Training

June 9, 2024

[1]:
```
import torch
import matplotlib.pyplot as plt
from model import GPT
from trainer import Trainer
from utils import set_seed
import json
from TransformerDataset import TransformerTrainDataset

DEBUG = True  # Set to True to enable debugging

print('starting ....')

# Configuration
config = {
    'num_samples': 10000000,
    'input_dim': 2,
    'function_class': 'linear',
    'noise_std': 0,
    'model_type': 'tiny',
    'batch_size': 5,
    'min_context_size': 1,
    'max_context_size': 10,
    'max_iters': 3,
    'learning_rate': 10e-4,
    'num_workers': 0
}

# Create the TransformerTrainDataset
transformer_train_dataset = ␣
 ↪TransformerTrainDataset(num_samples=config['num_samples'],␣
 ↪input_dim=config['input_dim'], max_context_size=config['max_context_size'],␣
 ↪function_class=config['function_class'], noise_std=config['noise_std'])

# Debug TransformerTrainDataset
if DEBUG:
    print("TransformerTrainDataset samples:")
    for i in range(3):
```

```
        inputs, targets = transformer_train_dataset[i]
        print(f"Sample {i} - Inputs: {inputs.numpy()}, Targets: {targets.
 ↪numpy()}")

# Model configuration
set_seed(42)

model_config = GPT.get_default_config()
model_config.model_type = config['model_type']
model_config.vocab_size = config['input_dim'] + 1  # Adding 1 for the target␣
 ↪dimension
model_config.block_size = (2 * config['max_context_size']) *␣
 ↪config['input_dim']  # Updated block_size based on max context size
model_config.input_dim = config['input_dim']  # Set the input dimension

# Create the model
model = GPT(model_config)
print("number of parameters: %.2fM" % (sum(p.numel() for p in model.
 ↪parameters()) / 1e6,))

# Training configuration
train_config = Trainer.get_default_config()
train_config.learning_rate = config['learning_rate']
train_config.max_iters = config['max_iters']
train_config.num_workers = config['num_workers']
train_config.batch_size = config['batch_size']

# Create the trainer
trainer = Trainer(train_config, model, transformer_train_dataset)
print("running on device", trainer.device)

# Initialize weights for tracking
initial_scalar_head_weight = model.scalar_head.weight.clone().detach()
initial_wte_weight = model.transformer.wte.weight.clone().detach()

# Lists to store weights and losses
trainer.scalar_head_weights = []
trainer.wte_weights = []
trainer.train_losses = []

# Define a callback for batch end to print training status and debug weights
def batch_end_callback(trainer):
    current_scalar_head_weight = model.scalar_head.weight.clone().detach()
    current_wte_weight = model.transformer.wte.weight.clone().detach()

    scalar_head_weight_changed = not torch.equal(initial_scalar_head_weight,␣
 ↪current_scalar_head_weight)
```

```
    wte_weight_changed = not torch.equal(initial_wte_weight, current_wte_weight)

    # Store weights for plotting
    trainer.scalar_head_weights.append(current_scalar_head_weight.cpu().numpy())
    trainer.wte_weights.append(current_wte_weight.cpu().numpy())
    trainer.train_losses.append(trainer.loss.item())

    if trainer.iter_num % 100 == 0:
        print(f"iter_dt {trainer.iter_dt * 1000:.2f}ms; iter {trainer.iter_num}:
 ↪ train loss {trainer.loss.item():.5f}; Scalar Head Weights Changed:␣
 ↪{scalar_head_weight_changed}; WTE Weights Changed: {wte_weight_changed}")

    # Update the initial weights
    initial_scalar_head_weight.copy_(current_scalar_head_weight)
    initial_wte_weight.copy_(current_wte_weight)

trainer.set_callback('on_batch_end', batch_end_callback)

# Run the training
trainer.run()

# Save the model and configurations
torch.save(model.state_dict(), 'trained_model.pth')
with open('config.json', 'w') as f:
    json.dump(config, f)
print("Model and configurations saved.")

# Plot the stored weight values
plt.figure(figsize=(12, 6))
