LARGE LANGUAGE MODELS

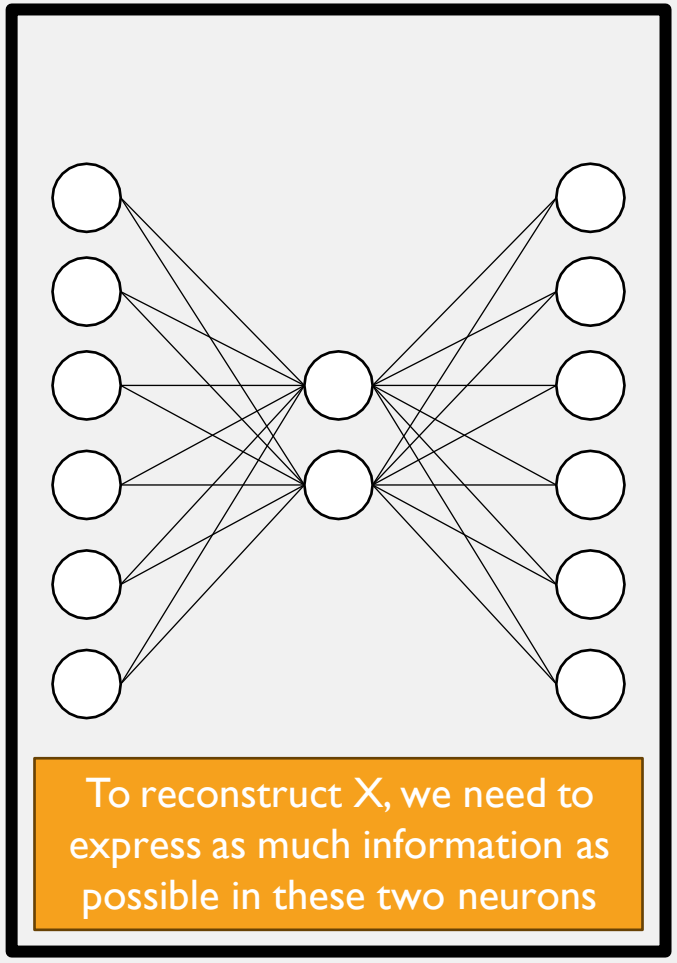
Lecture 8

MAL2, SPRING 2025

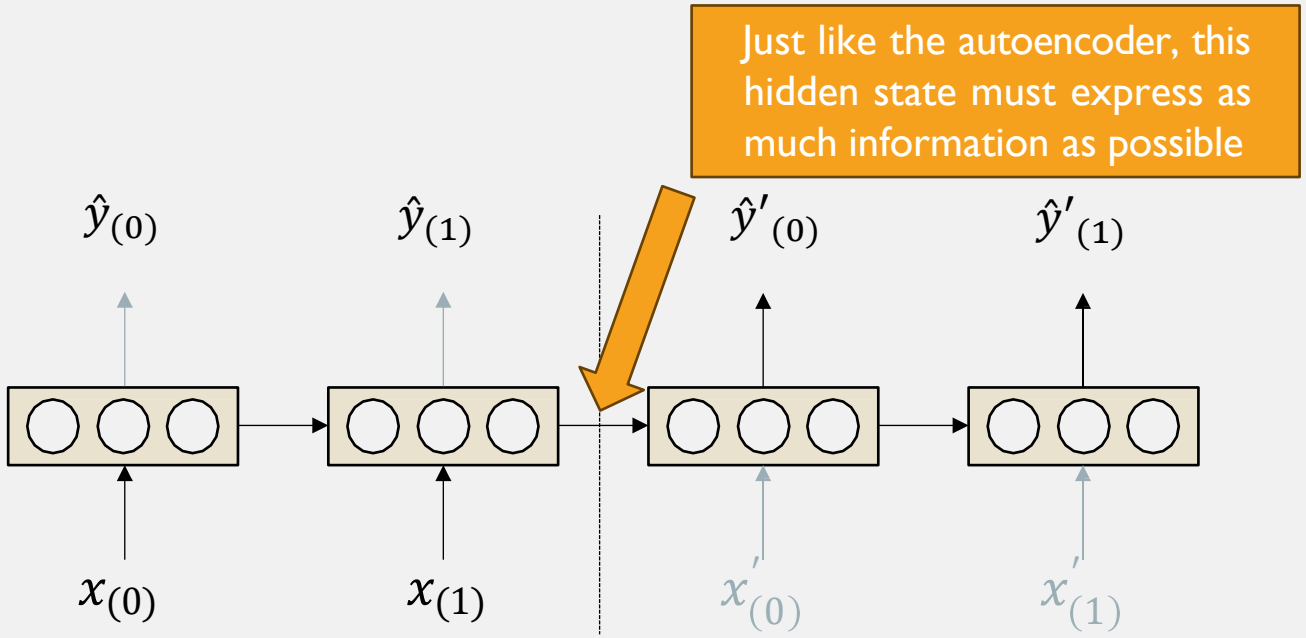
LARGE LANGUAGE MODELS

- Encoder-decoder networks
- Attention
- Transformers
- The Hugging Face Transformers Library

REMEMBER AUTOENCODERS?

