ECG 1786 Lecture #7

Last Day: Language Generation Using Transformers & Project Ideation

Work-in-flight: Assignment 3 - tonight.
Approval-In-Principle due Thurs but do sooner!
Proposal Document } due Oct 31 @ 9pm
Proposal Slides }

Today: Understanding Transformers
Tokenization
Assignment 4
Proposal Doc/Presentations next week.

no lecture week of Nov 8 (reading week)

Recall, last lecture discussion & demo of "zero-shot" prompting of GPT-3 - you tell it what you want it to do and it does what you asked! Remarable!

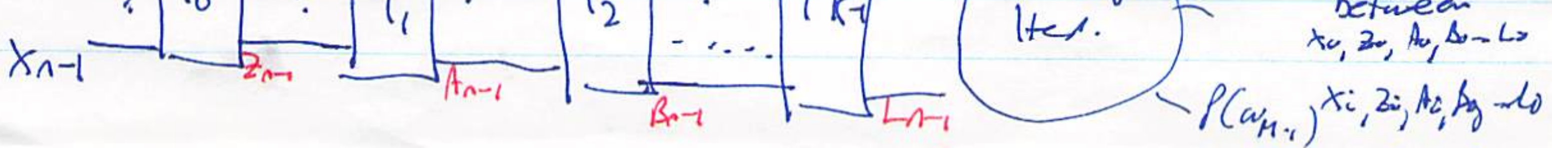
bad name → few-shot → "priming" → psychology

- a key goal for me in teaching this course was to dive in & try to understand what is happening in these models that give rise to these capabilities; and then to share it with you.

- for zero-shot generation, I feel that a part (not all) of the explanation was given last week:
  - the next word generated must be consistent with (or tries) all the words input → the initial context + everything generated. → so it does what you tell it to do
  → the "state" of this system is that full input context
  → I think that state "points" to higher level concepts represented in the model, and word-by-word generation brings those concepts out -eg. fear of injury in last week's hockey example

on now

with respect to word-by word generation

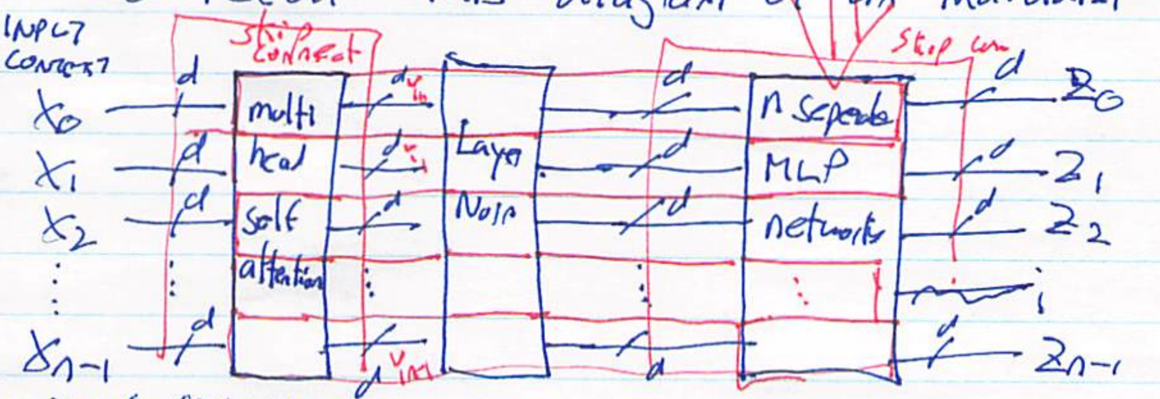


- this understanding is just part of the story; these models are somehow very good at learning these concepts, today want to look and see what we might understand about that, in part based on lecture 5's discussion of attention.

- but first, I need to correct a mistake from lecture 5 that is relevant to this discussion

TL;DR : I thought & presented the feed-forward fully connected MLP in the transformer block as one big MLP that fully-connected all $n \times d$ inputs to all $n \times d$ outputs
 - that's wrong; there are actually n separate MLPs, each with d-inputs and d outputs.
 → I thought the original way scrambled everything, but the correct way does not on board.

