SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.040 Natural Language Processing, Fall 2024

Homework 3

Due 29 November 2024, 23:59pm

Homework 3 will be graded by Chen Huang

Previous deep learning methods use architectures like multilayer perceptron, convolutional network, and recurrent network. In recent years, Transformer-based models are the leading approach for nearly all natural language processing tasks. The Transformer model is built on the attention mechanism, which was initially designed as an improvement for encoder-decoder RNNs in sequence-to-sequence tasks like machine translation. In this homework, we will begin with the attention mechanism and gradually progress to understanding Transformers.

1 Attention Mechanisms

Consider the following: denote by $\mathcal{D} = \{(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_m, \mathbf{v}_m)\}$ a database of m tuples of *keys* and *values*. Moreover, denote by $\mathbf{q}$ a *query*. Then we can define the *attention* over $\mathcal{D}$ as

$$\text{Attention}(\mathbf{q}, \mathcal{D}) = \sum_{i=1}^{m} \alpha(\mathbf{q}, \mathbf{k}_i)\mathbf{v}_i, \tag{1}$$

where $\alpha(\mathbf{q}, \mathbf{k}_i) \in \mathcal{R}$ $(i = 1, \ldots, m)$ are scalar attention weights. This operation is commonly known as *attention pooling*. The term *attention* reflects the mechanism's ability to focus on specific elements in the dataset, assigning higher weights α to the terms in $\mathcal{D}$ that are deemed more relevant or significant. Consequently, the attention mechanism produces a weighted linear combination of the values in the database, emphasizing the most important components.

A common strategy for ensuring that the weights sum up to 1 is to normalize them via

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{\alpha(\mathbf{q}, \mathbf{k}_i)}{\sum_j \alpha(\mathbf{q}, \mathbf{k}_j)}. \tag{2}$$

In particular, to ensure that the weights are also nonnegative, one can resort to exponentiation. This means that we can now pick any function $a(\mathbf{q}, \mathbf{k})$ and then apply the softmax operation used for multinomial models to it via

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{\exp(a(\mathbf{q}, \mathbf{k}_i))}{\sum_j \exp(a(\mathbf{q}, \mathbf{k}_j))}. \tag{3}$$

Then, we need to keep the order of magnitude of the arguments in the exponential function under control. Assume that all the elements of the query $\mathbf{q} \in \mathcal{R}^d$ and the key $\mathbf{k}_i \in \mathcal{R}^d$ are independent and identically drawn random variables with zero mean and unit variance. The dot product between both vectors has zero mean and a variance of d. To ensure that the variance of the dot product still remains 1 regardless of vector length, we use the *scaled dot product attention* scoring function. That is, we rescale the dot product by $1/\sqrt{d}$. We thus arrive at the first commonly used attention function that is used:

$$a(\mathbf{q}, \mathbf{k}_i) = \frac{\mathbf{q}^\top \mathbf{k}_i}{\sqrt{d}}. \tag{4}$$

Note that attention weights α still need normalizing. We can simplify this further via equation 3 by using the softmax operation:

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \text{softmax}(a(\mathbf{q}, \mathbf{k}_i)) = \frac{\exp(\mathbf{q}^\top \mathbf{k}_i/\sqrt{d})}{\sum_{j=1}^{m} \exp(\mathbf{q}^\top \mathbf{k}_j/\sqrt{d})}. \tag{5}$$

Question 1 [code] (5 points)

One of the most popular applications of the attention mechanism is to sequence models. For example, assume that we have the following three sentences with different length:

Study	about	Deep	Learning
Start	by	code	`<blank>`
Hello	world	`<blank>`	`<blank>`

Implement the function *masked_softmax* that deals with sequences of different lengths. Then, run the sanity check cell to check your implementation.

Question 2 (9 points)

In practice, we often think of minibatches for efficiency, such as computing attention for n queries and m key-value pairs, where queries and keys are of length d and values are of length v. The scaled dot product attention of queries $\mathbf{Q} \in \mathcal{R}^{n\times d}$, keys $\mathbf{K} \in \mathcal{R}^{m\times d}$, and values $\mathbf{V} \in \mathcal{R}^{m\times v}$ thus can be written as

$$\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \in \mathcal{R}^{n\times v}. \tag{6}$$

Question 2.1 [written] (4 points) Write the shape of queries, keys and values during the calculation of scaled dot product attention. You should fill in the shape inside the code box.

