Training

June 9, 2024

[3]:
```
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 = False  # Set to True to enable debugging

print('starting ....')

# Configuration
config = {
    'num_samples': 10000000,
    'input_dim': 10,
    'function_class': 'linear',
    'noise_std': 0,
    'model_type': 'standard',
    'batch_size': 64,
    'min_context_size': 1,
    'max_context_size': 40,
    'max_iters': 20000,
    '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 …
number of parameters: 9.55M
number of parameters: 9.55M
running on device cuda
running on device cuda
iter_dt 0.00ms; iter 0: train loss 8.75122; Scalar Head Weights Changed: True;
WTE Weights Changed: True
iter_dt 139.90ms; iter 100: train loss 0.51838; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 139.47ms; iter 200: train loss 0.69677; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 140.04ms; iter 300: train loss 0.25547; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 140.36ms; iter 400: train loss 0.24451; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.50ms; iter 500: train loss 0.28779; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 141.00ms; iter 600: train loss 0.41571; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 142.83ms; iter 700: train loss 0.49047; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 142.69ms; iter 800: train loss 0.23643; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.31ms; iter 900: train loss 0.26154; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.48ms; iter 1000: train loss 0.25476; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.34ms; iter 1100: train loss 0.24613; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 142.71ms; iter 1200: train loss 0.27886; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.37ms; iter 1300: train loss 0.31819; Scalar Head Weights Changed:
```

```
True; WTE Weights Changed: True
iter_dt 143.37ms; iter 1400: train loss 0.20167; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.30ms; iter 1500: train loss 0.23106; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 142.83ms; iter 1600: train loss 0.27816; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.48ms; iter 1700: train loss 0.16294; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.68ms; iter 1800: train loss 0.23595; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.85ms; iter 1900: train loss 0.22982; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.69ms; iter 2000: train loss 0.27854; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.87ms; iter 2100: train loss 0.27424; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.15ms; iter 2200: train loss 0.24976; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.80ms; iter 2300: train loss 0.24801; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.78ms; iter 2400: train loss 0.29324; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.21ms; iter 2500: train loss 0.18526; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.42ms; iter 2600: train loss 0.23814; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.98ms; iter 2700: train loss 0.26800; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.55ms; iter 2800: train loss 0.23148; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.19ms; iter 2900: train loss 0.21718; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.33ms; iter 3000: train loss 0.25639; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 143.92ms; iter 3100: train loss 0.29221; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.78ms; iter 3200: train loss 0.26791; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.06ms; iter 3300: train loss 0.33818; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.53ms; iter 3400: train loss 0.18278; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.84ms; iter 3500: train loss 0.26420; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.55ms; iter 3600: train loss 0.25753; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.54ms; iter 3700: train loss 0.36601; Scalar Head Weights Changed:
```

```
True; WTE Weights Changed: True
iter_dt 144.76ms; iter 3800: train loss 0.25839; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.46ms; iter 3900: train loss 0.25762; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.13ms; iter 4000: train loss 0.36375; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.53ms; iter 4100: train loss 0.16412; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.02ms; iter 4200: train loss 0.35662; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.79ms; iter 4300: train loss 0.24745; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.44ms; iter 4400: train loss 0.21593; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.95ms; iter 4500: train loss 0.26510; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.76ms; iter 4600: train loss 0.28419; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.49ms; iter 4700: train loss 0.26739; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.81ms; iter 4800: train loss 0.29740; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.45ms; iter 4900: train loss 0.18794; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.86ms; iter 5000: train loss 0.26791; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.31ms; iter 5100: train loss 0.26013; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.05ms; iter 5200: train loss 0.23921; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.84ms; iter 5300: train loss 0.30422; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.87ms; iter 5400: train loss 0.20679; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.73ms; iter 5500: train loss 0.18721; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.61ms; iter 5600: train loss 0.34455; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.30ms; iter 5700: train loss 0.29134; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.03ms; iter 5800: train loss 0.24686; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.68ms; iter 5900: train loss 0.23348; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.34ms; iter 6000: train loss 0.20089; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.69ms; iter 6100: train loss 0.29306; Scalar Head Weights Changed:
```

```
True; WTE Weights Changed: True
iter_dt 146.53ms; iter 6200: train loss 0.27909; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 149.31ms; iter 6300: train loss 0.32077; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.85ms; iter 6400: train loss 0.19783; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.94ms; iter 6500: train loss 0.24330; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.16ms; iter 6600: train loss 0.28777; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.45ms; iter 6700: train loss 0.29078; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.28ms; iter 6800: train loss 0.30941; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.26ms; iter 6900: train loss 0.32259; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.61ms; iter 7000: train loss 0.26304; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.45ms; iter 7100: train loss 0.25264; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.59ms; iter 7200: train loss 0.18717; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.64ms; iter 7300: train loss 0.20116; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 149.54ms; iter 7400: train loss 0.31620; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.43ms; iter 7500: train loss 0.24867; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.04ms; iter 7600: train loss 0.21200; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.25ms; iter 7700: train loss 0.17776; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.13ms; iter 7800: train loss 0.31013; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.24ms; iter 7900: train loss 0.24738; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.02ms; iter 8000: train loss 0.22992; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.84ms; iter 8100: train loss 0.27506; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.53ms; iter 8200: train loss 0.24645; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 148.86ms; iter 8300: train loss 0.20248; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.73ms; iter 8400: train loss 0.22164; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.05ms; iter 8500: train loss 0.30096; Scalar Head Weights Changed:
```

