

# CS 224N: Assignment 5: Self-Attention, Transformers, and Pretraining

**Note.** Here are some things to keep in mind as you plan your time for this assignment.

- There are math questions again!
- The total amount of PyTorch code to write, and code complexity, of this assignment is lower than Assignment 4. However, you're also given less guidance or scaffolding in how to write the code.
- This assignment involves a pretraining step that takes approximately 2 hours to perform on GCP, and you'll have to do it twice. Colab set-up notebook has been provided similar to Assignment 4. The 2 hour timeline is an upper bound on the training time assuming older/slower GPU. On faster GPUs, the pretraining can finish in around 30-40 minutes.

This assignment is an investigation into Transformer self-attention building blocks, and the effects of pre-training. It covers mathematical properties of Transformers and self-attention through written questions. Further, you'll get experience with practical system-building through repurposing an existing codebase. The assignment is split into a written (mathematical) part and a coding part, with its own written questions. Here's a quick summary:

1. **Mathematical exploration:** What kinds of operations can self-attention easily implement? Why should we use fancier things like multi-headed self-attention? This section will use some mathematical investigations to illuminate a few of the motivations of self-attention and Transformer networks. **Note:** for all questions, you should justify your answer with mathematical reasoning when required.
2. **Extending a research codebase:** In this portion of the assignment, you'll get some experience and intuition for a cutting-edge research topic in NLP: teaching NLP models facts about the world through pretraining, and accessing that knowledge through finetuning. You'll train a Transformer model to attempt to answer simple questions of the form "Where was person [x] born?" – without providing any input text from which to draw the answer. You'll find that models are able to learn some facts about where people were born through pretraining, and access that information during fine-tuning to answer the questions.

Then, you'll take a harder look at the system you built, and reason about the implications and concerns about relying on such implicit pretrained knowledge.

This assignment was originally created by John Hewitt, CS 224N Head TA in Winter 2021.

## 1. Attention exploration (20 points)

Multi-head self-attention is the core modeling component of Transformers. In this question, we'll get some practice working with the self-attention equations, and motivate why multi-headed self-attention can be preferable to single-headed self-attention.

Recall that attention can be viewed as an operation on a *query* vector  $q \in \mathbb{R}^d$ , a set of *value* vectors  $\{v_1, \dots, v_n\}, v_i \in \mathbb{R}^d$ , and a set of *key* vectors  $\{k_1, \dots, k_n\}, k_i \in \mathbb{R}^d$ , specified as follows:

$$c = \sum_{i=1}^n v_i \alpha_i \quad (1)$$

$$\alpha_i = \frac{\exp(k_i^\top q)}{\sum_{j=1}^n \exp(k_j^\top q)} \quad (2)$$

with  $\alpha = \{\alpha_1, \dots, \alpha_n\}$  termed the “attention weights”. Observe that the output  $c \in \mathbb{R}^d$  is an average over the value vectors weighted with respect to  $\alpha$ .

- (a) (5 points) **Copying in attention.** One advantage of attention is that it's particularly easy to “copy” a value vector to the output  $c$ . In this problem, we'll motivate why this is the case.

- i. (1 point) **Explain** why  $\alpha$  can be interpreted as a categorical probability distribution.

**Answer:**  $\alpha$  can be interpreted as a categorical probability distribution because we got:  
 $n > 0$  number of categories,  
 $\alpha_1, \dots, \alpha_n$  event probabilities  
and  $\alpha_i > 0, \sum \alpha_i = 1$   
which are the prerequisites for a categorical probability distribution

- ii. (2 points) The distribution  $\alpha$  is typically relatively “diffuse”; the probability mass is spread out between many different  $\alpha_i$ . However, this is not always the case. **Describe** (in one sentence) under what conditions the categorical distribution  $\alpha$  puts almost all of its weight on some  $\alpha_j$ , where  $j \in \{1, \dots, n\}$  (i.e.  $\alpha_j \gg \sum_{i \neq j} \alpha_i$ ). What must be true about the query  $q$  and/or the keys  $\{k_1, \dots, k_n\}$ ?

**Answer:** It must apply  $k_j^\top q \gg k_i^\top q \forall i \neq j$

- iii. (1 point) Under the conditions you gave in (ii), **describe** the output  $c$ .

**Answer:** From i. we know  $\alpha_i > 0, \sum \alpha_i = 1$  and since  $\alpha_j \gg \sum_{i \neq j} \alpha_i$  we can approximate  $\alpha_j \approx 1$  so that  $c = \sum_{i=1}^n v_i \alpha_i \approx v_j \alpha_j \approx v_j$

- iv. (1 point) **Explain** (in two sentences or fewer) what your answer to (ii) and (iii) means intuitively.

**Answer:** If one of all key vectors is very similar or almost identical to the given query, the attention weight will be put almost exclusively on its value. Therefore, the output equals that value and in that sense we “copied” the keys value.

- (b) (7 points) **An average of two.** Instead of focusing on just one vector  $v_j$ , a Transformer model might want to incorporate information from *multiple* source vectors. Consider the case where we instead want to incorporate information from **two** vectors  $v_a$  and  $v_b$ , with corresponding key vectors  $k_a$  and  $k_b$ .