for i in range(min(len(trainer.scalar_head_weights[0].flatten()), 5)):  # Plot␣
 ↪up to 5 scalar head weights
    plt.plot([w.flatten()[i] for w in trainer.scalar_head_weights],␣
 ↪label=f'Scalar Head Weight {i}')
plt.xlabel('Iterations')
plt.ylabel('Weight Value')
plt.title('Scalar Head Weights Over Time')
plt.legend()
plt.grid(True)
plt.show()

plt.figure(figsize=(12, 6))
for i in range(min(len(trainer.wte_weights[0].flatten()), 5)):  # Plot up to 5␣
 ↪wte weights
    plt.plot([w.flatten()[i] for w in trainer.wte_weights], label=f'WTE Weight␣
 ↪{i}')
plt.xlabel('Iterations')
plt.ylabel('Weight Value')
```

```
plt.title('Word Token Embedding Weights Over Time')
plt.legend()
plt.grid(True)
plt.show()

# Plot the training loss
plt.figure(figsize=(12, 6))
plt.plot(trainer.train_losses, label='Training Loss')
plt.xlabel('Iterations')
plt.ylabel('Loss')
plt.title('Training Loss Over Time')
plt.legend()
plt.grid(True)
plt.show()
```

```
starting …
TransformerTrainDataset samples:
Sample 0 - Inputs: [[ 0.69150347  1.813534  ]
 [-0.7195421   0.        ]
 [-1.1353933  -0.962374  ]
 [ 0.30963057  0.        ]
 [-0.66690356 -0.10963406]
 [-0.01523633  0.        ]
 [ 0.2379293  -1.5882341 ]
 [ 0.70941     0.        ]
 [-2.3150156   1.0578417 ]
 [-0.67513275  0.        ]
 [ 0.8866422  -1.7433792 ]
 [ 0.8374779   0.        ]
 [ 0.19102618  0.8615962 ]
 [-0.35476878  0.        ]
 [-0.80496424  0.81298614]
 [-0.42732477  0.        ]
 [ 0.66849196 -0.8589745 ]
 [ 0.43439582  0.        ]
 [-1.1372952  -0.7144748 ]
 [ 0.20221318  0.        ]
 [ 1.7278594  -1.2764744 ]
 [ 0.          0.        ]], Targets: [[ 0.          0.        ]
 [-0.7195421   0.        ]
 [ 0.          0.        ]
 [ 0.30963057  0.        ]
 [ 0.          0.        ]
 [-0.01523633  0.        ]
 [ 0.          0.        ]
 [ 0.70941     0.        ]
 [ 0.          0.        ]
```

```
 [-0.67513275  0.        ]
 [ 0.          0.        ]
 [ 0.8374779   0.        ]
 [ 0.          0.        ]
 [-0.35476878  0.        ]
 [ 0.          0.        ]
 [-0.42732477  0.        ]
 [ 0.          0.        ]
 [ 0.43439582  0.        ]
 [ 0.          0.        ]
 [ 0.20221318  0.        ]
 [ 0.          0.        ]
 [ 0.7145412   0.        ]]
Sample 1 - Inputs: [[ 2.1503491   0.04427584]
 [-2.7804537   0.        ]
 [-0.93444043  1.5775617 ]
 [ 1.0031055   0.        ]
 [ 1.5423133   0.02527745]
 [-1.9934163   0.        ]
 [-2.1957283   0.77417475]
 [ 2.73386     0.        ]
 [ 0.6674108  -0.37484822]
 [-0.8130544   0.        ]
 [ 0.19406931  0.12413976]
 [-0.2663718   0.        ]
 [ 0.4368768  -0.2980698 ]
 [-0.52544224  0.        ]
 [-1.052432   -0.5234039 ]
 [ 1.4252802   0.        ]
 [-0.15939331  1.6418892 ]
 [-0.00526345  0.        ]
 [ 0.7602862   1.3299204 ]
 [-1.1519185   0.        ]
 [-0.16865624  1.5033717 ]
 [ 0.          0.        ]], Targets: [[ 0.          0.        ]
 [-2.7804537   0.        ]
 [ 0.          0.        ]
 [ 1.0031055   0.        ]
 [ 0.          0.        ]
 [-1.9934163   0.        ]
 [ 0.          0.        ]
 [ 2.73386     0.        ]
 [ 0.          0.        ]
 [-0.8130544   0.        ]
 [ 0.          0.        ]
 [-0.2663718   0.        ]
 [ 0.          0.        ]
 [-0.52544224  0.        ]
```

```
 [ 0.          0.        ]
 [ 1.4252802   0.        ]
 [ 0.          0.        ]
 [-0.00526345  0.        ]
 [ 0.          0.        ]
 [-1.1519185   0.        ]
 [ 0.          0.        ]
 [ 0.02448523  0.        ]]
Sample 2 - Inputs: [[ 1.416772   -0.72487885]
 [ 0.4922129   0.        ]
 [-0.60053647 -0.00378863]
 [ 0.00606271  0.        ]
 [ 1.3371232   0.9741978 ]
 [-0.68011475  0.        ]
 [ 1.3592391  -0.62902737]
 [ 0.42638183  0.        ]
 [ 0.8015495  -1.3281966 ]
 [ 0.91218394  0.        ]
 [ 1.1125722  -0.83857816]
 [ 0.57244     0.        ]
 [ 1.4427917  -1.4251064 ]
 [ 0.9753944   0.        ]
 [ 0.23305182 -1.4527811 ]
 [ 1.0014418   0.        ]