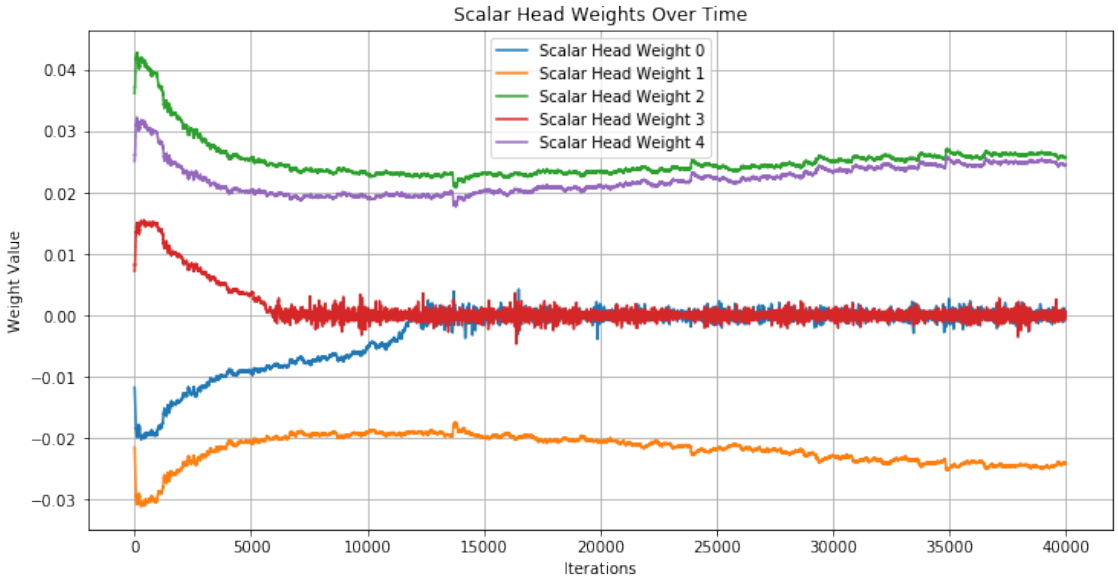
```
True; WTE Weights Changed: True
iter_dt 145.13ms; iter 8600: train loss 0.23493; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.72ms; iter 8700: train loss 0.17939; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.16ms; iter 8800: train loss 0.20200; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.54ms; iter 8900: train loss 0.25396; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.25ms; iter 9000: train loss 0.29652; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.66ms; iter 9100: train loss 0.30197; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.52ms; iter 9200: train loss 0.21102; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.01ms; iter 9300: train loss 0.32049; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.61ms; iter 9400: train loss 0.19587; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.85ms; iter 9500: train loss 0.28434; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.24ms; iter 9600: train loss 0.14894; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.25ms; iter 9700: train loss 0.27444; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.50ms; iter 9800: train loss 0.29637; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.42ms; iter 9900: train loss 0.17441; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.69ms; iter 10000: train loss 0.26540; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.38ms; iter 10100: train loss 0.22482; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.94ms; iter 10200: train loss 0.32613; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.02ms; iter 10300: train loss 0.17918; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.45ms; iter 10400: train loss 0.23811; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.66ms; iter 10500: train loss 0.22928; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.67ms; iter 10600: train loss 0.22943; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.13ms; iter 10700: train loss 0.21925; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.81ms; iter 10800: train loss 0.26445; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.77ms; iter 10900: train loss 0.22352; Scalar Head Weights Changed:
```