- i. (3 points) How should we combine two  $d$ -dimensional vectors  $v_a, v_b$  into one output vector  $c$  in a way that preserves information from both vectors? In machine learning, one common way to do so is to take the average:  $c = \frac{1}{2}(v_a + v_b)$ . It might seem hard to extract information about the original vectors  $v_a$  and  $v_b$  from the resulting  $c$ , but under certain conditions one can do so. In this problem, we'll see why this is the case.

Suppose that although we don't know  $v_a$  or  $v_b$ , we do know that  $v_a$  lies in a subspace  $A$  formed by the  $m$  basis vectors  $\{a_1, a_2, \dots, a_m\}$ , while  $v_b$  lies in a subspace  $B$  formed by the  $p$

basis vectors  $\{b_1, b_2, \dots, b_p\}$ . (This means that any  $v_a$  can be expressed as a linear combination of its basis vectors, as can  $v_b$ . All basis vectors have norm 1 and are orthogonal to each other.) Additionally, suppose that the two subspaces are orthogonal; i.e.  $a_j^\top b_k = 0$  for all  $j, k$ .

Using the basis vectors  $\{a_1, a_2, \dots, a_m\}$ , construct a matrix  $M$  such that for arbitrary vectors  $v_a \in A$  and  $v_b \in B$ , we can use  $M$  to extract  $v_a$  from the sum vector  $s = v_a + v_b$ . In other words, we want to construct  $M$  such that for any  $v_a, v_b$ ,  $Ms = v_a$ . Show that  $Ms = v_a$  holds for your  $M$ .

**Hint:** Given that the vectors  $\{a_1, a_2, \dots, a_m\}$  are both *orthogonal* and *form a basis* for  $v_a$ , we know that there exist some  $c_1, c_2, \dots, c_m$  such that  $v_a = c_1 a_1 + c_2 a_2 + \dots + c_m a_m$ . Can you create a vector of these weights  $c$ ?

**Answer:** First we figure out what  $M$  should look like.

To show:  $Ms = v_a$

$$\Leftrightarrow M(v_a + v_b) = v_a$$

$$\Leftrightarrow Mv_a + Mv_b = v_a \text{ For } M \text{ must apply: (i.) } Mv_a = v_a \text{ and (ii.) } Mv_b = 0$$

(i.) By deconstructing  $v_a$  we get:

$$Mv_a = M(c_1 a_1 + c_2 a_2 + \dots + c_m a_m) = MAc = v_a$$

Since all basis vectors have norm 1 and are orthogonal to each other (see text), we have  $a_j^\top a_i = 0$  with the exception  $a_j^\top a_i = 1$  if  $j = i$

Using this we can choose  $M = A^\top$  and get:

$$A^\top Ac = c_1 a_1^\top a_1 + \dots + c_m a_m^\top a_m = c$$

(Note:  $c$  is  $v_a$  expressed as a vector in  $\mathbb{R}^d$ )

(ii.) same procedure for  $v_b$ :

$$Mv_b = M(d_1 b_1 + d_2 b_2 + \dots + d_p b_p) = MBd = v_b$$

This time we take advantage of that the two subspaces  $A$  and  $B$  are orthogonal:  $a_j^\top b_k = 0 \forall j, k$

Choosing  $M = A^\top$ , we get:

$$A^\top Bd = d_1 a_1^\top b_1 + \dots + d_p a_p^\top b_p = 0$$

$M = A^\top$  satisfies both (i.) and (ii.). Inserting it in the equation  $Mv_a + Mv_b = v_a$  we get:  $A^\top Ac + A^\top Bd = c + 0 = c$

That way we created a vector of the weights  $c$  and showed that the  $M$  we are looking for is  $A^\top$ .

- ii. (4 points) As before, let  $v_a$  and  $v_b$  be two value vectors corresponding to key vectors  $k_a$  and  $k_b$ , respectively. Assume that (1) all key vectors are orthogonal, so  $k_i^\top k_j = 0$  for all  $i \neq j$ ; and (2) all key vectors have norm 1.<sup>1</sup> **Find an expression** for a query vector  $q$  such that  $c \approx \frac{1}{2}(v_a + v_b)$ , and justify your answer.<sup>2</sup>

**Answer:** From  $c \approx \frac{1}{2}(v_a + v_b)$  follows  $\alpha_a = \alpha_b = 1/2$

And from (a) we know that this means  $k_a^\top q = k_b^\top q \gg k_i^\top q \forall i \neq a, b$

Let  $k_a^\top q = k_b^\top q = \beta$ , then  $\frac{\exp(\beta)}{\sum_{j=1}^n \exp(\beta)} = \frac{\exp(\beta)}{n-2+2\exp(\beta)}$  and for  $\beta \gg 0 \Rightarrow \exp(\beta) \rightarrow \infty$  we get  $\approx \frac{\exp(\beta)}{2\exp(\beta)} = 1/2$  (see hint (2))

Therefore  $q = \beta(k_a + k_b)$  with  $\beta \gg 0$

- (c) (5 points) **Drawbacks of single-headed attention:** In the previous part, we saw how it was *possible* for a single-headed attention to focus equally on two values. The same concept could easily be extended to any subset of values. In this question we'll see why it's not a *practical* solution.

