Assignment 4: Induction Heads with Transformers

In-context learning in large language models (LLMs) refers to the process where a model learns to perform a task based on a few input-output examples provided within the context, using a prompt template [1]. Notably, no update to the model parameters is required during this process. The model leverages these examples to generalize and perform the task for a new test input.

Induction heads represent a synthetic experiment proposed to illustrate how in-context learning functions [3]. This mechanism allows models to recall patterns from earlier contexts and reproduce associated content.

Induction heads are name given to certain attention heads. Olsson et al [3] defined induction heads by analogy to inductive reasoning. In inductive reasoning, we might infer that if A is followed by B earlier in the context, A is more likely to be followed by B again later in the same context. Induction heads crystallize that inference. Induction heads search the context for previous instances of the present token, attend to the token which would come next if the pattern repeated, and increase its probability. Induction heads attend to tokens that would be predicted by basic induction (over the context, rather than over the training data).

For example in Figure 1, given the trigger token in black, the model is expected to reproduce the blue block because the context of the black token followed by the blue block has occurred earlier. When the model encounters a bigram such as *Harry Potter*, it should, upon seeing the token *Harry* again, predict *Potter* as the next token by copying it from the history.

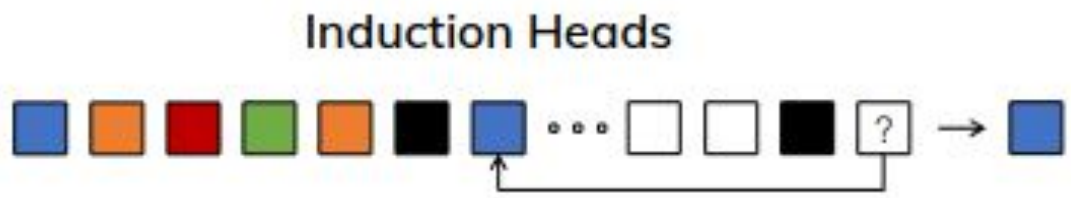

Figure 1: In the induction heads experiment, the model is tasked with reproducing content based on a previously encountered context. Figure reproduced from Mamba [2].

In next week's lecture, we're going to introduce an architecture called Mamba [2] that is particularly good at this kind of reasoning. But let's first try out how well the standard Transformer architecture performs at this task. During the lecture on state space models, we will have another assignment that involves running the same experiment using the Mamba model. In that assignment, we can compare the performance of the Transformer model with the Mamba model.

1 Dataset and Experimental Setup

We train a 2-layer Transformer model on the induction heads task using sequences of length 256. The dataset consists of synthetic sequences with a vocabulary size of 16 and vocab IDs ranging from 0 to 15.

For evaluation, we test the model on sequences of varying lengths, ranging from $2^6 = 64$ to $2^{20} = 1,048,576$. We use Transformer with multi-head attention (num_heads=8) and absolute position embeddings. The model has a dimensionality of 64.

2 Training and Evaluation

2.1 Loading Conda Environment

To set up the HPC for this assignment, run the following commands:

```
module load Anaconda3
source activate /opt/itu/condaenv/cs/rapu/synth_transformers
```

You can also add the commands to your `.bashrc` to make these load automatically upon logging in to the cluster.

2.2 Training Procedure

Follow the steps outlined in the `README.md` file located within the `induction_heads_assignment.tar` archive to train the Transformer model. The training process will generate a trained model for further inference.

2.3 Inference and Results

Once training is complete, run the inference script as described in the `README.md`. You are required to plot the accuracy of the model's predictions across different test sequence lengths. Compare the performance and analyze how the model generalizes across varying sequence lengths.

3 Conclusion

This assignment explores the ability of Transformers on the task of induction heads. By analyzing performance across different sequence lengths, we aim to better understand the mechanism of generalization of in-context learning performance within Transformers.

References

[1] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *CoRR* abs/2312.00752 (2023). DOI: 10.48550/ARXIV.2312.00752. arXiv: 2312.00752. URL: https://doi.org/10.48550/arXiv.2312.00752.

[3] Catherine Olsson et al. "In-context Learning and Induction Heads". In: *CoRR* abs/2209.11895 (2022). DOI: 10.48550/ARXIV.2209.11895. arXiv: 2209.11895. URL: https://doi.org/10.48550/arXiv.2209.11895.