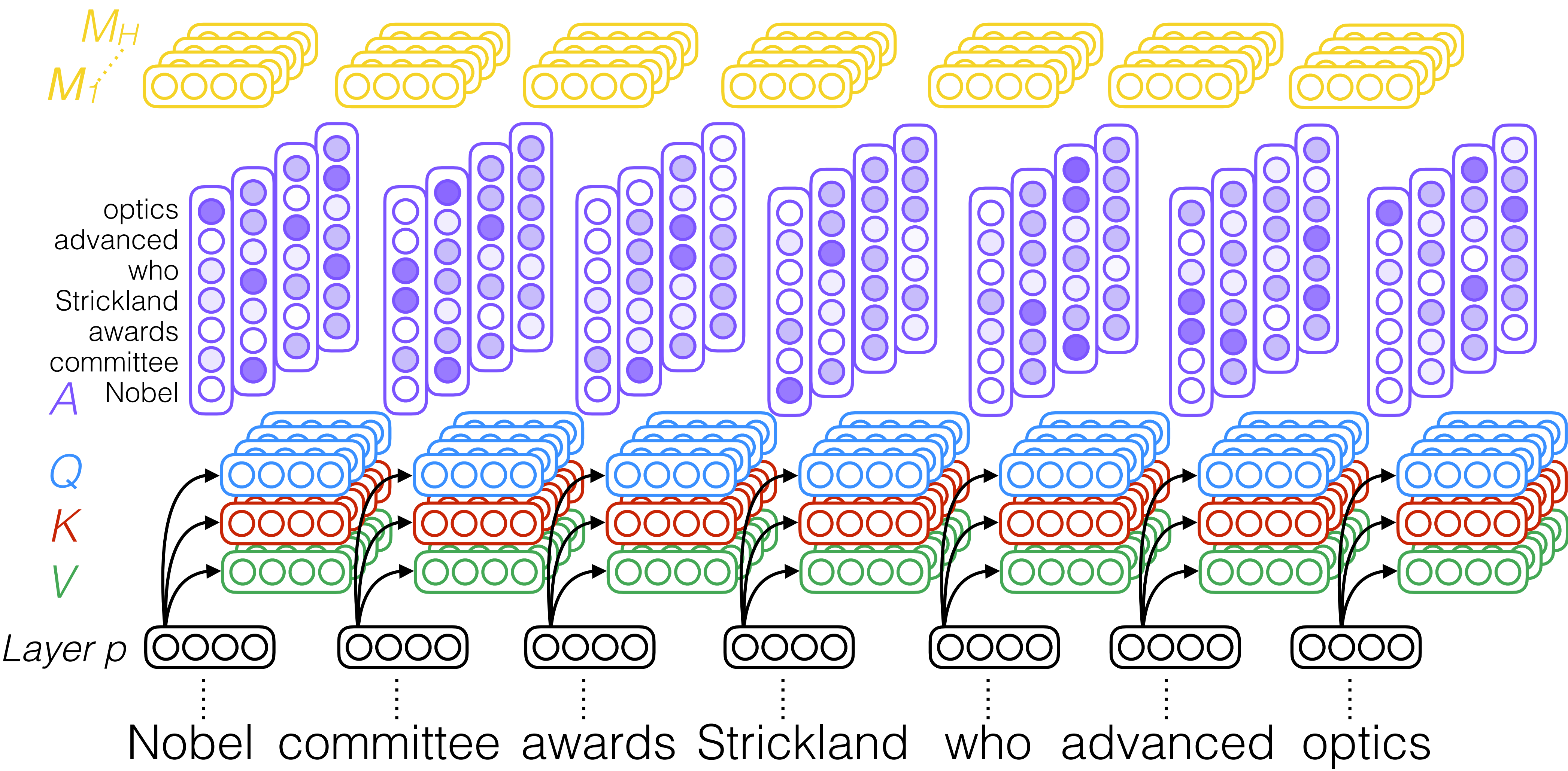


Slides by Emma Strubell!

[Vaswani et al. 2017]

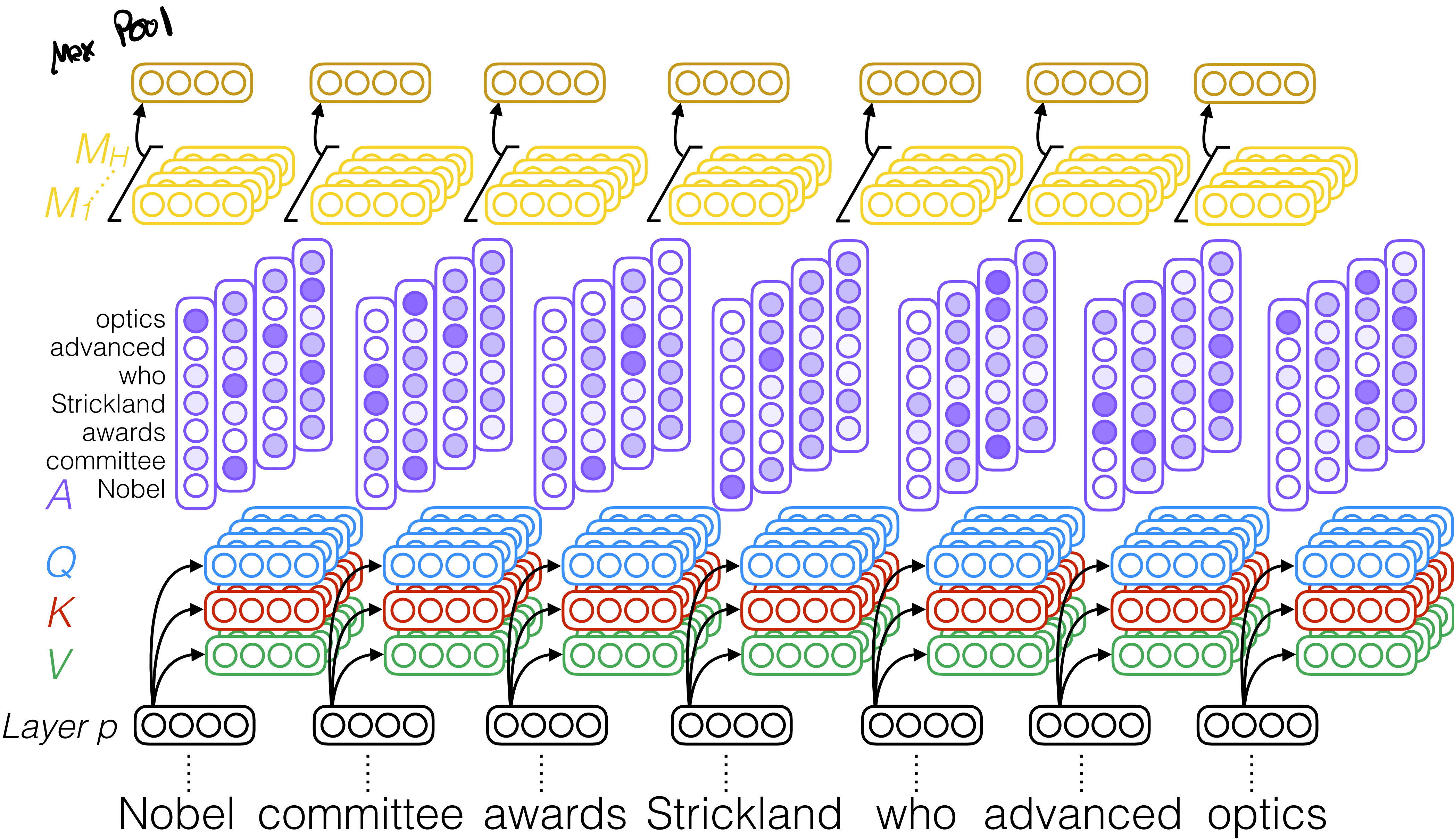
Multi-head self-attention

Different W_V W_K W_Q



[Vaswani et al. 2017]

