CS 601.471/671 NLP: Self-supervised Models

Homework 5: Seq-to-Seq models, Transformers and Self-Attention

For homework deadline and collaboration policy, check the calendar on the course website[*]

Name: __________________________

Collaborators, if any: ____

Sources used for your homework, if any: _______________________________________

This assignment focuses on on a variety of topics related Recurrent Neural Networks and Transformers, particularly in the context of language model.

How to hand in your written work: Via Gradescope as before.

1 Concepts, intuitions and big picture

Each question might have multiple correct answers. (Check all that apply).

1. True or false? A language model usually does not need labels annotated by humans for its pretraining.
   □ True □ False
   Answer: *TBD*

2. Select the sentence that best describes the terms "model", "architecture", and "weights".
   □ Model and weights are the same; they form a succession of mathematical functions to build an architecture.
   □ An architecture is a succession of mathematical functions to build a model and its weights are those functions parameters.
   Answer: *TBD*

3. Which of these types of models would you use for completing prompts with generated text?
   □ An encoder model
   □ A decoder model
   Answer: *TBD*

4. Which of these types of models would you use for classifying text inputs according to certain labels?
   □ An encoder model
   □ A decoder model
   Answer: *TBD*

5. To which of these tasks would you apply a many-to-one RNN architecture?
   □ Speech recognition (input an audio clip and output a transcript)
   □ Sentiment classification (input a piece of text and output a 0/1 to denote positive or negative sentiment)
   □ Gender recognition from speech (input an audio clip and output a label indicating the speaker's gender)
   Answer: *TBD*

6. What possible source can the bias observed in a model have?
   □ The model is a fine-tuned version of a pretrained model and it picked up its bias from it.
   □ The data the model was trained on is biased.
   □ The metric the model was optimizing for is biased.
   Answer: *TBD*

7. What is the order of the language modeling pipeline?
   □ First, the model, which handles text and returns raw predictions. The tokenizer then makes sense of these

[*] https://self-supervised.cs.jhu.edu/sp2023/

predictions and converts them back to text when needed.
☐ First, the tokenizer, which handles text and returns IDs. The model handles these IDs and outputs a prediction, which can be some text.
☐ The tokenizer handles text and returns IDs. The model handles these IDs and outputs a prediction. The tokenizer can then be used once again to convert these predictions back to some text.
Answer: *TBD*

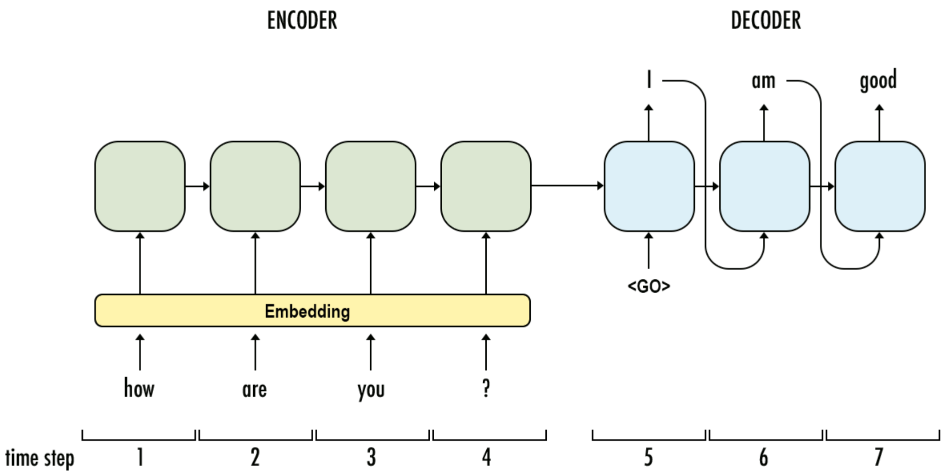
8. How many dimensions does the tensor output by a decoder Transformer model have, and what are they?
☐ 2: The sequence length and the batch size
☐ 2: The sequence length and the hidden size
☐ 3: The sequence length, the batch size, and the hidden size
Answer: *TBD*