<sup>1</sup>Recall that a vector  $x$  has norm 1 iff  $x^\top x = 1$ .

<sup>2</sup>Hint: while the softmax function will never *exactly* average the two vectors, you can get close by using a large scalar multiple in the expression.

Consider a set of key vectors  $\{k_1, \dots, k_n\}$  that are now randomly sampled,  $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ , where the means  $\mu_i \in \mathbb{R}^d$  are known to you, but the covariances  $\Sigma_i$  are unknown. Further, assume that the means  $\mu_i$  are all perpendicular;  $\mu_i^\top \mu_j = 0$  if  $i \neq j$ , and unit norm,  $\|\mu_i\| = 1$ .

- i. (2 points) Assume that the covariance matrices are  $\Sigma_i = \alpha I, \forall i \in \{1, 2, \dots, n\}$ , for vanishingly small  $\alpha$ . Design a query  $q$  in terms of the  $\mu_i$  such that as before,  $c \approx \frac{1}{2}(v_a + v_b)$ , and provide a brief argument as to why it works.

**Answer:** Since  $\alpha$  is vanishingly small the the covariance matrices diagonal is vanishingly small to and  $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$  will sample  $k_i \approx \mu_i$ .

Because the means  $\mu_i$  are all perpendicular, just like the keys in (b), it will result in the same expression for  $q = \beta(\mu_a + \mu_b)$  with  $\beta \gg 0$

- ii. (3 points) Though single-headed attention is resistant to small perturbations in the keys, some types of larger perturbations may pose a bigger issue. Specifically, in some cases, one key vector  $k_a$  may be larger or smaller in norm than the others, while still pointing in the same direction as  $\mu_a$ . As an example, let us consider a covariance for item  $a$  as  $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^\top)$  for vanishingly small  $\alpha$  (as shown in figure 1). This causes  $k_a$  to point in roughly the same direction as  $\mu_a$ , but with large variances in magnitude. Further, let  $\Sigma_i = \alpha I$  for all  $i \neq a$ .

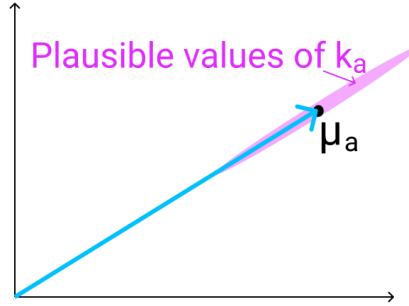


Figure 1: The vector  $\mu_a$  (shown here in 2D as an example), with the range of possible values of  $k_a$  shown in red. As mentioned previously,  $k_a$  points in roughly the same direction as  $\mu_a$ , but may have larger or smaller magnitude.

When you sample  $\{k_1, \dots, k_n\}$  multiple times, and use the  $q$  vector that you defined in part i., what do you expect the vector  $c$  will look like qualitatively for different samples? Think about how it differs from part (i) and how  $c$ 's variance would be affected.

**Answer:** From above we know  $\mu_i^\top \mu_i = 1$  and  $\alpha$  is vanishingly small, so considering the covariance for  $a$ ,  $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^\top)$  it follows that  $k_a \in [0.5\mu_a, 1.5\mu_a]$ . For all  $i \neq a$  nothing changes to the tasks before.

We can rewrite this as:

$$k_a = \gamma \mu_a \text{ where } \gamma \sim \mathcal{N}(1, 0.5)$$

$$k_i = \mu_i \quad \forall i \neq a$$

Using  $q$  we get:

$$k_a^\top q \approx \gamma \mu_a^\top \beta(\mu_a + \mu_b) \approx \gamma \beta \text{ where } \beta \gg 0$$

$$k_b^\top q \approx \mu_b^\top \beta(\mu_a + \mu_b) \approx \beta \text{ where } \beta \gg 0$$

$$k_i^\top q \approx \mu_i^\top \beta(\mu_a + \mu_b) = \beta(\mu_i^\top \mu_a + \mu_i^\top \mu_b) = \beta(0 + 0) = 0 \text{ where } \beta \gg 0$$

then the coefficients for  $v_a$  and  $v_b$  are calculated as follows

$$\text{For } v_a: \frac{\exp(k_a^\top q)}{\sum_{j=1}^n \exp(k_j^\top q)} = \frac{\exp(\gamma \beta)}{\exp(\gamma \beta) + \exp(\beta)} = \frac{1}{1 + \exp(\beta(1 - \gamma))}$$

$$\text{For } v_b: \frac{\exp(k_b^\top q)}{\sum_{j=1}^n \exp(k_j^\top q)} = \frac{\exp(\beta)}{\exp(\beta) + \exp(\gamma \beta)} = \frac{1}{1 + \exp(\beta(\gamma - 1))}$$

Next we have a look at the behaviour at the boundaries of  $\gamma \in [0.5, 1.5]$

We can see from the final term, that it is a shifted softmax function to the right by one. While  $\beta$

only controls the steepness,  $\gamma$  dictates if the value approaches 0 or 1. In our case the values of the coefficients will be mirrored because  $\gamma$  is sampled from  $[0.5, 1.5]$  and either subtracted by 1 or from 1.