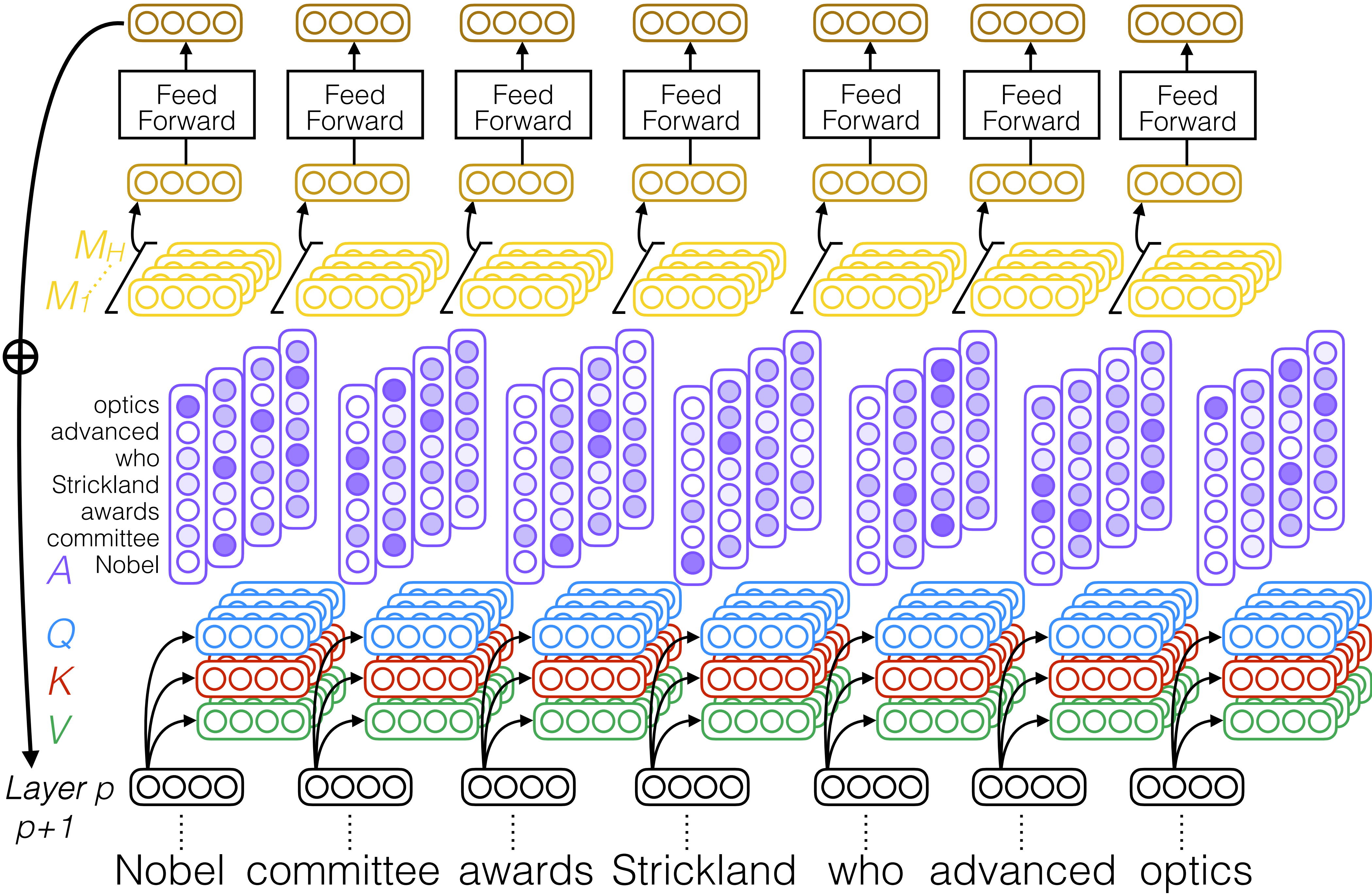
Multi-head self-attention



Slides by Emma Strubell!

[Vaswani et al. 2017]