9. You are training an RNN language model. At the the time step, what is the RNN doing? Choose the best answer.
☐ Estimating $P(y_1, y_2, ..., y_{t-1})$
☐ Estimating $P(y_1)$
☐ Estimating $P(y_t | y_1, y_2, ..., y_{t-1})$
☐ Estimating $P(y_t | y_1, y_2, ..., y_t)$
Answer: *TBD*

10. You are training an RNN, and find that your weights and activations are all taking on the value of `NaN` ("Not a Number"). Which of these is the most likely cause of this problem?
☐ Vanishing gradient problem.
☐ Exploding gradient problem.
☐ ReLU activation function g(.) used to compute g(z), where z is too large.
☐ Sigmoid activation function g(.) used to compute g(z), where z is too large.
Answer: *TBD*

11. You have a pet crab whose mood is heavily dependent on the current and past few days' weather. You've collected data for the past 42 days on the weather, which you represent as a sequence as $x_1, ..., x_{42}$. You've also collected data on your crab's mood, which you represent as $y_1, ..., y_{42}$. You'd like to build a model to map from $x \to y$. Should you use a Unidirectional RNN or Bidirectional RNN for this problem?
☐ Bidirectional RNN, because this allows the prediction of mood on day t to consider more information.
☐ Bidirectional RNN, because this allows backpropagation to compute more accurate gradients.
☐ Unidirectional RNN, because the value of y_t depends only on $x_1, ..., x_t$ but not on $x_{t+1}, ..., x_{42}$
☐ Unidirectional RNN, because the value of y_t depends only on x, and not other days' weather.
Answer: *TBD*

12. Consider this encoder-decoder model. This model is a "conditional language model" in the sense that the encoder portion (shown in green) is modeling the probability of the input sentence x.
☐ True ☐ False Answer: *TBD*



13. Compared to the RNN-LMs and fixed-window-LMs, we expect the attention-based LMs to have the greatest advantage when:
□ The input sequence length is large.
□ The input sequence length is small.
Answer: *TBD*

14. What is a model head?
□ A component of the base Transformer network that redirects tensors to their correct layers
□ Also known as the self-attention mechanism, it adapts the representation of a token according to the other tokens of the sequence
□ An additional component, usually made up of one or a few layers, to convert the transformer predictions to a task-specific output
Answer: *TBD*

15. What are the techniques to be aware of when batching sequences of different lengths together?
□ Truncating □ Returning tensors □ Padding □ Attention masking
Answer: *TBD*

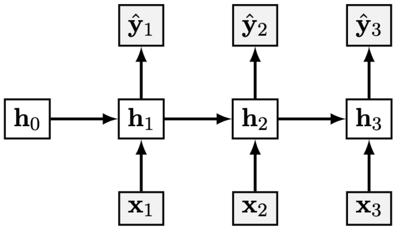
16. Is there something wrong with the following code?

```
from transformers import AutoTokenizer, AutoModel

tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
model = AutoModel.from_pretrained("gpt2")

encoded = tokenizer("Hey!", return_tensors="pt")
result = model(**encoded)
```

□ No, it seems correct.
□ The tokenizer and model should always be from the same checkpoint.
□ It's good practice to pad and truncate with the tokenizer as every input is a batch.
Answer: *TBD*

2 Training a Recurrent Neural Net

Consider the following simple RNN architecture: Where the layers and their corresponding weights are given below:



$$\begin{aligned} \mathbf{x}_t &\in \mathbb{R}^3 & \qquad \mathbf{W}_{hx} &\in \mathbb{R}^{4\times 3} \\ \mathbf{h}_t &\in \mathbb{R}^4 & \mathbf{W}_{yh} &\in \mathbb{R}^{2\times 4} \\ \mathbf{y}_t, \hat{\mathbf{y}}_t &\in \mathbb{R}^2 & \mathbf{W}_{hh} &\in \mathbb{R}^{4\times 4} \end{aligned}$$

$$\begin{aligned} J &= -\sum_{t=1}^{3}\sum_{i=1}^{2} y_{t,i}\log(\hat{y}_{t,i}) \\ \hat{\mathbf{y}}_t &= \sigma(\mathbf{o}_t) \\ \mathbf{o}_t &= \mathbf{W}_{yh}\mathbf{h}_t \\ \mathbf{h}_t &= \psi(\mathbf{z}_t) \\ \mathbf{z}_t &= \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{hx}\mathbf{x}_t \end{aligned}$$

Where σ is the softmax activation and ψ is the identity activation (i.e. no activation).