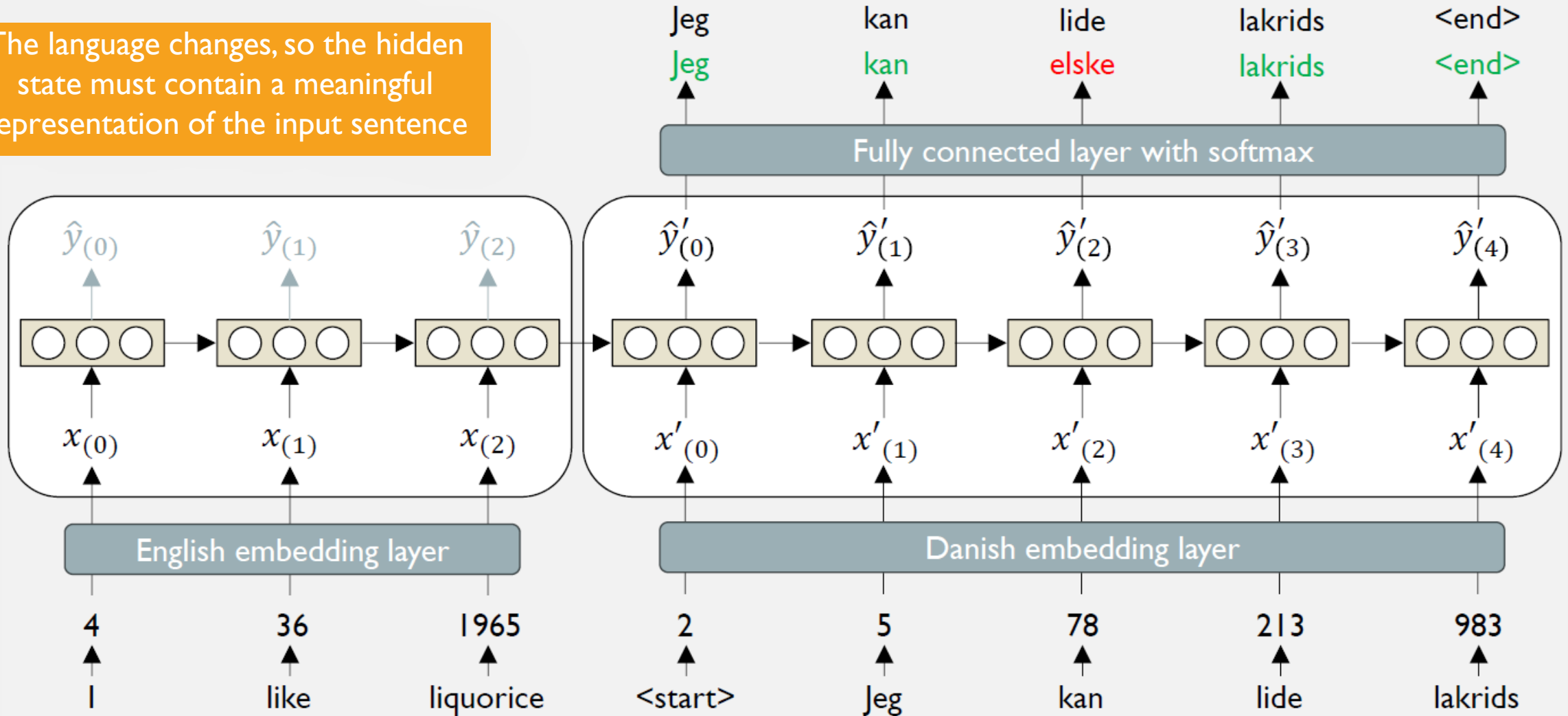


To reconstruct X, we need to express as much information as possible in these two neurons



ENCODER-DECODER NETWORKS FOR TRANSLATION

The language changes, so the hidden state must contain a meaningful representation of the input sentence



PROBLEMS WITH THIS APPROACH?

PROBLEMS WITH THIS APPROACH?

PROBLEMS WITH THIS APPROACH?

PROBLEMS WITH THIS APPROACH?

1. **The Causal Nature:**
2. Information Bottleneck:
3. **The Vanishing Gradient Problem:**
4. **Sequential Processing Limitation:**
   - RNNs process tokens one at a time, making parallelization difficult
   - Each token must wait for all previous tokens to be processed
   - This creates a computational bottleneck that limits training efficiency

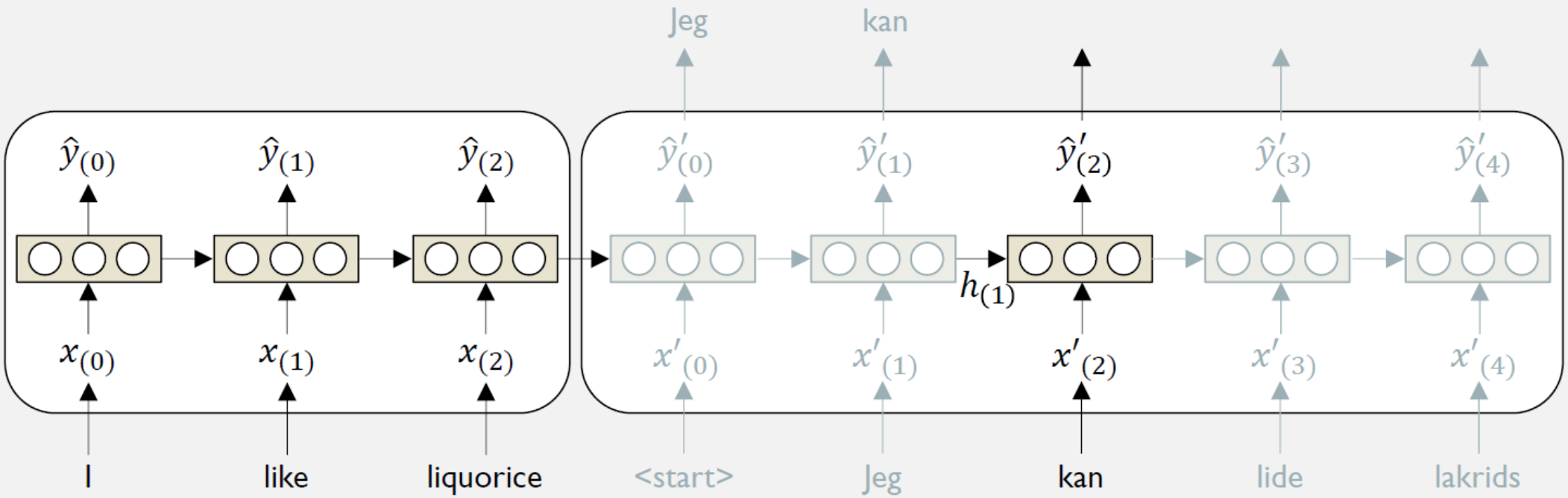
Many solutions have been proposed, but the one that beat them all was ***attention***

LARGE LANGUAGE MODELS

- Encoder-decoder networks
- Attention
- Transformers
- The Hugging Face Transformers Library

We modify the encoder-decoder network to ***pay attention to*** certain parts of the input

ATTENTION



Intuition Behind Self-Attention

Attending to the most important parts of an input.

