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 Azure, 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 pretraining. 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 $\{v_1, \ldots, v_n\}, v_i \in \mathbb{R}^d$, and a set of *key* vectors $\{k_1, \ldots, k_n\}, k_i \in \mathbb{R}^d$, specified as follows:

$$c = \sum_{i=1}^{n} v_i \alpha_i \tag{1}$$

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

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

(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 α can be interpreted as a categorical probability distribution.
Answer: There are n α scores - one for each value in a sequence. Each score is between 0 and 1 and it can be interpreted as probability for a category. It is a probability distribution because all scores are normalized (via softmax), i.e., they sum up to 1.

ii. (2 points) The distribution α is typically relatively "diffuse"; the probability mass is spread out between many different α_i. However, this is not always the case. **Describe** (in one sentence) under what conditions the categorical distribution α puts almost all of its weight on some α_j, where $j \in \{1, \ldots, n\}$ (i.e. $\alpha_j \gg \sum_{i \neq j} \alpha_i$). What must be true about the query q and/or the keys $\{k_1, \ldots, k_n\}$?
Answer: The categorical distribution α puts almost all of its weight on some α_j (i.e. $\alpha_j \gg \sum_{i \neq j} \alpha_i$) if the dot product of the key k_j and the query q gives much higher value than the other dot product. This means the key k_j is much closer to the query q in the d-dimension than other keys.

iii. (1 point) Under the conditions you gave in (ii), **describe** the output c.
Answer: The output c now will almost be a copy of the value v_j due to the fact that $\alpha_j \gg \sum_{i \neq j} \alpha_i$, i.e., $c \approx v_j$.

iv. (1 point) **Explain** (in two sentences or fewer) what your answer to (ii) and (iii) means intuitively.
Answer: If the dot product between some j_{th} words' *key* and *query* is very large compared to other words' *keys* and the same *query*, then the *attention output* for that j_{th} word will approach its value. It's as if the *value* is "copied" to the output.

(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 $\{a_1, a_2, \ldots, a_m\}$, while v_b lies in a subspace B formed by the p basis vectors $\{b_1, b_2, \ldots, b_p\}$. (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 $\{a_1, a_2, \ldots, a_m\}$, 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 $\{a_1, a_2, \ldots, a_m\}$ are both *orthogonal* and *form a basis* for v_a, we know that there exist some $c_1, c_2, \ldots, c_m$ such that $v_a = c_1 a_1 + c_2 a_2 + \cdots + c_m a_m$. Can you create a vector of these weights c?
Answer: Assume that $\mathbf{A}$ is a matrix of concatenated basis vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ and $\mathbf{B}$ is a matrix of concatenated basis vectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$. Linear combinations of vectors $\mathbf{v}_a$ and $\mathbf{v}_b$ can then be expressed as:

$$\mathbf{v}_a = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \ldots + c_m\mathbf{a}_m = \mathbf{Ac}$$

$$\mathbf{v}_b = d_1\mathbf{b}_1 + d_2\mathbf{b}_2 + \ldots + d_m\mathbf{b}_m = \mathbf{Bd}$$

We have to construct a matrix $\mathbf{M}$ such that when multiplied with $\mathbf{v}_b$, produces $\underline{0}$ and multiplied with $\mathbf{v}_a$, produces the same vector (in terms of its own space):

$$\mathbf{Ms} = \mathbf{v}_a$$

$$\mathbf{M}(\mathbf{v}_a + \mathbf{v}_b) = \mathbf{v}_a$$

$$\mathbf{Mv}_a + \mathbf{Mv}_b = \mathbf{v}_a$$

It is easy to see that, since $\mathbf{a}_j^T\mathbf{b}_k = 0$ for all j, k (two subspaces $\mathbf{A}$ and $\mathbf{B}$ are orthogonal to each other), $\mathbf{A}^T\mathbf{B} = \mathbf{0}$. Also, since $\mathbf{a}_i^T\mathbf{a}_j = 0$ for $i \neq j$ and $\mathbf{a}_i^T\mathbf{a}_j = 1$ whenever $i = j$ (all basis vectors have norm 1 and are orthogonal to each other), $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. If we substitute $\mathbf{M}$ with $\mathbf{A}^T$, $\mathbf{v}_a$ with $\mathbf{Ac}$ and $\mathbf{v}_b$ with $\mathbf{Bd}$:

$$\mathbf{A}^T\mathbf{Ac} + \mathbf{A}^T\mathbf{Bd} = \mathbf{Ic} + \mathbf{0d} = \mathbf{c}$$