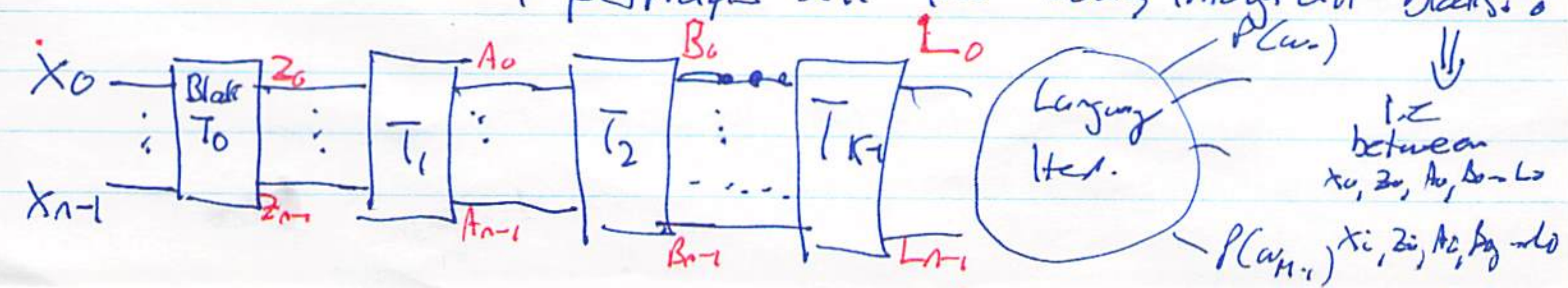
So recall this diagram of an individual Transformer Block:



- can see "mlpf" in mingpt (nano) Block module in model.py

$n \sim 2048$ for GPT-3; 1024 GPT-2

- this means that ① every red rectangle can be independently computed
 ② there is a relationship between z_i and x_i & perhaps all the way through all blocks?:



i.e between $x_0, z_0, A_0, B_0 \rightarrow L_0$

- In the literature & Jurafsky text the vectors $Z_i, A_i, B_i \ldots L_i$ are what are called "contextual" vectors → because the attention mechanism transforms the first set of vectors, x_i by looking at the surrounding x_i & merging them into x_i. ⇒ using <u>learned</u> matrices W^q, W^k, W^v for each head. + multiple heads

- these trained weights somehow contain/are able to make those good predictions, along with the MLP
- bertviz is one way to look at them. by Jesse Vig.

recall

1st layer

$$x_0 \xrightarrow{f} y_0$$
$$x_1 \rightarrow y_1$$
$$\vdots$$
$$x_{n-1} \rightarrow y_{n-1}$$

$$y_i = \sum_{j \le i} \alpha_{ij} v_j$$

$$\alpha_{ij} = \text{SOFTMAX}(q_i \cdot k_j)$$
$$q_i = W^q x_i$$
$$k_i = W^k x_i$$
$$v_i = W^v x_i$$

- Vig's code visualizes an individual α_{ij}, → I think showing, for a given head the link between x_i & x_j after being ~~trans~~ multiplied by W^q & W^k
- inputs are on the ~~right~~ left, outputs ~~left~~ right.
In demo → pull up colab demo.
- notice the original words are left ~~everywhere~~ in $x, y, z, A, B \ldots$ on later layers, but no guarantee that they survive intact.

- not exceedingly clear; too microscopic ⇒ perhaps other ways to analyze
- would like to find a way to uncover "concepts" that seem to be learned.

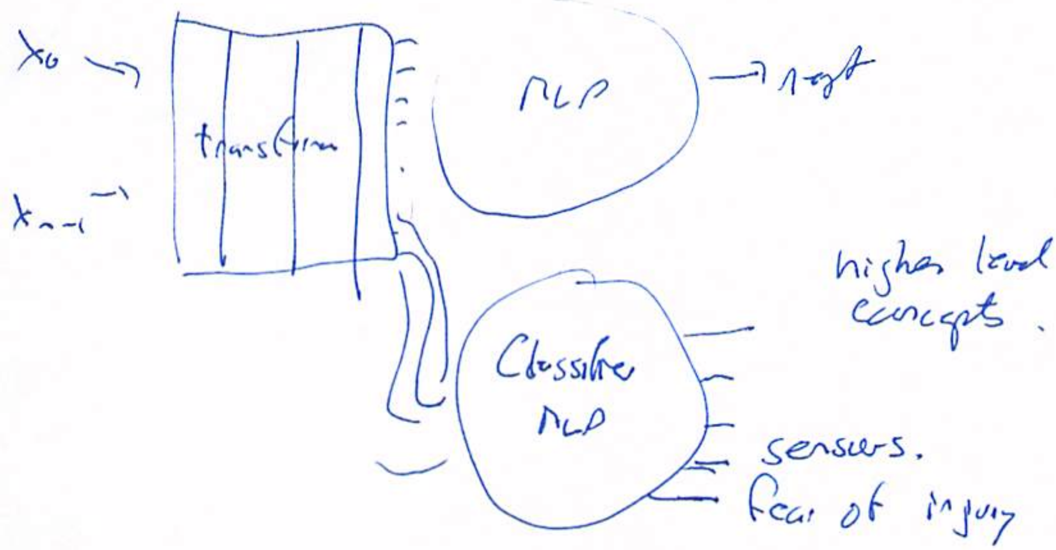
$\boxed{\text{Demo from Vig's collab - post link on Quercus}}$

Instead, let's think about higher-level concepts that are in play in GPT-3. 7-3a.

