SoME ablation runs v2

Series A: Stress-Testing Context Length (SEQ_LEN)

A1: Baseline

- Model Dimensions:
  - D_MODEL=384,
  - NUM_HEADS=6,
  - NUM_LAYERS=6
- SoME Parameters:
  - num_experts=64,
  - d_ffn=1024,
  - top_k=4
- Training Schedule:
  - 4 Epochs,
  - LR=8e-4
- Data Configuration:
  - SEQ_LEN=512,
  - BATCH_SIZE=128,
  - VOCAB_SIZE=8192,
  - train_subset_size=20000
  - Dataset = roneneldan/TinyStories
- Model & Training Scale:
  - Total Parameters: 313.27M
  - Trainable Parameters: 10.74M (3.4% of total)
  - Total Training Steps: 624
- Expert Initialization:
  - init_method="sparse”
- Results:
  - The Gini coefficient started high at 0.709 and remained high, finishing at 0.710.
  - The Shannon Entropy started low at 4.344 and remained low, finishing at 4.341.
  - Training Loss: Decreased steadily from 1.9960 at the end of Epoch 1 to 1.2650 at the end of Epoch 4.
  - Validation Loss: Decreased steadily from 1.4150 at the end of Epoch 1 to 1.2054 at the end of Epoch 4.
  - Validation Perplexity: Showed significant improvement, reducing from 4.12 to 3.34.

A2: Scale V1

- Model Dimensions:
  - D_MODEL=384,
  - NUM_HEADS=6,
  - NUM_LAYERS=6
- SoME Parameters:

    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=768,
    - BATCH_SIZE=42,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 1904
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient started very high at 0.742 and showed a slight but steady decrease to 0.719.
    - The Shannon Entropy started low at 3.862 and showed a slight but steady increase to 3.898.
    - Training Loss: Decreased steadily from 1.1475 at the end of Epoch 1 to 0.7356 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 0.8460 at the end of Epoch 1 to 0.7210 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 2.33 to 2.06.

A3: Scale v2

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,
    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:

    - SEQ_LEN=1024,
    - BATCH_SIZE=31,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 2580
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient started at an astronomical 0.787 and ended at 0.784 (essentially flat).
    - The Shannon Entropy started at a record low of 3.495 and ended at 3.503 (essentially flat).
    - Training Loss: Decreased steadily from 0.8477 at the end of Epoch 1 to 0.5387 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 0.6256 at the end of Epoch 1 to 0.5321 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 1.87 to 1.70.

New Series B: Stress-Testing Depth (with Long Context)

B1:

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=8,
    - NUM_LAYERS=8
- SoME Parameters:
    - num_experts=128,
    - d_ffn=1536,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=768,
    - BATCH_SIZE=64,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories

- Model & Training Scale:
    - Total Parameters: 1222.15M
    - Trainable Parameters: 12.23M (1.0% of total)
    - Total Training Steps: 1248
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient started at an absolutely massive 0.808 and remained flat.
    - The Shannon Entropy started at a record low of 4.092 and remained flat.
    - Training Loss: Decreased steadily from 1.1944 at the end of Epoch 1 to 0.7444 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 0.8615 at the end of Epoch 1 to 0.7243 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 2.37 to 2.06.

B2:

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=8,
    - NUM_LAYERS=10
- SoME Parameters:
    - num_experts=128,
    - d_ffn=1536,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=768,
    - BATCH_SIZE=64,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 1526.11M (1.53 Billion)
    - Trainable Parameters: 13.71M (0.9% of total)
    - Total Training Steps: 1248
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient started at a record-high 0.794, peaked at 0.798, and then showed a slight but clear decrease to 0.786.

    - The Shannon Entropy started at a record-low 4.134, dipped to 4.111, and then increased to 4.137.
    - Training Loss: Decreased steadily from 1.2537 at the end of Epoch 1 to 0.7264 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 0.8573 at the end of Epoch 1 to 0.7096 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 2.36 to 2.03.

Series C: Stress-Testing Width (with Long Context)

C1:

- Model Dimensions:
    - D_MODEL=512,
    - NUM_HEADS=8,
    - NUM_LAYERS=8
- SoME Parameters:
    - num_experts=128,
    - d_ffn=1536,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=768,
    - BATCH_SIZE=64,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 1631.63M (1.63 Billion)
    - Trainable Parameters: 18.92M (1.16% of total)
    - Total Training Steps: 1248
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient starts at a staggering 0.821—the highest initial concentration we have ever recorded. It then shows a healthy decrease to 0.793.
    - The Shannon Entropy starts at a record low of 4.058 and shows a healthy increase to 4.116.
    - Training Loss: Decreased steadily from 1.1968 at the end of Epoch 1 to 0.6864 at the end of Epoch 4.

    - Validation Loss: Decreased steadily from 0.8268 at the end of Epoch 1 to 0.6832 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 2.29 to 1.98.

C2:

- Model Dimensions:
    - D_MODEL=768,
    - NUM_HEADS=8,
    - NUM_LAYERS=8
- SoME Parameters:
    - num_experts=128,
    - d_ffn=1536,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=768,
    - BATCH_SIZE=64,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 2454.52M
    - Trainable Parameters: 36.24M
    - Total Training Steps: 1248
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient started at a mind-boggling 0.934 and rocketed to 0.957. A Gini of 0.957 means the workload is so concentrated that it's nearly a winner-take-all system. It's the mathematical equivalent of one expert doing almost all the work.
    - The Shannon Entropy started at a record-low 3.324 and plummeted to 2.764.
    - Training Loss: Decreased steadily from 2.3635 at the end of Epoch 1 to 2.3190 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 2.2214 at the end of Epoch 1 to 2.1578 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 9.22 to 8.65.