Assignment 6

Transformers and Next-Token Prediction

Yelp Review Generation

Instructions. This assignment focuses on how language models use next-token prediction to generate Yelp-style text. Explore the Week 6 web app and answer the questions below. Stick to 3 sentences maximum per question. Upload your responses to BruinLearn as a PDF.

Question 1. In the Next Token tab, write your own 4- or 5-word Yelp-style prompt and compare all three pretrained models: 10,000 reviews / 1 epoch, 500,000 reviews / 1 epoch, and 500,000 reviews / 20 epochs. Report the prompt you used, plus the top next token and probability for each model. What changes as the model sees more data or trains longer?

Question 2. In the Next Token tab, choose two different 4- or 5-word Yelp-style prompts using the 500,000 reviews / 20 epochs model. Report both prompts, then report the top 3 predicted next tokens for each prompt. How does the prompt context shape what the model expects next?

Question 3. In the Completion tab, use one 4- or 5-word Yelp-style prompt with the 500,000 reviews / 20 epochs model. Set Tokens to 20 and generate completions at temperatures 0.5, 0.8, and 1.3. Report the prompt you used. Which completion sounds most like a real Yelp review, and what does temperature appear to control?

Question 4. If you wanted to improve this tiny Yelp language model, what changes would you make and why? You may comment on data, architecture, vocabulary, training process, or evaluation. A strong answer should connect your recommendation to how the model works.