Introduction to Generative AI

2023.11.09

Dissecting a GPT2-Like Reference Implementation

I can‘t hand in my homework!

- Technical issues with JupyterHub: datalab@tuwien.ac.at
- Backup plan: Local Environment + Downloadable Notebooks
- Get help from other students: Forum
- Unsolvable Issues & organisational questions genai-ws23@ec.tuwien.ac.at

Agenda | Part 1

Dissecting a GPT2-Like Reference Implementation

01

Mini-Recap
Architecture

02

Recap
Homework 1

03

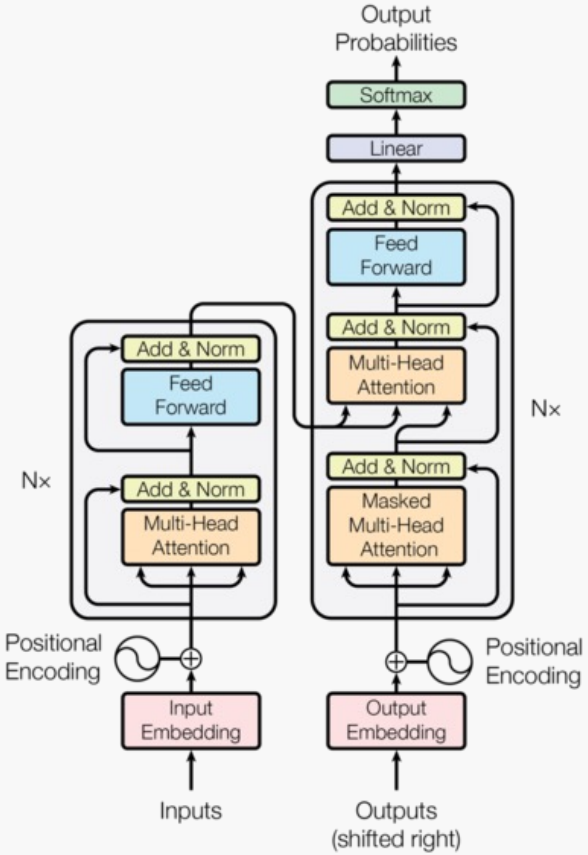
Attention
Mechanism

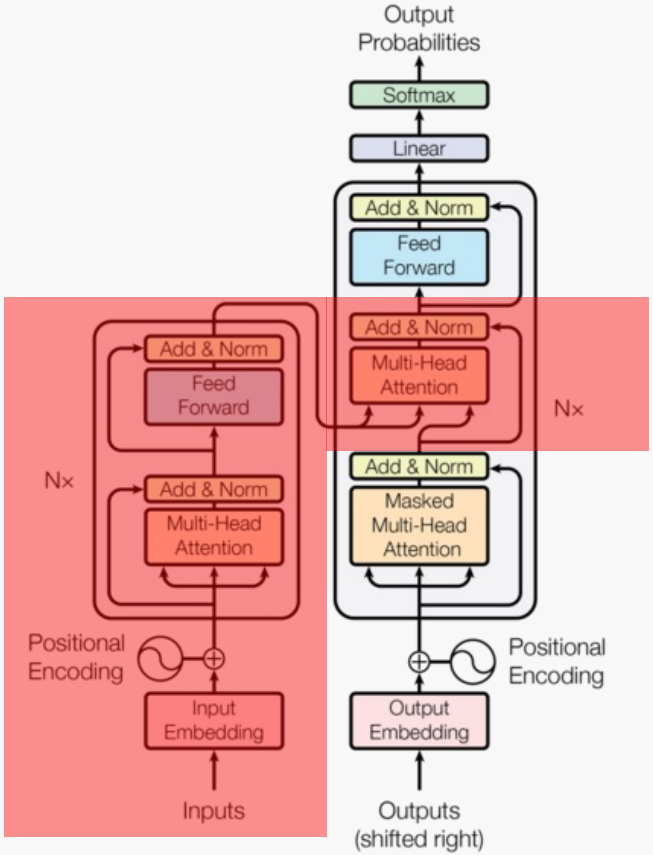
Code will be provided (after the homework deadlines)

