Report on Decoder-Only Transformer Training

1. Summary

This report details the training process and results of a 17.4 million parameter decoder-only transformer model. The model was trained for four epochs on the TinyStories dataset. The training was successful, demonstrating stable convergence and significant improvement in key metrics.

Over the four epochs, the **Training Loss** was reduced by 32% (from 3.2612 to 2.2134), and the **Validation Loss** was reduced by 18% (from 2.5748 to 2.1014). Crucially, **Perplexity** on the validation set improved by 38%, decreasing from 13.13 to 8.18. The training curves show no signs of overfitting, indicating that the model is generalizing well to unseen data.

2. Configuration

2.1. Model Architecture

The model is a 5-layer decoder-only transformer with the following specifications:

- **Total Parameters:** 17,422,500
- **Layers:** 5
- **Attention Heads:** 6
- **Model Dimension :** 300 (matching the FastText embedding dimension)
- **Feed-Forward Dimension (**d_{ff}**):** 1200
- **Max Sequence Length:** 64
- **Dropout:** 0.1

2.2. Training & Dataset

- **Dataset:** TinyStories (from HuggingFace)
- **Training Samples:** 2,119,719
- **Validation Samples:** 21,990
- **Vocabulary Size:** 10,004
- **Embeddings:** Pre-trained 300-dim FastText
- **Optimizer:** Adam
- **Learning Rate:** 3e-4
- **Batch Size:** 32
- **Epochs Completed:** 4

3. Results and Analysis

3.1. Performance Metrics by Epoch

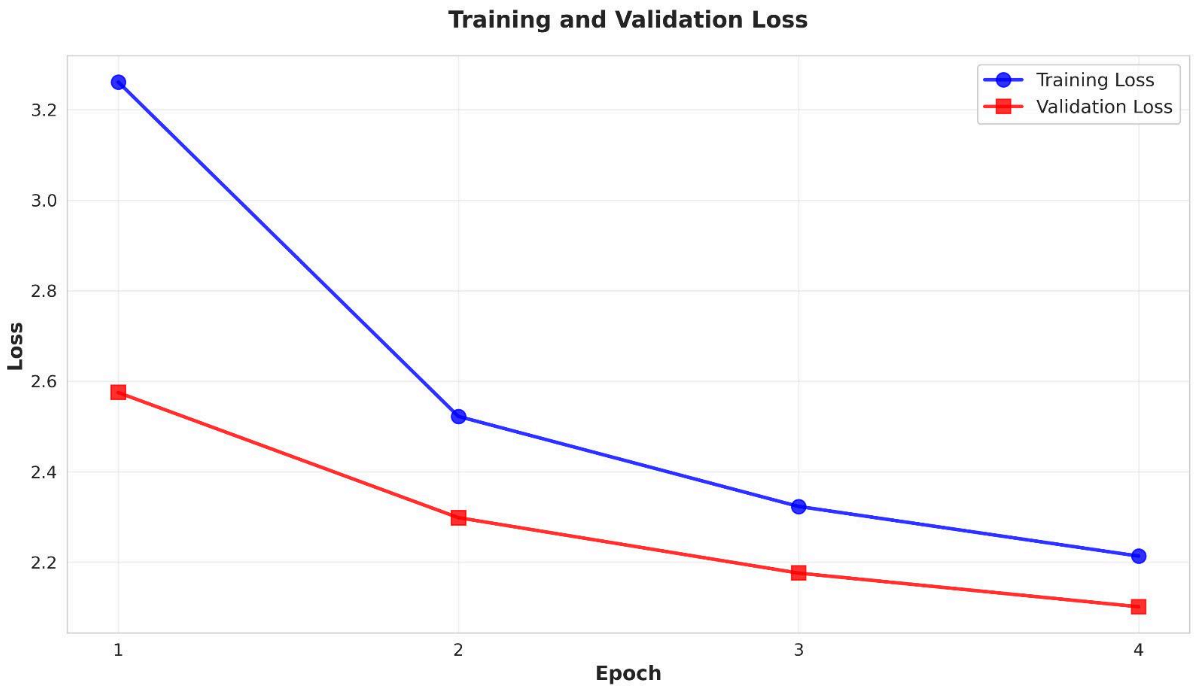
The model's performance improved consistently across all four epochs, as detailed in the table below.