 [-0.04643335  0.2470446 ]
 [-0.17025556  0.        ]
 [ 0.29267567  0.35035464]
 [-0.24351193  0.        ]
 [ 0.7920302  -0.84428716]
 [ 0.          0.        ]], Targets: [[ 0.          0.        ]
 [ 0.4922129   0.        ]
 [ 0.          0.        ]
 [ 0.00606271  0.        ]
 [ 0.          0.        ]
 [-0.68011475  0.        ]
 [ 0.          0.        ]
 [ 0.42638183  0.        ]
 [ 0.          0.        ]
 [ 0.91218394  0.        ]
 [ 0.          0.        ]
 [ 0.57244     0.        ]
 [ 0.          0.        ]
 [ 0.9753944   0.        ]
 [ 0.          0.        ]
 [ 1.0014418   0.        ]
 [ 0.          0.        ]
 [-0.17025556  0.        ]
 [ 0.          0.        ]
```

```
 [-0.24351193  0.        ]
 [ 0.          0.        ]
 [ 0.57822084  0.        ]]
number of parameters: 0.15M
number of parameters: 0.15M
running on device cuda
running on device cuda
inputs: tensor([[[ 0.6477,  1.5230],
         [ 0.1111,  0.0000],
         [-0.2341,  1.5792],
         [-0.3346,  0.0000],
         [-0.4695,  0.5426],
         [-0.3082,  0.0000],
         [-0.4657,  0.2420],
         [-0.2648,  0.0000],
         [-1.7249, -0.5623],
         [-0.7790,  0.0000],
         [ 0.3142, -0.9080],
         [ 0.2816,  0.0000],
         [ 1.4656, -0.2258],
         [ 0.7592,  0.0000],
         [-1.4247, -0.5444],
         [-0.6324,  0.0000],
         [-1.1510,  0.3757],
         [-0.6237,  0.0000],
         [-0.2917, -0.6017],
         [-0.0617,  0.0000],
         [-0.0135, -1.0577],
         [ 0.0000,  0.0000]],

        [[-1.9597, -1.3282],
         [ 2.1150,  0.0000],
         [ 0.7385,  0.1714],
         [-0.8658,  0.0000],
         [-0.3011, -1.4785],
         [ 0.0588,  0.0000],
         [-0.4606,  1.0571],
         [ 0.7832,  0.0000],
         [-1.7630,  0.3241],
         [ 2.2201,  0.0000],
         [-0.6769,  0.6117],
         [ 0.9542,  0.0000],
         [ 0.9313, -0.8392],
         [-1.3122,  0.0000],
         [ 0.3313,  0.9755],
         [-0.2007,  0.0000],
         [-0.1857, -1.1063],
         [-0.0044,  0.0000],
```

```
 [ 0.8125,  1.3562],
 [-0.7087,  0.0000],
 [ 1.0035,  0.3616],
 [ 0.0000,  0.0000]],

[[-0.0358,  1.5646],
 [ 2.3935,  0.0000],
 [ 0.8219,  0.0870],
 [ 0.4309,  0.0000],
 [ 0.0918, -1.9876],
 [-3.0238,  0.0000],
 [ 0.3571,  1.4779],
 [ 2.4021,  0.0000],
 [-0.8085, -0.5018],
 [-1.0639,  0.0000],
 [ 0.3288, -0.5298],
 [-0.6960,  0.0000],
 [ 0.0971,  0.9686],
 [ 1.5249,  0.0000],
 [-0.3277, -0.3921],
 [-0.7215,  0.0000],
 [ 0.2961,  0.2611],
 [ 0.5085,  0.0000],
 [-0.2346, -1.4154],
 [-2.2617,  0.0000],
 [-0.3427, -0.8023],
 [ 0.0000,  0.0000]],

[[ 0.1746,  0.2576],
 [ 0.5563,  0.0000],
 [-1.9188, -0.0265],
 [-0.8253,  0.0000],
 [ 2.4632, -0.1924],
 [ 0.6324,  0.0000],
 [-0.0347, -1.1687],
 [-2.2184,  0.0000],
 [ 0.7519,  0.7910],
 [ 1.7959,  0.0000],
 [ 1.4028, -1.4019],
 [-2.0774,  0.0000],
 [ 2.1905, -0.9905],
 [-0.9833,  0.0000],
 [ 0.0997, -0.5035],
 [-0.9094,  0.0000],
 [ 0.0686, -1.0623],
 [-1.9760,  0.0000],
 [-0.9194,  1.5499],
 [ 2.5520,  0.0000],
```

```
          [-0.3221,  0.8135],
          [ 0.0000,  0.0000]],

         [[-1.6075,  0.1846],
          [-0.1243,  0.0000],
          [ 0.7818, -1.2370],
          [-1.4390,  0.0000],
          [ 0.5219,  0.2970],
          [ 0.5069,  0.0000],
          [ 0.3464, -0.6800],
          [-0.8101,  0.0000],
          [ 0.2931, -0.7144],
          [-0.8671,  0.0000],
          [ 0.4738, -1.1913],
          [-1.4494,  0.0000],
          [-0.9747,  0.7871],
          [ 0.8071,  0.0000],
          [-0.8207,  0.9634],
          [ 1.0726,  0.0000],
          [ 0.8221,  1.8968],
          [ 2.6664,  0.0000],
          [-0.7537, -0.8895],
          [-1.3342,  0.0000],
          [-0.0771,  0.3412],
          [ 0.0000,  0.0000]]], device='cuda:0')
targets: tensor([[[ 0.0000,  0.0000],
          [ 0.1111,  0.0000],
          [ 0.0000,  0.0000],