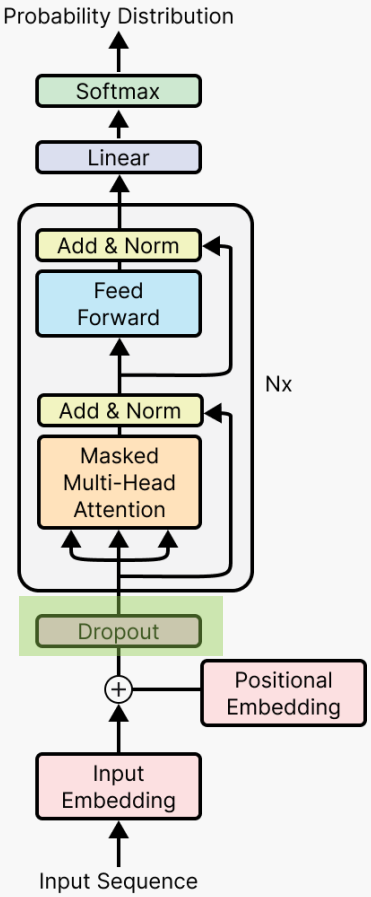
01

Mini-Reacp Architecture

Why actually a decoder is all you need.

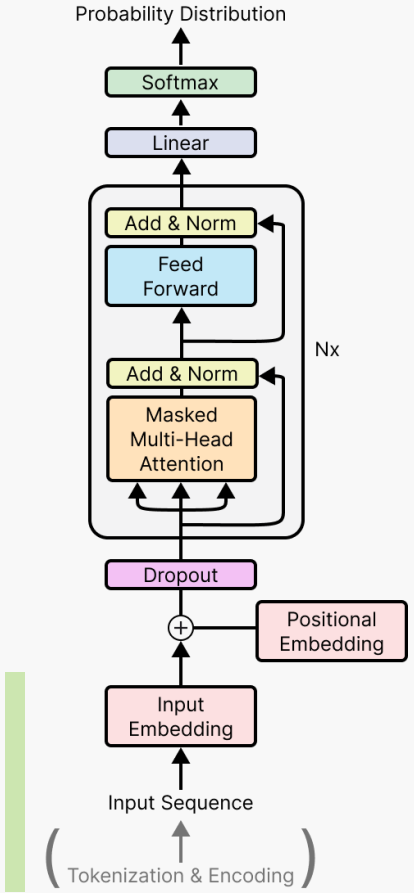

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


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


Probability Distribution
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Masked Multi-Head Attention
Nx
Dropout
Positional Embedding
Input Embedding
Input Sequence

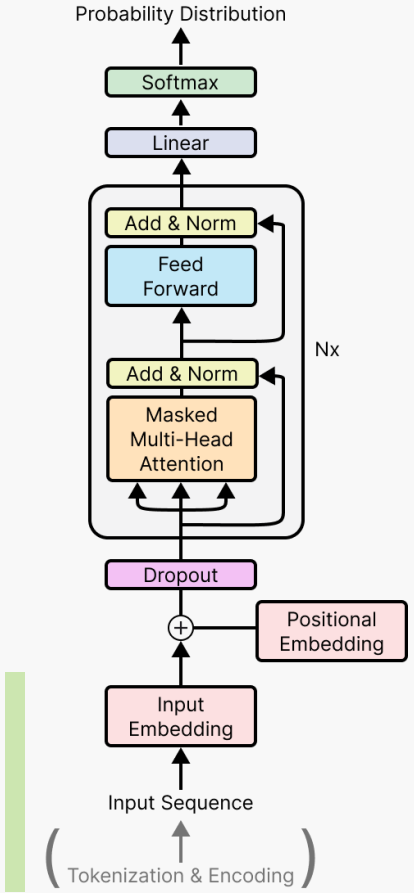
Tokenization / Encoding / Word Embeddings