Question 2.2 [code] (5 points) Implement function *DotProductAttention* that calculates the scaled dot product attention. Then, run the sanity check cell to check your implementation.

2 Attention Seq2Seq

Attention mechanisms can be effectively integrated into encoder-decoder architectures for sequence-to-sequence learning. Traditionally, in an RNN-based approach, all relevant information from the source sequence is encoded into a fixed-dimensional state representation by the encoder. However, rather than maintaining this state—represented by the context variable $\mathbf{c}$ that summarizes the source sentence—as a fixed value, it can be dynamically updated. This update is based on both the original text (encoder hidden states $\mathbf{h}_t$) and the previously generated text (decoder hidden states $\mathbf{s}_{t'-1}$). As a result, we obtain an updated context variable $\mathbf{c}_{t'}$ after each decoding time step t'. This approach allows the model to adapt the context dynamically, even for input sequences of length T, thereby improving the ability to handle long-range dependencies and capture more nuanced information from the source sequence. In this case, the context variable is the output of attention pooling:

$$\mathbf{c}_{t'} = \sum_{t=1}^{T} \alpha(\mathbf{s}_{t'-1}, \mathbf{h}_t)\mathbf{h}_t. \tag{7}$$

We used $\mathbf{s}_{t'-1}$ as the query, and $\mathbf{h}_t$ as both the key and the value. Note that $\mathbf{c}_{t'}$ is then used to generate the state $\mathbf{s}_{t'}$ and to generate a new token.

Question 3 [code] (6 points)

Implement the RNN decoder in the *Seq2SeqAttentionDecoder* class. The decoder's state is initialized using three components: (i) the hidden states of the encoder's last layer across all time steps, which are utilized as keys and values for the attention mechanism; (ii) the hidden state of the encoder's final time step at all

layers, which initializes the decoder's hidden state; and (iii) the valid length of the encoder to exclude padding tokens during attention pooling. During each decoding time step, the hidden state of the decoder's final layer from the previous step is used as the query for the attention mechanism. The attention mechanism's output is then concatenated with the input embedding to form the input for the RNN decoder, effectively guiding the generation process with context from both the source sequence and previous decoder outputs. Then, run the sanity check cell to check your implementation.

3 Multi-head attention

Rather than relying on a single attention pooling operation, the queries, keys, and values can be transformed through h independently learned linear projections. These h projected queries, keys, and values are then processed in parallel through attention pooling. Afterward, the h resulting attention outputs, known as *heads*, are concatenated and passed through another learned linear projection to generate the final output. This architecture, referred to as *multi-head attention*, allows each attention head to focus on different parts of the input, enabling the model to capture a wider range of information.

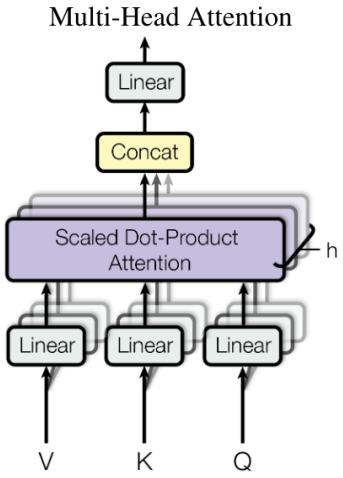

Figure 1: Multihead attention

Given a query $\mathbf{q} \in \mathcal{R}^{d_q}$, a key $\mathbf{k} \in \mathcal{R}^{d_k}$, and a value $\mathbf{v} \in \mathcal{R}^{d_v}$, each attention head $\mathbf{h}_i$ $(i = 1, \ldots, h)$ is computed as:

$$\mathbf{h}_i = f(\mathbf{W}_i^{(q)}\mathbf{q}, \mathbf{W}_i^{(k)}\mathbf{k}, \mathbf{W}_i^{(v)}\mathbf{v}) \in \mathcal{R}^{p_v}, \tag{8}$$

where $\mathbf{W}_i^{(q)} \in \mathcal{R}^{p_q \times d_q}$, $\mathbf{W}_i^{(k)} \in \mathcal{R}^{p_k \times d_k}$, and $\mathbf{W}_i^{(v)} \in \mathcal{R}^{p_v \times d_v}$ are learnable parameters and f is attention pooling. The multi-head attention output is another linear transformation via learnable parameters $\mathbf{W}_o \in \mathcal{R}^{p_o \times h p_v}$ of the concatenation of h heads:

$$\mathbf{W}_o \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_h \end{bmatrix} \in \mathcal{R}^{p_o}. \tag{9}$$

Based on this design, each head may attend to different parts of the input. More sophisticated functions than the simple weighted average can be expressed.

Question 4 [written] (4 points)

Please describe the benefits of using multi-head attention instead of single head attention.

Question 5 [code] (12 points)

In this implementation, we choose the scaled dot product attention for each head of the multi-head attention. To avoid significant growth of computational cost and parameterization cost, we set $p_q = p_k = p_v = p_o/h$. Note that h heads can be computed in parallel if we set the number of outputs of linear transformations for the query, key, and value to $p_q h = p_k h = p_v h = p_o$. In the following implementation, p_o is specified via the argument `num_hiddens`.

To allow for parallel computation of multiple heads, the *MultiHeadAttention* class uses two transposition methods *transpose_output* and *transpose_output*. Specifically, the *transpose_output* method reverses the operation of the *transpose_qkv* method.

Question 5.1 [code] (4 points) Implement function *transpose_qkv*, which is the transposition for parallel computation of multiple attention heads.

Question 5.2 [code] (4 points) Implement function *transpose_output* that reverse the operation of *transpose_qkv*.

Question 5.3 [code] (4 points) Complete *MultiHeadAttention* class. (Hint: you can use the two function you defined in question 5.1 and 5.2 .)

4 Self-Attention and Positional Encoding

In self-attention, the queries, keys, and values are represented as $n \times d$ matrices, where n is the sequence length and d is the feature dimension. The scaled dot-product attention operates by first multiplying an $n \times d$ query matrix by a $d \times n$ key matrix, producing an $n \times n$ output. This output is then multiplied by an $n \times d$ value matrix, resulting in another $n \times d$ matrix. Consequently, the self-attention mechanism has a computational complexity of $\mathcal{O}(n^2 d)$. Since each token can attend to every other token in the sequence, self-attention provides direct connections between all tokens, enabling parallel computation with $\mathcal{O}(1)$ sequential operations and a maximum path length of $\mathcal{O}(1)$. However, the quadratic complexity with respect to the sequence length (n^2) makes self-attention computationally expensive and impractical for very long sequences, significantly slowing down processing in such cases.

Unlike RNNs, which process tokens sequentially, self-attention eliminates the need for sequential operations by leveraging parallel computation. However, self-attention alone does not inherently capture the order of tokens in a sequence. So, what happens when the order of the input matters? The standard solution is to introduce *positional encodings*—additional information associated with each token to indicate its position in the sequence. These positional encodings can either be learned during training or predefined in advance. By incorporating this positional information, the model gains awareness of the token order, allowing it to preserve sequence structure while benefiting from the parallelism of self-attention.

Suppose that the input representation $\mathbf{X} \in \mathcal{R}^{n \times d}$ contains the d-dimensional embeddings for n tokens of a sequence. The positional encoding outputs $\mathbf{X} + \mathbf{P}$ using a positional embedding matrix $\mathbf{P} \in \mathcal{R}^{n \times d}$ of the same shape, whose element on the i^{th} row and the $(2j)^{\text{th}}$ or the $(2j+1)^{\text{th}}$ column is

$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right), \quad p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right). \tag{10}$$

Question 6 [code] (4 points)

Implement the *PositionalEncoding* class. Then, run the sanity check cell to check your implementation.

How to submit

1. Fill up your student ID and name in the Jupyter Notebook.
2. Click the Save button at the top of the Jupyter Notebook.
3. Select Cell - All Output - Clear. This will clear all the outputs from all cells (but will keep the content of all cells).
4. Select Cell Run All. This will run all the cells in order, and will take several minutes.
5. Once you've rerun everything, select File – Download as – PDF via LaTeX.

6. Look at the PDF file and make sure all your solutions are there, displayed correctly. **The PDF is the only thing your graders will see!**

7. Submit your PDF on eDimension.