2.1 Computation Graph

Draw the computational graph for the given model.
Answer: *TBD*

2.2 Backprop for RNN

Now you will derive the steps of the backpropagation algorithm that lead to the computation of $\frac{dJ}{dW_{hh}}$. For all parts of this question, please write your answer in terms of W_{hh}, W_{yh}, y, $\hat{y}$, h, and any additional terms specified in the question (note: this does not mean that every term listed shows up in every answer, but rather that you should simplify terms into these as much as possible when you can).

1. Let's define a cross-entropy for our RNN at time t: $J_t = -\sum_{i=1}^{2} y_{t,i}\log(\hat{y}_{t,i})$. What is $\frac{\partial J_t}{\partial o_t}$? Write your answer and show your work. **Hint:** First derive the partial derivatives with respective to everything after $\mathbf{h}_t$. After that, try to break up the objective into a term for each timestep.
   Answer: *TBD*

2. Suppose you have a variable $\mathbf{g}_{\mathbf{o}_t}$ that stores the value of $\frac{\partial J_t}{\partial \mathbf{o}_t}$. What is $\frac{\partial J_t}{\partial \mathbf{h}_i}$ for an arbitrary $i \in [1, 3]$? Write your solution in terms of the $\mathbf{g}_{\mathbf{o}_t}$ and the aforementioned variables and show your work.
   Answer: *TBD*

3. Suppose you have a variable $\mathbf{g}_{\mathbf{h}_i}$ that stores the value of $\frac{\partial J_t}{\partial \mathbf{h}_i}$. What is $\frac{\partial J_t}{\partial \mathbf{W}_{hh}}$? Write your solution in terms of the $\mathbf{g}_{\mathbf{h}_i}$ and the aforementioned variables. Show your work in the second.
   Answer: *TBD*

4. Suppose you have a variable $\mathbf{g}_{W_{hh},t}$ that stores the value of $\frac{\partial J_t}{\partial W_{hh}}$. What is $\frac{\partial J}{\partial \mathbf{W}_{hh}}$? Write your solution in terms of the $\mathbf{g}_{W_{hh},t}$ and the aforementioned variables. Show your work.
   Answer: *TBD*

3 Self-Attention and Transformers

In Transformers, self-attention reigns
Each word and phrase it carefully trains
To build a context, understanding deep
A powerful tool, to make meaning leap

And on this day of love and romance
Self-attention can also enhance
Our understanding of our inner selves
As we delve deep into our hearts and delve

We give ourselves the attention we need
To truly know our wants and heed
The call of our passions and desires
That set our hearts and souls on fire

So on this Valentine's Day, let us not forget
The power of self-attention, a tool we can't regret
May it help us love ourselves with full intention
And guide us to our truest heart's affection.

–ChatGPT Feb 14 2023

Recall that the transformer architecture uses scaled dot-product attention to compute *attention weights*:

$$\boldsymbol{\alpha}^{(t)} = \text{softmax}\left(\frac{\mathbf{q}_t \mathbf{K}^\top}{\sqrt{h}}\right) \in [0,1]^n$$

The resulting embedding in the output of attention at position t are:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})_t = \sum_{t'=1}^{n} \alpha_{t'}^{(t)} \mathbf{v}_{t'} \in \mathbb{R}^{1\times h},$$

where $\boldsymbol{\alpha}^{(t)} = [\alpha_0^{(t)}, \ldots, \alpha_n^{(t)}]$. The same idea can be stated in a matrix form,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{h}}\right) \mathbf{V} \in \mathbb{R}^{n\times h}.$$