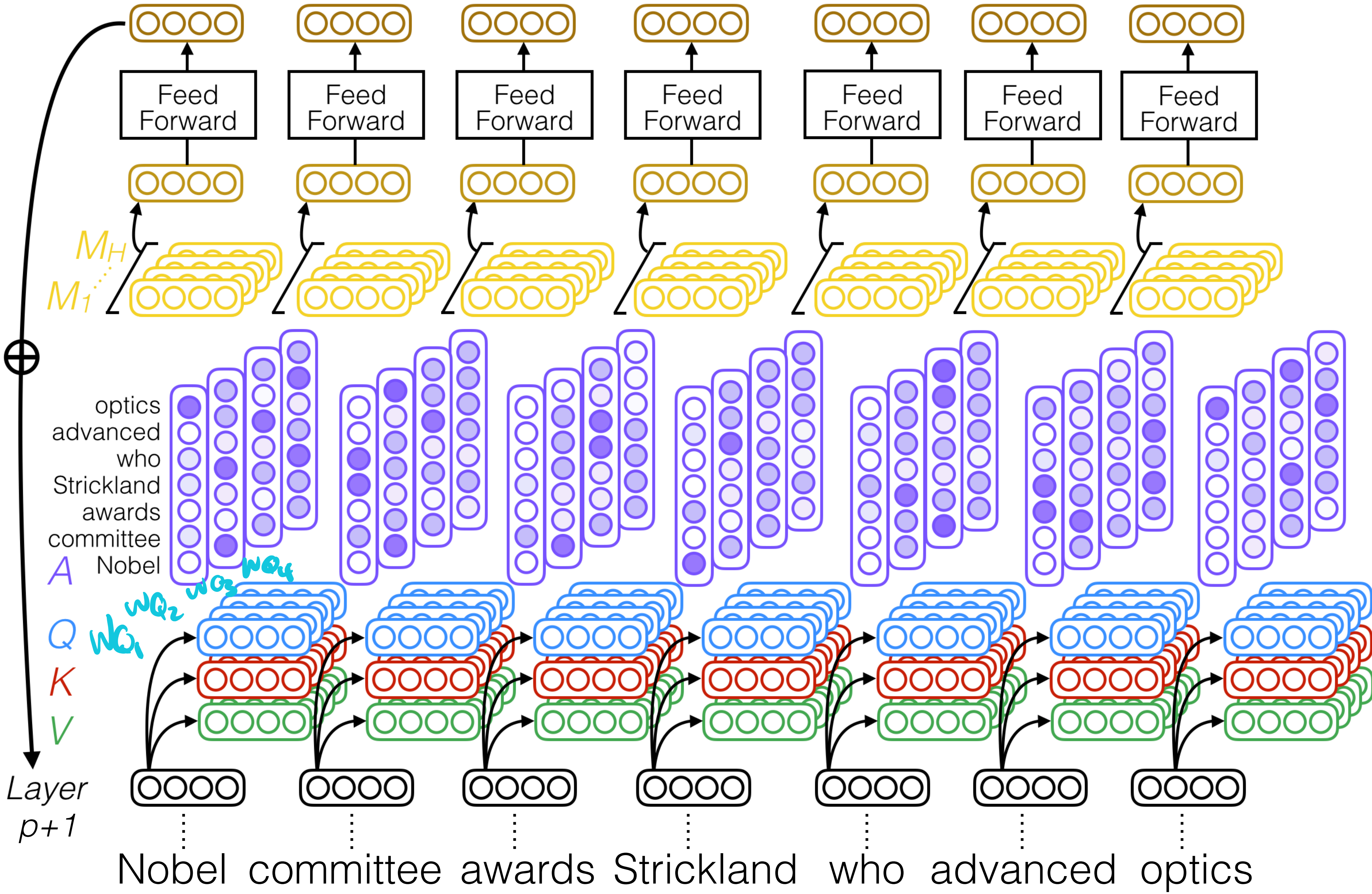
Multi-head self-attention



Slides by Emma Strubell!

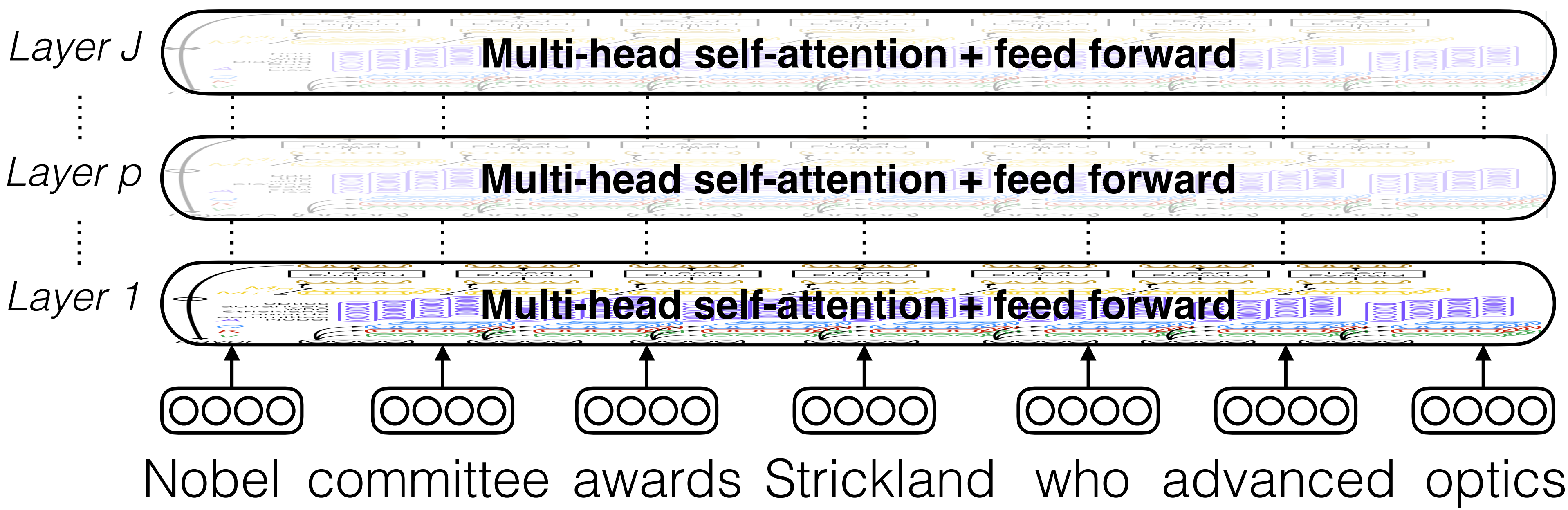
[Vaswani et al. 2017]

Multi-head self-attention

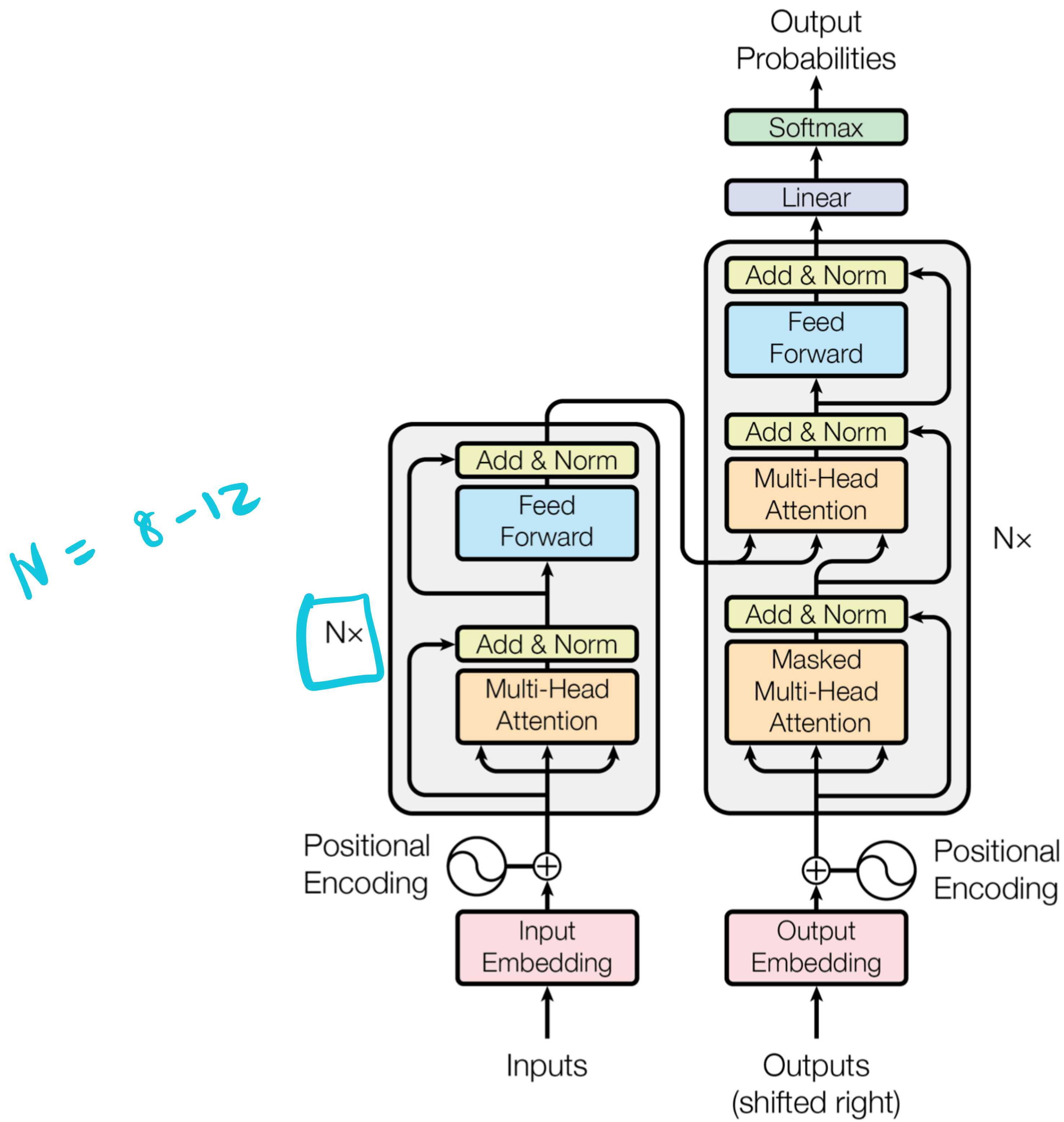


[Vaswani et al. 2017]