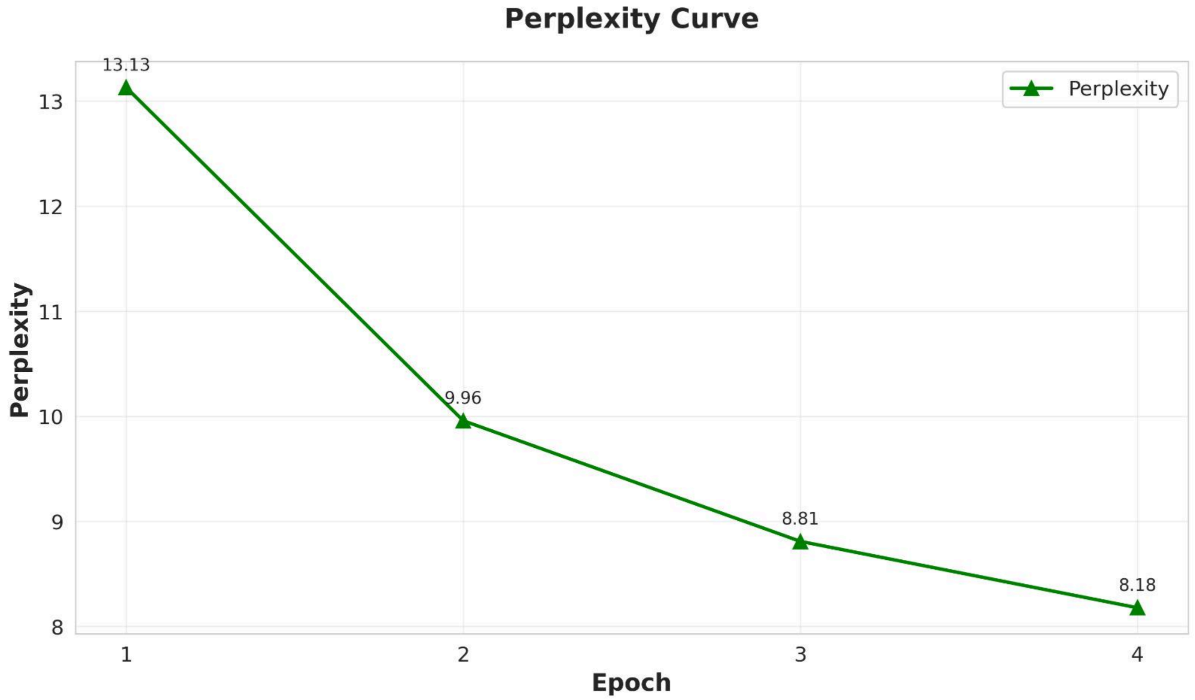
Epoch	Training Loss	Validation Loss	Perplexity
1	3.2612	2.5748	13.13
2	2.5221	2.2983	9.96
3	2.3232	2.1758	8.81
4	2.2134	2.1014	8.18

Final Metrics (Epoch 4):