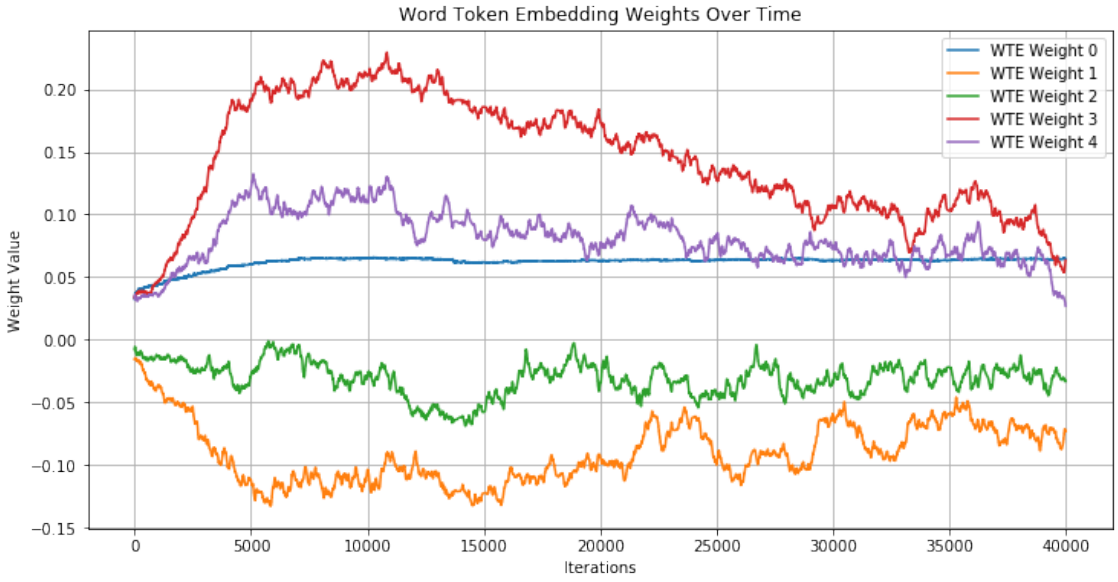
```
True; WTE Weights Changed: True
iter_dt 145.72ms; iter 11000: train loss 0.30188; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.09ms; iter 11100: train loss 0.31137; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.98ms; iter 11200: train loss 0.28759; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.95ms; iter 11300: train loss 0.27887; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 148.39ms; iter 11400: train loss 0.24962; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.82ms; iter 11500: train loss 0.24915; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.07ms; iter 11600: train loss 0.28209; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.80ms; iter 11700: train loss 0.27337; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.41ms; iter 11800: train loss 0.23158; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.33ms; iter 11900: train loss 0.21277; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.85ms; iter 12000: train loss 0.22222; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.16ms; iter 12100: train loss 0.22027; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.01ms; iter 12200: train loss 0.23486; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.01ms; iter 12300: train loss 0.30500; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.76ms; iter 12400: train loss 0.18727; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.54ms; iter 12500: train loss 0.24762; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.45ms; iter 12600: train loss 0.19285; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.96ms; iter 12700: train loss 0.27008; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.30ms; iter 12800: train loss 0.29738; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.19ms; iter 12900: train loss 0.17019; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.97ms; iter 13000: train loss 0.22995; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.59ms; iter 13100: train loss 0.19176; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.66ms; iter 13200: train loss 0.27051; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.94ms; iter 13300: train loss 0.15132; Scalar Head Weights Changed:
```

