CS672 Neural Network - Assignment 3

Submit by the Blackboard System

DUE: Monday, Nov 13, 2023, at 11:59 PM

- **For each problem, briefly explain/justify how you obtained your answer.** This will help us determine your understanding of the problem and whether or not you got the correct answer. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide.
- Work submitted after the due date may be graded for correctness, but not credited.
- The assignment includes three parts, i.e., Argmax Attention (1.5 points), Transformer (4.5 points), and Hand-Design Transformers (4 points).

1 Argmax Attention (1.5 points)

In this problem, we ask you to consider a hypothetical argmax version of attention where it returns exactly the value corresponding to the key that is most similar to the query, where similarity is measured using the traditional inner-product.

a. **Perform argmax attention with the following keys and values:**

Keys: $\left\{ \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$

Corresponding Values: $\left\{ \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \right\}$

using the following query:

$\mathbf{q} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$

What would be the output of the attention layer for this query?

b. Note that instead of using softmax we used argmax to generate outputs from the attention layer. **How does this design choice affect our ability to usefully train models involving attention?** (*Hint: think about how the gradients flow through the network in the backward pass. Can we learn to improve our queries or keys during the training process?*)

2 Transformer and Pretraining (4.5 points)

Transformer Architecture is illustrated in the schematic below.

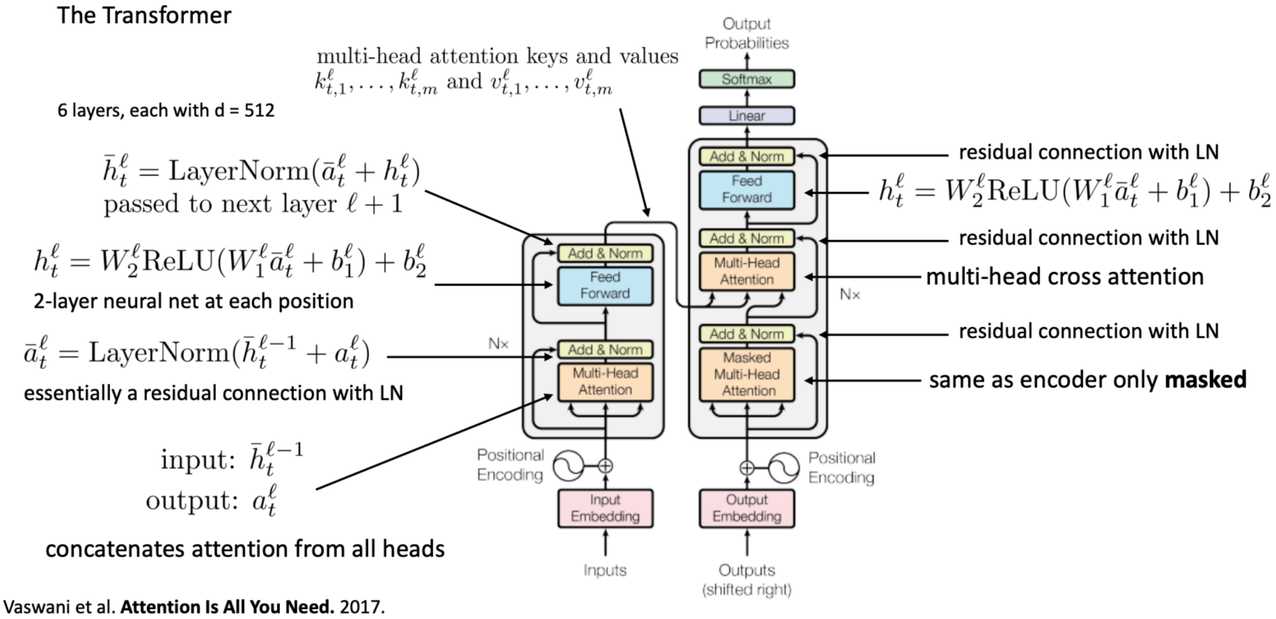
The Transformer
6 layers, each with d = 512
multi-head attention keys and values
$k_{t,1}^{\ell}, \dots, k_{t,m}^{\ell}$ and $v_{t,1}^{\ell}, \dots, v_{t,m}^{\ell}$
$\bar{h}_t^{\ell} = \mathrm{LayerNorm}(\bar{a}_t^{\ell} + h_t^{\ell})$
passed to next layer $\ell + 1$
$h_t^{\ell} = W_2^{\ell}\mathrm{ReLU}(W_1^{\ell}\bar{a}_t^{\ell} + b_1^{\ell}) + b_2^{\ell}$
2-layer neural net at each position
$\bar{a}_t^{\ell} = \mathrm{LayerNorm}(\bar{h}_t^{\ell-1} + a_t^{\ell})$
essentially a residual connection with LN
input: $\bar{h}_t^{\ell-1}$
output: a_t^{ℓ}
concatenates attention from all heads
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
residual connection with LN
$h_t^{\ell} = W_2^{\ell}\mathrm{ReLU}(W_1^{\ell}\bar{a}_t^{\ell} + b_1^{\ell}) + b_2^{\ell}$
residual connection with LN
multi-head cross attention
residual connection with LN
same as encoder only **masked**
Vaswani et al. **Attention Is All You Need.** 2017.

Figure 1: Overview of Transformer architecture

(a) Why do we need positional encoding? Describe a situation where word order information is necessary for the task performed.

(b) When using an absolute positional encoding (e.g. sinusoids at different frequencies like the hands of a clock), we can either add it to the input embedding or concatenate it. That is, if x_i is our word embedding and p_i is our positional embedding, we can either use $z = x_i + p_i$ or $z = [x_i, p_i]$. Consider a simple example where the query and key for the attention layer are both simply $q = k = z$. If we compute a dot-product of a query with another key in the attention layer, what would be the result in either case? Discuss the implications of this.

(c) It turns out we can extend the self-attention mechanics to have relative position matter without cross terms and without having to explicitly concatenate (and thereby increase the length of) two kinds of embeddings.

Relative position embedding explicitly adds a learnable set of biases π_{i-j} to the dot-product scores before the softmax operation. **For what $\pi_{i,j}$ would we get the same behavior from attention as concatenating the position embeddings $q_i^{(pos)}$, $k_j^{(pos)}$ to both the query q_i and the keys k_j?**

(d) What is the advantage of multi-headed attention? Give some examples of structures that can be found using multi-headed attention.

(e) A group of students are creating a language model, and one student suggests that they use random text from novels for pre-training. Another student says that this is just arbitrary text and is not useful because there are not any labels. **Who is right and why?**

(f) Would an encoder model or an encoder-decoder model be better suited for the following tasks?

Summarizing text in an article
Classify written restaurant reviews by their sentiment
Identifying useful pages when retrieving web search results
Translating one language to another

3 Hand-Design Transformers (4 points)

Please follow the instructions in the attached notebook "HW3.ipynb". You will implement a simple transformer model (with a single attention head) from scratch and then create a hand-designed attention head of the transformer model capable of solving a basic problem.

Once you finish with the notebook, answer the following questions in your submission of the written assignment and submit your code as well:

(a) Design a transformer that selects by contents. Compare the variables of your hand-designed Transformer with those of the learned Transformer. **Identify the similarities and differences between the two sets of variables and provide a brief explanation for each difference.**

(b) Design a transformer that selects by positions. Compare the variables of your hand-designed Transformers with those of the learned Transformer. **Identify the similarities and differences between the two sets of variables and provides a brief explanation for each difference.**