1. Identify which parts to attend to
2. Extract the features with high attention.

Similar to a search problem!

A Simple Example: Search

YouTube

Søg

11HRS of 4K Turtle Paradise - Undersea Nature Relaxation Film + Meditation Music by Jason Stephenson

15 mio. visninger • for 4 år siden

Nature Relaxation Films

... above behind and dive into a vibrant undersea world inhabited by the iconic **sea turtles**, for 11 full hours of Nature Relaxation.

4K

MIT Introduction to Deep Learning | 6.S191

219.479 visninger • for 4 uger siden

Alexander Amini

MIT **Introduction** to **Deep Learning** 6.S191: Lecture 1 *New 2025 Edition* Foundations of **Deep Learning** Lecturer: Alexander ...

4K

Messi Bicycle Kick Goal 4k

286.662 visninger • for 2 år siden

Jakob Clips8

Recent video Neymar Skills and **Goals** 2022 PSG | Billie Eilish | Armani White | 1080p 60fps https://youtu.be/Do-fg7Pvlkg ...

TAKING A STEP BACK

- ”The hedgehog, tired as it was, went to sleep”
- ”The hedgehog had been looking for food to eat all night, so it went to sleep”
- ”The hedgehog is a magnificent animal, and after a long night of foraging and hiding from any nocturnal predator, it went to sleep”

ATTENTION IS ALL YOU NEED

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

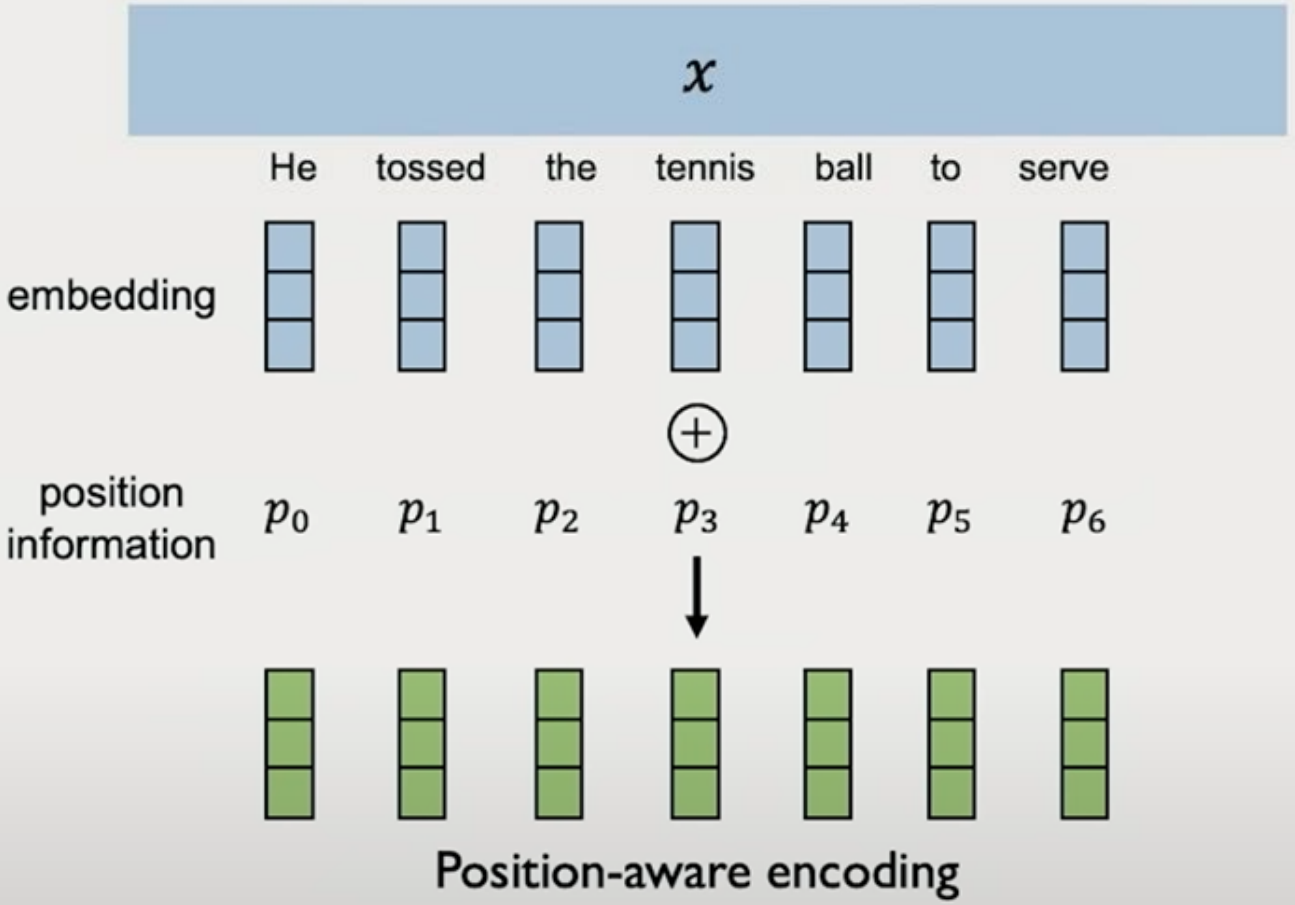
LARGE LANGUAGE MODELS

- Encoder-decoder networks
- Attention
- Transformers
- The Hugging Face Transformers Library

Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract **features with high attention**

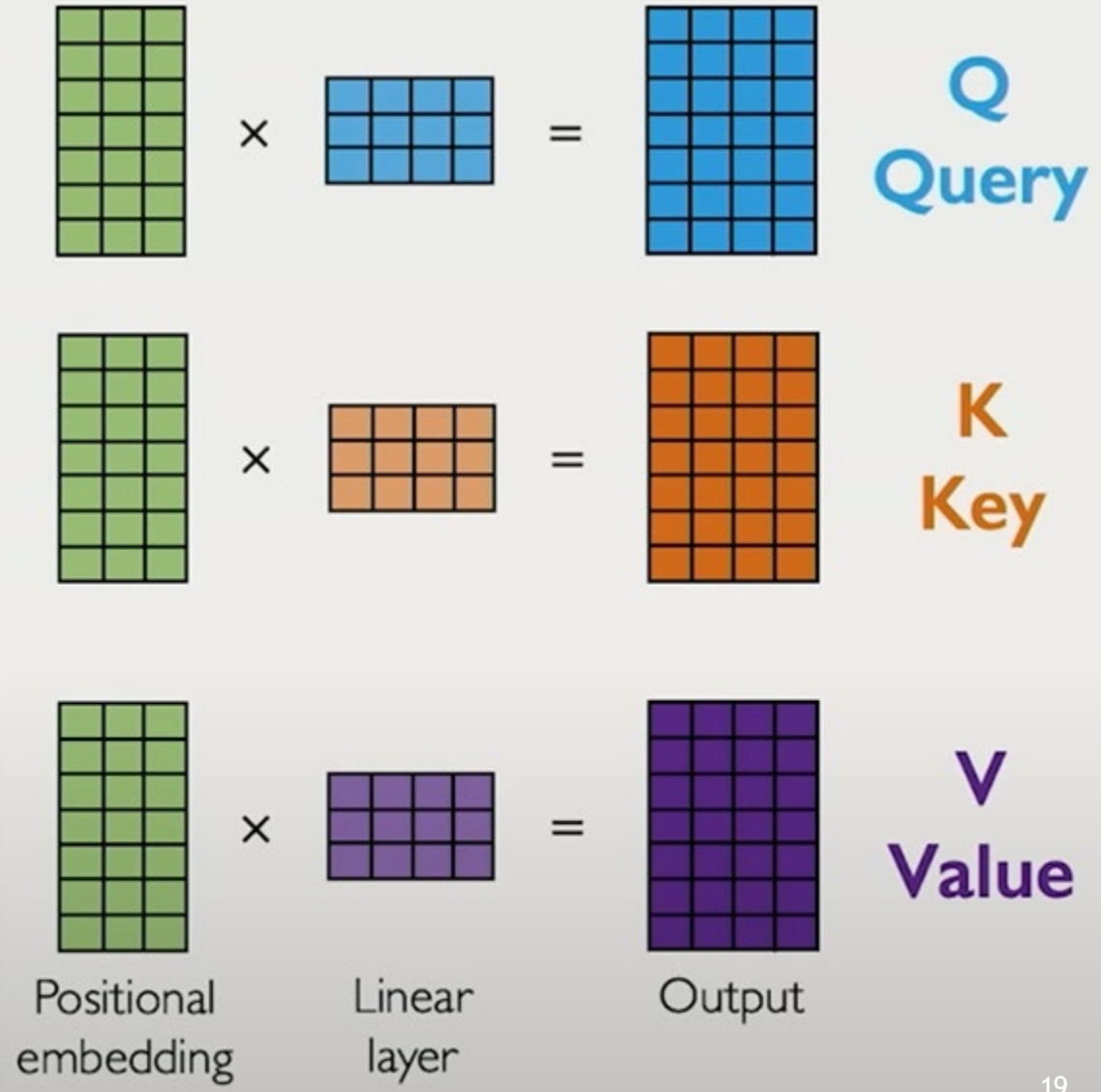


Data is fed in all at once! Need to encode position information to understand order.

Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute attention weighting
4. Extract features with high attention



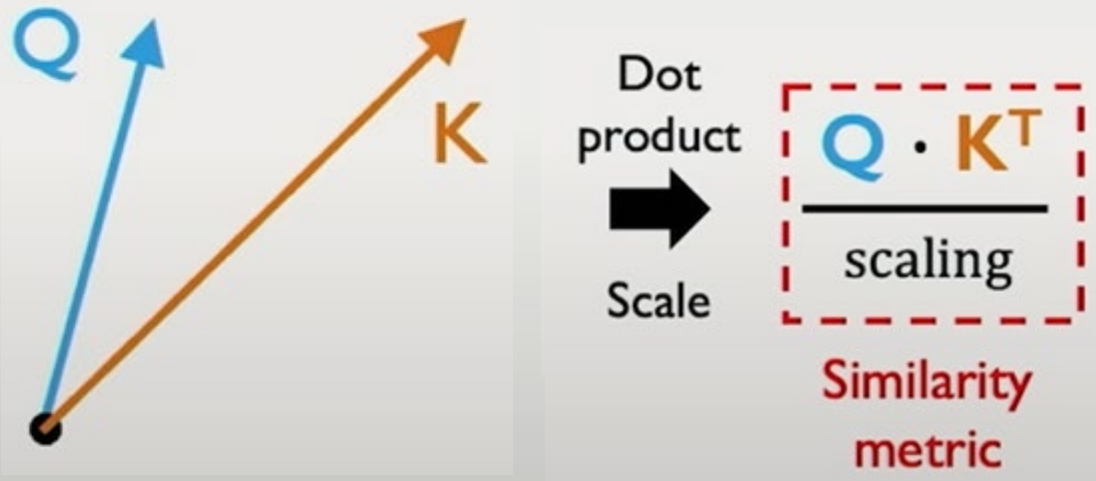
Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract features with high attention

Attention score: compute pairwise similarity between each query and key

How to compute similarity between two sets of features?