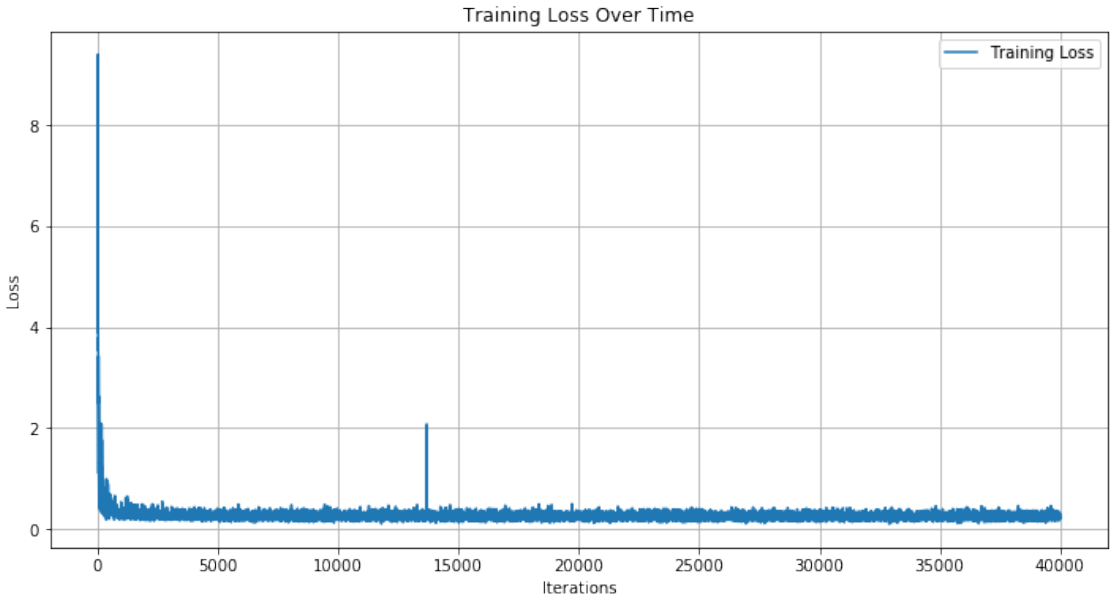
```
True; WTE Weights Changed: True
iter_dt 146.68ms; iter 13400: train loss 0.22904; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.87ms; iter 13500: train loss 0.24381; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.35ms; iter 13600: train loss 0.20844; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.76ms; iter 13700: train loss 0.23853; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.80ms; iter 13800: train loss 0.24535; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.15ms; iter 13900: train loss 0.21761; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.60ms; iter 14000: train loss 0.31673; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.40ms; iter 14100: train loss 0.21609; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.03ms; iter 14200: train loss 0.29542; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.11ms; iter 14300: train loss 0.22161; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.22ms; iter 14400: train loss 0.24779; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.68ms; iter 14500: train loss 0.21577; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.94ms; iter 14600: train loss 0.25874; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.94ms; iter 14700: train loss 0.29948; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.97ms; iter 14800: train loss 0.18149; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.84ms; iter 14900: train loss 0.20164; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.38ms; iter 15000: train loss 0.25049; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.54ms; iter 15100: train loss 0.24501; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.24ms; iter 15200: train loss 0.23974; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.60ms; iter 15300: train loss 0.29527; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.73ms; iter 15400: train loss 0.19430; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.45ms; iter 15500: train loss 0.29322; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.31ms; iter 15600: train loss 0.30744; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.20ms; iter 15700: train loss 0.14778; Scalar Head Weights Changed:
```

```
True; WTE Weights Changed: True
iter_dt 146.17ms; iter 15800: train loss 0.25453; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.00ms; iter 15900: train loss 0.19154; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.61ms; iter 16000: train loss 0.26600; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.15ms; iter 16100: train loss 0.25380; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.99ms; iter 16200: train loss 0.26352; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.62ms; iter 16300: train loss 0.22602; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.16ms; iter 16400: train loss 0.34143; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.98ms; iter 16500: train loss 0.30481; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.97ms; iter 16600: train loss 0.26380; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.12ms; iter 16700: train loss 0.27318; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.91ms; iter 16800: train loss 0.26120; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.30ms; iter 16900: train loss 0.19473; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.07ms; iter 17000: train loss 