Last Day: Language Generation Using Transformers & Project Ideation

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

Today: Understanding Transformers
Tokenization
Assignment 4
Proposal Doc/Presentations next 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! Remarkable!

- 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 generation, I feel that a part of the explanation was given last week:
  - The next word generated must be consistent with all the words that the initial context (everything generated up until the "state" of this system) is that full input context represented in the model, and word-by-word generation brings these concepts out - e.g., fear of injury in last week's lecture example.
- 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 these based on lecture 5's discussion of attention.

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

TL; DR: I thought I presented the feed-forward fully connected MLP in the transformer block as one big MLP that fully-connected all N inputs to all N 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.

- So recall this diagram of an individual Transformer Block:

- this means that 1) every red rectangle can be independently computed
  2) there is a relationship between $Z_i$ and $X_i$ maybe all the way through all blocks.

- can see "mlp" in mlp.py, Black module in model.py

- between $X_0, X_1, X_2, X_{n-1}$
- In the literature & Jurafsky text the vectors $\mathbf{Ri}$, $\mathbf{Ai}$, $\mathbf{Bi}$ ... $\mathbf{Li}$ are what are called "contextual" vectors $\Rightarrow$ because the attention mechanism transforms the first set of vectors, $\mathbf{x}_i$ by looking at the surrounding $\mathbf{x}_{i-1}$ merging them into $\mathbf{x}_i$. $\Rightarrow$ using learned matrices $W^a$, $W^k$, $W^v$ for each head $\Rightarrow$ multiple heads.

- these trained weights somehow contain x are able to make these good predictions along with the MCR.

- best viz is one way to look at them.

\[ \begin{align*}
\mathbf{x}_0 & \rightarrow \mathbf{v}_0 \\
\mathbf{x}_i & \rightarrow \mathbf{y}_i \\
& \vdots \\
\mathbf{x}_{n-1} & \rightarrow \mathbf{y}_{n-1}
\end{align*} \]

\[ \mathbf{y}_i = \sum_{j} \alpha_{ij} \mathbf{v}_j \]

$\alpha_{ij} = \text{softmax}(\mathbf{q}_i \cdot \mathbf{k}_j)$

$\mathbf{q}_i = W^a \mathbf{x}_i$

$\mathbf{k}_i = W^k \mathbf{x}_i$

$\mathbf{v}_i = W^v \mathbf{x}_i$

- Vig's code visualizes an individual $\alpha_{ij}$ showing for a given head the link between $\mathbf{x}_i \& \mathbf{x}_j$ after being multiplied by $W^a$.

- Inputs are on the left, outputs on the right. In demo $\Rightarrow$ pull up colab demo.

- Notice the original words are left on later layers, but no guarantee that they survive intact.

- Not exceedingly clear: too microscopic $\Rightarrow$ perhaps other ways to analyze "concepts" that seem to be learned.

- Demo from Vig's colab - post links on Grades
Instead, let's think about higher-level concepts that are in play in CPT-3.

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

Product: A system that warns the hockey player 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
- Train (fine tune) classifier
- Negative concepts

How do we 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 MLP architectures being based on RNNs, in the Transformer architecture the higher-level structure is described as:

  \[
  \text{input words} \rightarrow \text{encoder} \rightarrow \text{decoder} \rightarrow \text{P(w)} \quad \text{"decoding" is the goal or simply described as L6.}
  \]

- The famous BERT is only an encoder.

- Vaswani’s original Transformer was encoder-decoder.

- GP7-x is said to be decoder only (much less hard).

- 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—classifier, or \[ \text{language class} \text{,} \]

- Exactly as you did in Assigned 3.

- So why talk about both?:

  1. The RNN history that need encoder-decoder to do sequence to sequence

  2. The training is different? maybe, maybe

The essence of training BERT is to predict words missing (which are intentionally masked) in sentence.

- that actually is little different than predicting the next word, as described for LTN and done in \[ \text{A3。} \]
especially when context is >> 50 words, 
or like 2018 in GP7-3.

what really matters it seems is the model size
in parameters & how much data it is
trained on

⇒ See Hoffman et al. paper "Training Computationally
Optimal Large Language Models"
- bottom line GP7-3 is under-trained => needed more
data >> 1T.

- Churchill is 700 parameters vs 1.75B for GP7
but was trained on 4x data.
- is better across the board. ⇒ see paper linked
  on Reeves

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

Tokenization ⇒ look up BPE in Jursaerk 2.4.3
- have not really discussed ⇒ what is it?
- ok to begin to think of words as tokens
- but some subtleties are connected to tokenization.
- however, GPT-3 + GPT-x use byte-pair
encoding which not only encodes full
words, but parts of words and even characters.
- some tokenization is 13 vs 64?

- the latter allows unknown words to be
tokenized. ⇒ allows idk to work
character level tokenization may not bring meaning in its associated vectors—what could 'i' or 'd' or 'a' mean?

but it does allow *idle* to be represented.

if *idle* 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. *idle* is not a token in GP1-3
2. when asked GP1-3 says: the meaning of *idle* is I don't know. (source)

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 & due in 3 weeks
  Project
- next week's 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 to see if I say no! (which I have already if means you need to email me again)

- Proposal doc + slide due Friday 9 pm.
  -> 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