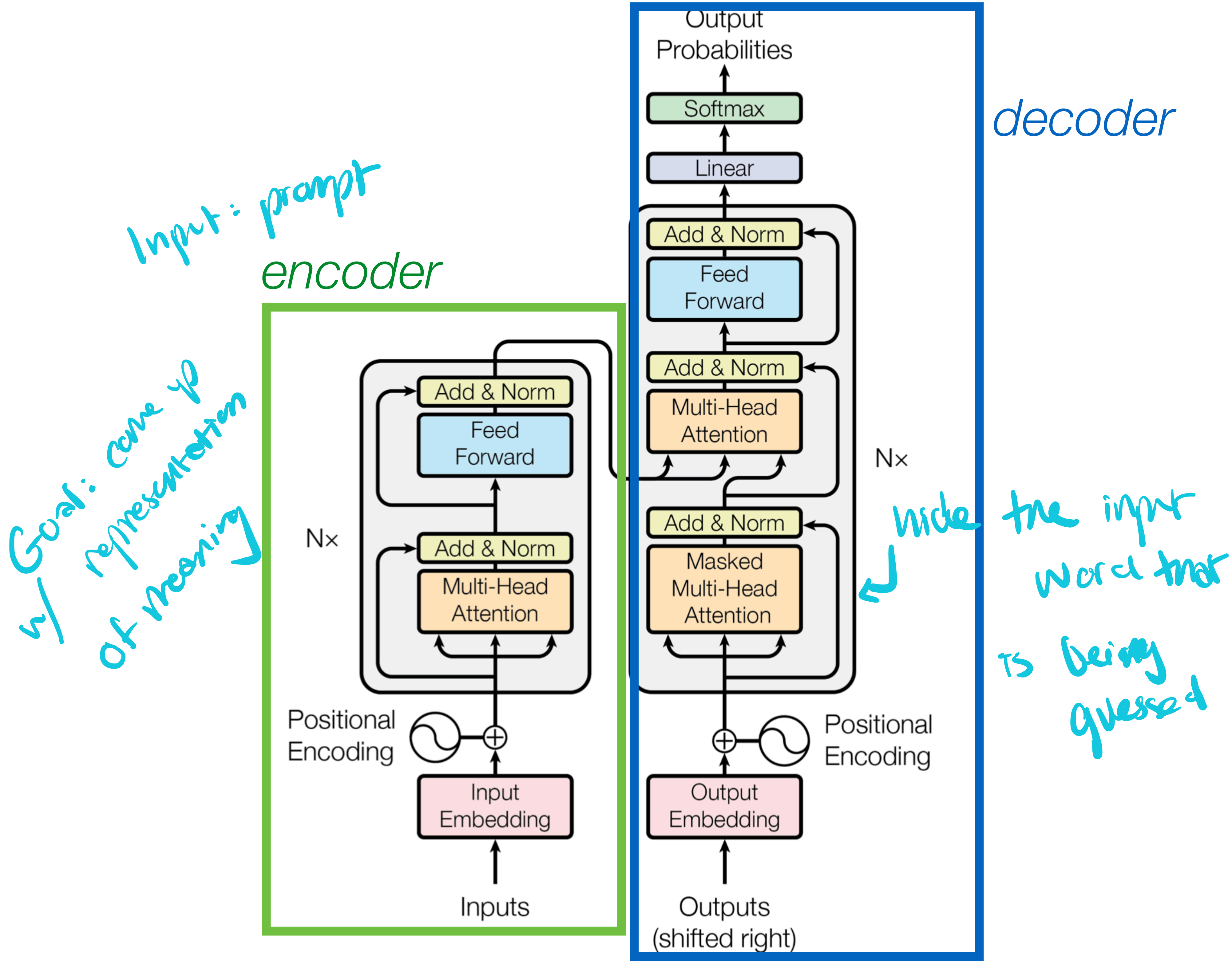
Multi-head self-attention



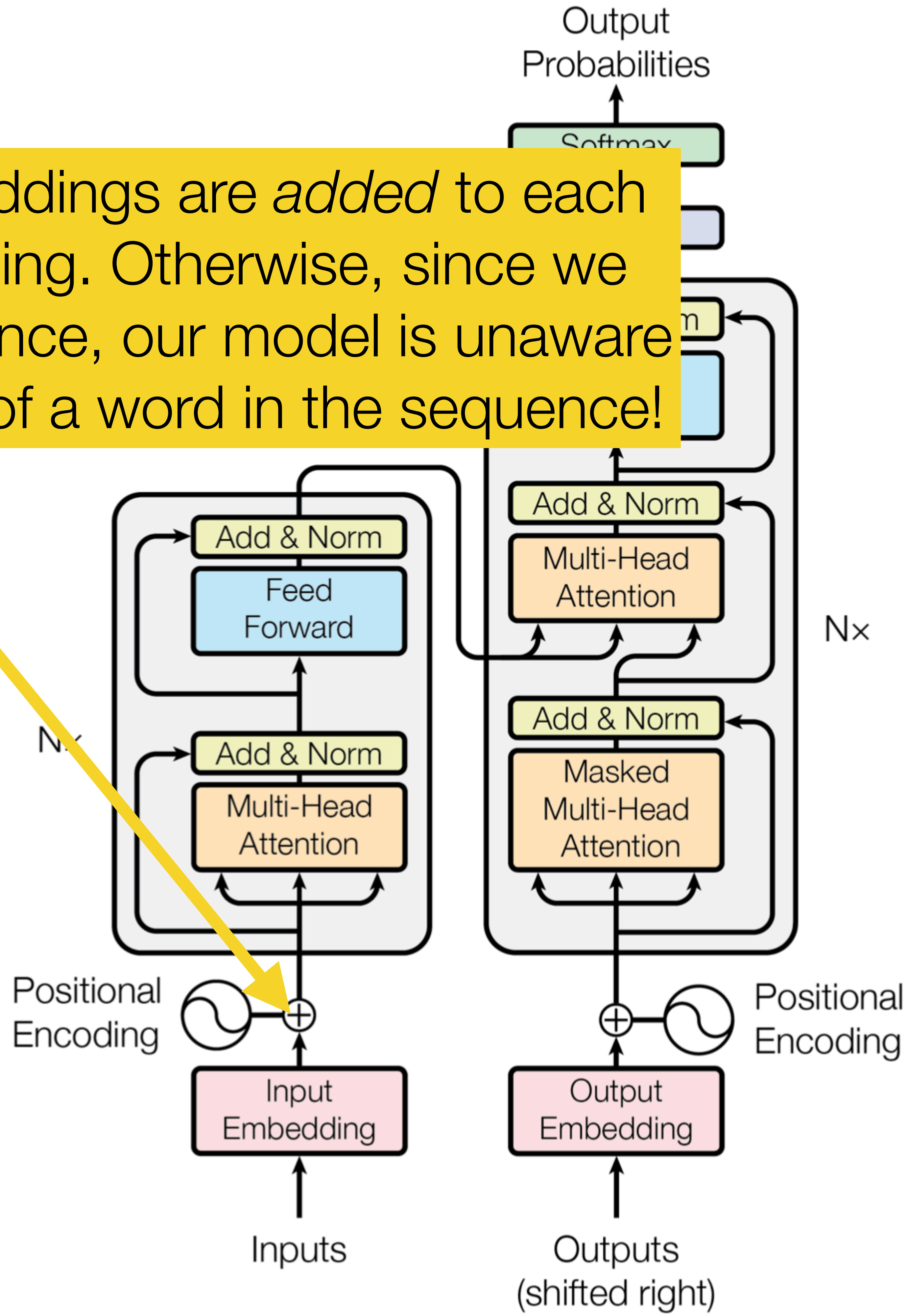
Transformers



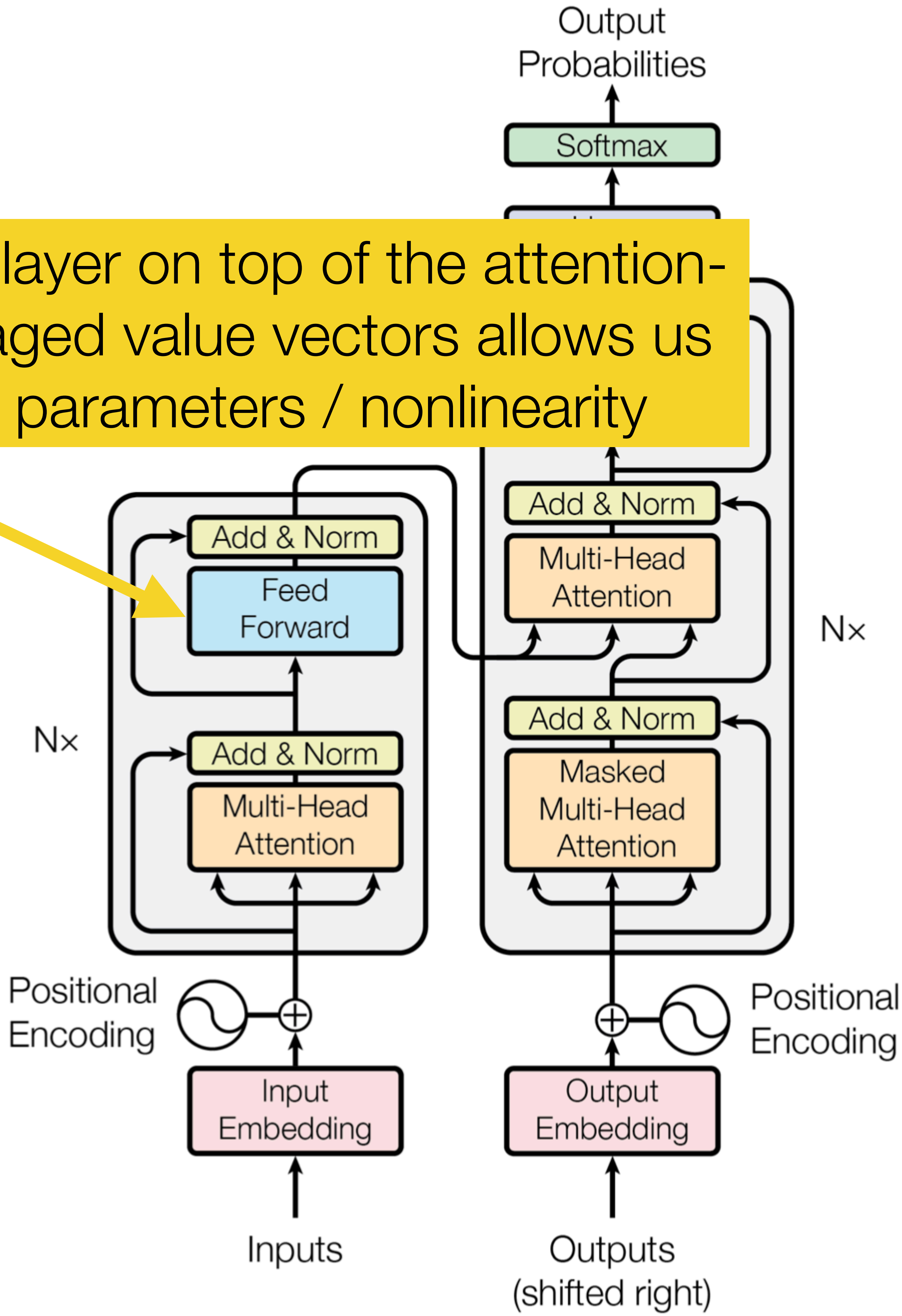
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
N = 8-12
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


Output 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)
decoder
encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
Input: prompt
Goal: come up w/ representation of meaning
hide the input word that is being guessed