- **Training Loss:** 2.2134
- **Validation Loss:** 2.1014
- **Perplexity:** 8.18

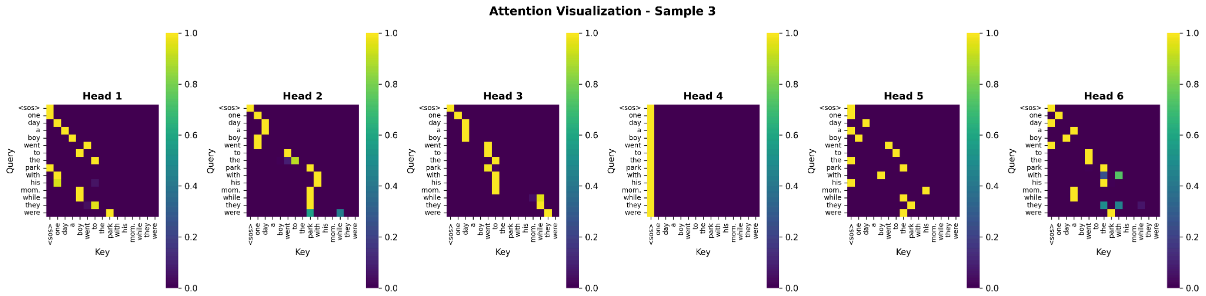
3.2. Visual Analysis

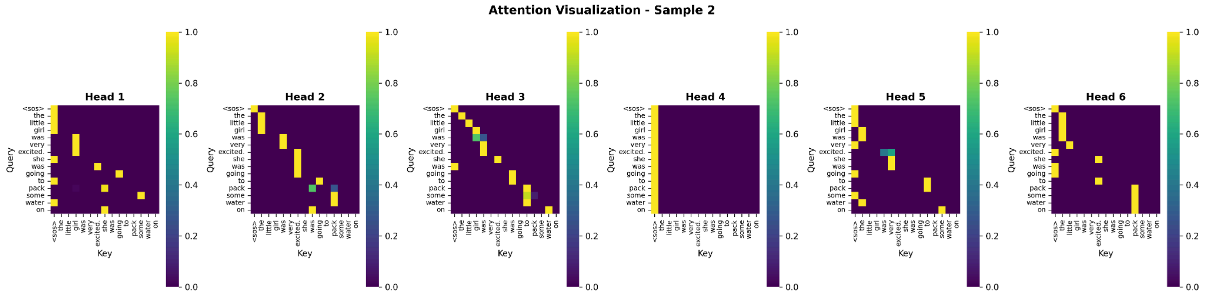


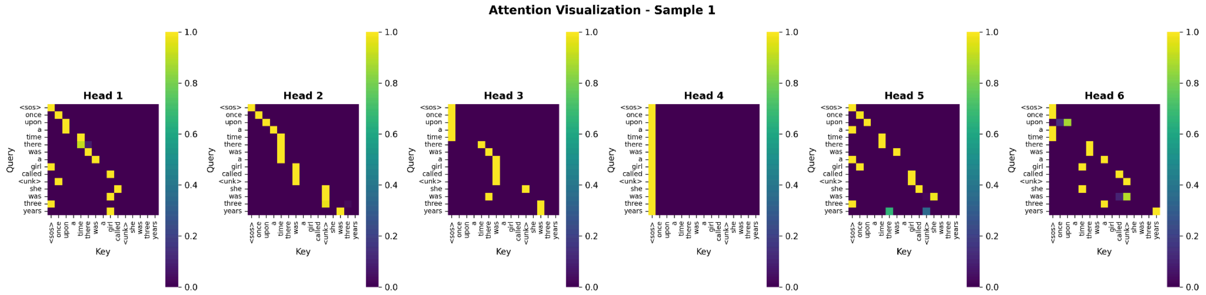


The provided plots visually confirm the positive training trend.

- **Training and Validation Loss:** This plot shows a steady, downward trend for both training loss (blue) and validation loss (red). This is the ideal behavior, as it indicates the model is continuously learning and improving its predictions.
- **Perplexity Curve:** This curve shows a sharp initial drop after Epoch 1, followed by a steady decrease. This directly corresponds to the model's improved ability to predict the next word in a sequence.







