ANLP Assignment 1

Report

Surisetty Tulasi Rushwik

2023111011

1. Translation Accuracy

Table 1 reports the final BLEU scores on the test set for both RoPE and Relative Bias configurations under three decoding strategies (Top-k, Greedy, and Beam Search).

Configuration	Top-k BLEU	Greedy BLEU	Beam Search BLEU
RoPE	0.038961401724655544	0.6484926914932718	0.6483546633168529
Relative Bias	0.032775816323831156	0.6484926914932718	0.6468082669666467

Table 1: BLEU scores of RoPE and Relative Bias on the test set under different decoding strategies.

Impact of Decoding Strategies

- **Top-k Sampling:** Both RoPE and Relative Bias show very poor BLEU scores (0.038961401724655544 and 0.032775816323831156, respectively). This is due to the randomness introduced by sampling, which increases diversity but harms accuracy and faithfulness under BLEU.
- **Greedy Decoding:** Both configurations achieve the highest BLEU (0.6484926914932718). Greedy deterministically selects the most probable token, which works well here because the model's probability distribution is sharp.
- **Beam Search:** RoPE (0.6483546633168529) performs almost identically to Greedy (0.6484926914932718), while Relative Bias (0.6468082669666467) is slightly worse. Beam search normally helps, but in this case, the peaked distributions leave little room for improvement, and extra exploration can introduce suboptimal sequences.

Why These Results Occur

1. **Sharp Probability Distributions:** The model assigns very high confidence to a few tokens. Greedy decoding captures this well, leaving little benefit for beam search.
2. **Sampling Noise:** Top-k introduces stochasticity, which lowers BLEU since it sacrifices exact n-gram matches in favor of diversity.
3. **Training Scale:** In smaller-scale training, beam search often provides minimal improvements, as the model lacks sufficient uncertainty for multiple candidates to matter.
4. **Metric Sensitivity:** BLEU rewards exact matches, favoring deterministic strategies like greedy decoding over stochastic ones.

2. Convergence Speed

The plot in Figure 1 shows the training loss curves for two different positional encoding methods: Rotary Positional Embeddings (RoPE) and Relative Bias.

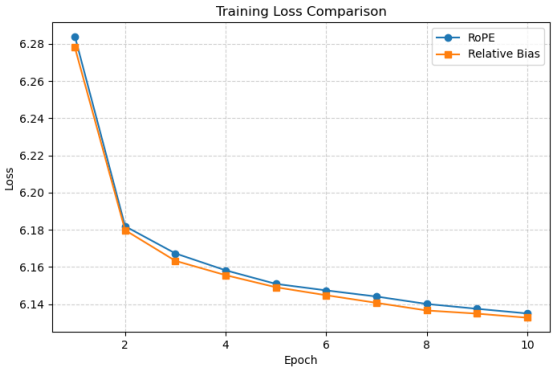

Figure 1: Comparison of training loss between RoPE and Relative Bias across 10 epochs.

Observations

- Both RoPE and Relative Bias show a steady decrease in training loss across epochs, indicating stable convergence.
- The Relative Bias configuration consistently achieves slightly lower losses compared to RoPE.
  - Epoch 3: RoPE = 6.1673, Relative Bias = 6.1633
  - Epoch 6: RoPE = 6.1474, Relative Bias = 6.1448
  - Epoch 10: RoPE = 6.1350, Relative Bias = 6.1327
- Although the difference is small, Relative Bias demonstrates **faster convergence** and reaches a marginally better final loss.

Why Relative Bias Performs Better

Relative Bias outperforms RoPE in this setup due to the following reasons:

1. **Trainability:** Relative Bias introduces learnable parameters for positional information, whereas RoPE is fixed and deterministic. This allows the model to adapt positional signals to the dataset.
2. **Gradient Flow:** Relative Bias provides direct gradients to position-related parameters, leading to smoother optimization and faster convergence.

3. **Local Dependency Modeling:** Many sequence tasks exhibit strong local patterns. Relative Bias helps the model capture these short-range dependencies more effectively than RoPE.

4. **Training Scale:** RoPE often shows greater benefits in long-context or large-scale training. In smaller-scale experiments (limited epochs or smaller datasets), Relative Bias tends to converge faster.

Conclusion

Relative Bias provides a slight optimization advantage over RoPE in this setup, leading to more efficient convergence and marginally improved training loss. For decoding strategies, Greedy decoding emerges as the most effective method, while Top-k drastically reduces accuracy and Beam Search offers no clear improvement.