          [-0.3346,  0.0000],
          [ 0.0000,  0.0000],
          [-0.3082,  0.0000],
          [ 0.0000,  0.0000],
          [-0.2648,  0.0000],
          [ 0.0000,  0.0000],
          [-0.7790,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.2816,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.7592,  0.0000],
          [ 0.0000,  0.0000],
          [-0.6324,  0.0000],
          [ 0.0000,  0.0000],
          [-0.6237,  0.0000],
          [ 0.0000,  0.0000],
          [-0.0617,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.1395,  0.0000]],
```

```
[[ 0.0000,  0.0000],
 [ 2.1150,  0.0000],
 [ 0.0000,  0.0000],
 [-0.8658,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0588,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.7832,  0.0000],
 [ 0.0000,  0.0000],
 [ 2.2201,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.9542,  0.0000],
 [ 0.0000,  0.0000],
 [-1.3122,  0.0000],
 [ 0.0000,  0.0000],
 [-0.2007,  0.0000],
 [ 0.0000,  0.0000],
 [-0.0044,  0.0000],
 [ 0.0000,  0.0000],
 [-0.7087,  0.0000],
 [ 0.0000,  0.0000],
 [-1.1496,  0.0000]],

[[ 0.0000,  0.0000],
 [ 2.3935,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.4309,  0.0000],
 [ 0.0000,  0.0000],
 [-3.0238,  0.0000],
 [ 0.0000,  0.0000],
 [ 2.4021,  0.0000],
 [ 0.0000,  0.0000],
 [-1.0639,  0.0000],
 [ 0.0000,  0.0000],
 [-0.6960,  0.0000],
 [ 0.0000,  0.0000],
 [ 1.5249,  0.0000],
 [ 0.0000,  0.0000],
 [-0.7215,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.5085,  0.0000],
 [ 0.0000,  0.0000],
 [-2.2617,  0.0000],
 [ 0.0000,  0.0000],
 [-1.3578,  0.0000]],

[[ 0.0000,  0.0000],
 [ 0.5563,  0.0000],
```

```
          [ 0.0000,  0.0000],
          [-0.8253,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.6324,  0.0000],
          [ 0.0000,  0.0000],
          [-2.2184,  0.0000],
          [ 0.0000,  0.0000],
          [ 1.7959,  0.0000],
          [ 0.0000,  0.0000],
          [-2.0774,  0.0000],
          [ 0.0000,  0.0000],
          [-0.9833,  0.0000],
          [ 0.0000,  0.0000],
          [-0.9094,  0.0000],
          [ 0.0000,  0.0000],
          [-1.9760,  0.0000],
          [ 0.0000,  0.0000],
          [ 2.5520,  0.0000],
          [ 0.0000,  0.0000],
          [ 1.4043,  0.0000]],

         [[ 0.0000,  0.0000],
          [-0.1243,  0.0000],
          [ 0.0000,  0.0000],
          [-1.4390,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.5069,  0.0000],
          [ 0.0000,  0.0000],
          [-0.8101,  0.0000],
          [ 0.0000,  0.0000],
          [-0.8671,  0.0000],
          [ 0.0000,  0.0000],
          [-1.4494,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.8071,  0.0000],
          [ 0.0000,  0.0000],
          [ 1.0726,  0.0000],
          [ 0.0000,  0.0000],
          [ 2.6664,  0.0000],
          [ 0.0000,  0.0000],
          [-1.3342,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.4284,  0.0000]]], device='cuda:0')
target_predictions: tensor([ 0.0325, -0.0074, -0.0177, -0.0123, -0.0352,
0.0111,  0.0350, -0.0230,
        -0.0097,  0.0213, -0.0129,  0.0248, -0.0289,  0.0443,  0.0108,  0.0244,
         0.0176, -0.0239, -0.0377, -0.0393, -0.0468, -0.0886,  0.0234,  0.0262,
        -0.0243,  0.0222, -0.0248, -0.0254,  0.0210, -0.0382,  0.0085, -0.0237,
```

```
        -0.0031,  0.0281, -0.0147,  0.0134, -0.0218,  0.0272, -0.0174, -0.0145,
        -0.0279, -0.0262,  0.0251, -0.0347, -0.0074, -0.0145, -0.0081, -0.0107,
        -0.0359, -0.0237, -0.0031,  0.0156,  0.0264, -0.0232,  0.0267],
       device='cuda:0', grad_fn=<IndexBackward0>)
non_zero_targets: tensor([ 0.1111, -0.3346, -0.3082, -0.2648, -0.7790,  0.2816,
0.7592, -0.6324,
        -0.6237, -0.0617,  0.1395,  2.1150, -0.8658,  0.0588,  0.7832,  2.2201,
         0.9542, -1.3122, -0.2007, -0.0044, -0.7087, -1.1496,  2.3935,  0.4309,
        -3.0238,  2.4021, -1.0639, -0.6960,  1.5249, -0.7215,  0.5085, -2.2617,
