Assignment 2: Building a Small-Scale Foundation Model from Scratch

Yazhen Han 002950305
Northeastern University

1 Model Architecture and Parameters

This implementation features a transformer-based language model (mini-GPT) for next-token prediction with a decoder-only transformer architecture.

Table 1: Model Configuration

Component	Value
Embedding Dimension	64
Transformer Layers	1
Attention Heads	2
Feed-Forward Dimension	256
Max Sequence Length	64 tokens
Dropout Rate	0.3
Vocabulary Size	49,805
Total Parameters	**6,479,053**

The model implements scaled dot-product attention with causal masking: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ where $d_k = 32$. Key components include: (1) multi-head attention with proper causal masking for autoregressive generation, (2) position-wise feed-forward networks (64→256→64) with GELU activation, (3) pre-normalization layer normalization, (4) learned positional embeddings up to 64 tokens, and (5) residual connections around all sublayers.

2 Dataset Details

The dataset contains 10 text chunks (approximately 5,120 tokens) preprocessed with GPT-2 BPE tokenizer, producing 5,056 training sequences with batch size 4.

Table 2: Dataset Statistics

Metric	Value
Text Chunks	10
Total Tokens	5,120
Vocabulary Size	49,805
Training Sequences	5,056
Batches per Epoch	1,264

Critical Limitation: This sample dataset is 1000× smaller than production foundation models (100K+ sequences). This creates severe overfitting conditions where the model can memorize all training data.

Observed Impact: Loss decreased from 4.24 to 0.23 (94.6% reduction) in 10 epochs; perplexity fell from 69.53 to 1.26. These metrics indicate memorization rather than generalization. With a full dataset, final loss would stabilize around 2-4 with gradual convergence over 20-50 epochs.

Mitigation Applied: High dropout (0.3), strong weight decay (0.1), small model capacity (64 dims, 1 layer), and limited epochs (10) to delay overfitting.

3 Training Setup and Results

Table 3: Training Configuration

Parameter	**Value**	**Rationale**
Optimizer	AdamW	Decoupled weight decay
Learning Rate	5e-4	Standard for small models
Batch Size	4	Limited data
Weight Decay	0.1	Strong regularization
Gradient Clipping	1.0	Stability
LR Scheduler	CosineAnnealing	Smooth convergence
Loss Function	Cross-Entropy	Token prediction

Table 4: Training Results by Epoch

Epoch	**Loss**	**Perplexity**
1	4.24	69.53
2	1.55	4.72
3	0.88	2.40
5	0.40	1.49
7	0.28	1.32
10	0.23	1.26

Final Metrics: Training loss 0.2274, perplexity 1.26, training time 9 minutes (50 seconds per epoch). The rapid 94.6% loss reduction with steep initial drops in epochs 1-3 demonstrates quick convergence on the limited dataset.

4 Observations and Challenges

4.1 Key Findings

1. Data Scale is Critical: The sample dataset (5,120 tokens) is fundamentally insufficient for meaningful language modeling. The model memorizes all training samples by epoch 5-6. Production foundation models require minimum 100K+ sequences with diverse sources.

2. Training Behavior: Three distinct phases emerged: (1) Phase 1 (epochs 1-3) with 79% loss reduction showing rapid initial learning, (2) Phase 2 (epochs 4-7) with diminishing returns, and (3) Phase 3 (epochs 8-10) with minimal improvement approaching memorization limits.

3. Implementation Validation: Successful training demonstrates correct implementation of multi-head attention, positional embeddings, layer normalization, residual connections, and gradient clipping. Smooth loss curves without spikes indicate stable training dynamics.

4.2 Technical Challenges

Challenge 1 - Limited Dataset: Model achieves very low training loss (0.23) and perplexity (1.26), indicating memorization rather than generalization. Mitigation through high dropout and weight decay only delayed overfitting.

Challenge 2 - Training Stability: Gradient clipping (max norm 1.0) and layer normalization prevented gradient explosions across 1,264 batches per epoch, maintaining stable convergence.

Challenge 3 - Hyperparameter Selection: Conservative architecture (1 layer, 64 dims) with strong regularization balances learning capability against overfitting. With full dataset, would scale to 2-6 layers and 128-512 dimensions.

5 Conclusion

This implementation successfully demonstrates a complete transformer-based language model with proper attention mechanisms, normalization, and training procedures. Key learnings include: (1) technical mastery of self-attention and transformer components, (2) recognition that foundation models are fundamentally data-driven, and (3) understanding the critical distinction between memorization and genuine learning. The code is production-ready and scales to larger datasets with minimal modifications. Future work includes training on the full 1GB+ corpus, implementing validation-based early stopping, and scaling to larger architectures.