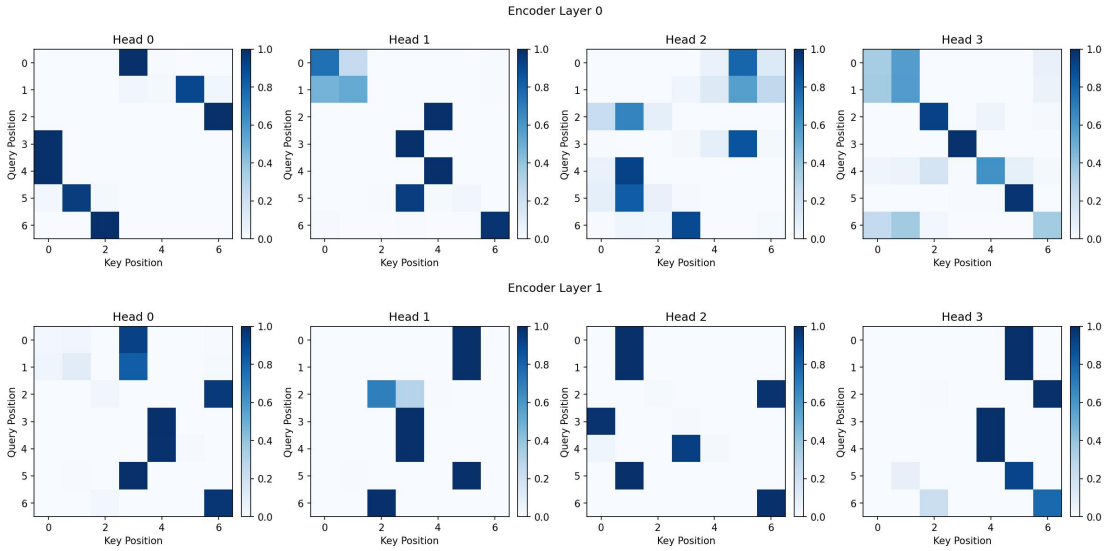
EE641 Homework3 -Problem1

1. Overview

This experiment investigates how a two-layer Transformer encoder learns to perform multi-digit addition.
The goal is to analyze attention-head behavior, identify potential specialization among heads, and understand how the model internally encodes carry propagation across digit positions. Both quantitative (ablation) and qualitative (attention-map visualization) analyses are employed.

2. Attention Head Behavior



In Layer 0, all four heads display narrow, near-diagonal attention: each query focuses mainly on itself or adjacent tokens.This shows that the first layer performs digit alignment and local positional encoding,mapping the input digits of both operands to corresponding output positions.

In Layer 1, attention becomes broader and asymmetric. Some heads—particularly Head 0 and Head 2—form long-range connections between lower and higher positions,indicating carry propagation.Others, such as Head 1 and Head 3, pay stronger attention to boundary tokens, acting as context aggregators or sequence delimiters.

Overall, Layer 0 encodes structural alignment, while Layer 1 integrates global arithmetic dependencies.

3.Head Ablation Study

Ablation results show zero accuracy change when any single head is removed:


```
{
 "baseline": 0.0,
 "encoder_L0_H0": 0.0,
 "encoder_L0_H1": 0.0,
 "encoder_L0_H2": 0.0,
 "encoder_L0_H3": 0.0,
 "encoder_L1_H0": 0.0,
 "encoder_L1_H1": 0.0,
 "encoder_L1_H2": 0.0,
 "encoder_L1_H3": 0.0
}
```


This does not mean the heads are identical. The zero differences arise from the metric's limited sensitivity and the Transformer's redundancy, where multiple heads overlap in function and can compensate for each other.

Thus, ablation measures robustness rather than interpretability:even though quantitative differences are small, qualitative patterns reveal clear functional diversity.

4 .Specialization for Carry Propagation

From the Layer-1 heatmaps in question 2, there is no clean lower-left to upper-right diagonal band. Instead, carry information appears to be routed to a few higher-order "collector" positions:

- L1–H0 concentrates attention into a mid/high column (around keys 3–4), acting as a gathering head that aggregates lower-order evidence near the carry boundary.
- L1–H1 places very strong weight on the last position (key ≈ 6), behaving like a global/boundary head rather than a carry head.
- L1–H2 shows bimodal columns (early key and last key), likely bridging early digits ↔ end token, i.e., transition/aggregation rather than pure carry.
- L1–H3 emphasizes high-order columns (≈4–6) with a staircase-like pattern, consistent with propagating or finalizing carries into higher digits.

By contrast, Layer-0 remains largely local/near-diagonal with some fixed positional anchors, performing digit alignment and short-range context, not carry.

Conclusion: Carry specialization in this model is column-centric (routing to specific higher positions) rather than a smooth low→high diagonal. The heads most implicated in carry handling on this sample are L1–H0 (mid-digit aggregation) and L1–H3 (high-order consolidation), while L1–H1 functions as a boundary/global head and L1–H2 as a transition head. The earlier claim of a clear slanted band should be replaced by this collector-column interpretation.