        -1.3578,  0.5563, -0.8253,  0.6324, -2.2184,  1.7959, -2.0774, -0.9833,
        -0.9094, -1.9760,  2.5520,  1.4043, -0.1243, -1.4390,  0.5069, -0.8101,
        -0.8671, -1.4494,  0.8071,  1.0726,  2.6664, -1.3342,  0.4284],
       device='cuda:0')
iter_dt 0.00ms; iter 0: train loss 1.67970; Scalar Head Weights Changed: True;
WTE Weights Changed: True
inputs: tensor([[[ 1.4535, -0.2647],
         [ 1.1989,  0.0000],
         [ 0.6257, -0.8572],
         [ 0.5064,  0.0000],
         [ 0.4825, -0.2235],
         [ 0.3962,  0.0000],
         [ 0.4732, -0.0728],
         [ 0.3905,  0.0000],
         [-1.5148, -0.4465],
         [-1.2589,  0.0000],
         [ 0.2141, -1.2457],
         [ 0.1609,  0.0000],
         [ 0.3853, -0.8839],
         [ 0.3072,  0.0000],
         [ 0.0582, -1.1430],
         [ 0.0333,  0.0000],
         [ 0.5608,  1.0831],
         [ 0.4780,  0.0000],
         [-1.3777, -0.9378],
         [-1.1518,  0.0000],
         [ 0.5138,  0.5150],
         [ 0.0000,  0.0000]],

        [[ 0.9540,  0.6514],
         [ 1.2843,  0.0000],
         [ 0.7590, -0.7728],
         [-0.4443,  0.0000],
         [-0.4854,  0.0819],
         [-0.1841,  0.0000],
         [-1.8673,  0.6863],
         [-0.2867,  0.0000],
         [-0.4719,  1.0890],
         [ 0.9672,  0.0000],
```

```
 [-1.0777, -0.7153],
 [-1.4275,  0.0000],
 [-0.7304,  0.2165],
 [-0.1712,  0.0000],
 [-0.6516,  2.1439],
 [ 2.0626,  0.0000],
 [-2.0251,  0.1865],
 [-0.9444,  0.0000],
 [ 0.8524, -0.7925],
 [-0.4133,  0.0000],
 [ 0.5050,  0.8658],
 [ 0.0000,  0.0000]],

[[-0.6533,  1.7655],
 [-0.6200,  0.0000],
 [-1.2609,  0.9179],
 [-0.0142,  0.0000],
 [ 1.0325, -1.5194],
 [ 0.3763,  0.0000],
 [ 1.2669, -0.7077],
 [-0.0877,  0.0000],
 [ 0.7746, -0.9269],
 [ 0.1811,  0.0000],
 [-3.2413, -1.0244],
 [ 1.5707,  0.0000],
 [-1.2478,  1.6324],
 [-0.3579,  0.0000],
 [-0.4400,  0.1307],
 [ 0.0851,  0.0000],
 [-1.4359,  1.1632],
 [-0.0721,  0.0000],
 [-0.9815,  0.4621],
 [ 0.1088,  0.0000],
 [-0.6002,  0.0698],
 [ 0.0000,  0.0000]],

[[ 1.5860, -1.2378],
 [-0.6396,  0.0000],
 [-1.9521, -0.1518],
 [-0.3221,  0.0000],
 [ 0.2810, -0.6227],
 [-0.3804,  0.0000],
 [-0.4930, -0.5894],
 [-0.4462,  0.0000],
 [ 0.3570, -0.6929],
 [-0.4183,  0.0000],
 [ 0.3073,  0.8129],
 [ 0.5731,  0.0000],
```

```
          [-0.8290, -0.5602],
          [-0.4650,  0.0000],
          [ 0.6104, -0.0209],
          [ 0.0554,  0.0000],
          [ 1.2777, -0.5916],
          [-0.2467,  0.0000],
          [-0.2022, -0.2177],
          [-0.1671,  0.0000],
          [ 0.8254,  0.8135],
          [ 0.0000,  0.0000]],

         [[-0.3103,  0.3242],
          [ 0.2145,  0.0000],
          [ 0.0970,  0.5952],
          [ 0.4079,  0.0000],
          [ 2.0924, -1.0060],
          [-0.6421,  0.0000],
          [ 1.1581,  0.7917],
          [ 0.5642,  0.0000],
          [ 0.6283, -0.0122],
          [ 0.0048,  0.0000],
          [ 0.0758, -0.6772],
          [-0.4602,  0.0000],
          [-0.1471, -0.8255],
          [-0.5660,  0.0000],
          [ 0.4129, -0.5637],
          [-0.3758,  0.0000],
          [ 0.2437,  0.2450],
          [ 0.1722,  0.0000],
          [-0.4710,  0.2320],
          [ 0.1484,  0.0000],
          [-1.4075, -0.7184],
          [ 0.0000,  0.0000]]], device='cuda:0')
targets: tensor([[[ 0.0000,  0.0000],
          [ 1.1989,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.5064,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.3962,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.3905,  0.0000],
          [ 0.0000,  0.0000],
          [-1.2589,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.1609,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.3072,  0.0000],
          [ 0.0000,  0.0000],
```

```
 [ 0.0333,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.4780,  0.0000],
 [ 0.0000,  0.0000],
 [-1.1518,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.4317,  0.0000]],

