Homework 1

Anshul Sawant

Due date: January 24 at 11:59 PM

Question 1.1 (5 points)

Weight tying makes the input embedding projection (1-hot → embedding) and the output unembedding projection (embedding dim → vocab size logits) transpose of each other by sharing weights between corresponding matrices. This leads to shared weights being updated at each time step (because gradients flow back from all unsaturated logits) instead of only the input rows for current input tokens being updated. In practice, it leads to faster convergence of input embeddings and consequently faster overall convergence.

Question 1.2 (5 points)

```
self.token_logits = nn.Linear(n_embd, vocab_size)
self.token_logits.weight = self.token_embeddings.weight
```