In the above equations:

- h is the hidden state dimension and n is the input sequence length;
- $\mathbf{X} \in \mathbb{R}^{n\times h}$ is the input to the attention;
- $\mathbf{x}_t \in \mathbb{R}^{1\times h}$ is the slice of $\mathbf{X}$ at position t, i.e. vector representation (embedding) of the input token at position t;
- $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{h\times h}$ are the projection matrices to build query, key and value representations;
- $\mathbf{Q} = \mathbf{X}\mathbf{W}_q \in \mathbb{R}^{n\times h}, \mathbf{K} = \mathbf{X}\mathbf{W}_k \in \mathbb{R}^{n\times h}, \mathbf{V} = \mathbf{X}\mathbf{W}_v \in \mathbb{R}^{n\times h}$ are the query, key and value representations;
- $\mathbf{q}_t = \mathbf{x}_t\mathbf{W}_q \in \mathbb{R}^{1\times h}$ is the slice of $\mathbf{Q}$ at position t. Similarly, $\mathbf{k}_t = \mathbf{x}_t\mathbf{W}_k \in \mathbb{R}^{1\times h}$ and $\mathbf{v}_t = \mathbf{x}_t\mathbf{W}_v \in \mathbb{R}^{1\times h}$.

Now answer the following questions:

3.1 Complexity

What is the computational complexity of self-attention layer in terms of n and h? In particular, show that the complexity of a self-attention layer at test-time scales quadratically with the sequence length n.
Answer: *TBD*

3.2 Extra Credit: Linear Attention with SVD

It has been empirically shown in Transformer models that the context mapping matrix $P = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{h}}\right)$ often has a low rank. Show that if the rank of P is k and we already have access to the SVD of P, then it is possible to compute self-attention in $O(nkd)$ time.
Answer: *TBD*

3.3 Extra Credit: Linear Attention by Projection

Suppose we ignore the Softmax and scaling and let $\mathbf{P} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{h}} \in \mathbb{R}^{n\times n}$. Assume $\mathbf{P}$ is rank k. Show that there exist two linear projection matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{k\times n}$ such that $\mathbf{P}\mathbf{V} = \mathbf{Q}(\mathbf{C}\mathbf{K})^\top\mathbf{D}\mathbf{V}$ and the right hand side can be computed in $O(nkd)$ time. **Hint:** Consider using SVD in your proof.
Answer: *TBD*

3.4 Masking in Self-Attention

Suppose we are are using token-making objective to create a language generation model that uses self-attention. For example, suppose we want to mask $t = 3$ position. Then, it does not make sense for $\mathbf{q}_t : \forall t \in [0 \ldots n]$ to look at $\mathbf{k}_3$ and $\mathbf{v}_3$. Describe one way we can go about implementing such masking.
Answer: *TBD*

3.5 Importance of Scaled Dot Product Attention

In practice, we scale each dot product $\mathbf{QK}^\top$ by a factor of $\sqrt{h}$. This is called *scaled dot product attention*. In this part, we will prove why we perform this scaling. Suppose we are performing a dot product between a key $\mathbf{k}$ and query $\mathbf{q}$, where $\mathbf{k}, \mathbf{q} \in \mathbb{R}^h$ and $\mathbf{k}, \mathbf{q} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$.

1. Compute $\mathbf{E}[\mathbf{qk}^\top]$ in terms of $\boldsymbol{\mu}, d, \sigma$.
   Answer: *TBD*

2. Compute $\mathbf{Var}[\mathbf{qk}^\top]$ in terms of $\boldsymbol{\mu}, d, \sigma$.
   Answer: *TBD*

3. Based on the variance computed above, explain why we need to scale the dot product by $\sqrt{h}$.
   Answer: *TBD*

4 Programming

See the course website for the link to Google Colab: Colab

Optional Feedback

Have feedback for this assignment? Found something confusing? We'd love to hear from you!