0.16520; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.13ms; iter 17100: train loss 0.22897; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.63ms; iter 17200: train loss 0.22575; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.21ms; iter 17300: train loss 0.27649; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.63ms; iter 17400: train loss 0.25912; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.42ms; iter 17500: train loss 0.29166; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.18ms; iter 17600: train loss 0.29829; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.48ms; iter 17700: train loss 0.35260; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.91ms; iter 17800: train loss 0.22697; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 147.25ms; iter 17900: train loss 0.23180; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.90ms; iter 18000: train loss 0.24580; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.41ms; iter 18100: train loss 0.25081; Scalar Head Weights Changed:
```

```
True; WTE Weights Changed: True
iter_dt 146.63ms; iter 18200: train loss 0.24469; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 148.09ms; iter 18300: train loss 0.25784; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.96ms; iter 18400: train loss 0.29341; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.69ms; iter 18500: train loss 0.25522; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.66ms; iter 18600: train loss 0.29307; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.92ms; iter 18700: train loss 0.17734; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.35ms; iter 18800: train loss 0.23508; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.42ms; iter 18900: train loss 0.27623; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.42ms; iter 19000: train loss 0.29385; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.08ms; iter 19100: train loss 0.28543; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.70ms; iter 19200: train loss 0.27464; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.53ms; iter 19300: train loss 0.23107; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.65ms; iter 19400: train loss 0.20498; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 144.93ms; iter 19500: train loss 0.26461; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.33ms; iter 19600: train loss 0.20086; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 145.32ms; iter 19700: train loss 0.24804; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.37ms; iter 19800: train loss 0.27853; Scalar Head Weights Changed:
True; WTE Weights Changed: True
iter_dt 146.53ms; iter 19900: train loss 0.29121; Scalar Head Weights Changed:
True; WTE Weights Changed: True
Model and configurations saved.
```


Scalar Head Weights Over Time
Scalar Head Weight 0
Scalar Head Weight 1
Scalar Head Weight 2
Scalar Head Weight 3
Scalar Head Weight 4
Weight Value
Iterations
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
5000
10000
15000
20000
25000
30000
35000
40000


Word Token Embedding Weights Over Time
WTE Weight 0
WTE Weight 1
WTE Weight 2
WTE Weight 3
WTE Weight 4
Weight Value
Iterations
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
5000
10000
15000
20000
25000
30000
35000
40000


Training Loss Over Time
Training Loss
Loss
Iterations
0
2
4
6
8
0
5000
10000
15000
20000
25000
30000
35000
40000

[ ]: