Homework 4: Transformer

Deep Learning

Fall 2025

The goal of homework 4 is to test your understanding of attention and Transformers.

In part 1, you should submit all your answers in a pdf file. As before, we recommend using LaTeX.

For part 2, you will implement some neural networks by adding your code to the provided ipynb file.

The due date of homework 4 is 11:55pm 11/2. Submit the following files in a zip file `your_net_id.zip` through NYU classes:

- `hw4_theory.pdf`
- `hw4_impl.ipynb`

The following behaviors will result in penalty of your final score:

1. 10% penalty for submitting your file without using the correct naming format (including naming the zip file, PDF file or python file wrong, adding extra files in the zip folder, like the testing scripts in your zip file).
2. 10% penalty for every extra day of lateness. Up to 4 days max (after that we won't accept submission)
3. 20% penalty for code submission that cannot be executed following the steps we mentioned.

1 Theory (50pt)

1.1 Attention (13pts)

This question tests your intuitive understanding of attention and its property.

(a) (1pts) Given queries $\boldsymbol{Q} \in \mathbb{R}^{d \times n}$, keys $\boldsymbol{K} \in \mathbb{R}^{d \times m}$ and values $\boldsymbol{V} \in \mathbb{R}^{t \times m}$, describe the operations needed to calculate the output $\boldsymbol{H}$ of the standard dot-product

attention. What is the output dimension? (You can use the softargmax$_\beta$ function directly. It is applied to the column of each matrix).

(b) (2pts) Explain how the scale β influence the output of the attention? And what β is conveniently to use?

(c) (2pts) One advantage of the attention operation is that it is really easy to preserve a value vector $\boldsymbol{v}$ to the output $\boldsymbol{h}$. Explain in what situation, the outputs preserves the value vectors. Also, what should the scale β be if we just want the attention operation to preserve value vectors. Which of the four types of attention we are referring to? How can this be done when using fully connected architectures?

(d) (2pts) On the other hand, the attention operation can also dilute different value vectors $\boldsymbol{v}$ to generate new output $\boldsymbol{h}$. Explain in what situation the outputs is spread version of the value vectors. Also, what should the scale β be if we just want the attention operation to diffuse as much as possible. Which of the four types of attention we are referring to? How can this be done when using fully connected architectures?

(e) (2pts) If we have a small perturbation to one of the $\boldsymbol{k}_i$ (you could assume the perturbation is a zero-mean Gaussian with small variance, so the new $\hat{\boldsymbol{k}}_i = \boldsymbol{k}_i + \boldsymbol{\epsilon}$), how will the output of the $\boldsymbol{H}$ change?

(f) (2pts) If we have a small perturbation to one of the queries $\boldsymbol{q_i}$, how will the output of the $\boldsymbol{H}$ change? How would this differ from the previous case?

(g) (2pts) If we have a large perturbation that it scales one key so the $\hat{\boldsymbol{k}} = \alpha \boldsymbol{k}$ for $\alpha > 1$, how will the output of the $\boldsymbol{H}$ change?

1.2 Multi-headed Attention (3pts)

This question tests your intuitive understanding of Multi-headed Attention and its property.

(a) (1pts) Given queries $\boldsymbol{Q} \in \mathbb{R}^{d \times n}$, $\boldsymbol{K} \in \mathbb{R}^{d \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{t \times m}$, describe the operations for calculating the output $\boldsymbol{H}$ of the standard multi-headed scaled dot-product attention? Assume we have h heads.

(b) (2pts) Is there anything similar to multi-headed attention for convolutional networks? Explain why do you think they are similar.

1.3 Self Attention (11pts)

This question tests your intuitive understanding of Self Attention and its property.

(a) (2pts) Given an input $\boldsymbol{C} \in \mathbb{R}^{e \times n}$, what is the queries $\boldsymbol{Q}$, the keys $\boldsymbol{K}$ and the values $\boldsymbol{V}$ and the output $\boldsymbol{H}$ of the standard multi-headed scaled dot-product self-attention? Assume we have h heads. (You can name and define the weight matrices by yourself)

(b) (2pts) Explain what is positional encoding. What is the difference between absolute and relative positional encoding. When is it appropriate to use absolute positional encoding? When is it more appropriate to use relative encoding?

(c) (2pts) Show us one situation that the self attention layer behaves like an identity layer or permutation layer.

(d) (3pts) Show us one situation that the self attention layer behaves like an convolution layer with a kernel larger than 1. You can assume we use positional encoding.

(e) (2pts) Suppose we are training a transformer architecture for real time automatic speech recognition. Do we need to do anything special to the attention mechanism? How do we achieve this?

1.4 Transformer (15pts)

Read the original paper on the Transformer model: "Attention is All You Need" by Vaswani et al. (2017).

(a) (3pts) Explain the primary differences between the Transformer architecture and previous sequence-to-sequence models (such as RNNs and LSTMs).

(b) (3pts) Explain the concept of self-attention and its importance in the Transformer model.

(c) (3pts) Describe the multi-head attention mechanism and its benefits.

(d) (3pts) Explain the feed-forward neural networks used in the model and their purpose.

(e) (3pts) Name two techniques used in the paper to improve training stability of the transformer model, in particular regards to the issue of exploding / vanishing gradients. And briefly explain how they do so.

1.5 Vision Transformer (8pts)

Read the paper on the Transformer model: "An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale".

(a) (2pts) What is the key difference between the Vision Transformer (ViT) and traditional convolutional neural networks (CNNs) in terms of handling input images? Can you spot a convolution layer in the ViT architecture?

(b) (2pts) What is the role of positional embeddings in the Vision Transformer model, and how do they differ from positional encodings used in the original Transformer architecture?

(c) (2pts) How does the Vision Transformer model generate the final classification output? Describe the process and components involved in this step.

(d) (2pts) How does ViT compare with CNN in terms of performance across different data regimes? What explains this trend?

2 Implementation (50pt)

Please add your solutions to this notebook `HW4-VIT-Student.ipynb` . **Plase use your NYU account to access the notebook.** The notebook contains parts marked as `TODO`, where you should put your code or explanations. The notebook is a Google Colab notebook, you should copy it to your drive, add your solutions, and then download and submit it to NYU Classes. You're also free to run it on any other machine, as long as the version you send us can be run on Google Colab.