Mathematically:

For  $\gamma \rightarrow 0.5 \wedge \beta \gg 0$ :  $\frac{1}{1+\exp(\beta(1-0.5))} \approx \frac{1}{1+\infty} \approx 0$  while  $\frac{1}{1+\exp(\beta(0.5-1))} \approx \frac{1}{1+0} \approx 1$

For  $\gamma \rightarrow 1.5 \wedge \beta \gg 0$ :  $\frac{1}{1+\exp(\beta(1-1.5))} \approx \frac{1}{1+0} \approx 1$  while  $\frac{1}{1+\exp(\beta(1.5-1))} \approx \frac{1}{1+\infty} \approx 0$

Therefore  $c = v_a$  when  $\gamma \rightarrow 1.5 \wedge \beta \gg 0$  and  $c = v_b$  when  $\gamma \rightarrow 0.5 \wedge \beta \gg 0$

In (i.)  $c = 0.5v_a + 0.5v_b$  and is therefore always an evenly weighted combination of both. Now  $c$  oscillates between  $v_a$  and  $v_b$ .

- (d) (3 points) **Benefits of multi-headed attention:** Now we'll see some of the power of multi-headed attention. We'll consider a simple version of multi-headed attention which is identical to single-headed self-attention as we've presented it in this homework, except two query vectors ( $q_1$  and  $q_2$ ) are defined, which leads to a pair of vectors ( $c_1$  and  $c_2$ ), each the output of single-headed attention given its respective query vector. The final output of the multi-headed attention is their average,  $\frac{1}{2}(c_1 + c_2)$ . As in question 1(c), consider a set of key vectors  $\{k_1, \dots, k_n\}$  that are randomly sampled,  $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ , where the means  $\mu_i$  are known to you, but the covariances  $\Sigma_i$  are unknown. Also as before, assume that the means  $\mu_i$  are mutually orthogonal;  $\mu_i^\top \mu_j = 0$  if  $i \neq j$ , and unit norm,  $\|\mu_i\| = 1$ .

- i. (1 point) Assume that the covariance matrices are  $\Sigma_i = \alpha I$ , for vanishingly small  $\alpha$ . Design  $q_1$  and  $q_2$  such that  $c$  is approximately equal to  $\frac{1}{2}(v_a + v_b)$ . Note that  $q_1$  and  $q_2$  should have different expressions.

**Answer:** In (c) (i.) we designed  $q = \beta(\mu_a + \mu_b)$  such that  $c = \frac{1}{2}(v_a + v_b)$

We exploit that the final  $c$  is again the average of the terms  $c_1$  and  $c_2$ :

Let  $c_1 = \frac{1}{2}(v_a + v_b)$  for  $q_1 = \beta(\mu_a + \mu_b)$  and  $c_2 = \frac{1}{2}(v_a + v_b)$  for  $q_2 = \beta(\mu_a + \mu_b)$  we get:  
 $c = \frac{1}{2}(c_1 + c_2) = \frac{1}{2}(\frac{1}{2}(v_a + v_b) + \frac{1}{2}(v_a + v_b)) = \frac{1}{4}(v_a + v_b) + \frac{1}{4}(v_a + v_b) = \frac{1}{2}(v_a + v_b)$

- ii. (2 points) Assume that the covariance matrices are  $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^\top)$  for vanishingly small  $\alpha$ , and  $\Sigma_i = \alpha I$  for all  $i \neq a$ . Take the query vectors  $q_1$  and  $q_2$  that you designed in part i. What, qualitatively, do you expect the output  $c$  to look like across different samples of the key vectors? Explain briefly in terms of variance in  $c_1$  and  $c_2$ . You can ignore cases in which  $k_a^\top q_i < 0$ .

**Answer:** Because we chose  $q = \beta(\mu_a + \mu_b)$  before, we get  $c = \frac{1}{1+\exp(\beta(1-\gamma))} v_a + \frac{1}{1+\exp(\beta(\gamma-1))} v_b$  just like in (c) (ii.). By adding more attention heads,  $c$  won't oscillate as much between  $v_a$  and  $v_b$  as before because  $\gamma \sim \mathcal{N}(1, 0.5)$  approaches its mean 1.

So  $c \approx \frac{1}{1+\exp(\beta(1-1))} v_a + \frac{1}{1+\exp(\beta(1-1))} v_b = \frac{1}{1+1} v_a + \frac{1}{1+1} v_b = \frac{1}{2}(v_a + v_b)$  for  $\gamma \rightarrow 1$

## 2. Pretrained Transformer models and knowledge access (35 points)

You'll train a Transformer to perform a task that involves accessing knowledge about the world — knowledge which isn't provided via the task's training data (at least if you want to generalize outside the training set). You'll find that it more or less fails entirely at the task. You'll then learn how to pretrain that Transformer on Wikipedia text that contains world knowledge, and find that finetuning that Transformer on the same knowledge-intensive task enables the model to access some of the knowledge learned at pretraining time. You'll find that this enables models to perform considerably above chance on a held out development set.

The code you're provided with is a fork of Andrej Karpathy's [minGPT](#). It's nicer than most research code in that it's relatively simple and transparent. The "GPT" in minGPT refers to the Transformer language model of OpenAI, originally described in [this paper](#) [2].

As in previous assignments, you will want to develop on your machine locally, then run training on Azure/Colab. You can use the same conda environment from previous assignments for local development,

and the same process for training on a GPU.<sup>3</sup> You'll need around 5 hours for training, so budget your time accordingly! We have provided a sample Colab with the the commands that require GPU training. **Note that dataset multi-processing can fail on local machines without GPU, so to debug locally, you might have to change `num_workers` to 0.**

Your work with this codebase is as follows:

- (a) (0 points) **Check out the demo.**

In the `mingpt-demo/` folder is a Jupyter notebook `play_char.ipynb` that trains and samples from a Transformer language model. Take a look at it (locally on your computer) to get somewhat familiar with how it defines and trains models. Some of the code you're writing below will be inspired by what you see in this notebook.

Note that you do not have to write any code or submit written answers for this part.

- (b) (0 points) **Read through `NameDataset` in `src/dataset.py`, our dataset for reading name-birthplace pairs.**

The task we'll be working on with our pretrained models is attempting to access the birth place of a notable person, as written in their Wikipedia page. We'll think of this as a particularly simple form of question answering:

*Q: Where was [person] born?*

*A: [place]*

From now on, you'll be working with the `src/` folder. **The code in `mingpt-demo/` won't be changed or evaluated for this assignment.** In `dataset.py`, you'll find the the class `NameDataset`, which reads a TSV (tab-separated values) file of name/place pairs and produces examples of the above form that we can feed to our Transformer model.

To get a sense of the examples we'll be working with, if you run the following code, it'll load your `NameDataset` on the training set `birth_places_train.tsv` and print out a few examples.

```
python src/dataset.py namedata
```

Note that you do not have to write any code or submit written answers for this part.

- (c) (0 points) **Implement finetuning (without pretraining).**

Take a look at `run.py`. It has some skeleton code specifying flags you'll eventually need to handle as command line arguments. In particular, you might want to *pretrain*, *finetune*, or *evaluate* a model with this code. For now, we'll focus on the finetuning function, in the case without pretraining.

Taking inspiration from the training code in the `play_char.ipynb` file, write code to finetune a Transformer model on the name/birthplace dataset, via examples from the `NameDataset` class. For now, implement the case without pretraining (i.e. create a model from scratch and train it on the birthplace prediction task from part (b)). You'll have to modify two sections, marked [part c] in the code: one to initialize the model, and one to finetune it. Note that you only need to initialize the model in the case labeled "vanilla" for now (later in section (g), we will explore a model variant). Use the hyperparameters for the `Trainer` specified in the `run.py` code.

Also take a look at the *evaluation* code which has been implemented for you. It samples predictions from the trained model and calls `evaluate_places()` to get the total percentage of correct place predictions. You will run this code in part (d) to evaluate your trained models.

This is an intermediate step for later portions, including Part d, which contains commands you can run to check your implementation. No written answer is required for this part.

- (d) (5 points) **Make predictions (without pretraining).**

Train your model on `birth_places_train.tsv`, and evaluate on `birth_dev.tsv`. Specifically, you should now be able to run the following three commands:

---

<sup>3</sup>See [CS224n Azure Guide](#) for a refresher on Azure.

```
# Train on the names dataset
python src/run.py finetune vanilla wiki.txt \
    --writing_params_path vanilla.model.params \
    --finetune_corpus_path birth_places_train.tsv

# Evaluate on the dev set, writing out predictions
python src/run.py evaluate vanilla wiki.txt \
    --reading_params_path vanilla.model.params \
    --eval_corpus_path birth_dev.tsv \
    --outputs_path vanilla.nopretrain.dev.predictions

# Evaluate on the test set, writing out predictions
python src/run.py evaluate vanilla wiki.txt \
    --reading_params_path vanilla.model.params \
    --eval_corpus_path birth_test_inputs.tsv \
    --outputs_path vanilla.nopretrain.test.predictions
```

Training will take less than 10 minutes (on Azure). Report your model's accuracy on the dev set (as printed by the second command above). Similar to assignment 4, we also have Tensorboard logging in assignment 5 for debugging. It can be launched using `tensorboard --logdir expt/`. Don't be surprised if it is well below 10%; we will be digging into why in Part 3. As a reference point, we want to also calculate the accuracy the model would have achieved if it had just predicted "London" as the birth place for everyone in the dev set. Fill in `london.baseline.py` to calculate the accuracy of that approach and report your result in your write-up. You should be able to leverage existing code such that the file is only a few lines long.

**Answer:**

Model accuracy: Correct: 10.0 out of 500.0: 2.0%

"London" accuracy: Correct: 25.0 out of 500.0: 5.0%

(e) (10 points) **Define a *span corruption* function for pretraining.**

In the file `src/dataset.py`, implement the `__getitem__()` function for the dataset class `CharCorruptionDataset`. Follow the instructions provided in the comments in `dataset.py`. Span corruption is explored in the [T5 paper](#) [3]. It randomly selects spans of text in a document and replaces them with unique tokens (noising). Models take this noised text, and are required to output a pattern of each unique sentinel followed by the tokens that were replaced by that sentinel in the input. In this question, you'll implement a simplification that only masks out a single sequence of characters.

This question will be graded via autograder based on whether your span corruption function implements some basic properties of our spec. We'll instantiate the `CharCorruptionDataset` with our own data, and draw examples from it.

To help you debug, if you run the following code, it'll sample a few examples from your `CharCorruptionDataset` on the pretraining dataset `wiki.txt` and print them out for you.

```
python src/dataset.py charcorruption
```

No written answer is required for this part.

(f) (10 points) **Pretrain, finetune, and make predictions. Budget 2 hours for training.**

Now fill in the *pretrain* portion of `run.py`, which will pretrain a model on the span corruption task. Additionally, modify your *finetune* portion to handle finetuning in the case *with* pretraining. In particular, if a path to a pretrained model is provided in the bash command, load this model before finetuning it on the birthplace prediction task. Pretrain your model on `wiki.txt` (which should



take approximately two hours), finetune it on NameDataset and evaluate it. Specifically, you should be able to run the following four commands: (Don't be concerned if the loss appears to plateau in the middle of pretraining; it will eventually go back down.)

```
# Pretrain the model
python src/run.py pretrain vanilla wiki.txt \
    --writing_params_path vanilla.pretrain.params

# Finetune the model
python src/run.py finetune vanilla wiki.txt \
    --reading_params_path vanilla.pretrain.params \
    --writing_params_path vanilla.finetune.params \
    --finetune_corpus_path birth_places_train.tsv

# Evaluate on the dev set; write to disk
python src/run.py evaluate vanilla wiki.txt \
    --reading_params_path vanilla.finetune.params \
    --eval_corpus_path birth_dev.tsv \
    --outputs_path vanilla.pretrain.dev.predictions

# Evaluate on the test set; write to disk
python src/run.py evaluate vanilla wiki.txt \
    --reading_params_path vanilla.finetune.params \
    --eval_corpus_path birth_test_inputs.tsv \
    --outputs_path vanilla.pretrain.test.predictions
```

Report the accuracy on the dev set (printed by the third command above). We expect the dev accuracy will be at least 10%, and will expect a similar accuracy on the held out test set.

**Answer:**

- (g) (10 points) **Research! Write and try out a more efficient variant of Attention (Budget 2 hours for pretraining!)**

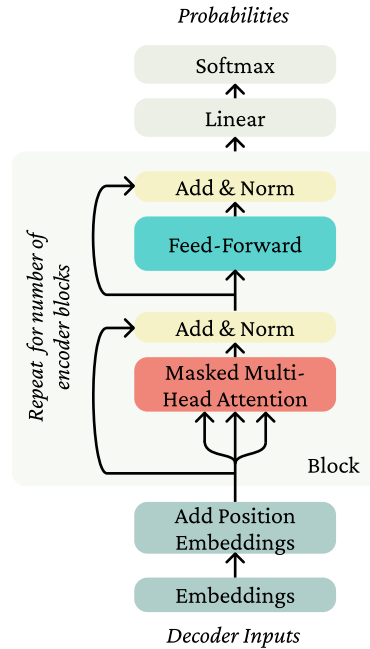
We'll now go to changing the Transformer architecture itself – specifically the first and last transformer blocks. The transformer model uses a self-attention scoring function based on dot products, this involves a rather intensive computation that's quadratic in the sequence length. This is because the dot product between  $\ell^2$  pairs of word vectors is computed in each computation, where  $\ell$  is the sequence length. If we can reduce the length of the sequence passed on the self-attention module, we should observe significant reduction in compute. For example, if we develop a technique that can reduce the sequence length to half, we can save around 75% of the compute time!

PerceiverAR [1] proposes a solution to make the model more efficient by reducing the sequence length of the input to self-attention for the intermediate layers. In the first layer, the input sequence is projected onto a lower-dimensional basis. Subsequently, all self-attention layers operate in this smaller subspace. The last layer projects the output back to the original input sequence length. In this assignment, we propose a simpler version of the PerceiverAR transformer model.

The provided CausalSelfAttention layer implements the following attention for each head of the multi-headed attention: Let  $X \in \mathbb{R}^{\ell \times d}$  (where  $\ell$  is the block size and  $d$  is the total dimensionality,  $d/h$  is the dimensionality per head).<sup>4</sup>

<sup>4</sup>Note that these dimensionalities do not include the minibatch dimension.





### Transformer Decoder

Figure 2: Illustration of the transformer block.

Let  $Q_i, K_i, V_i \in \mathbb{R}^{d \times d/h}$ . Then the output of the self-attention head is

$$Y_i = \text{softmax}\left(\frac{(XQ_i)(XK_i)^\top}{\sqrt{d/h}}\right)(XV_i) \quad (3)$$

where  $Y_i \in \mathbb{R}^{\ell \times d/h}$ . Then the output of the self-attention is a linear transformation of the concatenation of the heads:

$$Y = [Y_1; \dots; Y_h]A \quad (4)$$

where  $A \in \mathbb{R}^{d \times d}$  and  $[Y_1; \dots; Y_h] \in \mathbb{R}^{\ell \times d}$ . The code also includes dropout layers which we haven't written here. We suggest looking at the provided code and noting how this equation is implemented in PyTorch.

Our model uses this self-attention layer in the transformer block as shown in Figure 2. As discussed in the lecture, the transformer block contains residual connections and layer normalization layers. If we compare this diagram with the `Block` code provided in `model.py`, we notice that the implementation does not perform layer normalization on the output of the MLP (Feed-Forward), but on the input of the `Block`. This can be considered equivalent since we have a series of transformer blocks on top of each other.

In the Perceiver model architecture, we replace the first transformer `Block` in the model with the `DownProjectBlock`. This block reduces the length of the sequence from  $\ell$  to  $m$ . This is followed by a series of regular transformer blocks, which would now perform self-attention on the reduced sequence length of  $m$ . We replace the last block of the model with the `UpProjectBlock`, which takes in the  $m$  length output of the previous block, and projects it back to the original sequence length of  $\ell$ .

You need to implement the `DownProjectBlock` in `model.py` that reduces the dimensionality of the sequence in the first block. To do this, perform cross-attention on the input sequence with a

learnable basis  $C \in \mathbb{R}^{m \times d}$  as the query, where  $m < \ell$ . Consequently, Equation 3 becomes:

$$Y_i^{(1)} = \text{softmax}\left(\frac{(CQ_i)(XK_i)^\top}{\sqrt{d/h}}\right)(XV_i) \quad (5)$$

resulting in  $Y_i^{(1)} \in \mathbb{R}^{m \times d}$ , with  $(1)$  denoting that the output corresponds to the first layer. With this dimensionality reduction, the subsequent `CausalSelfAttention` layers operate on inputs  $\in \mathbb{R}^{m \times d}$  instead of  $\mathbb{R}^{\ell \times d}$ . We refer to  $m$  as the `bottleneck_dim` in code. Note that for implementing Equation 5, we need to perform cross attention between the learnable basis  $C$  and the input sequence. This has been provided to you as the `CausalCrossAttention` layer. We recommend reading through `attention.py` to understand how to use the cross-attention layer, and map which arguments correspond to the key, value and query inputs. Initialize the basis vector matrix  $C$  using Xavier Uniform initialization.

To get back to the original dimensions, the last block in the model is replaced with the `UpProjectBlock`. This block will bring back the output sequence length to be the same as input sequence length by performing cross-attention on the previous layer's output  $Y^{L-1}$  with the original input vector  $X$  as the query:

$$Y_i^{(L)} = \text{softmax}\left(\frac{(XQ_i)(Y^{(L-1)}K_i)^\top}{\sqrt{d/h}}\right)(Y^{(L-1)}V_i) \quad (6)$$

where  $L$  is the total number of layers. This results in the final output vector having the same dimension as expected in the original `CausalSelfAttention` mechanism. Implement this functionality in the `UpProjectBlock` in `model.py`.

We provide the code to assemble the model using your implemented `DownProjectBlock` and `UpProjectBlock`. The model uses these blocks when the `variant` parameter is specified as `perceiver`.

Below are bash commands that your code should support in order to pretrain the model, finetune it, and make predictions on the dev and test sets. Note that the pretraining process will take approximately 2 hours.

```

# Pretrain the model
python src/run.py pretrain perceiver wiki.txt --bottleneck_dim 64 \
    --pretrain_lr 6e-3 --writing_params_path perceiver.pretrain.params

# Finetune the model
python src/run.py finetune perceiver wiki.txt --bottleneck_dim 64 \
    --reading_params_path perceiver.pretrain.params \
    --writing_params_path perceiver.finetune.params \
    --finetune_corpus_path birth_places_train.tsv

# Evaluate on the dev set; write to disk
python src/run.py evaluate perceiver wiki.txt --bottleneck_dim 64 \
    --reading_params_path perceiver.finetune.params \
    --eval_corpus_path birth_dev.tsv \
    --outputs_path perceiver.pretrain.dev.predictions

# Evaluate on the test set; write to disk
python src/run.py evaluate perceiver wiki.txt --bottleneck_dim 64 \
    --reading_params_path perceiver.finetune.params \
    --eval_corpus_path birth_test_inputs.tsv \
    --outputs_path perceiver.pretrain.test.predictions

```

Report the accuracy of your perceiver attention model on birthplace prediction on `birth_dev.tsv` after pretraining and fine-tuning.

Save the predictions of the model on `birth_test_inputs.tsv` to `perceiver.pretrain.test.predictions`.

For this section, you'll submit: `perceiver.finetune.params`, `perceiver.pretrain.dev.predictions`, and `perceiver.pretrain.test.predictions`. Your model should get at least 6% accuracy on the dev set.

- i. (8 points) We'll score your model as to whether it gets at least 5% accuracy on the test set, which has answers held out.
- ii. (2 points) Provide an expression for the time complexity of the Perceiver model and the vanilla model, in terms of number of layers ( $L$ ), input sequence length ( $\ell$ ) and basis bottleneck dimension ( $m$ ).

**Answer:** The Attention layers are the crucial part of the perceiver approach. If we denote the input dimensionality as  $d$ , the complexity of the QKV attention operation is  $\mathcal{O}(d \times \ell)$ . However, if we use cross-attention,  $K$  and  $V$  are just projections of the input, while the  $Q$  is a projection of a learned latent vector reduced by  $m$ . The dimension of this latent vector is  $m$ , and we consider that  $m \ll \ell$ . That way the complexity of attention operation is reduced to  $\mathcal{O}(d \times m)$ . In the same way the complexity of self attentions in the latent transformer blocks will reduce to  $\mathcal{O}(m^2)$ .

So while multi-headed attention has a time complexity of  $\mathcal{O}(\ell^2 d + \ell d^2)$  the time complexity of the perceiver model is  $\mathcal{O}(dm + Lm^2)$ , where  $d$  is the dimensionality of the byte array,  $m$  is the dimensionality of the latent array, and  $L$  is the depth of the transformer.

### 3. Considerations in pretrained knowledge (5 points)

Please type the answers to these written questions (to make TA lives easier).

- (a) (1 point) Succinctly explain why the pretrained (vanilla) model was able to achieve an accuracy of above 10%, whereas the non-pretrained model was not.

**Answer:** The pretrained model achieved higher accuracy because it had prior knowledge from being trained on a large dataset. It learned patterns and features during pretraining, enabling it to generalize better on new examples. In contrast, the non-pretrained model started from scratch, lacking the initial knowledge and taking longer to learn meaningful representations, resulting in lower accuracy.

- (b) (2 points) Take a look at some of the correct predictions of the pretrain+finetuned vanilla model, as well as some of the errors. We think you'll find that it's impossible to tell, just looking at the output, whether the model *retrieved* the correct birth place, or *made up* an incorrect birth place. Consider the implications of this for user-facing systems that involve pretrained NLP components. Come up with two **distinct** reasons why this model behavior (i.e. unable to tell whether it's retrieved or made up) may cause concern for such applications, and an example for each reason.

**Answer:**

1. **Misinformation and Trust Issues:** If the model cannot reliably distinguish between correct and fabricated information, it can inadvertently provide users with incorrect or misleading data. This can lead to the spread of misinformation and erode trust in the system. For example, in a news recommendation system, if the model generates false birthplaces for public figures, users may unknowingly consume and share incorrect information, leading to confusion and a loss of credibility.
  2. **Legal and Ethical Implications:** In certain scenarios, the inability to verify the accuracy of retrieved information can have legal and ethical consequences. For instance, in a legal research application that relies on the model to retrieve historical facts, if the model cannot distinguish between actual birthplaces and made-up ones, it may inadvertently provide inaccurate information to lawyers, researchers, or judges.
- (c) (2 points) If your model didn't see a person's name at pretraining time, and that person was not seen at fine-tuning time either, it is not possible for it to have "learned" where they lived. Yet, your model will produce *something* as a predicted birth place for that person's name if asked. Concisely describe a strategy your model might take for predicting a birth place for that person's name, and one reason why this should cause concern for the use of such applications. (You do not need to submit the same answer for 3c as for 3b.)

**Answer:** The model might use associative inference by leveraging patterns and correlations it learned during pretraining. For example, if the model has seen other individuals from a particular country or city who share similar attributes as the person in question, it may associate that birthplace with the given name. This strategy can be problematic because the model's predictions are based on associations rather than actual knowledge and cause concerns as mentioned in 3b.

## Submission Instructions

You will submit this assignment on GradeScope as two submissions – one for **Assignment 5 [coding]** and another for **Assignment 5 [written]**:

1. Verify that the following files exist at these specified paths within your assignment directory:
  - The no-pretraining model and predictions: `vanilla.model.params`, `vanilla.nopretrain.dev.predictions`, `vanilla.nopretrain.test.predictions`
  - The pretrain-finetune model and predictions: `vanilla.finetune.params`, `vanilla.pretrain.dev.predictions`, `vanilla.pretrain.test.predictions`
  - The Perceiver model and predictions: `perceiver.finetune.params`, `perceiver.pretrain.dev.predictions`, `perceiver.pretrain.test.predictions`
2. Run the `collect_submission.sh` script to produce your `assignment5.zip` file.

3. Upload your assignment5.zip file to GradeScope to **Assignment 5 [coding]**.
4. Check that the public autograder tests passed correctly.
5. Upload your written solutions, for questions 1, parts of 2, and 3, to GradeScope to **Assignment 5 [written]**. Tag it properly!

## References

- [1] HAWTHORNE, C., JAEGLE, A., CANGA, C., BORGEAUD, S., NASH, C., MALINOWSKI, M., DIELEMAN, S., VINYALS, O., BOTVINICK, M. M., SIMON, I., SHEAHAN, H., ZEGHIDOUR, N., ALAYRAC, J., CARREIRA, J., AND ENGEL, J. H. General-purpose, long-context autoregressive modeling with perceiver AR. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA* (2022), vol. 162 of *Proceedings of Machine Learning Research*, pp. 8535–8558.
- [2] RADFORD, A., NARASIMHAN, K., SALIMANS, T., AND SUTSKEVER, I. Improving language understanding with unsupervised learning. *Technical report, OpenAI* (2018).
- [3] RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W., AND LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.