Transformers (LLM Notes cont)

- Transformers can be used for img gen, text to speech translation of language etc but in our case ✗ transformers are next word predictors they take in a peice of text as input (along with image or sound for context) and outputs the next word in that peice of text. This predicts all possible next words as a probability distribution and the word with highest prob is chosen. ✗ LLM are a transformer based Model trained on a large ammount of text to understand and generate language like GPT. So while transformers can be used to generate many things LLM are used to generate text then the LLM repeatedly prompts the transformer with the additional new predicted word and does this until a stoping point is reached like a stopword "</s>" or "[End]" is reached or a token limit. This is autoregressive generation

- Step by step inside a transformer

(Embbeding → Attention → MLP → unembedding) → general transformer architecture

- Embedding: first the input is broken up into individual peices:

Ex: To date, the smartest person is ⇒ Tokenization ⇒ [To] [date] [,] [the] [smartest] [person] [is]
prompt

in LLM these peices are the words that make up the prompt = □ these peices are called tokens. in img's they can be pixels in voice, fractals then each token is associated with a vector. a list of numbers which is meant to encode the meaning of that peice

Ex:

To ↓ date ↓ , ↓ the ↓ ... ?

$\begin{bmatrix}5.2\\7.1\\3.1\\\vdots\\0\end{bmatrix}$ $\begin{bmatrix}3.5\\2.1\\7\\\vdots\\2\end{bmatrix}$ $\begin{bmatrix}5\\7\\11\\\vdots\\41\end{bmatrix}$ $\begin{bmatrix}12\\13\\15\\\vdots\\1\end{bmatrix}$

- also encode position

- these vectors are like cordinates in high dimention space and similler words tend to land on vectors that are close to each other in that space and vise versa