Week 4 Questions

Karpathy GPT

1. In the forward function of the GPT class, why is there not a call written to apply the softmax function to the logits?
2. What does torch.no_grad() on line 282 in model.py do and why is it a good idea to use this?
3. Why do they transpose the tensors on lines 57-59 in model.py?
4. What does the self.register_buffer call on line 47 of model.py do? What would go wrong if instead we added the tensor in this line to the model parameters?
5. Try training and evaluating the transformer on some data of your choice. You may want to use Google Colab, which gives gpu access for free, to run the code. Note that using a gpu will make training and evaluation go much more quickly than cpu.

Language Models are Few-Shot Learners

1. What do the authors mean by ‘zero-shot’, ‘one-shot’, and ‘few-shot’?
2. In what ways is few-shot learning different from fine-tuning?
3. They use a filtered version of Common crawl for most of their training data. What advantages and disadvantages are there for using this data?
4. The mention in section 2.4 that larger numbers of in-context examples are not always better. Why might this be the case?
5. In section 3.9.4, they show that news articles generated by the 175B version of GPT-3 are nearly indistinguishable from real news articles. What limitations might there be in this section?
6. In section 6.2 the authors do a short analysis on some of the biases GPT-3 has. What are some reasons the model has these biases?

Finetuned Language Models Are Zero-Shot Learners

1. What is the main advantage of the Adafactor optimizer compared to Adam?
2. In the paper, they use instruction tuning to improve performance on certain benchmarks. What are some reasons this improves performance compared to the model with only pre-training?
3. They find that instruction tuning smaller models hurts performance on unseen tasks. However, nowadays you can find many small open-source models for which instruction tuning improves performance compared to the non-instruction tuned variant. What are some possible explanations for this difference?