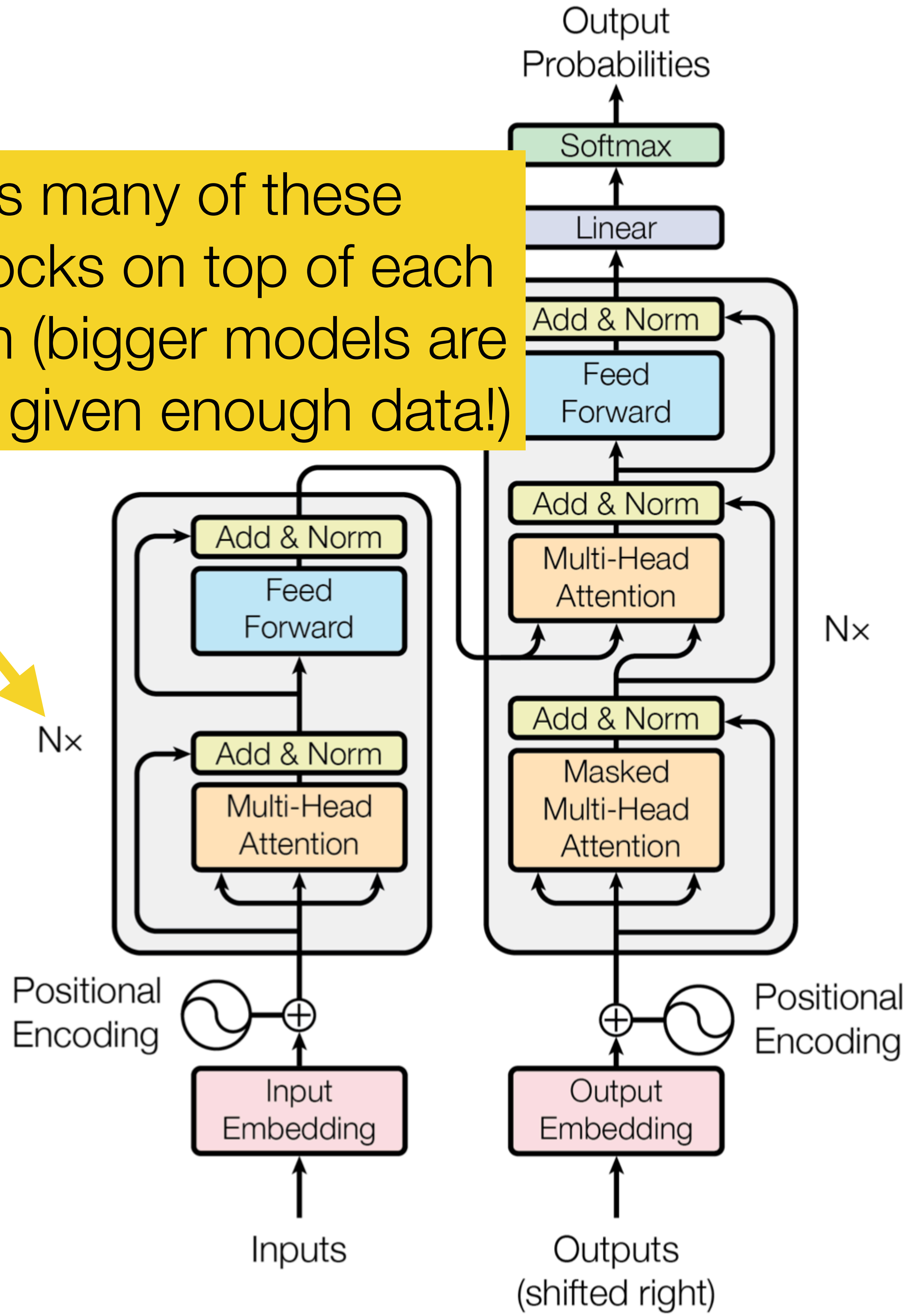
Position embeddings are *added* to each word embedding. Otherwise, since we have no recurrence, our model is unaware of the position of a word in the sequence!



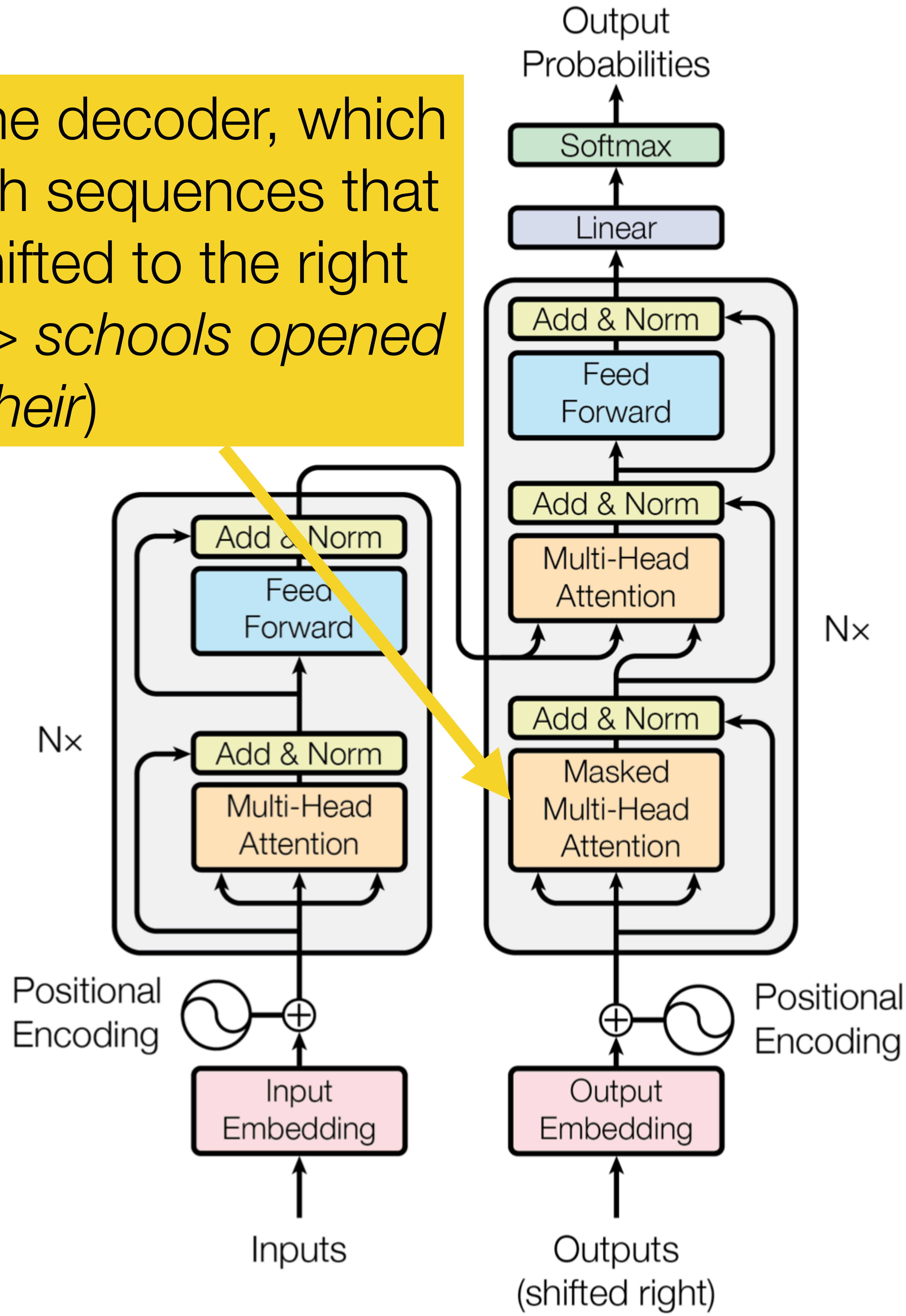
A feed-forward layer on top of the attention-weighted averaged value vectors allows us to add more parameters / nonlinearity



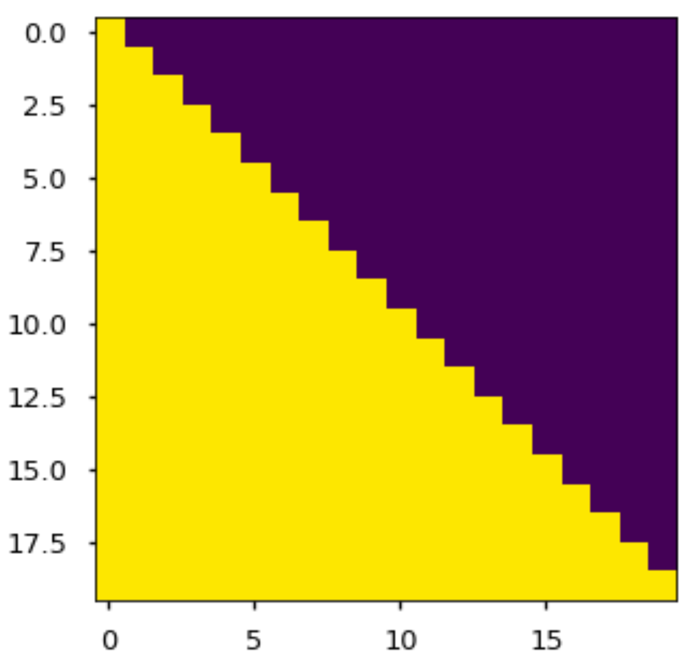
We stack as many of these *Transformer* blocks on top of each other as we can (bigger models are generally better given enough data!)

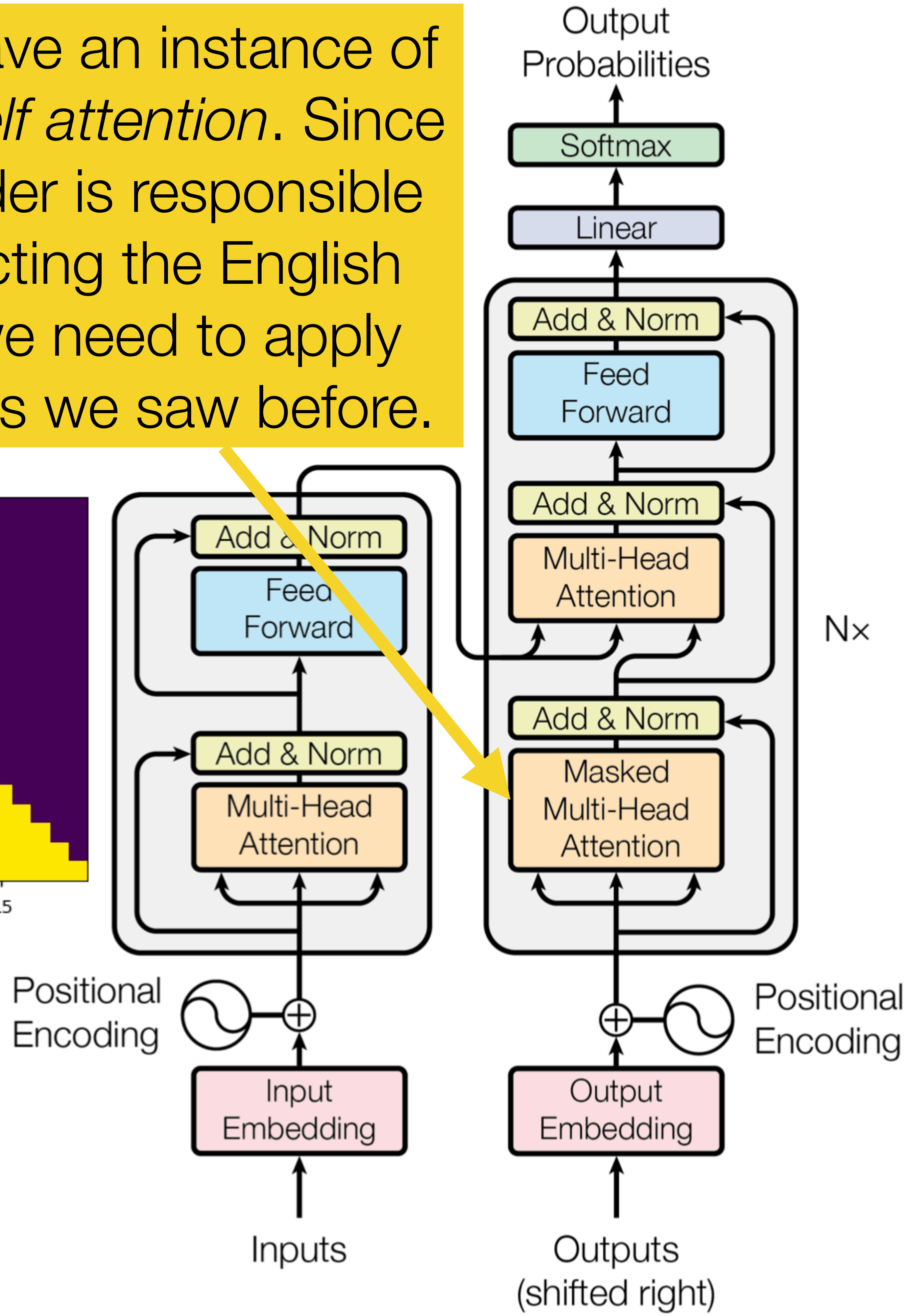


Moving onto the decoder, which takes in English sequences that have been shifted to the right (e.g., *<START> schools opened their*)

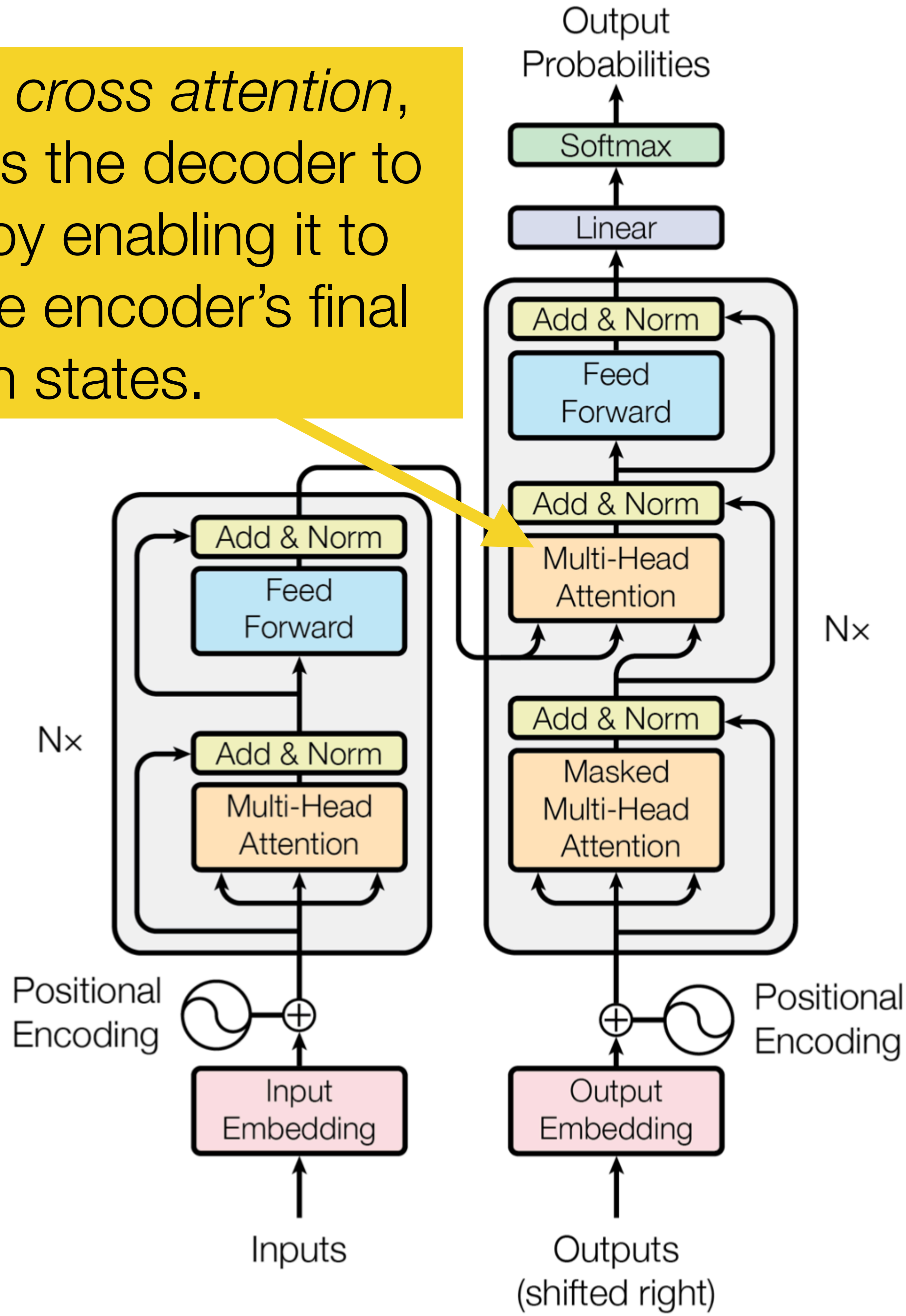


We first have an instance of *masked self attention*. Since the decoder is responsible for predicting the English words, we need to apply masking as we saw before.

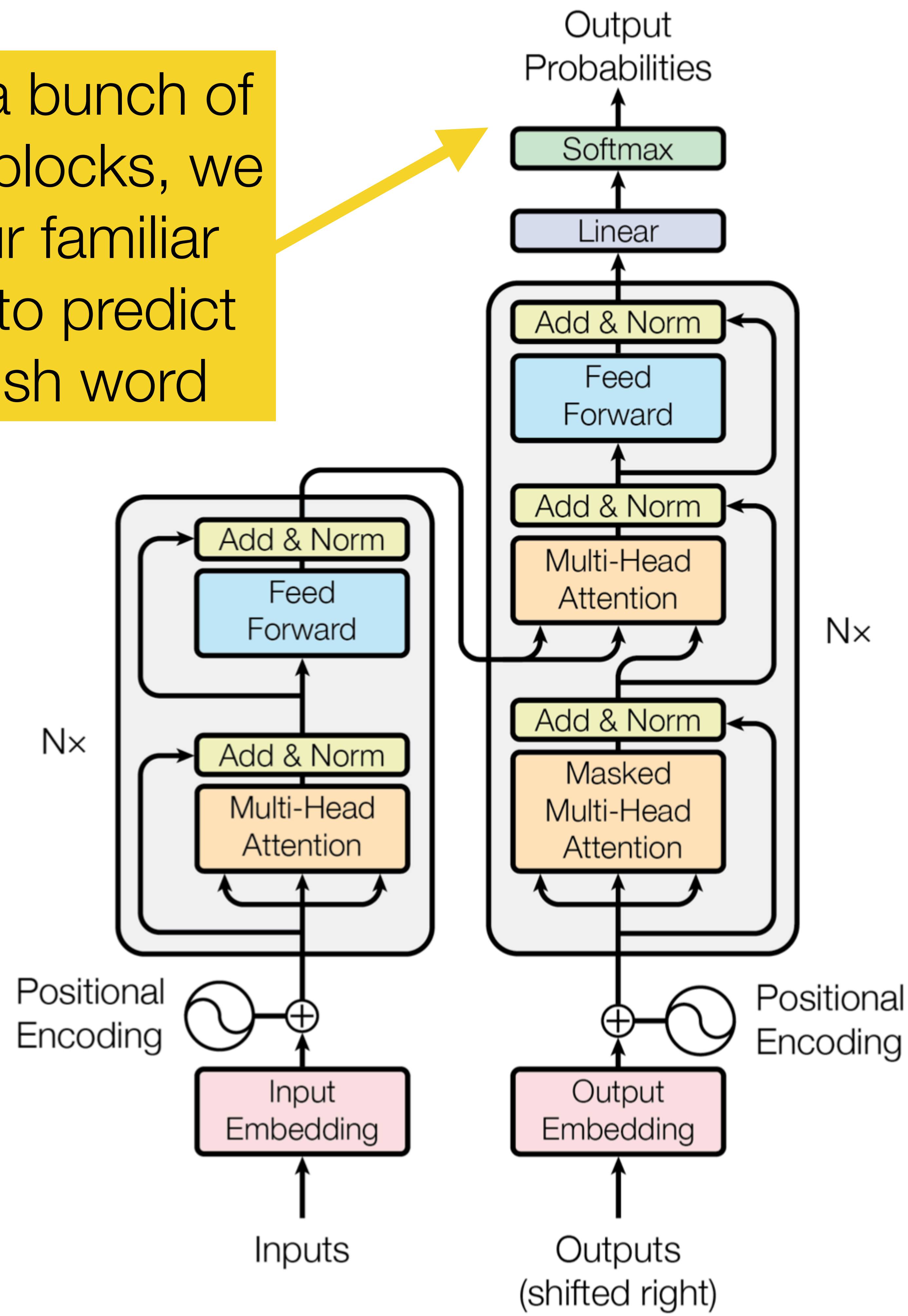




Now, we have *cross attention*, which connects the decoder to the encoder by enabling it to attend over the encoder's final hidden states.



After stacking a bunch of these decoder blocks, we finally have our familiar Softmax layer to predict the next English word



Homework Tour