Write a creative ad for the following product on Google aimed at hockey players:

Product: A system that warns the hockey player that that they are about to run into the boards.

Hockey players are always at risk of running into the boards and sustaining serious injuries. But with our new board-warning system, they can avoid these dangerous collisions! This system uses sensors to detect when a player is getting close to the boards and sends out a warning signal, giving them time to adjust their course and avoid a dangerous collision. Don't let your next game be your last! Protect yourself with our board-warning system.



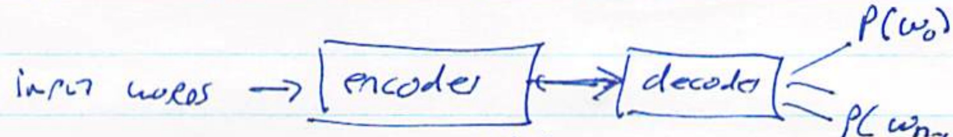
- how would we know what concepts to look for?
  - → we can label examples with concepts. (create positive & negative concepts)
  - → train (fine tune) classifier MLP.

- how to test the result?
  - with other inputs that do or do not contain the concepts

→ project?

Now, an important side point about Transformer architectures as they are commonly described.

- because of historical NLP architectures being based on RNNs, in the transformer architecture world^(for seq-to-seq) the higher-level structure is described as:



"decoding" is the greedy or sampling described in L6

- the famous BERT ^(architecture) is only an encoder.
- Vaswani's original transformer was encoder-decoder.
- GPT-x ~~is said to be~~ is decoder only (words fed itself)

- however, I believe one only needs 1 block (either encoder or decoder they are the same).
  - you can for sure stick whatever head you want on it – classification or – language gen. | as shown in A3
  - exactly as you did in Assignment 3.

- So why talk about both?..
  1. The RNN history, that need encoder → decoder to do sequence → sequence
  2. The training is different? maybe, maybe not

↳ The essence of training BERT is to predict words missing (which are intentionally masked) in sentences
- that actually is little different than predicting the next word, as described for LM & done in A3

- especially when context is >> 50 words, or like 2048 in GPT-3.

- what really matters, it seems is the model size (in # parameters) & how much data it is trained on

March 2022

⟹ See Hoffman et. al. paper "Training Compute-Optimal Large Language Models"
  - bottom line GPT-3 is under-trained → needed more data >1T.
  - Chinchilla is 70B parameters vs 175B for GPT3 but was trained on 4x data.
    - is better across the board. - See paper linked on Quercus

⟹ I think once you've got a good LM, you can use it with whichever "head" you want to do generation or classification.

Tokenization

→ look up BPE in Jurafsky 2.4.3
byte-pair encoding

- have not really discussed - what is it?
- OK to begin to think of words as tokens.
- but some subtleties are connected to tokenization.
- however, Assign 3 + GPT-x use byte-pair encoding which not only encodes full words, but parts of words and even characters.
  - same tokenization in A3 as GPT-2
- the latter allows unknown words to be tokenized. → allows idk to work ~~somehow~~

chop words in pieces
↓
tokens
↓
embeddings
↓
meaning

→ character level tokenization may not bring meaning in its associated vectors
  - what could 'i' or 'd' or 'k' mean?

- but it does allow idk to be represented.
- if idk shows up enough in training data then the sequence id-k will have meaning encoded in the trained network maybe?

  → experimentally seems to be true:
  1. idk is not a token in GPT-3
  2. when asked GPT-3 says: the maning of idk is I don't know ← generated ↑ context

- by contrast, lol is a token, and it means laugh out loud, and is also "understood"

- other sub-words are produced in byte-pair encoding - e.g. hold, which could be used to tokenize "threshold"
  - it seems that sub-words are distinct from full words in tokenization.

- discuss Assignment 4 due in 3 weeks
- next week's Project Proposal → doc + presentation
- week after is reading week
  → no lecture.

→ Approval-in-Principle due by Thursday
  → means I've said yes.
  → you should email me well before then in case I say no! (which I have already & means you need to email me again)

- Proposal doc + slide due Monday @ 9pm.
  - don't be late.

- Peer review: see next page

Recall: Peer Review of Proposals

- You will be assigned another to another group to provide feedback on their written proposal and in-person presentation.

- The goal is to give you the opportunity to think critically about others' work
  - Typical and important in engineering workplaces

Peer Review Questions

1. Answer the following questions about their proposal:
2. State the goal of the project in your own words.
3. What was the best part of the proposal's technical contribution?
4. What is one suggestion you have (that was not already made in class) that might help the project?
5. Provide feedback on the quality of the oral presentation - what was good, and what could be improved.

- Submit PDF document of about 300 words
- Due Friday November 4 at 6pm