[[ 0.0000,  0.0000],
 [ 1.2843,  0.0000],
 [ 0.0000,  0.0000],
 [-0.4443,  0.0000],
 [ 0.0000,  0.0000],
 [-0.1841,  0.0000],
 [ 0.0000,  0.0000],
 [-0.2867,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.9672,  0.0000],
 [ 0.0000,  0.0000],
 [-1.4275,  0.0000],
 [ 0.0000,  0.0000],
 [-0.1712,  0.0000],
 [ 0.0000,  0.0000],
 [ 2.0626,  0.0000],
 [ 0.0000,  0.0000],
 [-0.9444,  0.0000],
 [ 0.0000,  0.0000],
 [-0.4133,  0.0000],
 [ 0.0000,  0.0000],
 [ 1.2714,  0.0000]],

[[ 0.0000,  0.0000],
 [-0.6200,  0.0000],
 [ 0.0000,  0.0000],
 [-0.0142,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.3763,  0.0000],
 [ 0.0000,  0.0000],
 [-0.0877,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1811,  0.0000],
 [ 0.0000,  0.0000],
 [ 1.5707,  0.0000],
 [ 0.0000,  0.0000],
 [-0.3579,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0851,  0.0000],
 [ 0.0000,  0.0000],
```

```
 [-0.0721,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1088,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1676,  0.0000]],

[[ 0.0000,  0.0000],
 [-0.6396,  0.0000],
 [ 0.0000,  0.0000],
 [-0.3221,  0.0000],
 [ 0.0000,  0.0000],
 [-0.3804,  0.0000],
 [ 0.0000,  0.0000],
 [-0.4462,  0.0000],
 [ 0.0000,  0.0000],
 [-0.4183,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.5731,  0.0000],
 [ 0.0000,  0.0000],
 [-0.4650,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0554,  0.0000],
 [ 0.0000,  0.0000],
 [-0.2467,  0.0000],
 [ 0.0000,  0.0000],
 [-0.1671,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.6323,  0.0000]],

[[ 0.0000,  0.0000],
 [ 0.2145,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.4079,  0.0000],
 [ 0.0000,  0.0000],
 [-0.6421,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.5642,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0048,  0.0000],
 [ 0.0000,  0.0000],
 [-0.4602,  0.0000],
 [ 0.0000,  0.0000],
 [-0.5660,  0.0000],
 [ 0.0000,  0.0000],
 [-0.3758,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1722,  0.0000],
 [ 0.0000,  0.0000],
```

```
         [ 0.1484,  0.0000],
         [ 0.0000,  0.0000],
         [-0.5195,  0.0000]]], device='cuda:0')
target_predictions: tensor([ 0.1955,  0.1766,  0.1799,  0.1792, -0.1969,
0.1374,  0.1666, -0.0556,
         0.1901, -0.2000, -0.0839,  0.1997, -0.1638, -0.1811, -0.1919,  0.2014,
        -0.2067, -0.1889,  0.1954, -0.2103, -0.1950, -0.1398, -0.1937, -0.1558,
         0.1895, -0.1714,  0.1891,  0.2006, -0.1957,  0.1193, -0.1648,  0.1089,
        -0.1426, -0.2025, -0.1973, -0.2062, -0.2033, -0.2078,  0.1853, -0.2030,
        -0.0471, -0.2044, -0.1938, -0.1383,  0.1208,  0.1989, -0.1919,  0.1973,
        -0.0305, -0.1846, -0.1891, -0.1771,  0.1602,  0.1623, -0.1422],
       device='cuda:0', grad_fn=<IndexBackward0>)
non_zero_targets: tensor([ 1.1989,  0.5064,  0.3962,  0.3905, -1.2589,  0.1609,
0.3072,  0.0333,
         0.4780, -1.1518,  0.4317,  1.2843, -0.4443, -0.1841, -0.2867,  0.9672,
        -1.4275, -0.1712,  2.0626, -0.9444, -0.4133,  1.2714, -0.6200, -0.0142,
         0.3763, -0.0877,  0.1811,  1.5707, -0.3579,  0.0851, -0.0721,  0.1088,
         0.1676, -0.6396, -0.3221, -0.3804, -0.4462, -0.4183,  0.5731, -0.4650,
         0.0554, -0.2467, -0.1671,  0.6323,  0.2145,  0.4079, -0.6421,  0.5642,
         0.0048, -0.4602, -0.5660, -0.3758,  0.1722,  0.1484, -0.5195],
       device='cuda:0')
inputs: tensor([[[ 8.5766e-01, -1.5994e-01],
         [ 3.0687e-02,  0.0000e+00],
         [-1.0025e+00, -1.8513e-02],
         [-3.3901e-01,  0.0000e+00],
         [ 3.2272e-01, -8.2723e-01],
         [-1.1201e+00,  0.0000e+00],
         [ 1.5327e+00, -1.0876e-01],
         [ 3.1608e-01,  0.0000e+00],
         [ 6.9014e-01, -4.0122e-01],
         [-3.7737e-01,  0.0000e+00],
         [ 1.2592e-02,  9.7676e-02],
         [ 1.4802e-01,  0.0000e+00],
         [ 2.4510e-02,  4.9800e-01],