We know that in terms of $\mathbb{R}^d$ (not in terms of $\mathbf{A}$ or $\mathbf{B}$), $\mathbf{v}_a$ is just a collection of $\mathbf{c}$ (or we can think of $\mathbf{v}_a$ as $\mathbf{c}$ expressed as a vector in $\mathbb{R}^d$). Thus, $\mathbf{M} = \mathbf{A}^T$

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.[1] **Find an expression** for a query vector q such that $c \approx \frac{1}{2}(v_a + v_b)$, and justify your answer. [2]
Answer: Assume that $\mathbf{c}$ is approximated as follows:

$$\mathbf{c} \approx 0.5\mathbf{v}_a + 0.5\mathbf{v}_b$$

This means that $\alpha_a \approx 0.5$ and $\alpha_b \approx 0.5$, which can be achieved when (whenever $i \neq a$ and $i \neq b$):

$$\mathbf{k}_a^\top\mathbf{q} \approx \mathbf{k}_b^\top\mathbf{q} \gg \mathbf{k}_i^\top\mathbf{q}$$

Like explained in the previous question, if the dot product is big, the probability mass will also be big and we want a balanced mass between α_a and α_b. $\mathbf{q}$ will be largest for $\mathbf{k}_a$ and $\mathbf{k}_b$

[1] Recall that a vector x has norm 1 iff $x^\top x = 1$.
[2] 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.

when it is a large multiplicative of a vector that contains a component in $\mathbf{k}_a$ direction and in $\mathbf{k}_b$ direction:

$$\mathbf{q} = \beta(\mathbf{k}_a + \mathbf{k}_b), \quad \text{where } \beta \gg 0$$

Now, since the keys are orthogonal to each other, it is easy to see that:

$$\mathbf{k}_a^\top \mathbf{q} = \beta \mathbf{k}_a^\top \mathbf{k}_a + \beta \mathbf{k}_a^\top \mathbf{k}_b = \beta \times 1 + \beta \times 0 = \beta$$

$$\mathbf{k}_b^\top \mathbf{q} = \beta \mathbf{k}_b^\top \mathbf{k}_a + \beta \mathbf{k}_b^\top \mathbf{k}_b = \beta \times 0 + \beta \times 1 = \beta$$

$$\mathbf{k}_i^\top \mathbf{q} = \beta \mathbf{k}_i^\top \mathbf{k}_a + \beta \mathbf{k}_i^\top \mathbf{k}_b = \beta \times 0 + \beta \times 0 = 0$$

Thus when we exponentiate, only $\exp(\beta)$ will be matter, because $\exp(0) = 1$ will be insignificant to the probability mass. We get that:

$$\alpha_a = \alpha_b = \frac{\exp(\beta)}{n - 2 + 2\exp(\beta)} \approx \frac{\exp(\beta)}{2\exp(\beta)} \approx \frac{1}{2}, \quad \text{for } \beta \gg 0$$

(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. Consider a set of key vectors $\{k_1, \ldots, k_n\}$ 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 Σ_i are unknown. Further, assume that the means μ_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 $\Sigma_i = \alpha I, \forall i \in \{1, 2, \ldots, n\}$, for vanishingly small α. Design a query q in terms of the μ_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 the variances (diagonal covariance values) for $i \in \{1, 2, \ldots, n\}$ are vanishingly small, we can assume each key vector will be close to its mean vector:

$$\mathbf{k}_i \approx \mu_i$$

Because all the mean vectors are perpendicular, the problem reduces to the previous case when all keys were perpendicular to each other. $\mathbf{q}$ can now be expressed as:

$$\mathbf{q} = \beta(\mu_a + \mu_b), \quad \text{where } \beta \gg 0$$

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 μ_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 α (as shown in figure 1). This causes k_a to point in roughly the same direction as μ_a, but with large variances in magnitude. Further, let $\Sigma_i = \alpha I$ for all $i \neq a$.

When you sample $\{k_1, \ldots, k_n\}$ 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: Since $\mu_i^\top \mu_i = 1$, $\mathbf{k}_a$ varies between $(\alpha + 0.5)\mu_a$ and $(\alpha + 1.5)\mu_a$. All other $\mathbf{k}_i$, whenever $i \neq a$, almost don't vary at all. Noting that α is vanishingly small:

$$\mathbf{k}_a \approx \gamma \mu_a, \quad \text{where } \gamma \sim \mathcal{N}(1, 0.5)$$

$$\mathbf{k}_i \approx \mu_i, \quad \text{whenever } i \neq a$$

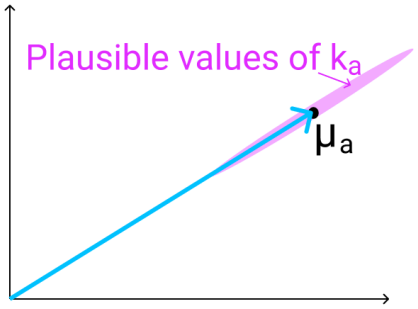


Figure 1: The vector μ_a (shown here in 2D as an example), with the range of possible values of k_a shown in red. As mentioned previously, k_a points in roughly the same direction as μ_a, but may have larger or smaller magnitude.

Since $\mathbf{q}$ is most similar in directions $\mathbf{k}_a$ and $\mathbf{k}_b$, we can assume that the dot product between $\mathbf{q}$ and any other key vector is 0 (since all key vectors are orthogonal). Thus there are 2 cases to consider (note that means are normalized and orthogonal to each other):

$$\mathbf{k}_a^\top \mathbf{q} \approx \gamma \mu_a^\top \beta(\mu_a + \mu_b) \approx \gamma\beta, \quad \text{where } \beta \gg 0$$

$$\mathbf{k}_b^\top \mathbf{q} \approx \mu_b^\top \beta(\mu_a + \mu_b) \approx \beta, \quad \text{where } \beta \gg 0$$

We can now directly solve for coefficients α_a and α_b, remembering that for large β values $\exp(0) = 1$ are significant (note how $\frac{\exp(a)}{\exp(a)+\exp(b)} = \frac{\exp(a)}{\exp(a)+\exp(b)}\frac{\exp(-a)}{\exp(-a)} = \frac{1}{1+\exp(b-a)}$):

$$\alpha_a \approx \frac{\exp(\gamma\beta)}{\exp(\gamma\beta) + \exp(\beta)} \approx \frac{1}{1 + \exp(\beta(1-\gamma))}$$

$$\alpha_b \approx \frac{\exp(\beta)}{\exp(\gamma\beta) + \exp(\beta)} \approx \frac{1}{1 + \exp(\beta(\gamma-1))}$$

Since γ varies between 0.5 and 1.5, and since $\beta \gg 0$, we have that:

$$\alpha_a \approx \frac{1}{1+\infty} \approx 0; \quad \alpha_b \approx \frac{1}{1+0} \approx 1; \quad \text{when } \gamma = 0.5$$

$$\alpha_a \approx \frac{1}{1+0} \approx 1; \quad \alpha_b \approx \frac{1}{1+\infty} \approx 0; \quad \text{when } \gamma = 1.5$$

Since $\mathbf{c} \approx \alpha_a \mathbf{v}_a + \alpha_b \mathbf{v}_b$ because other terms are insignificant when β is large, we can see that $\mathbf{c}$ oscillates between $\mathbf{v}_a$ and $\mathbf{v}_b$:

$$\mathbf{c} \approx \mathbf{v}_b, \quad \text{when } \gamma \to 0.5; \quad \mathbf{c} \approx \mathbf{v}_a, \quad \text{when } \gamma \to 0.5$$

(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 $\{k_1, \ldots, k_n\}$ that are randomly sampled, $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, where the means μ_i are known to you, but the covariances Σ_i are unknown. Also as before, assume that the means μ_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 α. 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: With the same assumptions as before, we can design $\mathbf{q}_1$ and $\mathbf{q}_2$ such that one of them copies $\mathbf{v}_a$ and the other copies $\mathbf{v}_b$. Since all keys are similar to their means and following the explanation in above question, we express the queries as:

$$\mathbf{q}_1 = \beta\mu_a, \quad \mathbf{q}_2 = \beta\mu_b, \quad \text{for } \beta \gg 0$$

This gives us (since means are orthogonal):

$$\mathbf{c}_1 \approx \mathbf{v}_a; \quad \mathbf{c}_2 \approx \mathbf{v}_b$$

And since multi-headed attention is just an average of the 2 values, we can see that:

$$\mathbf{c} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$$

Note extra answers:

1. It is also possible to set $\mathbf{q}_1$ to $\beta\mu_b$ and $\mathbf{q}_2$ to $\beta\mu_a$ which would yield the same answer, just that $\mathbf{v}_a$ and $\mathbf{v}_b$ would be swapped, i.e., $\mathbf{c}_1 \approx \mathbf{v}_b$ and $\mathbf{c}_2 \approx \mathbf{v}_a$.
2. It is even possible to use the same query designed in the previous question which would be the same for both queries in this question, i.e., $\mathbf{q}_1 = \mathbf{q}_2 = \beta(\mathbf{v}_a + \mathbf{v}_b)$. Then $\mathbf{c}_1 = \mathbf{c}_2 = \mathbf{c}$, i.e., an average of equal averages are the same average.

ii. (2 points) Assume that the covariance matrices are $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a\mu_a^\top)$ for vanishingly small α, 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 $c2$. You can ignore cases in which $k_a^\top q_i < 0$.

Answer: With regards to question **(c) ii.**, if we choose $\mathbf{q}_1 = \beta\mu_a$ and $\mathbf{q}_2 = \beta\mu_b$, we get that (note that all other key-query dot products will be insignificant):

$$\mathbf{k}_a^\top \mathbf{q} \approx \gamma\mu_a^\top \beta\mu_a \approx \gamma\beta, \quad \text{where } \beta \gg 0$$

$$\mathbf{k}_b^\top \mathbf{q} \approx \mu_b^\top \beta\mu_b \approx \beta, \quad \text{where } \beta \gg 0$$

We can solve for α values (again, note that all other key-query dot products will be insignificant when β is large):

$$\alpha_{a1} \approx \frac{\exp(\gamma\beta)}{\exp(\gamma\beta)} \approx 1, \quad \alpha_{b2} \approx \frac{\exp(\beta)}{\exp(\beta)} \approx 1$$

Since we can say that $\alpha_{i1} \approx 0$ for all $i \neq a$ and $\alpha_{i2} \approx 0$ for all $i \neq b$, it is easy to see that:

$$\mathbf{c}_1 \approx \mathbf{v}_a, \quad \mathbf{c}_2 \approx \mathbf{v}_b$$

Which means that the final output will always approximately be an average of the values:

$$\mathbf{c} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$$

Extra answers:

1. Now if we choose $\mathbf{q}_1 = \beta\mu_b$ and $\mathbf{q}_2 = \beta\mu_a$, a similar conclusion should be shown, just that the outputs would swap places, i.e., $\mathbf{c}_1 \approx \mathbf{v}_b$ and $\mathbf{c}_2 \approx \mathbf{v}_a$.
2. If we choose $\mathbf{q}_1 = \mathbf{q}_2 = \beta(\mathbf{v}_a + \mathbf{v}_b)$ then the problem should be similar to question **(c) ii.** as it is easy to show that, when $\gamma \to 0.5$, then $\alpha_{a1} = \alpha_{a2} \approx 0$ and $\alpha_{b1} = \alpha_{b2} \approx 1$, and when $\gamma \to 1.5$, then $\alpha_{a1} = \alpha_{a2} \approx 1$ and $\alpha_{b1} = \alpha_{b2} \approx 0$. Then it is clear that $\mathbf{c}$ will approach $\frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$ when $\gamma \to 1$ (i.e., when $\mathbf{k}_a$ is close to its mean μ_a).

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.[3] 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.

[3] See CS224n Azure Guide for a refresher on Azure.

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:

```
# 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's accuracy**: Correct: 10.0 out of 500.0: 2.0%
- **If the model only predicts "London"**: 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.

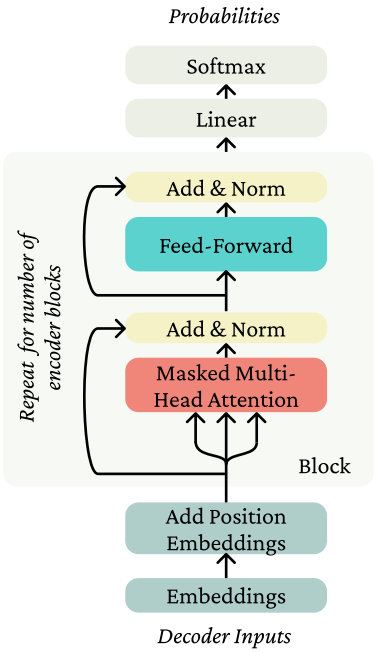
Answer: Correct: 86.0 out of 500.0: 17.2%

(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 ℓ^2 pairs of word vectors is computed in each computation, where ℓ 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.



Transformer Decoder

Figure 2: Illustration of the transformer block.

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 ℓ is the block size and d is the total dimensionality, d/h is the dimensionality per head.).[4]

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

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

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:

$$Y = [Y_1; \ldots; Y_h]A \tag{4}$$

where $A \in \mathbb{R}^{d \times d}$ and $[Y_1; \ldots; Y_h] \in \mathbb{R}^{\ell \times d}$. 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.

[4]Note that these dimensionalities do not include the minibatch dimension.

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 ℓ 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 ℓ.

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:

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

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 $\mathbb{R}^{l \times d}$. 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:

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

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 (ℓ) and basis bottleneck dimension (m).

Answer:

1. Correct: 24.0 out of 500.0: 4.8%
2. The Attention layer is the crucial part of the PerceiverAR. If we denote the input dimension 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 a 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 PerceiverAR model is $\mathcal{O}(dm + Lm^2)$, where d is the dimension of the byte array, m is the dimension of the latent array, and L is the depth of the transformer (number of layers).

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 (vanilla) model was able to learn the relationships between words. It learned how parts of a sentence depend on one another whereas the finetuned model was only trained to extract certain parts of a sentence without really "knowing" how they relate to the given input. It is also worth to note that the dataset for the pretraining was a lot larger thus the model could have "memorized" relevant parts of the questions.

(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. Such behaviour may cause users to refer to false information in their work unintentionally. For instance, if a user wants to mention the birthplace of a famous person, they might cite the wrong place if the the model does not provide the true information. This could affect user's work quality.
2. Such behaviour may cause users spread false information around the subject. For example, if the user learns something about a famous person, such as their birthplace, from a model that retrieves information, but that information is false, the person might unintentionally spread false facts about famous people and others may believe it. This could cause confusion and arguments in society.

(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, given a name, will try to maximize its relevance with its parameters (it will look for information that belongs to people with similar names to the provided name). If there is a typo in the name, then such functionality is desirable, however, if the new name is only similar to one of the names the model has learnt about, then retrieved information will be false and might cause quality and social 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., Cangea, 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.