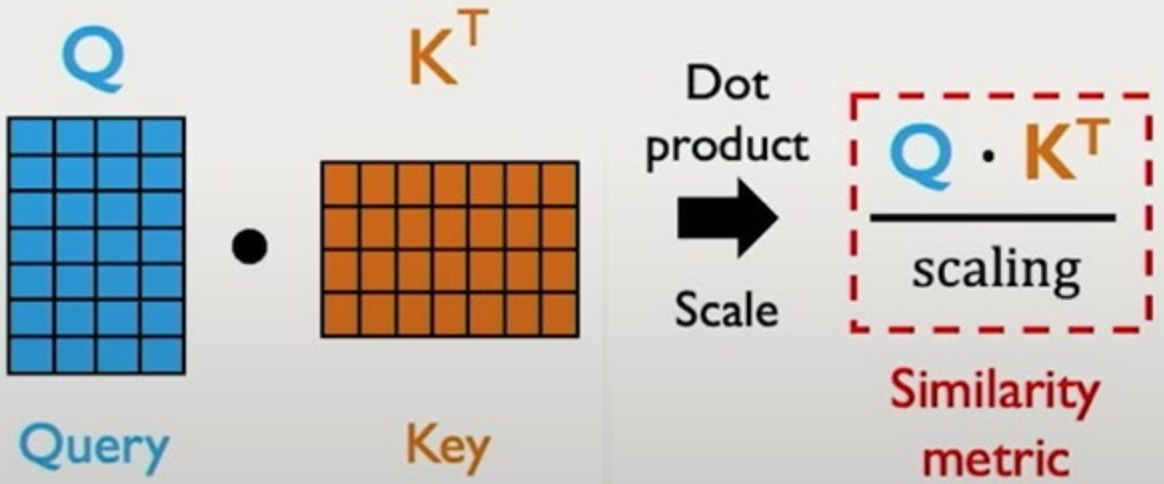
Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract features with high attention

Attention score: compute pairwise similarity between each query and key

How to compute similarity between two sets of features?

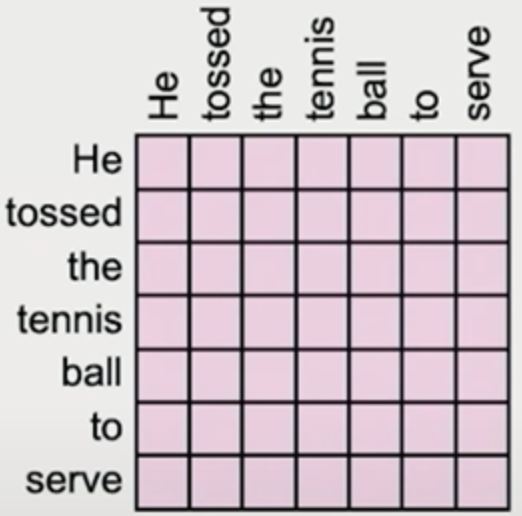


Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract features with high attention

Attention weighting: where to attend to! How similar is the key to the query?

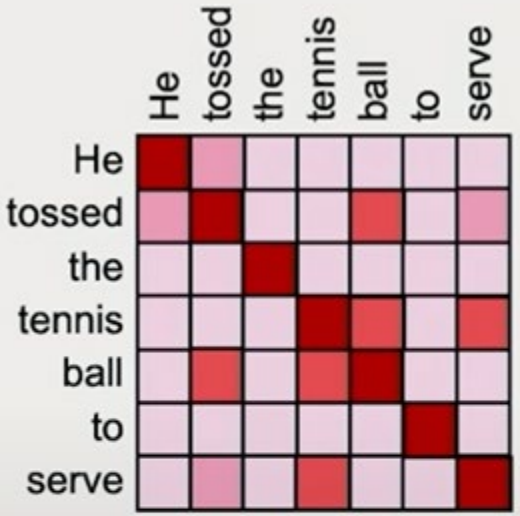


Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract features with high attention

Attention weighting: where to attend to! How similar is the key to the query?



$$softmax\left(\frac{Q \cdot K^T}{scaling}\right)$$

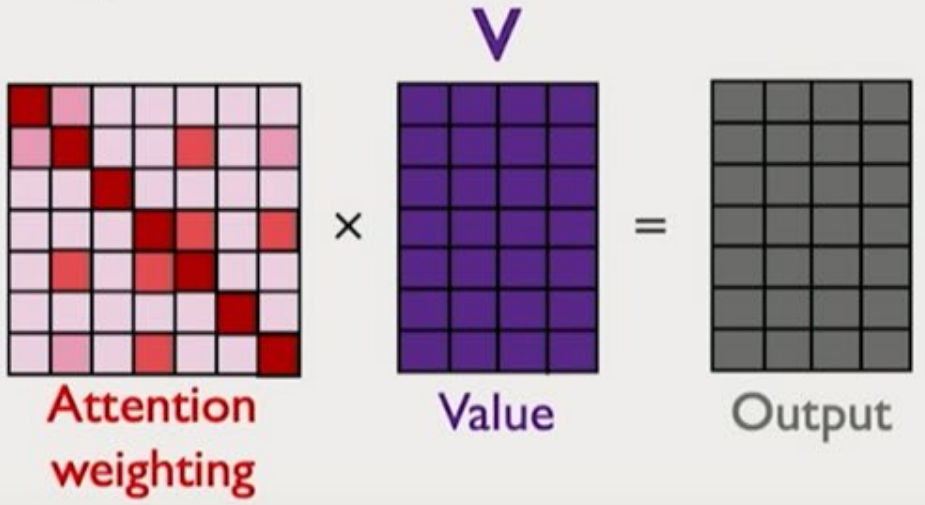
Attention weighting

Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract **features with high attention**

Last step: self-attend to extract features

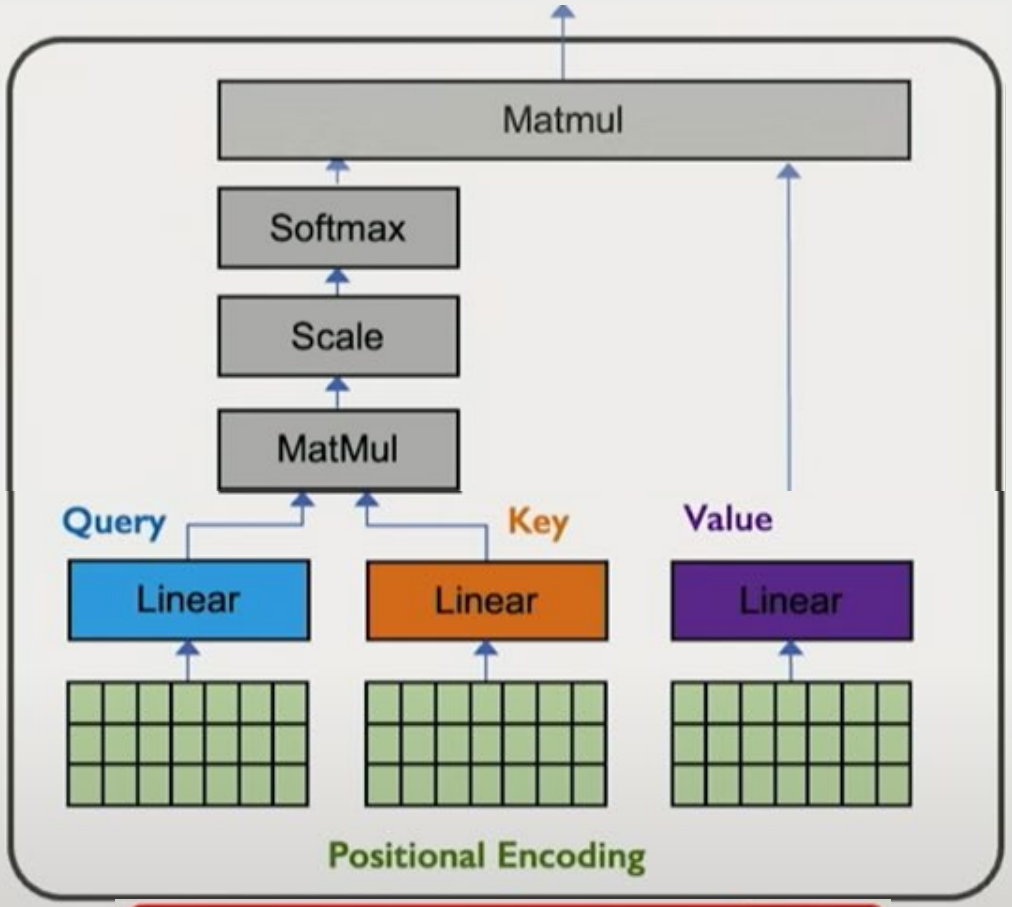


$$softmax\left(\frac{Q \cdot K^T}{scaling}\right) \cdot V = A(Q,K,V)$$

Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

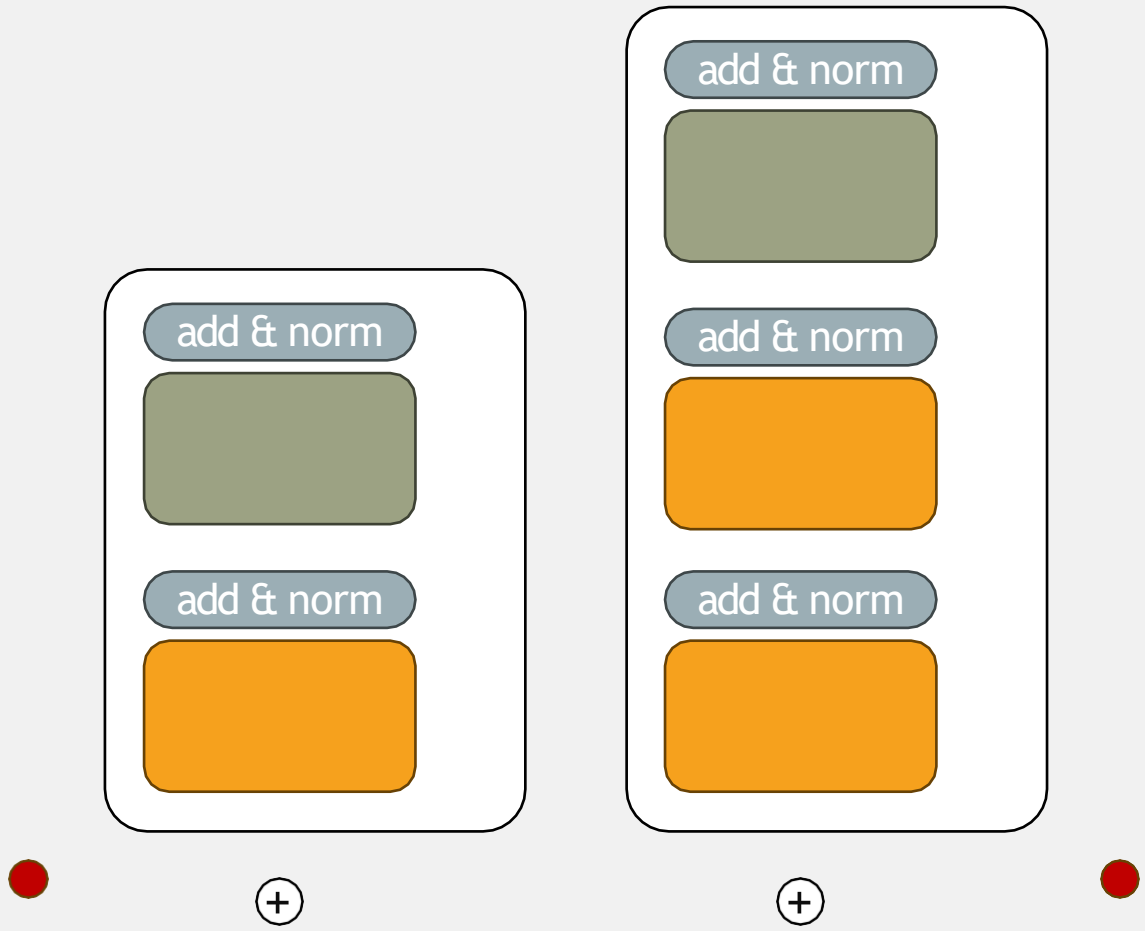
1. Encode **position** information
2. Extract **query, key, value** for search
3. Compute **attention weighting**
4. Extract **features with high attention**



$$softmax\left(\frac{Q \cdot K^T}{scaling}\right) \cdot V$$

These operations form a self-attention head that can plug into a larger network.
Each head attends to a different part of input.

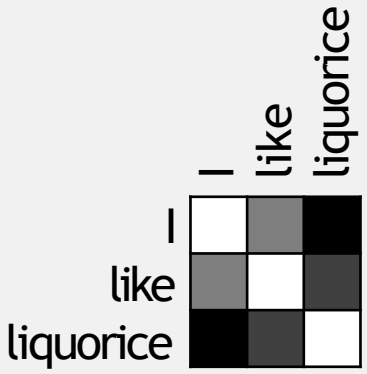
TRANSFORMERS



THREE TYPES OF ATTENTION

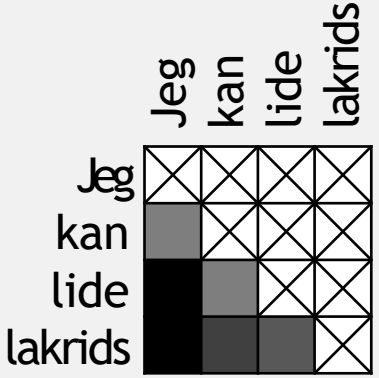
Self-attention in the encoder

Learns relationships between input tokens



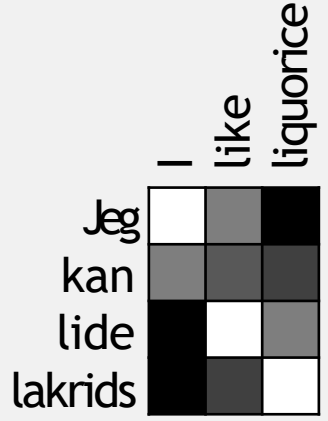
Masked self-attention in the decoder

Learns relationships between output tokens - but only those we've seen so far

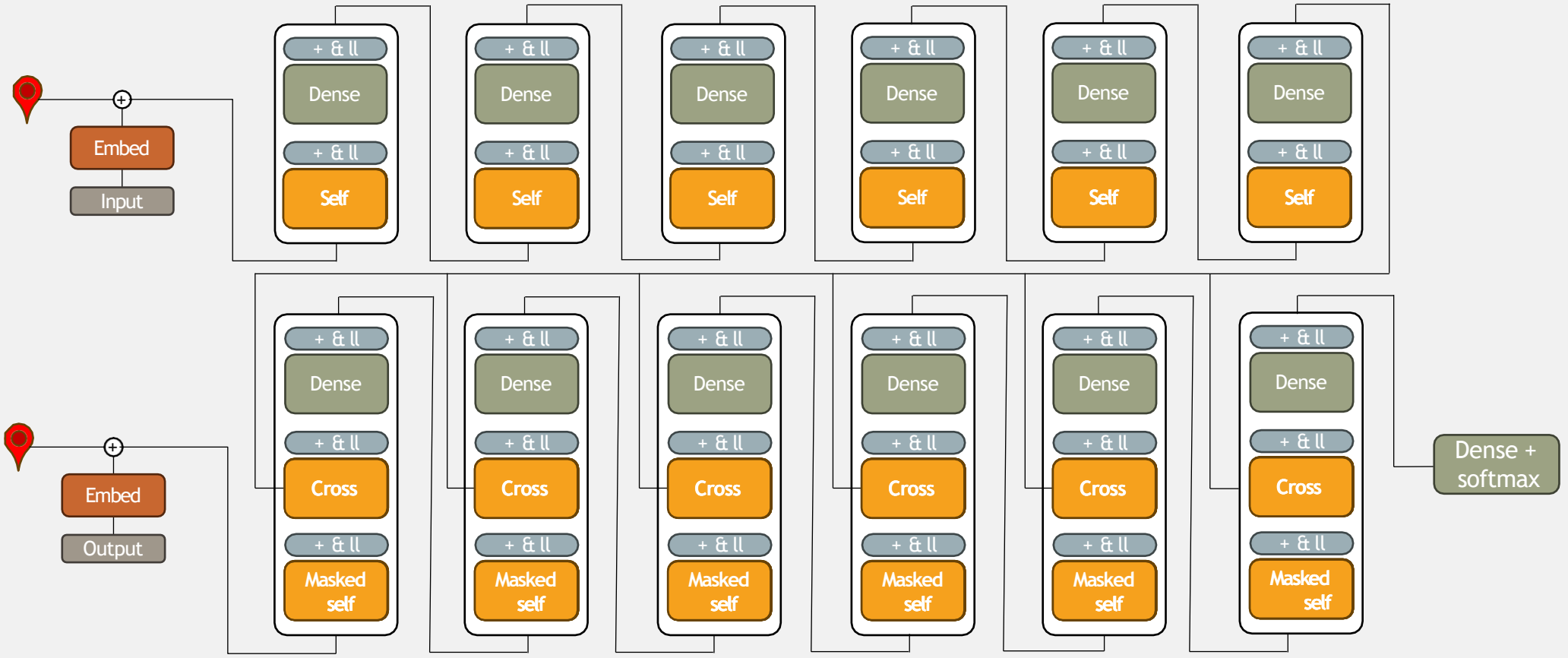


Cross-attention in the decoder

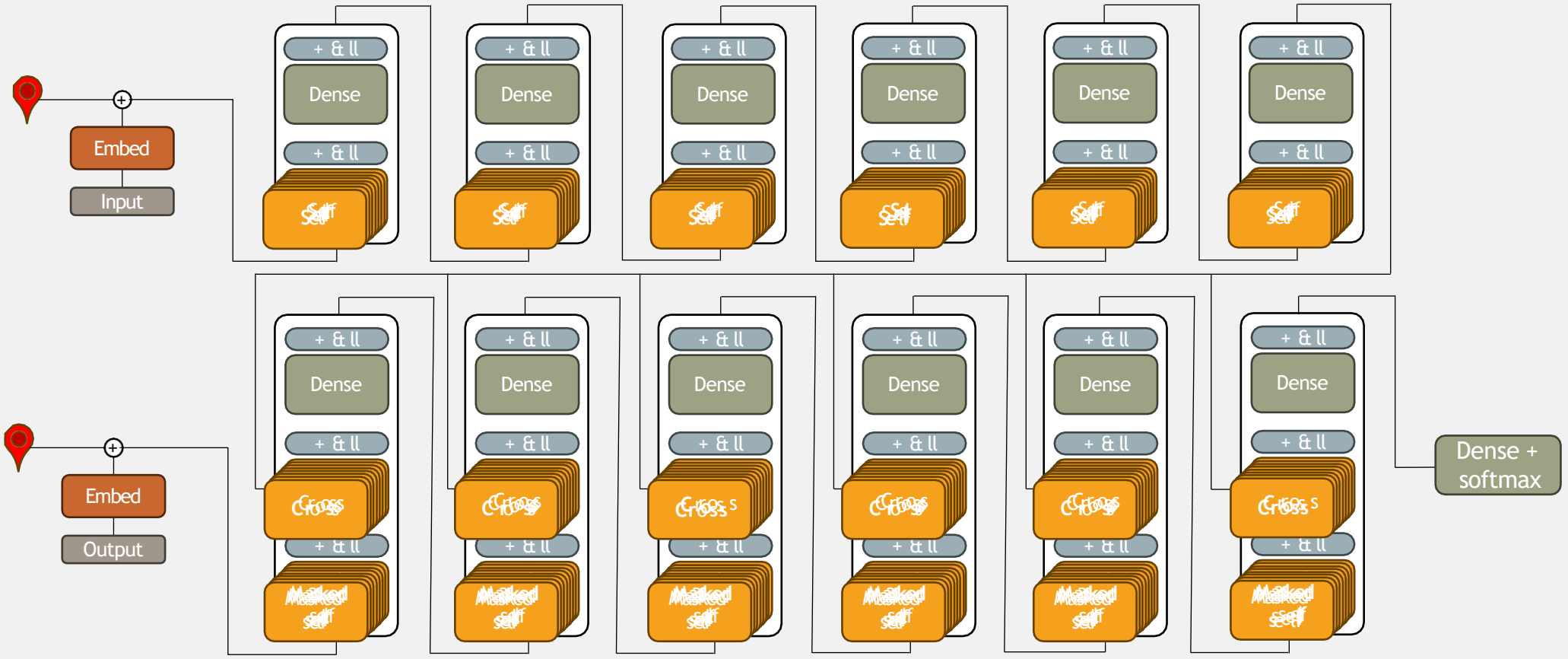
Learns which part of input relevant for current output



THE ORIGINAL TRANSFORMER ARCHITECHTURE



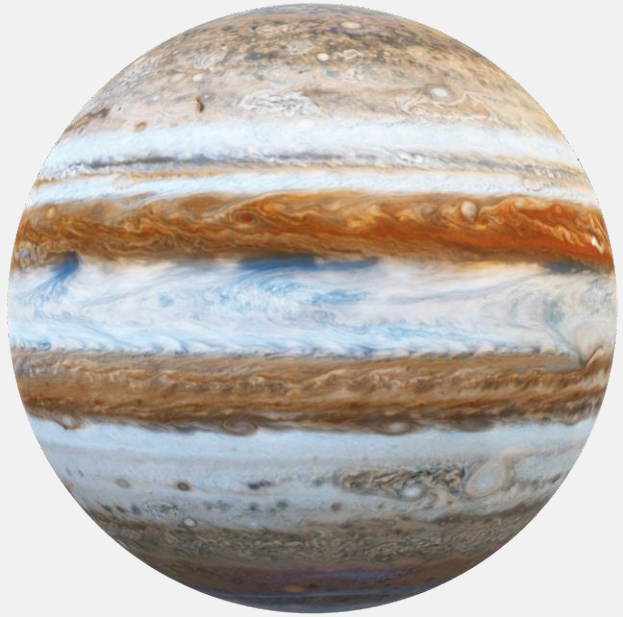
THE ORIGINAL TRANSFORMER ARCHITECHTURE



LARGE LANGUAGE MODELS

- Encoder-decoder networks
- Attention
- Transformers
- The Hugging Face Transformers Library

HUGGING FACE



HUGGING FACE

Explore the transformers library

Go to huggingface.co/docs/transformers and find a tutorial or task you want to experiment with

You have 20 minutes