         [ 7.4235e-01,  0.0000e+00],
         [ 9.5927e-01,  2.1532e+00],
         [ 3.4750e+00,  0.0000e+00],
         [ 8.7232e-01,  1.8334e-01],
         [ 5.4171e-01,  0.0000e+00],
         [-8.0830e-01, -8.3972e-01],
         [-1.4902e+00,  0.0000e+00],
         [-2.1239e+00, -5.2576e-01],
         [ 0.0000e+00,  0.0000e+00]],

        [[ 1.8762e+00,  9.5042e-01],
         [ 6.0698e-01,  0.0000e+00],
         [-8.9841e-01,  4.9192e-01],
         [ 3.3000e-02,  0.0000e+00],
```

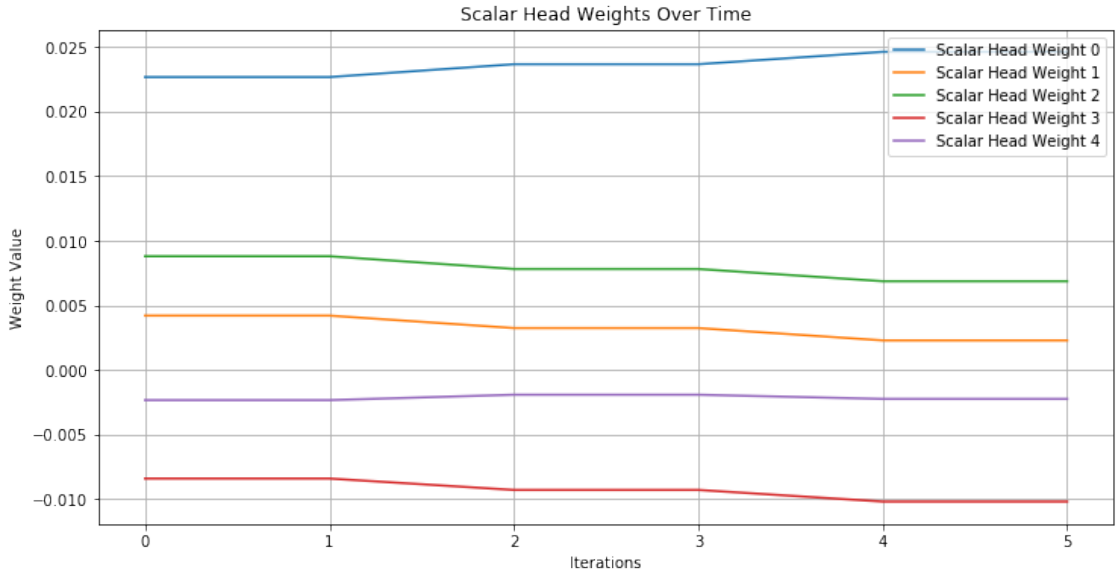
```
 [ 1.8315e+00,  1.1794e+00],
 [ 6.7852e-01,  0.0000e+00],
 [-1.7131e+00,  1.3539e+00],
 [ 2.0505e-01,  0.0000e+00],
 [ 1.2378e+00, -1.5944e+00],
 [-3.5875e-01,  0.0000e+00],
 [ 5.2437e-03,  4.6981e-02],
 [ 1.6845e-02,  0.0000e+00],
 [ 6.2285e-01, -1.0676e+00],
 [-2.7119e-01,  0.0000e+00],
 [ 1.2030e-01,  5.1444e-01],
 [ 1.9390e-01,  0.0000e+00],
 [-1.1246e+00, -1.5341e+00],
 [-6.9343e-01,  0.0000e+00],
 [ 3.3231e-01, -7.4849e-01],
 [-2.0582e-01,  0.0000e+00],
 [ 1.1567e-01,  1.1793e+00],
 [ 0.0000e+00,  0.0000e+00]],

[[-2.4896e-01,  9.7157e-01],
 [ 1.1924e+00,  0.0000e+00],
 [ 1.3686e+00, -9.6492e-01],
 [ 1.1266e+00,  0.0000e+00],
 [ 1.0584e+00, -1.7587e+00],
 [-9.0604e-01,  0.0000e+00],
 [-2.0392e+00, -2.6941e-01],
 [-4.6752e+00,  0.0000e+00],
 [ 1.5024e+00,  7.4095e-02],
 [ 3.2260e+00,  0.0000e+00],
 [-1.3801e+00, -1.7034e+00],
 [-5.8341e+00,  0.0000e+00],
 [ 3.8407e-01, -3.2695e-02],
 [ 7.3407e-01,  0.0000e+00],
 [-8.9120e-02, -1.3045e+00],
 [-2.4734e+00,  0.0000e+00],
 [ 3.6660e-01, -9.3988e-01],
 [-8.9434e-01,  0.0000e+00],
 [-1.0592e+00, -6.2679e-02],
 [-2.2928e+00,  0.0000e+00],
 [-9.8573e-01,  5.0405e-01],
 [ 0.0000e+00,  0.0000e+00]],

[[-1.0352e+00, -5.5365e-01],
 [ 8.8007e-01,  0.0000e+00],
 [ 1.9647e+00,  3.5264e-02],
 [-1.5616e+00,  0.0000e+00],
 [ 2.1398e-01, -1.1233e-01],
 [-1.5764e-01,  0.0000e+00],
```

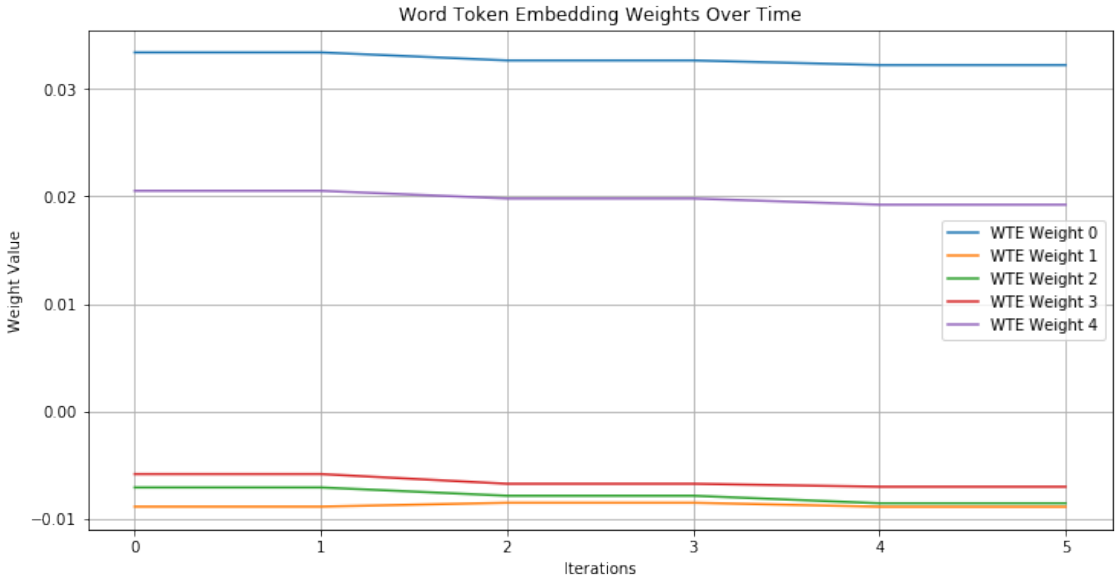
```
          [ 6.1417e-01,  7.5751e-01],
          [-5.6803e-01,  0.0000e+00],
          [-5.7582e-01, -2.7505e-01],
          [ 4.8599e-01,  0.0000e+00],
          [-1.5152e+00,  1.3669e+00],
          [ 1.0551e+00,  0.0000e+00],
          [-2.4904e-01,  5.7656e-01],
          [ 1.3574e-01,  0.0000e+00],
          [ 3.0789e+00,  1.1196e+00],
          [-2.5610e+00,  0.0000e+00],
          [-9.5554e-01, -1.6064e+00],
          [ 9.2956e-01,  0.0000e+00],
          [-7.5635e-01, -1.4223e+00],
          [ 7.5191e-01,  0.0000e+00],
          [-1.0815e+00,  1.6871e+00],
          [ 0.0000e+00,  0.0000e+00]],

         [[ 7.7368e-02, -8.6128e-01],
          [-1.2753e+00,  0.0000e+00],
          [ 5.3891e-01, -1.0372e+00],
          [-1.5394e+00,  0.0000e+00],
          [-8.7562e-01, -1.3828e+00],
          [-2.0395e+00,  0.0000e+00],
          [ 1.9094e+00, -1.3986e+00],
          [-2.0850e+00,  0.0000e+00],
          [-6.5064e-01, -4.8713e-01],
          [-7.1573e-01,  0.0000e+00],
          [-8.6399e-01,  4.8522e-02],
          [ 7.8698e-02,  0.0000e+00],
          [ 2.7046e-01, -5.0238e-02],
          [-7.6506e-02,  0.0000e+00],
          [-9.0756e-01, -5.7677e-01],
          [-8.4635e-01,  0.0000e+00],
          [ 5.0092e-01, -9.7756e-01],
          [-1.4507e+00,  0.0000e+00],
          [ 7.5139e-01, -1.6694e+00],
          [-2.4766e+00,  0.0000e+00],
          [-6.6262e-01,  5.7060e-01],
          [ 0.0000e+00,  0.0000e+00]]], device='cuda:0')
targets: tensor([[[ 0.0000,  0.0000],
          [ 0.0307,  0.0000],
          [ 0.0000,  0.0000],
          [-0.3390,  0.0000],
          [ 0.0000,  0.0000],
          [-1.1201,  0.0000],
          [ 0.0000,  0.0000],
          [ 0.3161,  0.0000],
          [ 0.0000,  0.0000],
```

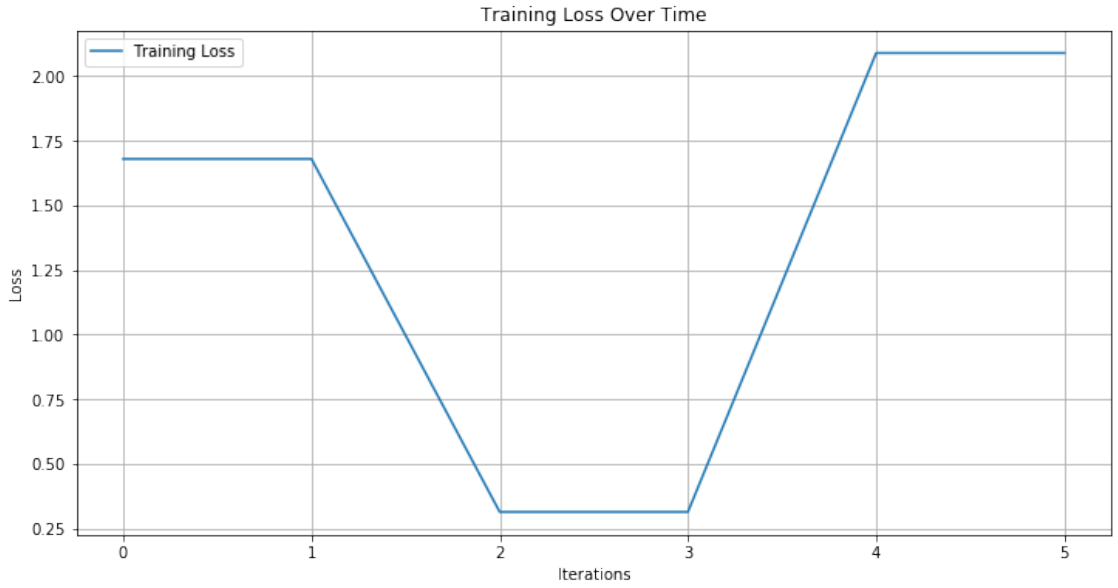
```
 [-0.3774,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1480,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.7423,  0.0000],
 [ 0.0000,  0.0000],
