



**CORNELL
TECH**

Spring 2024

Practical Deep Learning

Week 4

Transformers & Attention

Project proposals

Propose a project, if you want :)

bit.ly/pdl24projectproposal

Take a minute to read through and I'll answer any questions you might have!

Notes

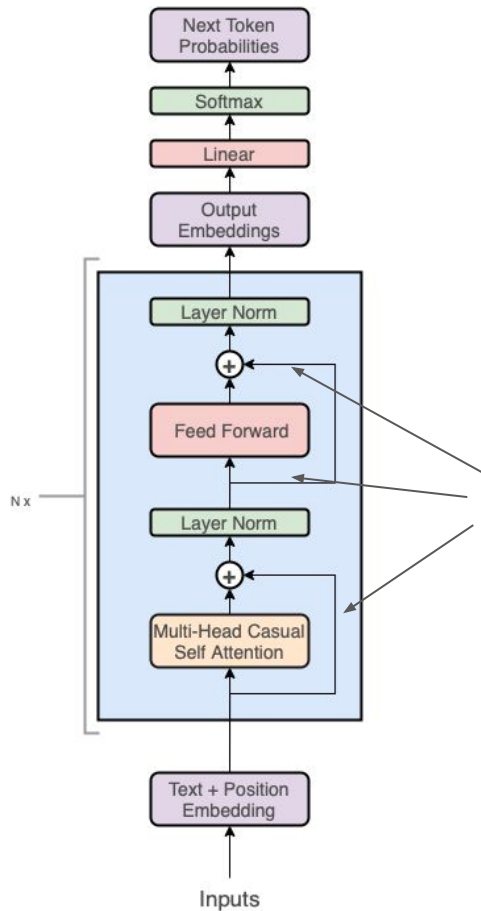
- Anonymous feedback link: bit.ly/pdl24feedback
- Laptops are allowed (but please be respectful!)
- Will put my slides on the course website
 - <https://jxmo.io/deep-learning-workshop/>

A review of softmax.

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

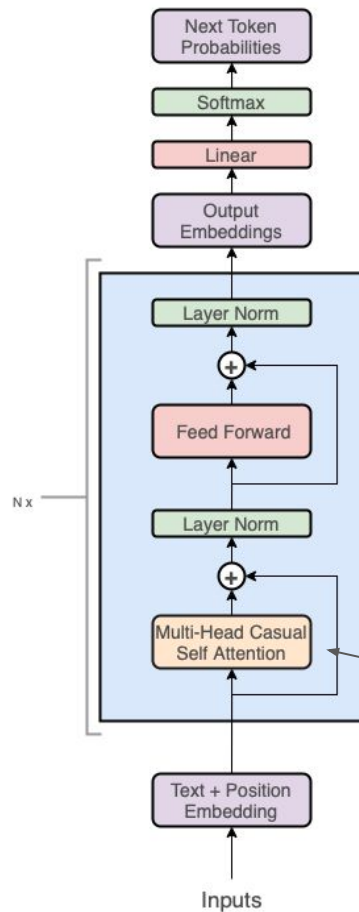
This is a transformer.

This part happens in a loop



Every layer has the same input and output shape: a sequence of embeddings

This is a transformer.



Today we're going
to talk about this
part

Attention

Input: x . A sequence of embeddings.

shape $[s, d_k]$

$W_{q,k,v}$ are shape $[d_k, d_k]$


$Q = x @ W_q \leftarrow \text{queries}$

$K = x @ W_k \leftarrow \text{keys}$

$V = x @ W_v \leftarrow \text{value}$

d_k is the hidden dimension / embedding size

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



This is possibly the
most important
equation in deep
learning rn



Puzzle

bit.ly/cs5787-mar29-puzzle