Head To Notebook of Homework 1



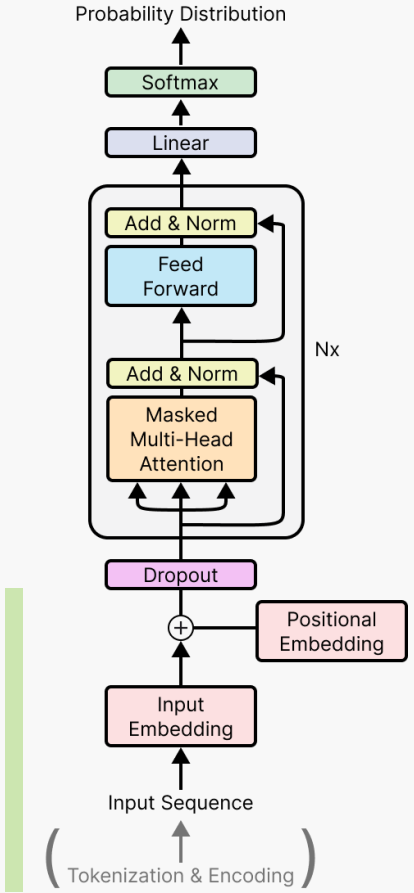
Word Embedding Visualization

Head To Notebook of Homework 1



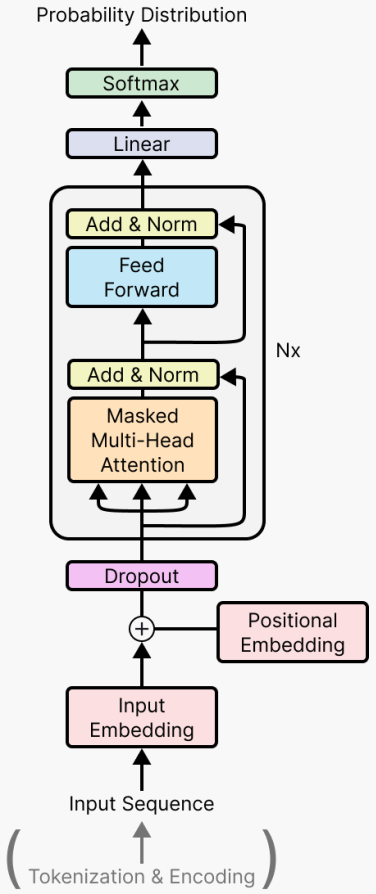
Positional Encoding & Putting it together

Head To Lecture Notebook



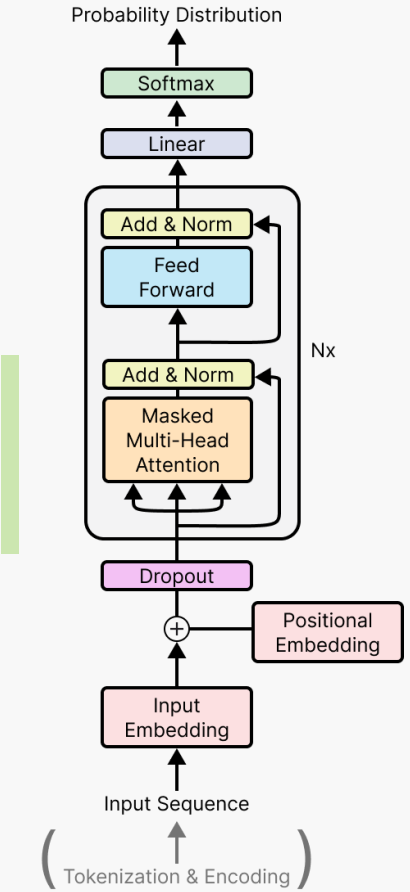
Multi-Head Masked Scaled Dot Product Self-Attention

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



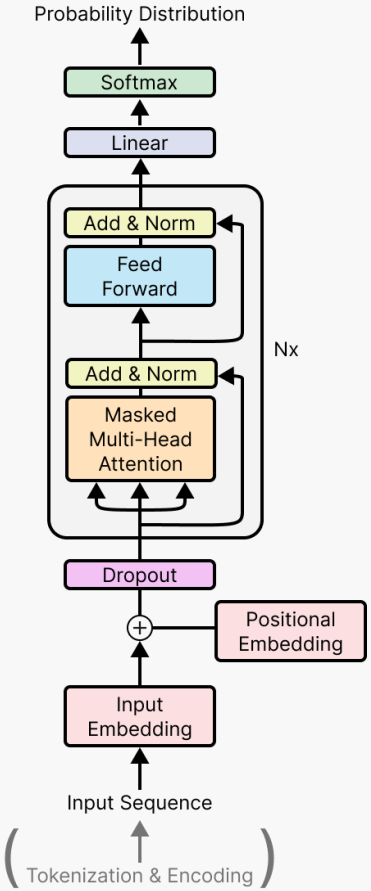
Multi-Head Masked Scaled Dot Product Self-Attention

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad \text{softmax}(\frac{QK^T}{\sqrt{d_k}} + M)V$$



QKV What?

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad \text{softmax}(\frac{QK^T}{\sqrt{d_k}} + M)V$$



Multi-Head Masked Scaled Dot Product Self-Attention

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad \text{softmax}(\frac{QK^T}{\sqrt{d_k}} + M)V$$

Head To Lecture Notebook