3.3. Analysis of Observations

The training log and plots support several key conclusions:

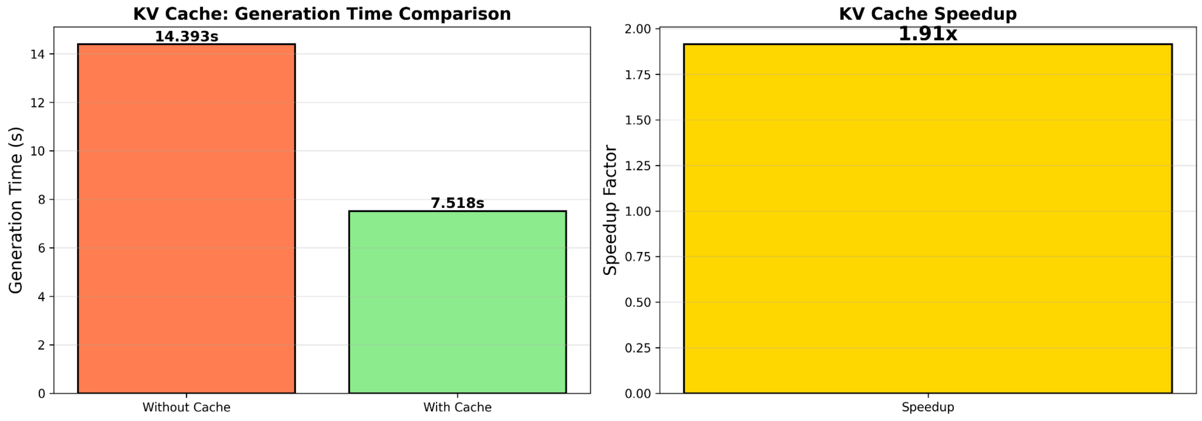
1. **Consistent Learning and Convergence:** The simultaneous decrease in both training and validation loss (as seen in the plot above) confirms the model is effectively learning the underlying patterns of the TinyStories dataset. The convergence is stable, without erratic spikes, suggesting the learning rate (3e-4) and other hyperparameters are well-chosen for this architecture.
2. **No Signs of Overfitting:** A critical finding is the absence of overfitting. Overfitting occurs when a model learns the training data "too well" and loses its ability to generalize to new, unseen data. This would be visually represented by the validation loss *increasing* while the training loss continues to decrease. Here, the validation loss consistently *decreases* alongside the training loss, proving the model is generalizing well.
3. **Validation Loss Lower than Training Loss:** A notable observation is that the validation loss is consistently lower than the training loss at each epoch. This is a common and expected outcome when using dropout.
   - **Explanation:** Dropout (set to 0.1) is a regularization technique that randomly "turns off" a fraction of neurons *during training* to prevent co-adaptation. This makes the training task harder for the model, resulting in a higher loss. During validation, dropout is *disabled*, allowing the model to use its full, trained

network. This "full-strength" model naturally performs better on the validation set, leading to a lower loss.

4. **Significant Perplexity Reduction:** Perplexity (PPL) is a core metric for language models. It measures how "surprised" a model is by the test data. A PPL of 13.13 means that, on average, the model was unsure what the next word was, with its "guess" being roughly equivalent to choosing between 13 words. By Epoch 4, the PPL dropped to 8.18. This 38% reduction indicates a substantial improvement in the model's predictive accuracy and its understanding of the language.

Part 2

KV_cache:



KV_cache on 10 prompts has shown a 1.91 speed up and in some other cases it is showing a speed up of 2.25x and on an average it is increasing the speed by 2 times while prompting.

Beam_search:

1. Greedy Decoding:

○ Fastest method (0.849s per sample)

○ Lowest BLEU score (0.0148)

○ Simply selects the highest probability token at each step

○ Suffers from lack of diversity and inability to recover from early mistakes

2. Beam Search (k=5):

○ 4.9x slower than greedy (4.5031s)

○ Best BLEU score (0.0460) - 3.1x better than greedy

○ Maintains 5 hypothesis sequences, providing good balance

○ Optimal choice for quality-time trade-off

3. Beam Search (k=10):

○ 10.3x slower than greedy (7.476s)

○ Lower BLEU (0.0328) than k=5, suggesting over-searching

○ Increased beam size causes the model to prefer overly generic sequences

○ Diminishing returns beyond k=5

4. Overall BLEU Scores: The absolute BLEU scores are quite low (< 0.05), indicating the model's

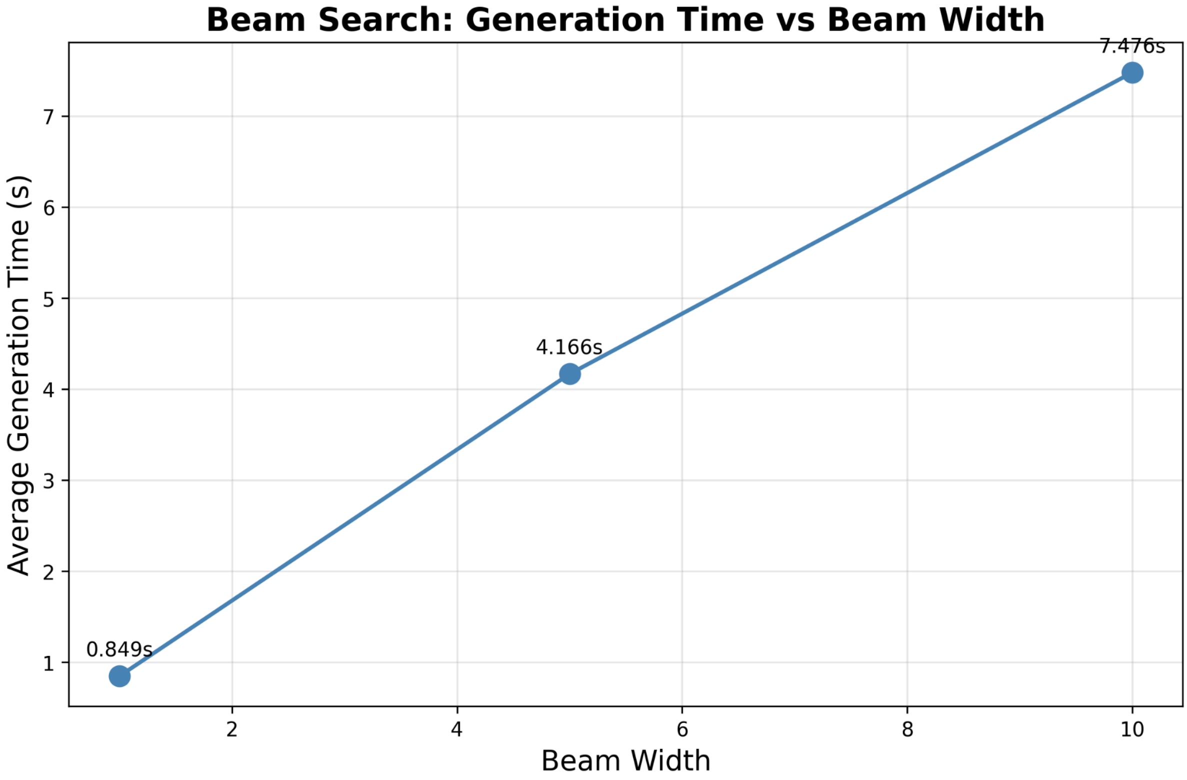
generations differ significantly from references. This suggests either:

○ The model needs more training

○ The task is challenging

○ The evaluation dataset is difficult

While using beam search we can see an increase of time



4. Conclusion

The training of the 17.4M parameter transformer was highly effective. The model achieved a final validation perplexity of 8.18 after only four epochs, with no evidence of overfitting. The learning curves are stable and show that the model could likely be trained for even more epochs to achieve further performance gains. The use of pre-trained FastText embeddings and a 0.1 dropout rate proved to be an effective configuration.