 [ 3.4750,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.5417,  0.0000],
 [ 0.0000,  0.0000],
 [-1.4902,  0.0000],
 [ 0.0000,  0.0000],
 [-1.4360,  0.0000]],

[[ 0.0000,  0.0000],
 [ 0.6070,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0330,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.6785,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.2050,  0.0000],
 [ 0.0000,  0.0000],
 [-0.3587,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0168,  0.0000],
 [ 0.0000,  0.0000],
 [-0.2712,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1939,  0.0000],
 [ 0.0000,  0.0000],
 [-0.6934,  0.0000],
 [ 0.0000,  0.0000],
 [-0.2058,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.4204,  0.0000]],

[[ 0.0000,  0.0000],
 [ 1.1924,  0.0000],
 [ 0.0000,  0.0000],
 [ 1.1266,  0.0000],
 [ 0.0000,  0.0000],
 [-0.9060,  0.0000],
 [ 0.0000,  0.0000],
 [-4.6752,  0.0000],
 [ 0.0000,  0.0000],
 [ 3.2260,  0.0000],
 [ 0.0000,  0.0000],
```

```
 [-5.8341,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.7341,  0.0000],
 [ 0.0000,  0.0000],
 [-2.4734,  0.0000],
 [ 0.0000,  0.0000],
 [-0.8943,  0.0000],
 [ 0.0000,  0.0000],
 [-2.2928,  0.0000],
 [ 0.0000,  0.0000],
 [-1.1466,  0.0000]],

[[ 0.0000,  0.0000],
 [ 0.8801,  0.0000],
 [ 0.0000,  0.0000],
 [-1.5616,  0.0000],
 [ 0.0000,  0.0000],
 [-0.1576,  0.0000],
 [ 0.0000,  0.0000],
 [-0.5680,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.4860,  0.0000],
 [ 0.0000,  0.0000],
 [ 1.0551,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.1357,  0.0000],
 [ 0.0000,  0.0000],
 [-2.5610,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.9296,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.7519,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.6770,  0.0000]],

[[ 0.0000,  0.0000],
 [-1.2753,  0.0000],
 [ 0.0000,  0.0000],
 [-1.5394,  0.0000],
 [ 0.0000,  0.0000],
 [-2.0395,  0.0000],
 [ 0.0000,  0.0000],
 [-2.0850,  0.0000],
 [ 0.0000,  0.0000],
 [-0.7157,  0.0000],
 [ 0.0000,  0.0000],
 [ 0.0787,  0.0000],
 [ 0.0000,  0.0000],
```

```
         [-0.0765,  0.0000],
         [ 0.0000,  0.0000],
         [-0.8464,  0.0000],
         [ 0.0000,  0.0000],
         [-1.4507,  0.0000],
         [ 0.0000,  0.0000],
         [-2.4766,  0.0000],
         [ 0.0000,  0.0000],
         [ 0.8497,  0.0000]]], device='cuda:0')
target_predictions: tensor([ 0.2420, -0.0308, -0.2532,  0.2613, -0.1273,
0.2640,  0.2688,  0.2672,
         0.2617, -0.2628,  0.2553,  0.2594,  0.2511,  0.2707,  0.2564, -0.2257,
         0.2488, -0.1802,  0.2582, -0.2405,  0.0656,  0.2322,  0.2713,  0.2740,
        -0.2575, -0.2574,  0.2696, -0.2558,  0.2685, -0.2683, -0.2602, -0.2723,
         0.2355,  0.2641, -0.2701,  0.1876, -0.2398,  0.2656,  0.2752,  0.2604,
        -0.2697,  0.2639,  0.2711,  0.2570, -0.2620, -0.2713, -0.2723, -0.2707,
        -0.2362,  0.2287,  0.2157, -0.2514, -0.2627, -0.2707,  0.2441],
       device='cuda:0', grad_fn=<IndexBackward0>)
non_zero_targets: tensor([ 0.0307, -0.3390, -1.1201,  0.3161, -0.3774,  0.1480,
0.7423,  3.4750,
         0.5417, -1.4902, -1.4360,  0.6070,  0.0330,  0.6785,  0.2050, -0.3587,
         0.0168, -0.2712,  0.1939, -0.6934, -0.2058,  0.4204,  1.1924,  1.1266,
        -0.9060, -4.6752,  3.2260, -5.8341,  0.7341, -2.4734, -0.8943, -2.2928,
        -1.1466,  0.8801, -1.5616, -0.1576, -0.5680,  0.4860,  1.0551,  0.1357,
        -2.5610,  0.9296,  0.7519,  0.6770, -1.2753, -1.5394, -2.0395, -2.0850,
        -0.7157,  0.0787, -0.0765, -0.8464, -1.4507, -2.4766,  0.8497],
       device='cuda:0')
Model and configurations saved.
```




Word Token Embedding Weights Over Time
Weight Value
Iterations
WTE Weight 0
WTE Weight 1
WTE Weight 2
WTE Weight 3
WTE Weight 4
0.03
0.02
0.01
0.00
−0.01
0
1
2
3
4
5


Training Loss Over Time
Training Loss
Loss
Iterations
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
1
2
3
4
5

[ ]:

[ ]: