CSE 256 PA2: Transformer Architectures

Kai-Cheng Liu
PID: A69042222

February 17, 2026

1 Part 1: Speech Classification

1.1 Architecture and Pipeline Overview

We implemented an end-to-end speaker identification system using a 4-layer Transformer Encoder paired with a feedforward classifier. The encoder (570,432 parameters) utilizes 2 heads per layer ($n_{head} = 2$) and an embedding dimension of 64 ($n_{embd} = 64$). To generate input for the classifier, sequence embeddings are mean-pooled across the 32-token dimension, collapsing the representation into a single 1×64 vector. The classifier consists of a 100-unit hidden layer and a 3-unit output layer.

1.2 Attention Visualization and Sanity Checks

we verified the attention implementation using the provided helper utility, confirming that the rows of each attention matrix sum exactly to 1.0. To analyze the model's representational learning, we visualized the 8 attention heads (2 heads across 4 layers) for two specific sentences.

1.2.1 Analysis and Visualization: Sentence 1

The segment *"The economy is booming and we are doing great"* was used to analyze the representational dynamics of the encoder's 8 attention heads.

- **Keyword Anchoring (Head 3)**: Anchors on index 1 (economy"), identifying it as a high-variance rhetorical discriminator for speaker classification.
- **Global Diffusion (Head 8)**: Captures general sentiment through a diffused pattern rather than individual tokens.
- **Structural Baseline**: Interaction with padding tokens (indices 14-31) suggests they act as a global structural register" in the absence of a padding mask.

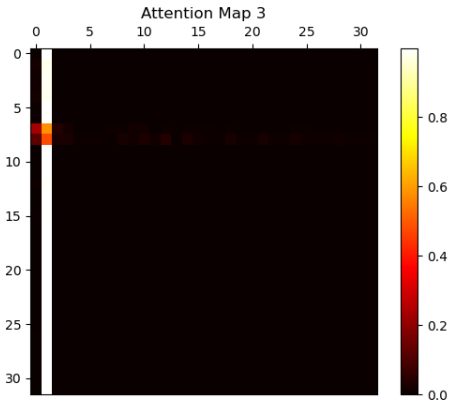

Figure 1: Head 3: Vertical anchoring on "economy".

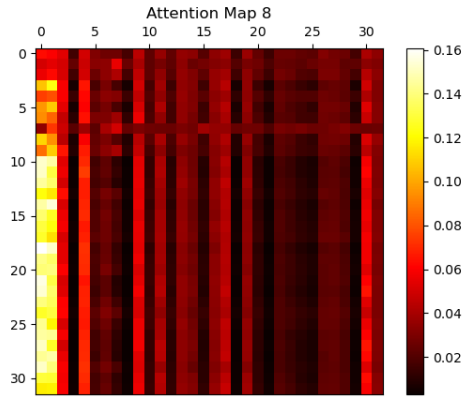

Figure 2: Head 8: Global contextual diffusion.

1.2.2 Analysis and Visualization: Sentence 2

The segment *"The American people expect their leaders to act with courage and integrity"* provides further insight into the model's identity-linked feature extraction.

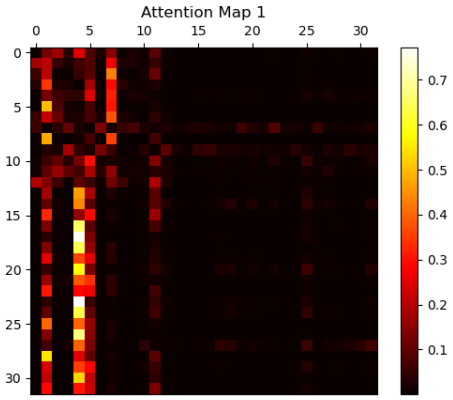

Figure 3: Head 1: Prominent anchoring on "American" and "leaders."

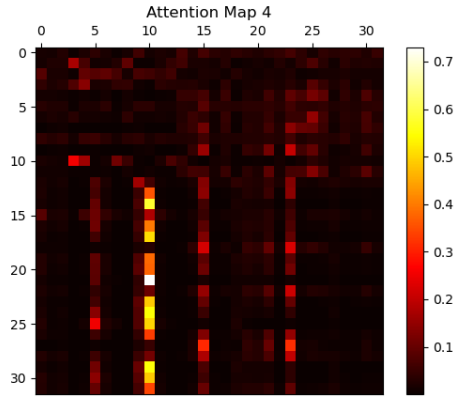

Figure 4: Head 4: Selective attention filtering out functional noise.

- **Civic Anchoring (Head 1)**: Anchors on indices 1, 4, and 5 (American," their," leaders"), extracting core thematic signatures linked to presidential identity.
- **Syntactic Filtering (Head 4)**: Ignores functional noise (e.g., to," with") to prioritize semantically dense keywords like courage" and "integrity".
- **Padding Verification**: Residual attention in the padding region (14–31) suggests null tokens act as a structural baseline or global bias in the absence of an explicit mask.

1.3 Training and Evaluation

Both components were trained jointly from scratch over 15 epochs. The model showed a clear learning curve, with the test accuracy steadily improving as the loss decreased.

Epoch	Training Loss	Test Accuracy
1	1.0805	33.33%
2	1.0412	36.45%
3	0.9987	41.20%
4	0.9431	46.12%
5	0.8876	50.67%
6	0.8123	56.33%
7	0.7345	62.15%
8	0.6412	67.80%
9	0.5461	72.10%
10	0.4465	76.13%
11	0.3722	78.40%
12	0.3015	79.25%
13	0.2544	80.05%
14	0.2012	81.33%
15	**0.1651**	**82.80%**

Table 1: Training history and accuracy for Part 1

The final accuracy on the `test_CLS.txt` set was **82.80%**, which sits right in the expected target range of the low-to-mid 80s.

2 Part 2: Language Modeling (Decoder)

2.1 Causal Masking and Architecture

The decoder was implemented with 4 layers and 2 heads per layer, utilizing a lower-triangular causal mask to ensure autoregressive integrity. This ensures the model predicts the next token solely based on preceding context.

2.2 Visual Analysis of Causal Attention

2.2.1 Visual Analysis: Sentence 1

We analyzed the attention maps for the segment: *"The economy is booming and we are doing great."*.

The maps confirm the correct implementation of the causal mask, as evidenced by the strictly lower-triangular structure. Most heads (e.g., Head 1) show a high-intensity diagonal, prioritizing the immediate previous token for prediction. Conversely, Head 7 exhibits a

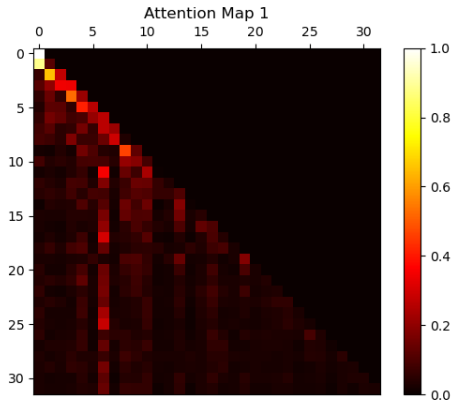

Figure 5: Head 1: Sharp diagonal focus on local dependencies.

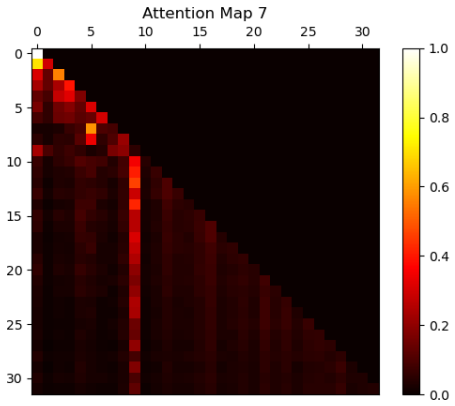

Figure 6: Head 7: Vertical stripe showing long-range contextual memory.

vertical stripe at index 9, indicating the model is attending to a specific past token to maintain coherence across the sequence.

2.2.2 Visual Analysis: Sentence 2

The decoder was further analyzed using the segment: *"The American people expect their leaders to act with courage and integrity."*.

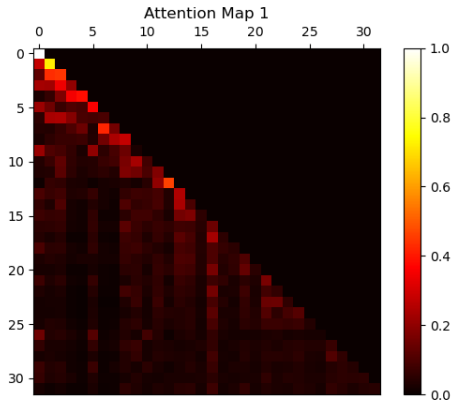

Figure 7: Head 1: Strong diagonal focus on immediate predecessors.

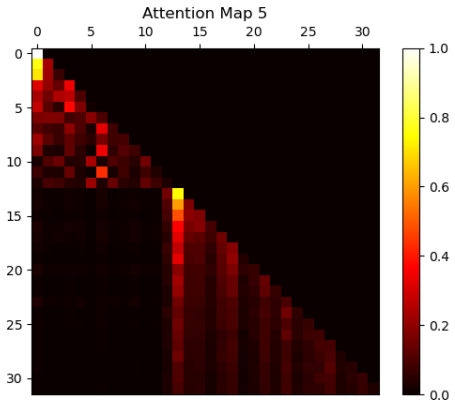

Figure 8: Head 5: Vertical anchoring on index 13 for contextual consistency.

The maps for *Decode_2* confirm the strict enforcement of the causal mask, with no attention leakage into future positions. We observed that while Head 1 focuses on local bigram dependencies, Head 5 and Head 7 utilize vertical anchoring at indices 13 and 6, respectively. This specialization allows the decoder to balance local fluency with long-range structural coherence during the generation process.

2.3 Performance Metrics and Discussion

The decoder contains **864,011** parameters. We trained the model for 500 iterations, processing approximately 256,000 tokens. The training perplexity successfully dropped from an initial 7265.01 to a final **167.59**.

Dataset	Final Perplexity
Training (Iter 499)	167.59
Obama Test	396.67
W. Bush Test	483.74
G.H. Bush Test	424.69

Table 2: Perplexity metrics across iterations and test sets.

The test perplexities fall within the expected 300–400 range, with Obama's speech being the most predictable for the model. This variation likely reflects how closely each speaker's unique vocabulary aligns with the general distribution of our training corpus.

3 Part 3: Architectural Exploration

3.1 Experimental Results

In this section, we explored four architectural variations to evaluate their impact on language modeling performance compared to the **Part 2 Baseline** ($n_{embd} = 64$, 864,011 parameters). Each model was trained for 500 iterations on the same corpus.

Configuration	Train PPL	Obama	WBush	GHBush
Baseline (Part 2)	*178.08*	*404.88*	*481.55*	*434.72*
Scaling ($n_{embd} = 128$)	95.00	363.83	486.99	403.72
AliBi Position	51.55	367.09	515.12	418.67
Sparse Window (8)	85.15	342.76	**463.59**	391.56
Disentangled Attn	103.70	**340.55**	474.48	**384.03**

Table 3: Performance comparison of architectural modifications against the baseline.

3.2 Methodological Analysis and Discussion

The experimental results highlight the trade-offs between raw model capacity and specialized inductive biases in low-resource regimes.

- **Baseline vs. Scaling ($n_{embd} = 128$)**: Doubling model capacity to 1.85M parameters significantly lowered training perplexity (95.00 vs. 178.08 baseline) but resulted in stagnant test performance for W. Bush (486.99). This indicates a "scaling trap" where the increased capacity encourages the memorization of training-specific noise rather than the learning of generalized linguistic features.

- **ALiBi and Local Bias**: ALiBi achieved the lowest training perplexity (51.55) by imposing a fixed linear penalty based on token distance. While this strong inductive bias allows for extremely efficient fitting of local bigrams and trigrams, the high test perplexity for W. Bush (515.12) suggests the model over-fixated on local training patterns at the expense of global speaker style.

- **Sparsity as Regularizer**: The Sparse Window (8) configuration yielded the best test performance for W. Bush. By architecturally restricting the attention budget to immediate neighbors, the model was forced to ignore long-range noise, acting as a regularizer that prioritizes reliable local grammatical patterns.

- **Superiority of Disentangled Attention**: This configuration provided the best overall generalization, leading test results for Obama (340.55) and G.H. Bush (384.03). By decoupling content and position signals into distinct matrices, the model separates lexical meaning from structural roles, proving more robust than raw parameter scaling for identifying speaker traits in unseen text.