

Class core values

1. Be **respectful** to yourself and others
2. Be **confident** and believe in yourself
3. Always do your **best**
4. Be **cooperative**
5. Be **creative**
6. Have **fun**
7. Be **patient** with yourself while you learn
8. Don't be shy to **ask "stupid" questions**
9. Be **inclusive** and **accepting**

Three Transformers robots are shown in the background. On the left is Bumblebee, a yellow and black robot. In the center is Optimus Prime, a blue and red robot. On the right is Megatron, a large, dark, and menacing robot. They are all standing on a white surface.

Week 6, Lecture 1

Learning on sequences: Transformers

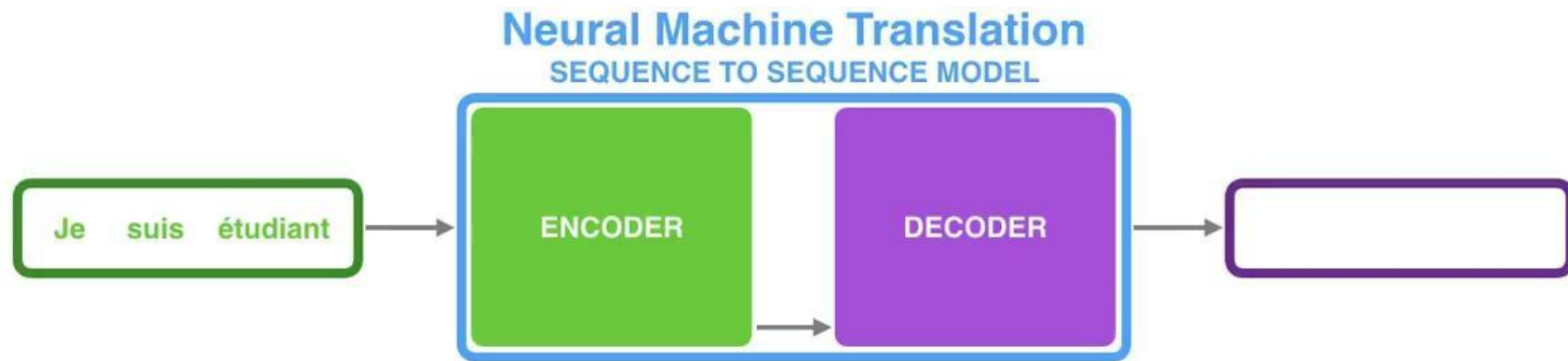
Learning Objectives

1. Describe the basic concept of a transformer unit
2. Explain positional embedding and its application
3. Explain attention matrix and the difference between its variant
4. Describe decoding and encoding
5. Explain the inferring process in transformers

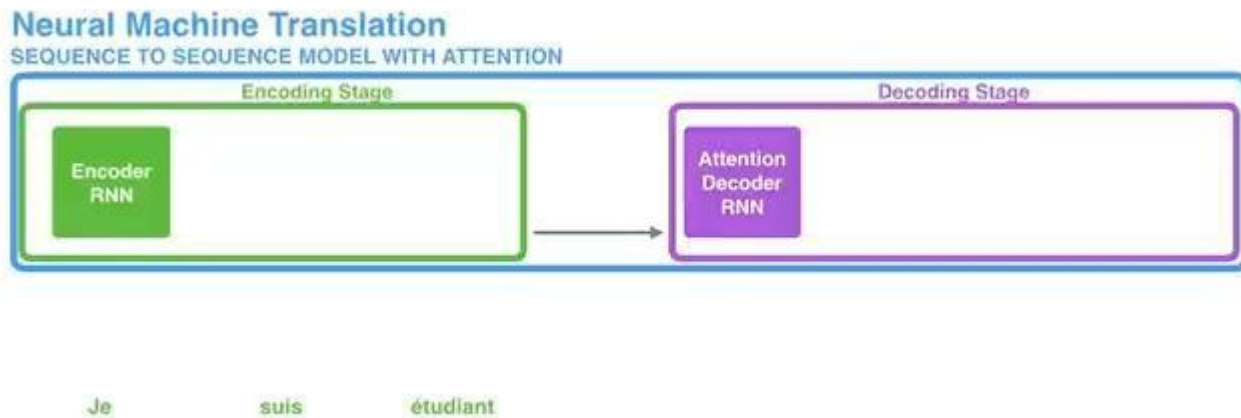
Transformers were first developed for Seq2Seq tasks



Seq2Seq models use a simple model of encoding and decoding



Attention helps seq2seq models to remember the context



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

Attention is all you need!

Task: Hotel location

you get what you pay for . not the **cleanest rooms** but bed was **clean** and so was bathroom . bring your own towels though as very **thin** . service was **excellent** , let us book in at 8:30am ! for **location and price** , **this ca n't be beaten** , but it is **cheap** for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel cleanliness

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own **towels** though as very **thin** . service was **excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is **cheap** for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

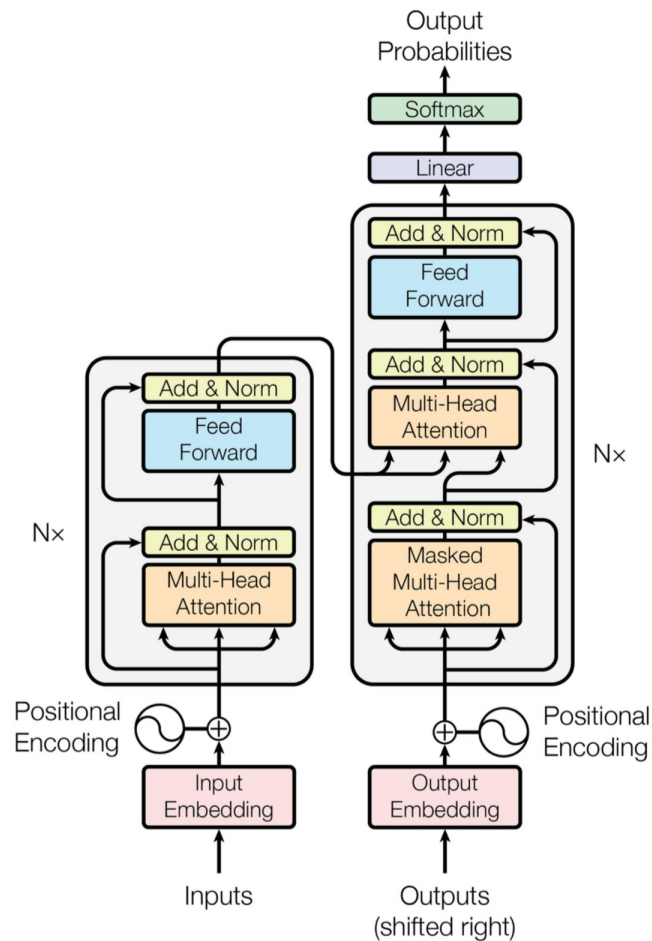
Task: Hotel service

you get what you pay for . not the cleanest rooms but bed was **clean** and so was bathroom . bring your own **towels** though as very **thin** . **service was excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is **cheap** for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

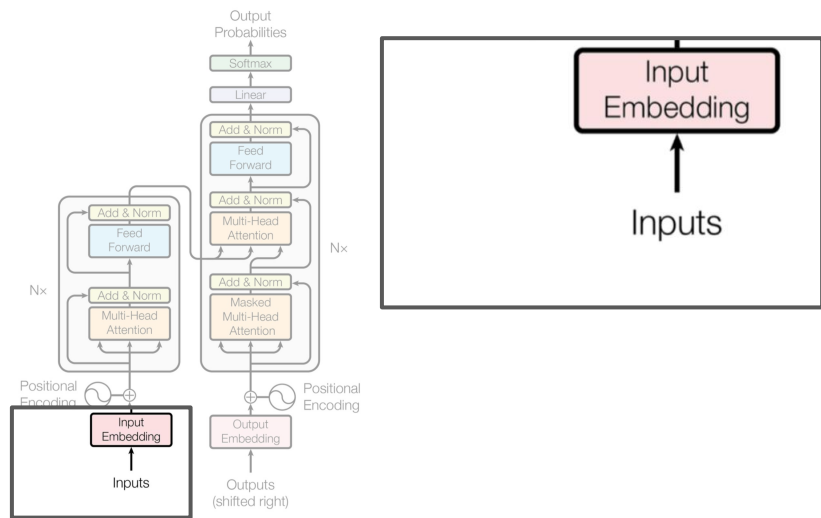
What it can do ...

<https://transformer.huggingface.co/doc/distil-gpt2>

Basic architecture of a transformer



The second step is word embedding



original
text

"hello world!"

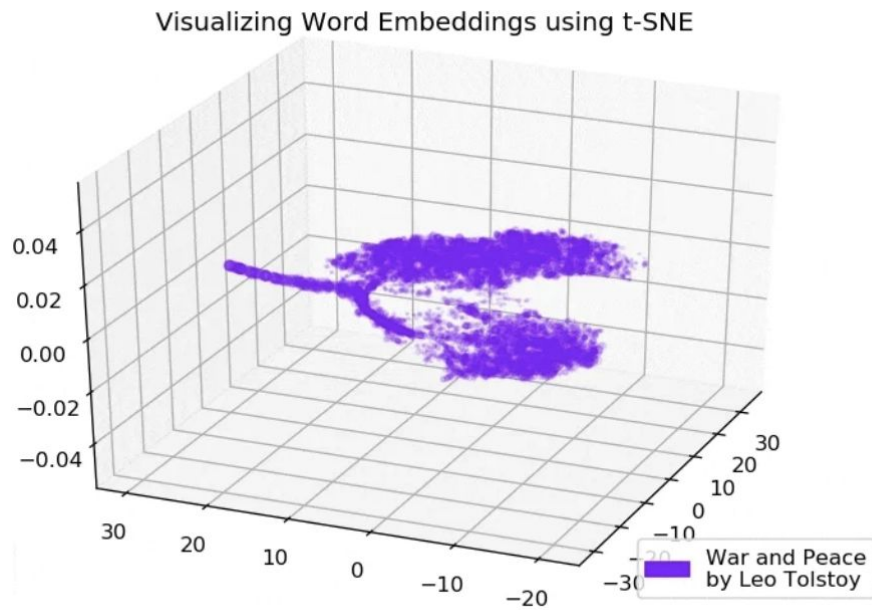
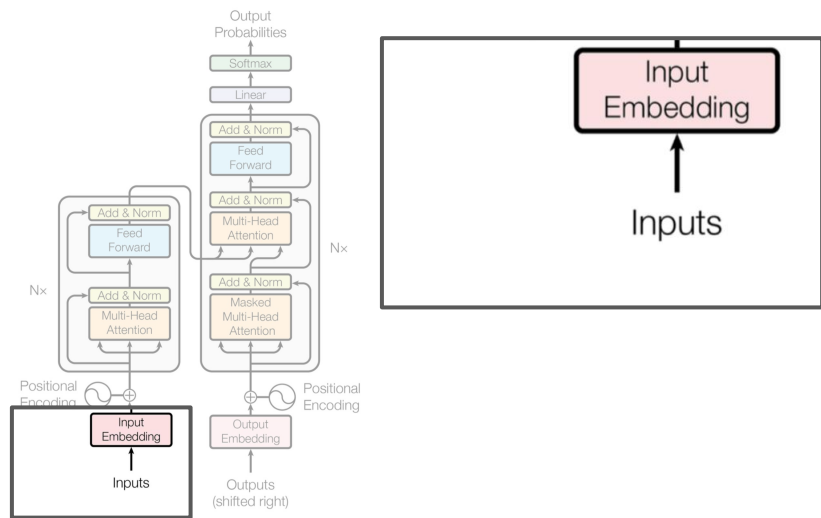
tokens

['hello', 'world', '!']

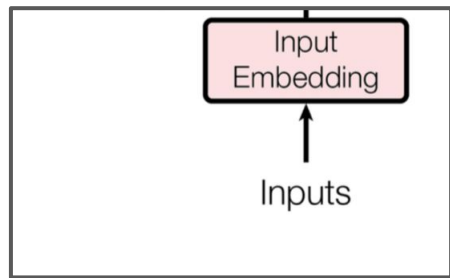
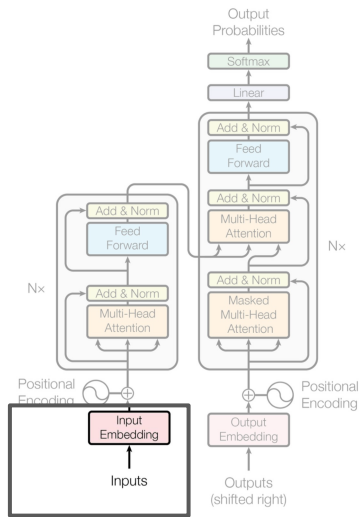
token
IDs

[7592, 2088, 999]

The first step of the process is word embedding



During the tokenization and embedding, we lose the order of words



original
text

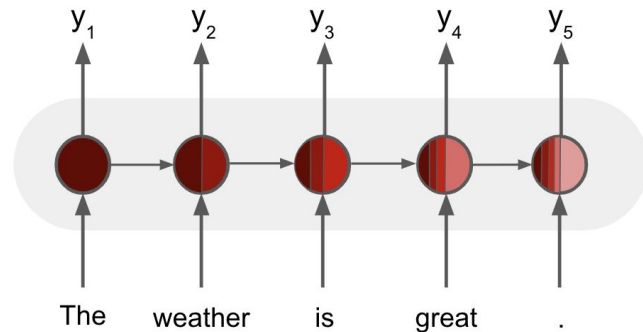
"hello world!"

tokens

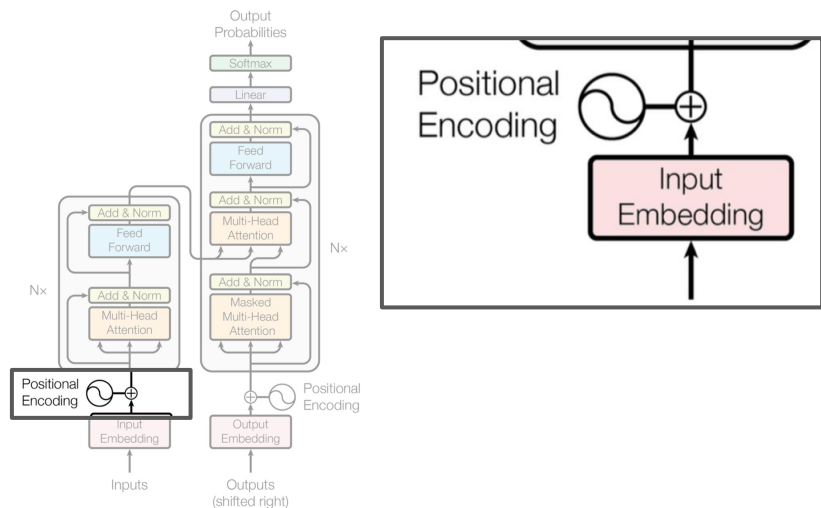
['hello', 'world', '!']

token
IDs

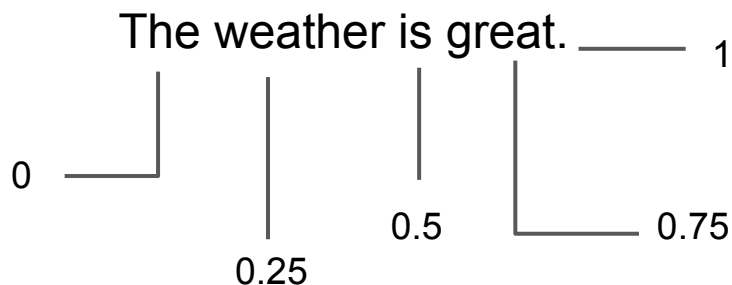
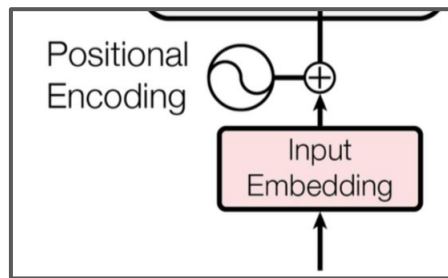
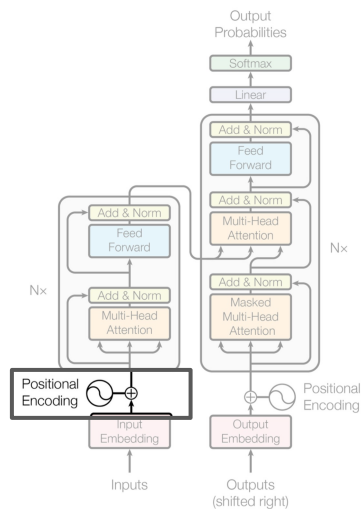
[7592, 2088, 999]



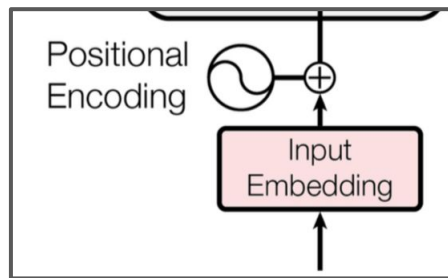
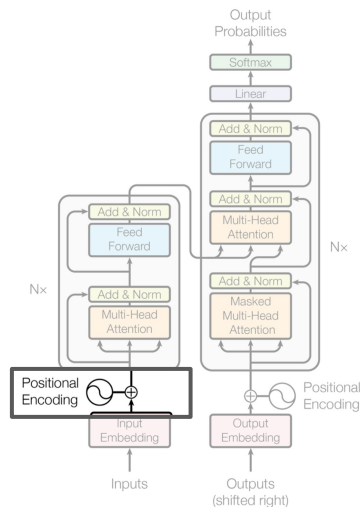
Positional encoding allows for some notion of order



Positional encoding allows for some notion of order



A good positional encoding should have consistent distance



The weather is great.

0, 0.25, 0.5, 0.75, 1

The weather is really great!

0, 0.2, 0.4, 0.6, 0.8, 1

A good positional encoding should have

1. Unique stamps for each position in the sentence
2. Distances should be consistent
3. A way to easily generalize to larger sentences
4. Deterministic assignment

Sinusoidal functions are good embeddings

$$\vec{p}_t^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k \cdot t), & \text{if } i = 2k \\ \cos(\omega_k \cdot t), & \text{if } i = 2k + 1 \end{cases}$$

where

$$\omega_k = \frac{1}{10000^{2k/d}}$$

Sinusoidal functions are good embeddings

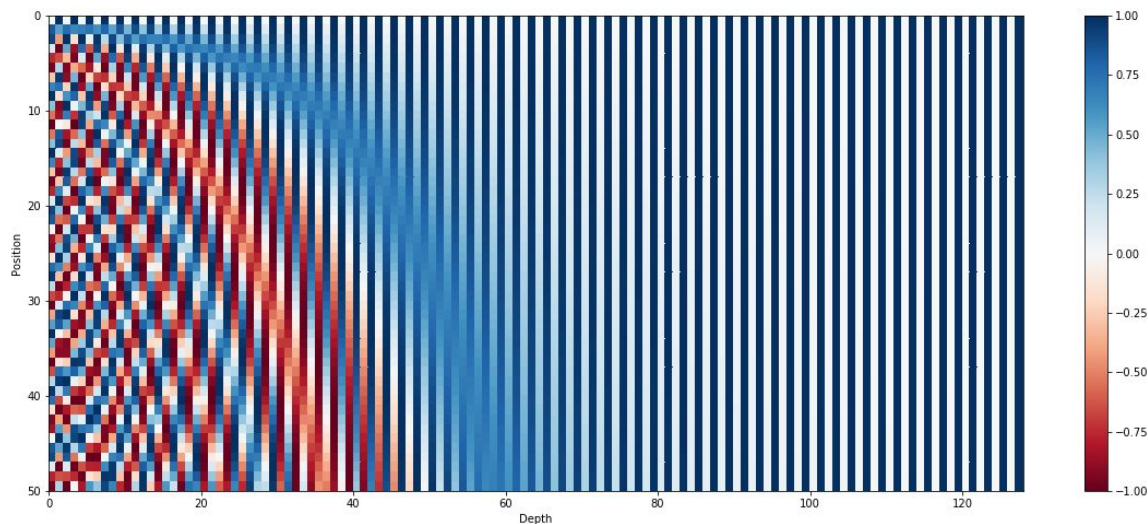
$$\vec{p}_t^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k \cdot t), & \text{if } i = 2k \\ \cos(\omega_k \cdot t), & \text{if } i = 2k + 1 \end{cases}$$

where

$$\omega_k = \frac{1}{10000^{2k/d}}$$

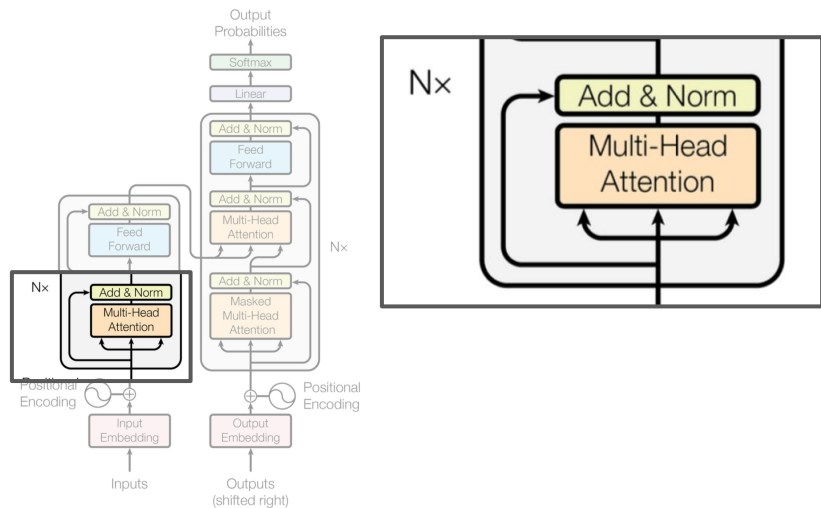
$$\vec{p}_t = \begin{bmatrix} \sin(\omega_1 \cdot t) \\ \cos(\omega_1 \cdot t) \\ \\ \sin(\omega_2 \cdot t) \\ \cos(\omega_2 \cdot t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} \cdot t) \\ \cos(\omega_{d/2} \cdot t) \end{bmatrix}_{d \times 1}$$

Sinusoidal functions are good embeddings



$$\vec{p}_t = \begin{bmatrix} \sin(\omega_1 \cdot t) \\ \cos(\omega_1 \cdot t) \\ \\ \sin(\omega_2 \cdot t) \\ \cos(\omega_2 \cdot t) \\ \\ \vdots \\ \sin(\omega_{d/2} \cdot t) \\ \cos(\omega_{d/2} \cdot t) \end{bmatrix}_{d \times 1}$$

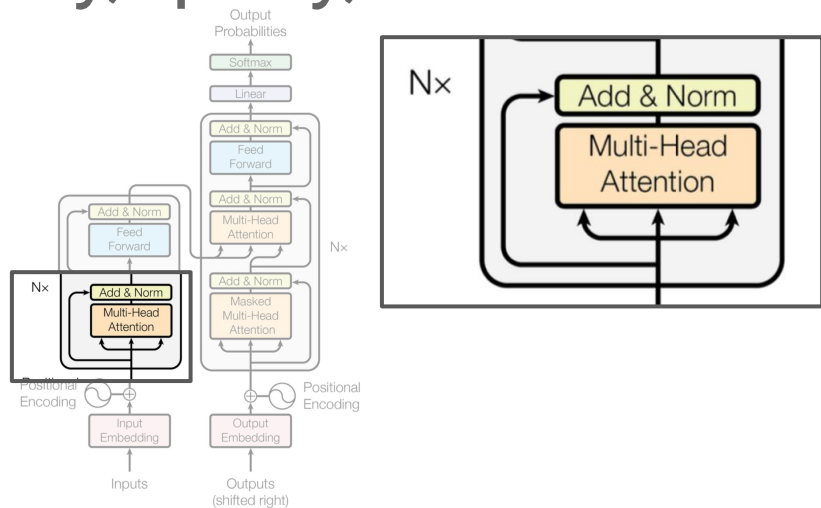
The main module of a transformer is an *attention* unit



Self-attention allows for finding correlation between words

	the	weather	is	great
the	0.8	0.1	0.05	0.05
weather	0.1	0.6	0.2	0.1
is	0.05	0.2	0.65	0.1
great	0.2	0.1	0.1	0.6

To understand attention, we need to understand key, query, and value



Feature-based attentions

The image shows the PubMed.gov search interface. It features a dark blue header with the PubMed.gov logo on the left. Below the logo is a white search bar containing the text "transformer protein design 2020". To the right of the search bar is a green button with the word "Search" in white. Below the search bar, the word "Advanced" is visible. The background of the header has a faint molecular structure pattern.

PubMed.gov

transformer protein design 2020

Search

Advanced

Feature-based attentions

Query

PubMed.gov

transformer protein design 2020



Search

Advanced

Feature-based attentions



Feature-based attentions

6 results

« < Page 1 of 1 > »

☐ 1 [Wrangling Shape-Shifting Morpheeins to Tackle Disease and Approach Drug Discovery.](#)

Cite Jaffe EK.

Front Mol Biosci. 2020 Nov 27;7:582966. doi: 10.3389/fmolb.2020.582966. eCollection 2020.

Share

PMID: 33330623 [Free PMC article.](#) [Review.](#)

Homo-multimeric **proteins** that can come apart, change shape, and reassemble differently with functional consequences have been called morpheeins and/or **transformers**; these provide a largely unexplored context for understanding disease and developing allosteric therap ...

Values



☐ 2 [Signal Peptides Generated by Attention-Based Neural Networks.](#)

Wu Z, Yang KK, Liszka MJ, Lee A, Batzilla A, Wernick D, Weiner DP, Arnold FH.

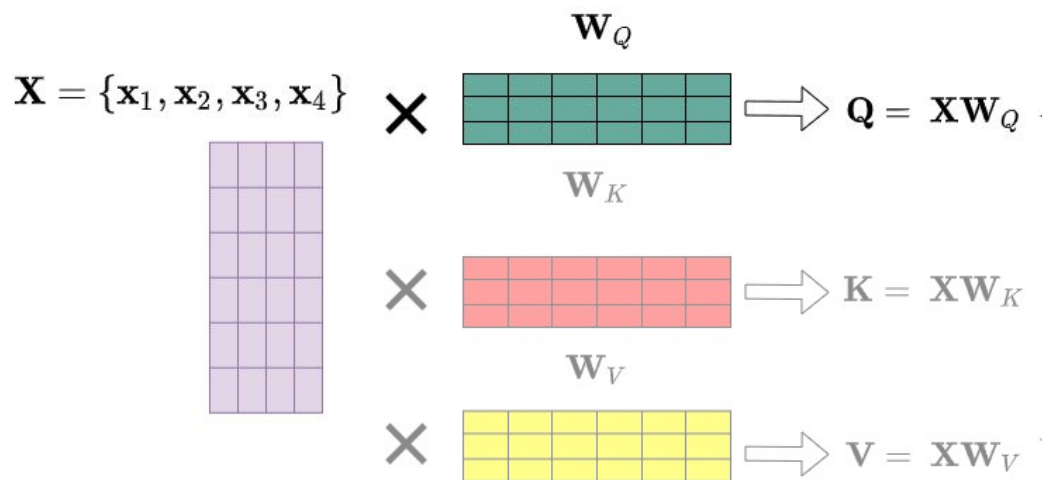
Cite ACS Synth Biol. 2020 Aug 21;9(8):2154-2161. doi: 10.1021/acssynbio.0c00219. Epub 2020 Jul 27.

Share

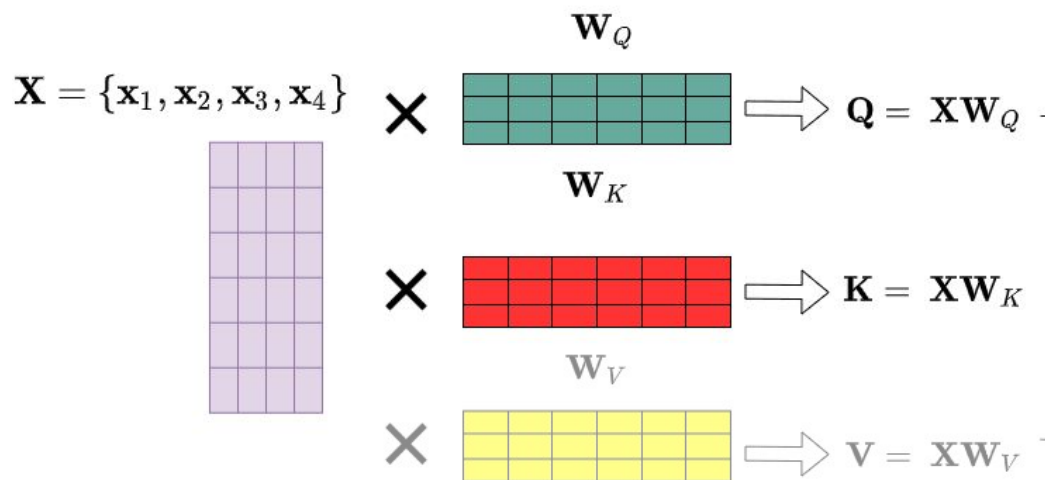
PMID: 32649182 [Free article.](#)

Short (15-30 residue) chains of amino acids at the amino termini of expressed **proteins** known as signal peptides (SPs) specify secretion in living cells. We trained an attention-based neural network, the **Transformer** model, on data from all available organisms in Swis ...

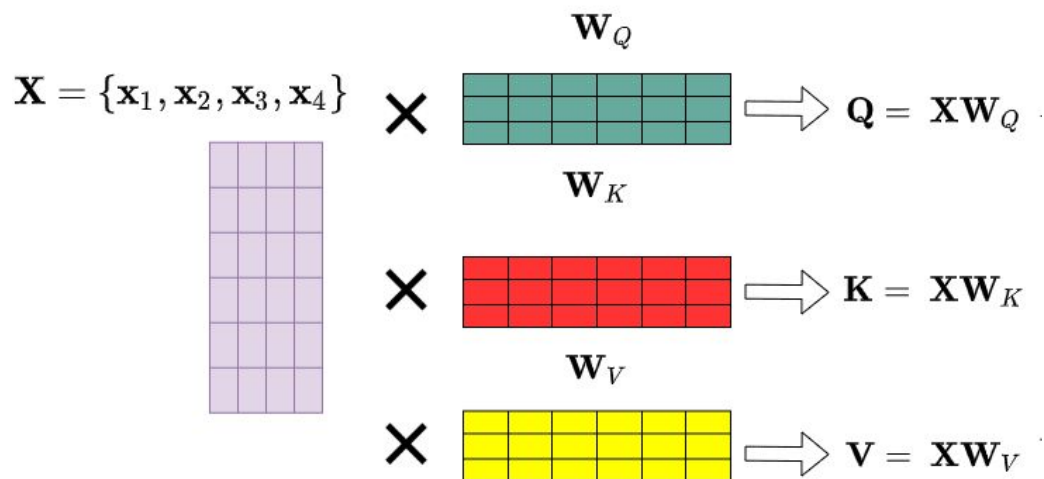
To calculate attention, query, key, and value matrices are generated



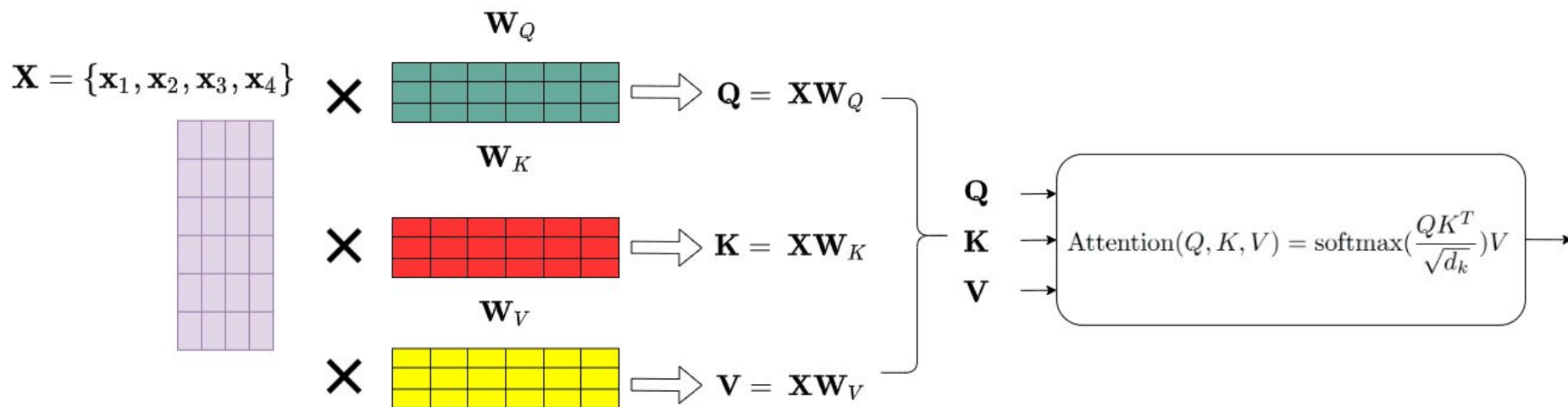
To calculate attention, query, key, and value matrices are generated



To calculate attention, query, key, and value matrices are generated



To calculate attention, query, key, and value matrices are generated



Softmax takes in values and return a number between (0,1) → great for generating probabilities or as activation for multi categories

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

σ = softmax

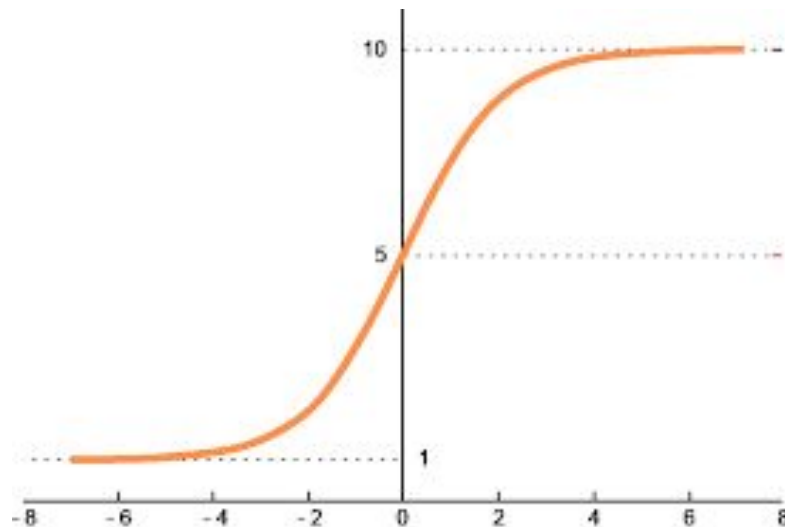
\vec{z} = input vector

e^{z_i} = standard exponential function for input vector

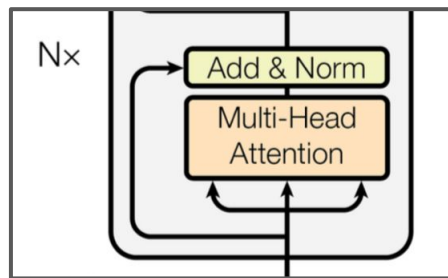
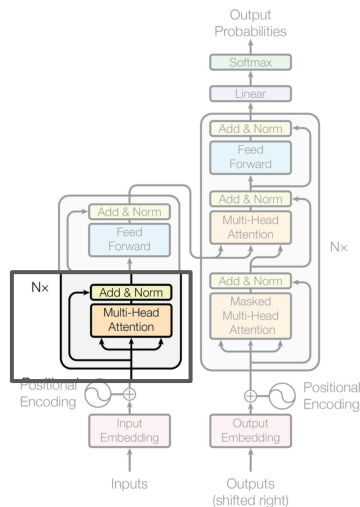
K = number of classes in the multi-class classifier

e^{z_j} = standard exponential function for output vector

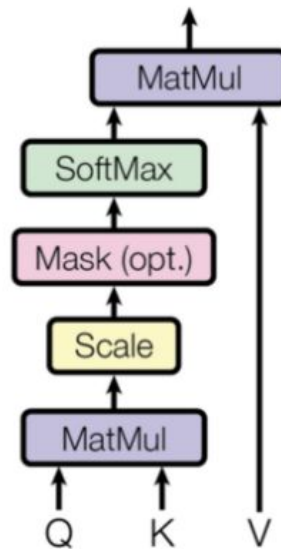
e^{z_j} = standard exponential function for output vector



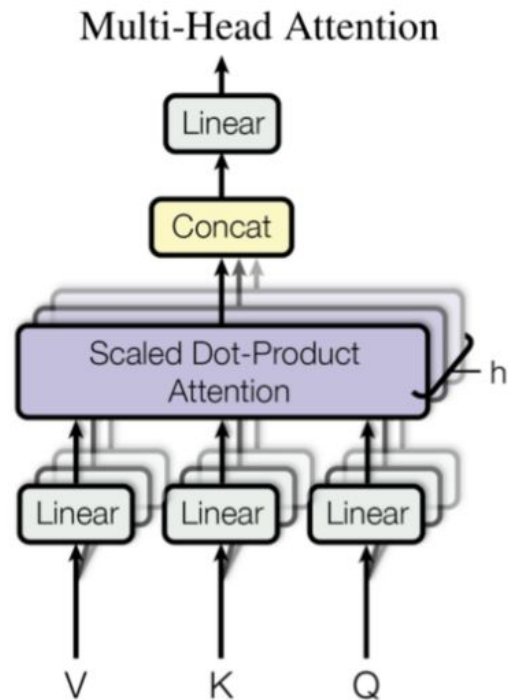
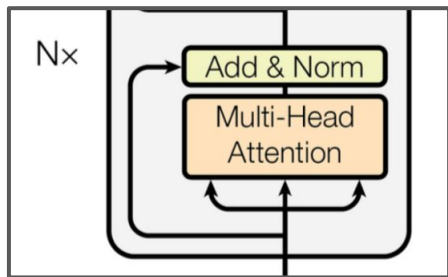
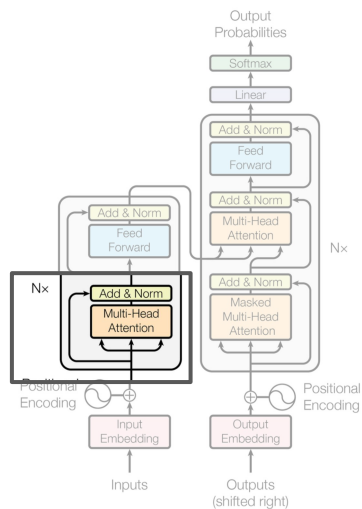
The simplest attention is scaled dot product attention



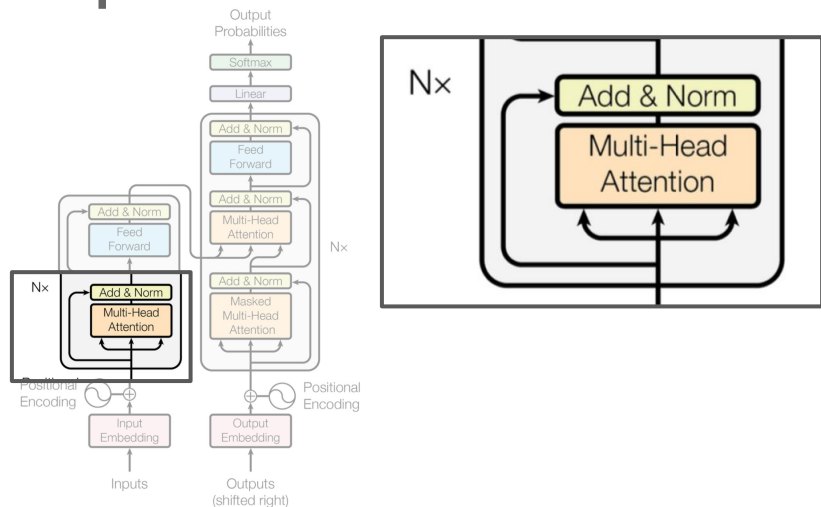
Scaled Dot-Product Attention



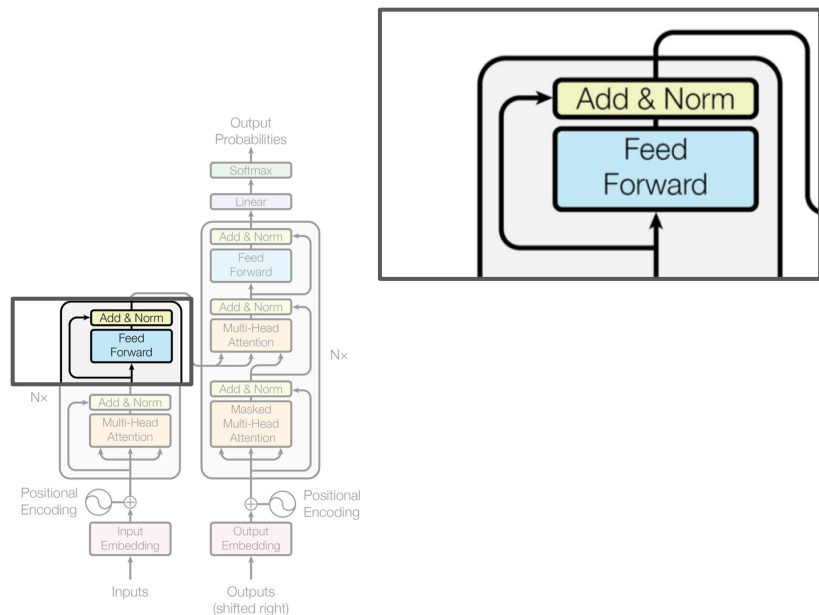
In most cases multi-head attentions are used



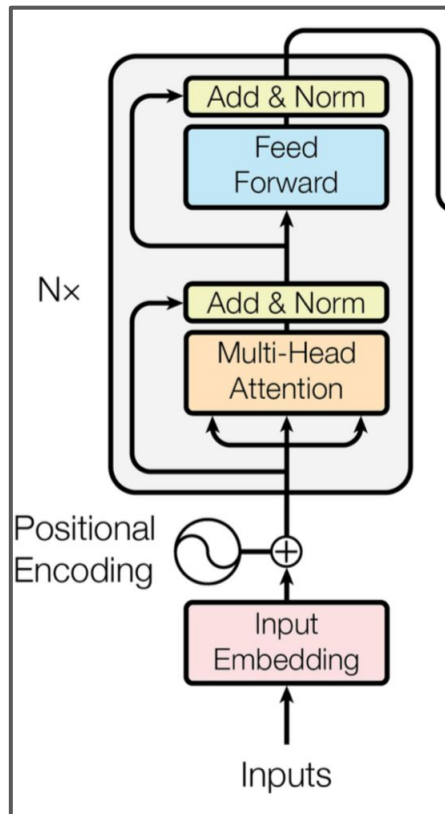
The attention layer is followed by short residual skip connections and a normalization layer



The output of the attention is then fed into a feed forward layer

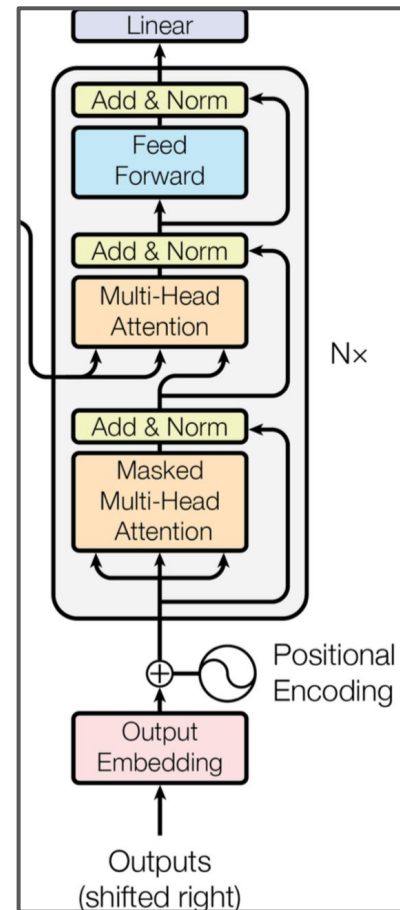


The encoder is now complete!



The decoder is in charge of creating the translation

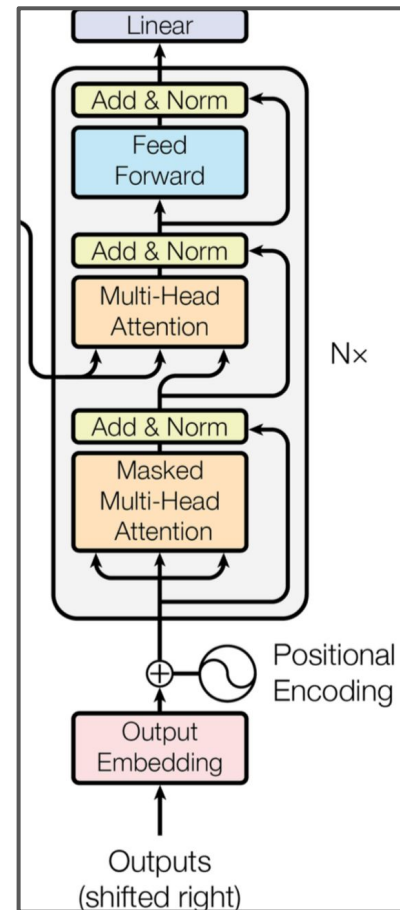
The weather is great.



The decoder is in charge of creating the translation

The weather is great.

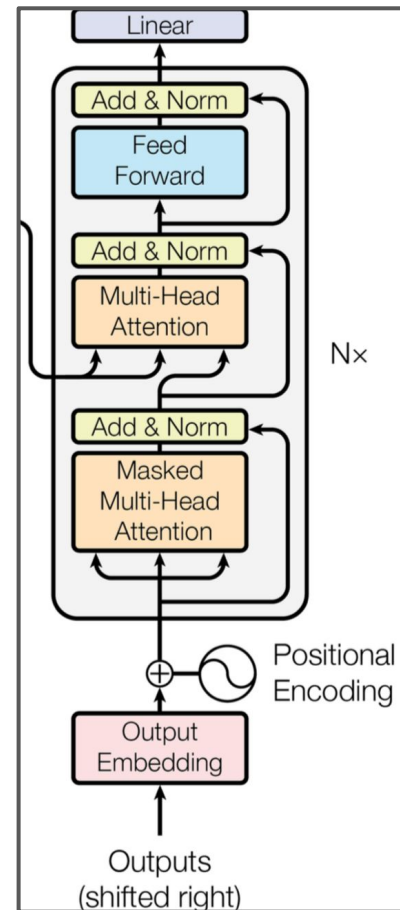
هوا



The decoder is in charge of creating the translation

The weather is great.

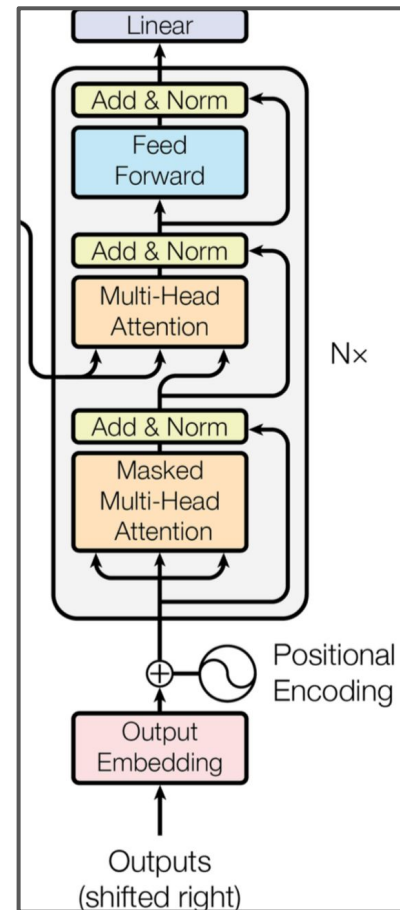
هوا عالی



The decoder is in charge of creating the translation

The weather is great.

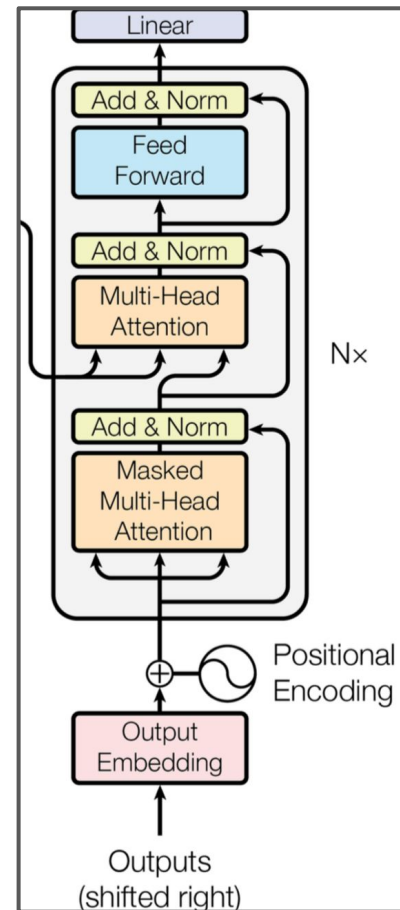
هوا عالی است



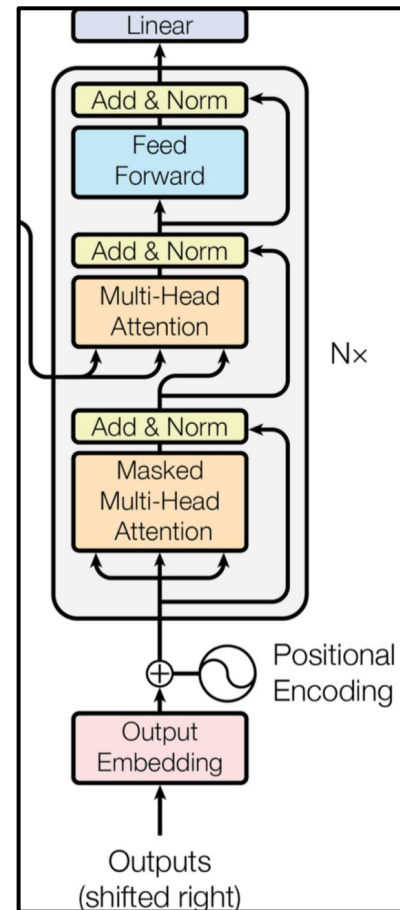
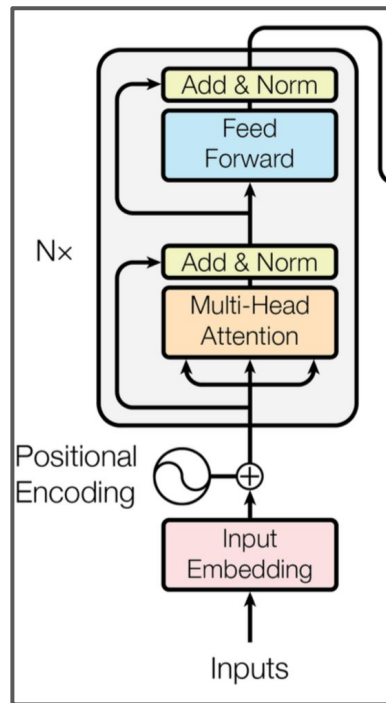
The decoder is in charge of creating the translation

The weather is great.

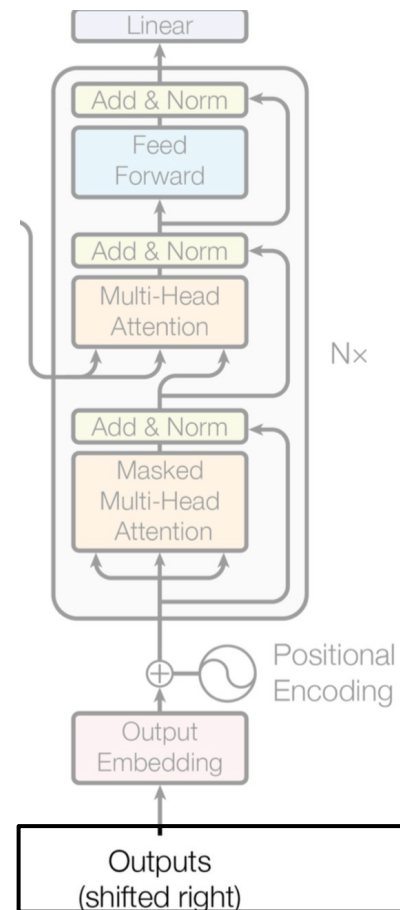
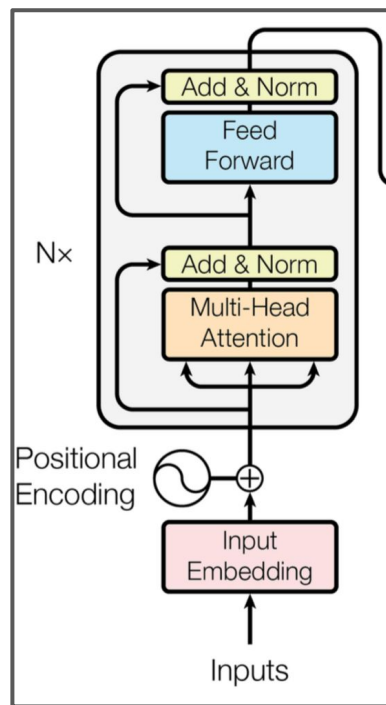
هوا عالی است.



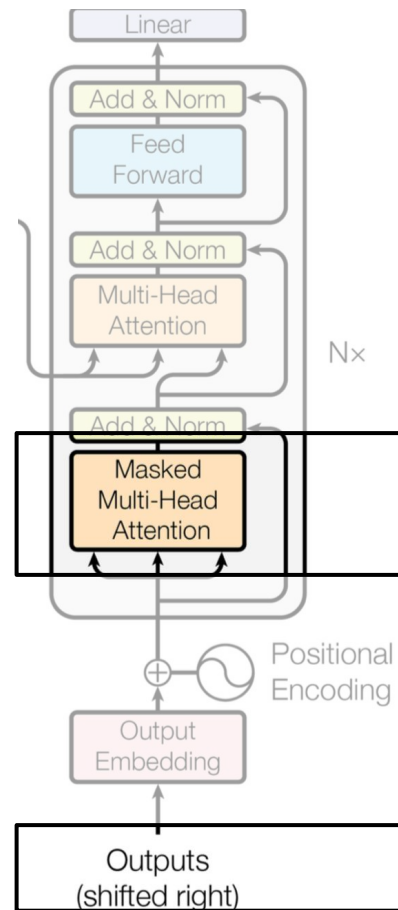
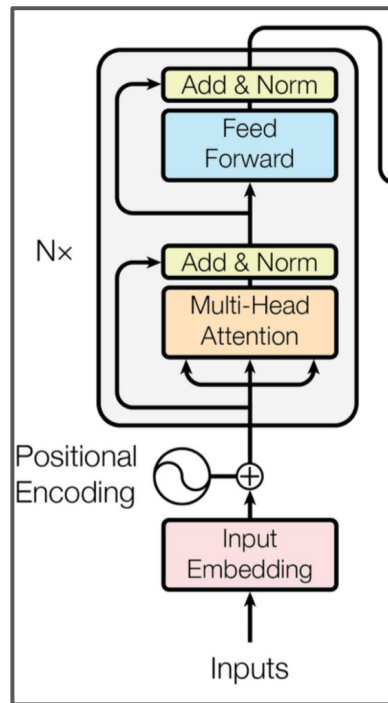
The decoder works in a way similar to the encoder



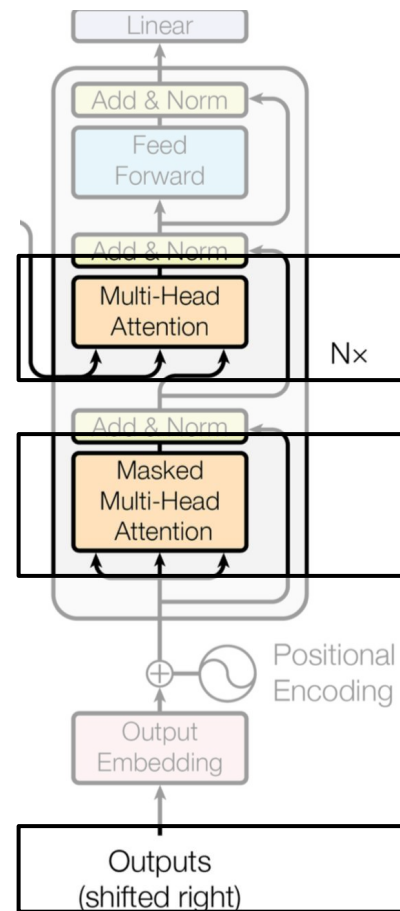
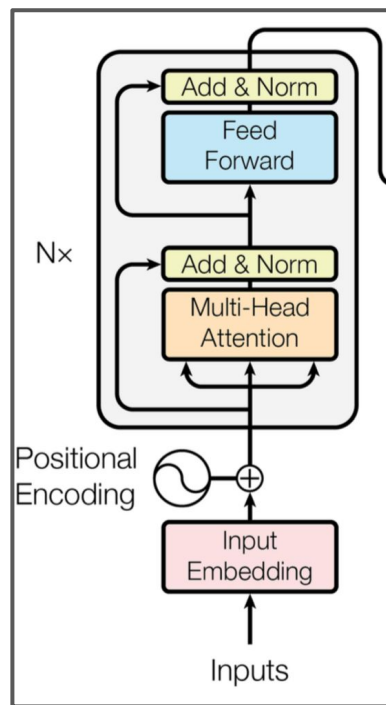
The decoder works in a way similar to the encoder



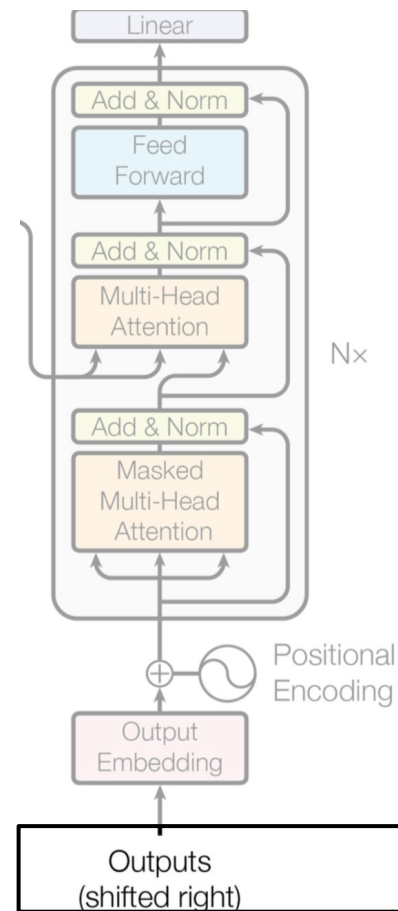
The decoder works in a way similar to the encoder



The decoder works in a way similar to the encoder



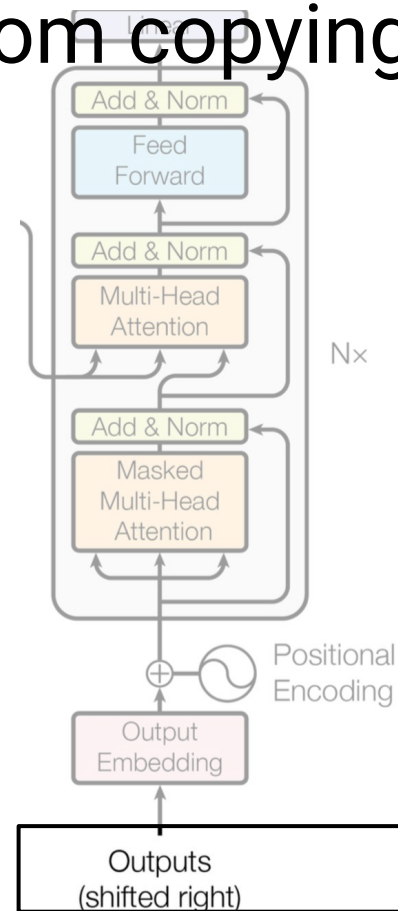
Shifting outputs ...



Shifting outputs prevent network from copying the decoder input

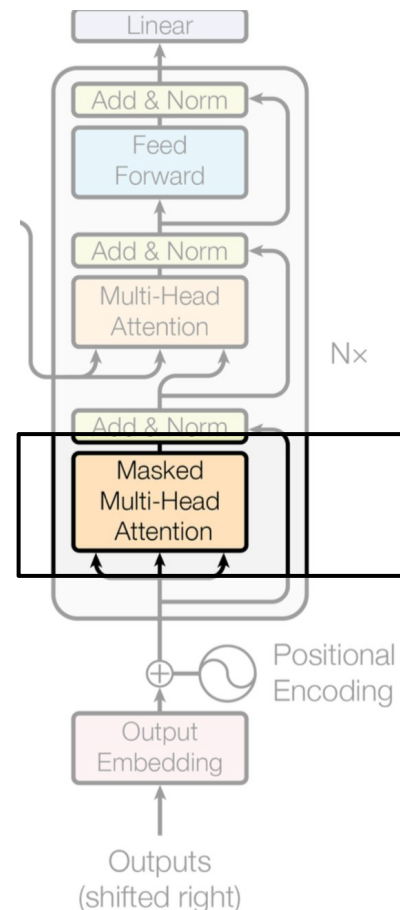
The weather is great.

هوا ← هوا



Masked layers ensure that we're only seeing information before the word to predict

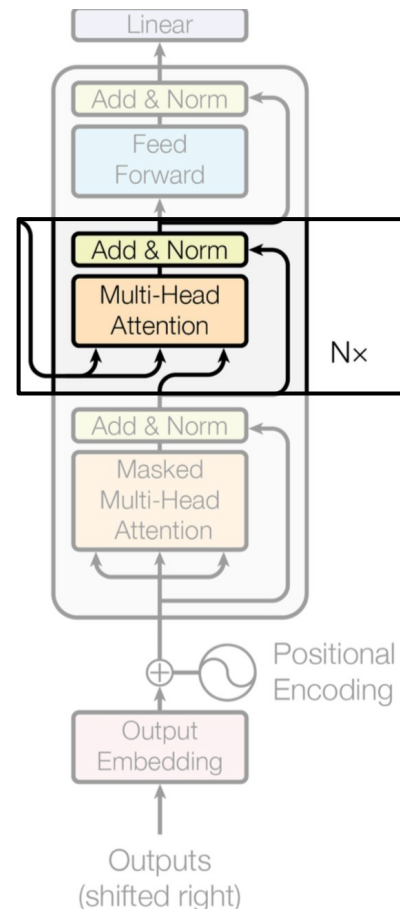
$$\text{MaskedAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{d_k}}\right) \mathbf{V}$$



Encoder-decoder attention is where the magic happens

The output of encoder:
Keys and values

The decoder's latest sentence:
The query



The final step is another linear layer and a softmax

Which word in our vocabulary
is associated with this index?

am

Get the index of the cell
with the highest value
(argmax)

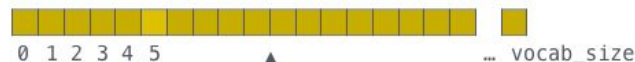
5

log_probs



Softmax

logits

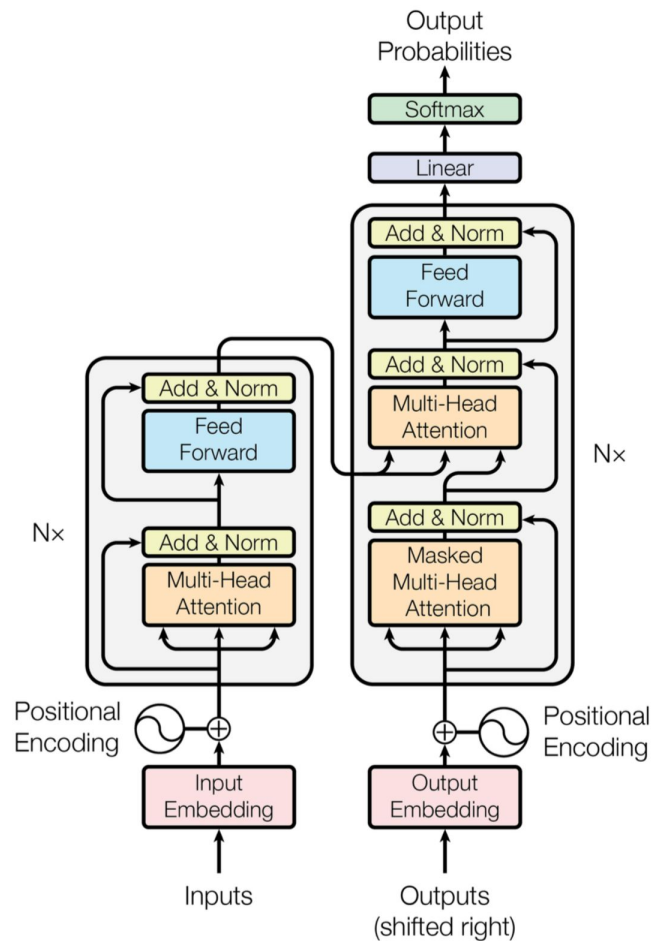
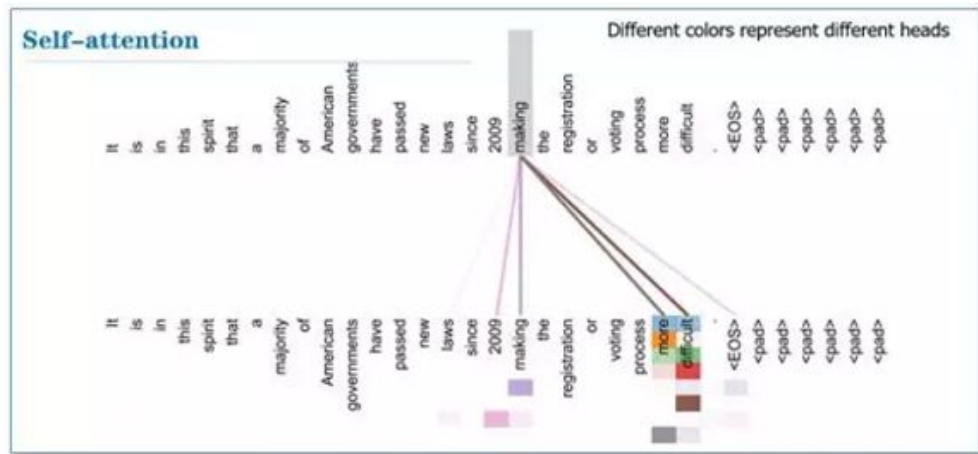


Linear

Decoder stack output



Basic architecture of a transformer

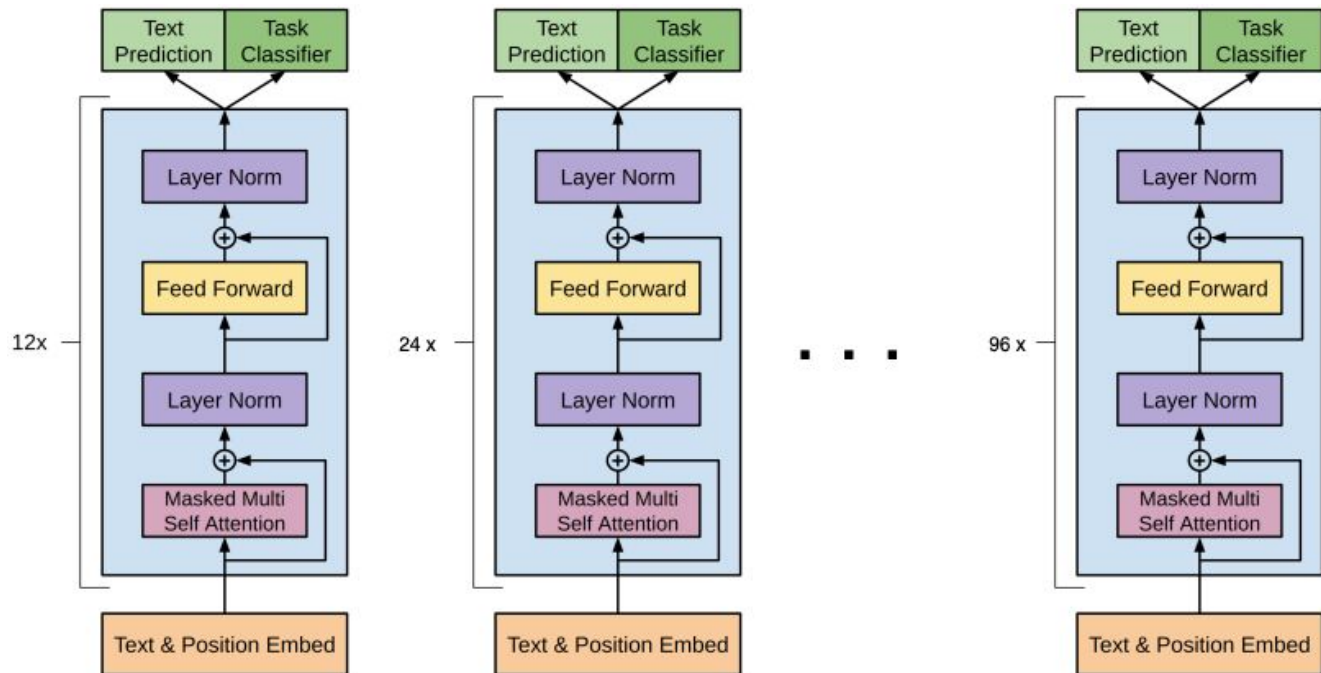


Why do transformers work so well?

Why do transformers work so well?

1. Distributed and independent representation at each block
2. Meaning heavily depends on the context
3. Multiple encoder and decoder units
4. Combination of high- and low-level information

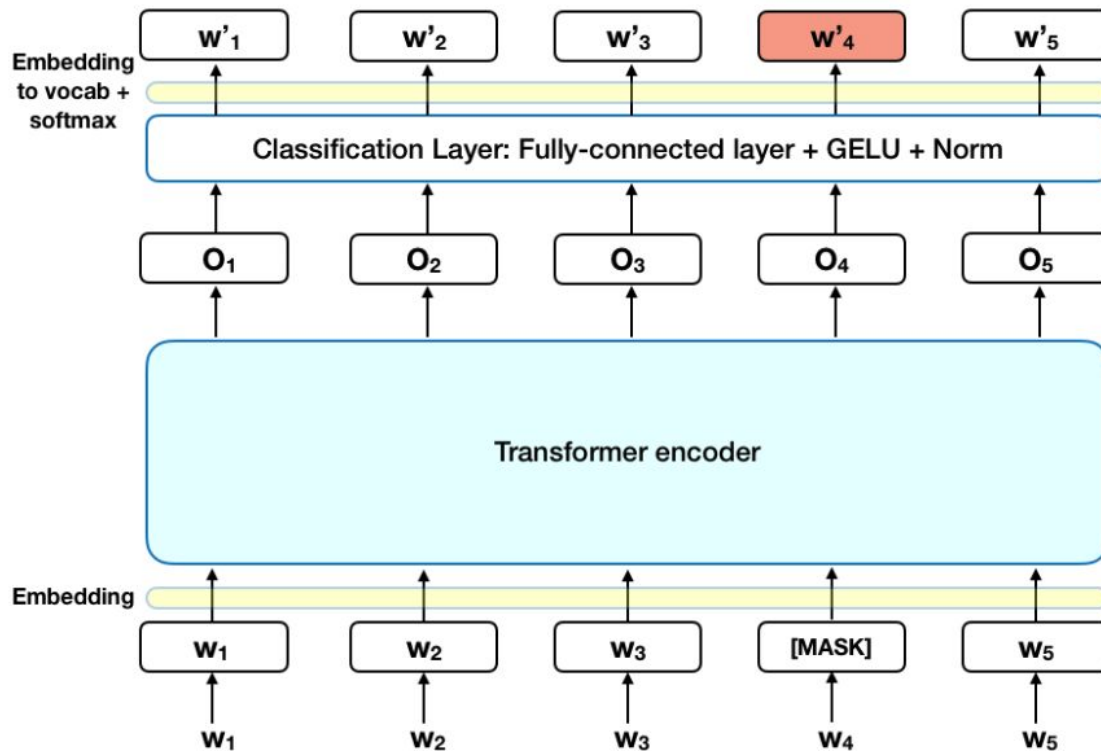
Famous transformers



Famous transformers



BERT
(Bidirectional Encoder
Representations from
Transformers)



Next lecture:

AlphaFold2 and DL in protein engineering



Kristine Deibler · 1st

Computational Design Scientist at Novo Nordisk



Layne Price · 1st

Sr Machine Learning Scientist at Amazon



Nikhil Naik · 2nd

Senior Research Manager | Machine learning, Computer Vision,
NLP, AI for Biology



Jack Maguire · 1st

Senior Scientist at Genentech