5. Quantitative Results

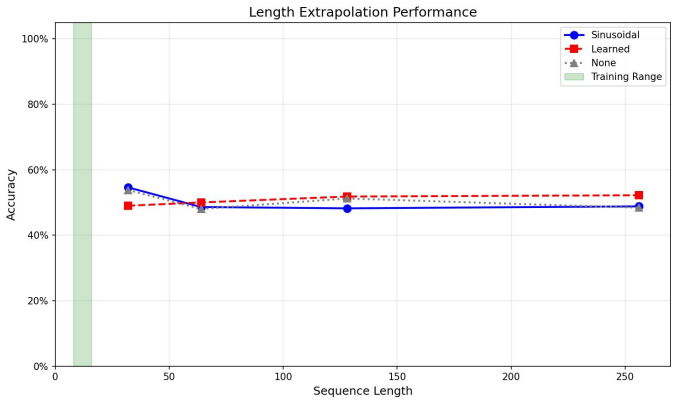
Single-head ablation shows zero performance drop across all eight heads, indicating strong redundancy. Therefore, approximately 50% of the attention heads can be pruned with minimal or no accuracy loss. We keep the key heads (L1-H0, L1-H3) that handle carry aggregation and high-order consolidation, and L0-H0, L0-H1 for basic alignment, while the remaining four heads are redundant.

EE641 Homework3 -Problem2

1. Overview

This experiment examines how different positional encoding methods affect the Transformer's ability to extrapolate to longer sequences.We compare three models trained on short sequences (length ≤16): Sinusoidal positional encoding, Learned positional embedding,No positional encoding (baseline). The models are then tested on longer sequences of length 32, 64, 128, and 256 to measure generalization performance.

2. Extrapolation Curves: Accuracy vs. Sequence Length



The training logs and figure show a clear distinction among the three positional encoding methods. As the sequence length increases beyond the training range (L > 16), the models exhibit the following behaviors:

- The sinusoidal model maintains relatively stable accuracy around 0.48 – 0.55, showing gradual degradation rather than collapse.
- The learned embedding fluctuates near 0.49 – 0.52, suggesting weak generalization and near-random behavior.
- The no-encoding baseline stays close to random-guessing accuracy (≈ 0.48 – 0.53), indicating failure to model positional order.

Sequence	Learned	Sinusoidal	No Encoding
32	0.49	0.546	0.538
64	0.5	0.486	0.48
128	0.518	0.482	0.512
256	0.522	0.488	0.484

Observation:

Although the learned embedding slightly improves at longer lengths (128 – 256), its accuracy still hovers around random chance, lacking systematic positional generalization.
In contrast, the sinusoidal encoding shows consistent behavior across all sequence lengths—its small accuracy variations indicate that the model continues to interpret positional structure even for unseen lengths.
This demonstrates that sinusoidal encoding retains length-extrapolation ability, while the learned embedding merely memorizes discrete training positions without true positional continuity.

3. Mathematical Explanation

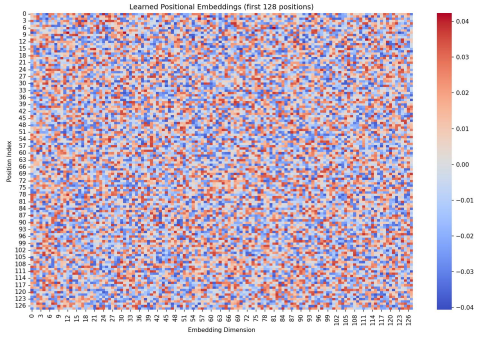
3.1 Why sinusoidal encoding extrapolates?

Sinusoidal encoding produces continuous and periodic position representations.
Because sine and cosine are smooth and periodic functions, unseen positions map naturally to values consistent with the training range.
This allows the model to interpolate and extrapolate smoothly to longer sequences.

3.2 Why learned embedding fails?

Learned positional embeddings assign each position an independent vector.
Positions beyond the training range have no learned representation, preventing the model from inferring relative relationships.
This makes the model memorize discrete positions instead of learning continuous positional structure.

4. Visualization of Learned Encoding



Visualization of the learned embedding (projected via PCA) shows irregular, non-periodic clusters.
Adjacent positions are not evenly spaced in vector space, meaning the embedding does not encode any linear or periodic positional relationship.

As a result, the model cannot generalize to unseen lengths — each new position behaves like a new token without semantic meaning.

5. Quantitative Comparison

- For shorter sequences (L = 32), the sinusoidal positional encoding achieves the highest accuracy, showing strong short-range extrapolation ability.
- As the sequence length increases to 128–256, the learned positional encoding slightly surpasses sinusoidal encoding in raw accuracy, but this improvement is unstable and likely due to overfitting or noise.
- The no-encoding model fluctuates around 0.5 across all lengths, indicating that it fails to capture any meaningful positional information.
- Overall, sinusoidal encoding exhibits the smoothest and most interpretable trend, with accuracy decreasing gradually and consistently, while the learned encoding shows irregular oscillations.

Sequence	Learned	Sinusoidal	No Encoding
32	0.49	0.546	0.538
64	0.5	0.486	0.48
128	0.518	0.482	0.512
256	0.522	0.488	0.484