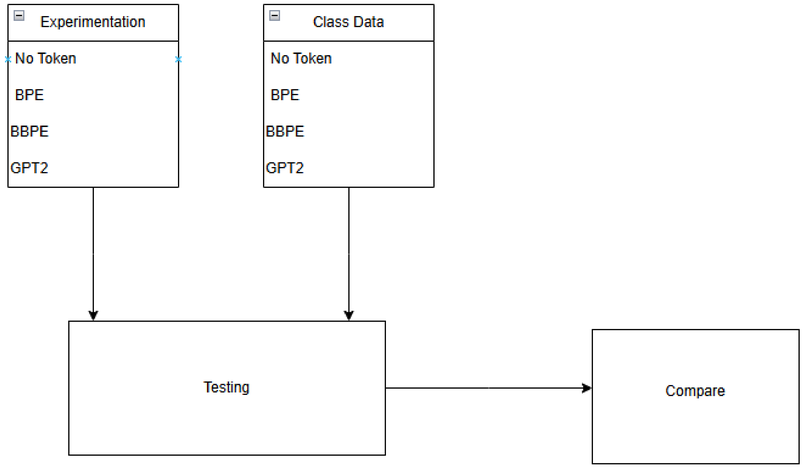
AI First Course Assistance LLM

Romil Patel
Georgia Institute of Technology
rpatel778@gatech.edu

Motivation



Objective: Experiment with different hyperparameters and see how it effects a model, then train using class data.

Training Models

```
=== MODEL: ./models/no token/model_learning_rate_.0001.pth ===

Step 0: train loss 4.6677, val loss 4.6675
Step 50: train loss 3.3149, val loss 3.3527
Step 100: train loss 3.3100, val loss 3.3469
Step 150: train loss 3.2997, val loss 3.3358
Step 200: train loss 3.2536, val loss 3.2861
Step 250: train loss 3.1440, val loss 3.1459
Step 300: train loss 3.0864, val loss 3.0823
Step 350: train loss 2.9268, val loss 2.9204
Step 400: train loss 2.7829, val loss 2.7775
Step 450: train loss 2.7386, val loss 2.7349
Step 500: train loss 2.6784, val loss 2.6684
Step 550: train loss 2.6476, val loss 2.6358
Step 600: train loss 2.6012, val loss 2.5871
Step 650: train loss 2.5679, val loss 2.5562
Step 700: train loss 2.5411, val loss 2.5291
Step 750: train loss 2.5054, val loss 2.4938
Step 800: train loss 2.4729, val loss 2.4665
Step 850: train loss 2.4511, val loss 2.4459
Step 900: train loss 2.4522, val loss 2.4439
Step 950: train loss 2.4126, val loss 2.4054
```

Figure 1. Example Train and val loss of learning rate model

For the first part:

- Experiment with different hyperparameters such as learning rate, batch size, tokenizers, etc. (In total 8-9 parameters changed 4 times over 20 iterations)
- Using val loss and train loss, determine if models were overfitting or had bias and prep for part2.

Testing Outcomes

```
What is regularization? b of heardome,
There my cien ging unisheavother: lefares;
My the sontell this sat mer that tchany in litan
To der with, and ther
```

```
Wnoch, ate thy to shayou Wl
Bum t feck wond rshearehard.

That Mar'EFO:
Shay d. notlld pf For ld.

For my wit ho
```

```
Wtona bpd
Ling cer tof sawo.

Gas Went hen:
What ding-sse ye l her daun-
Mund suthacoully as lo hyscav, Fostnds gtristur
T
```

Some Responses from Tokenizer Models w/o Class Data

```
What is overfitting in machine learning and how can it be avoided?
3. Dasuly
2-t) the parameted frecifint to the och se the lab to nperesent the mearted imate for diffferent functed works rerations mace. This Man
compute labooution whangen notlot calesing dathe alse import orive.
• Selar of ithow s: Anot its the Jelariouti0ns the poace differed a to daf the
perffing the bhe conveces Po 4 the simpovation.
(c) Bust = ue hnelfe.

(n d -f) aling dataseticed, the it a we canch Heassity as dine F and probiath of the the Dargo ystemactin is mosily that se necll off perocted
the oGpherabit
```

```
What is padding in CNNs?
a. Cppporation Caner
bartived
9. Pe feal
4. Doser in the laxerl simputed wein nof model the an of the meas mus. The have sthan und reprobabile nodel the a deacy vallue, that the
virsely or roculaters the
```

Responses from Tokenizer Models w/ Class Data

Challenges of Experimentation

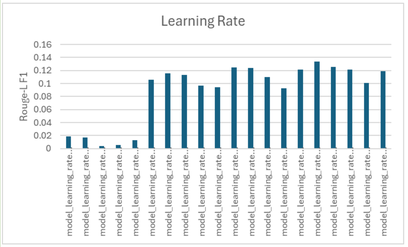
```
Failed to load model ./models/class_data/model_bbpe.pth due to weight size mismatch.
Error(s) in loading state_dict for GPTLanguageModel:
        size mismatch for token_embedding_table.weight: copying a param with shape torch.Size([400, 384]) from checkpoint, th
e shape in current model is torch.Size([8000, 384]).
        size mismatch for lm_head.weight: copying a param with shape torch.Size([400, 384]) from checkpoint, the shape in cur
rent model is torch.Size([8000, 384]).
        size mismatch for lm_head.bias: copying a param with shape torch.Size([400]) from checkpoint, the shape in current mo
del is torch.Size([8000]).
```

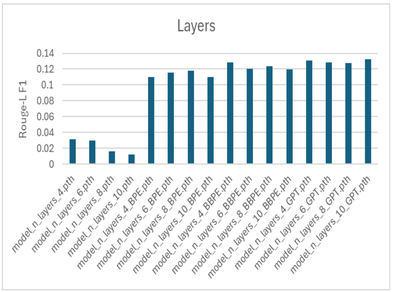
```
Failed to load model ./models/class_data/model_bbpe.pth due to weight size mismatch.
Error(s) in loading state_dict for GPTLanguageModel:
        size mismatch for token_embedding_table.weight: copying a param with shape torch.Size([400, 384]) from checkpoint, th
e shape in current model is torch.Size([8000, 384]).
        size mismatch for lm_head.weight: copying a param with shape torch.Size([400, 384]) from checkpoint, the shape in cur
rent model is torch.Size([8000, 384]).
        size mismatch for lm_head.bias: copying a param with shape torch.Size([400]) from checkpoint, the shape in current mo
del is torch.Size([8000]).
```

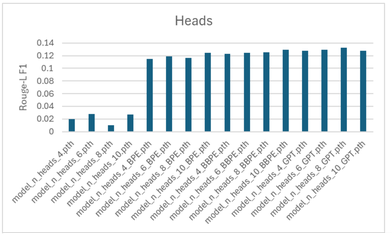
a) Hardware Limitations
b) Model Hallucinations
c) Hyperparameter Training vs Testing errors
d) Tokenizer Decoding errors

Experiments Enabled by Dataset

Results of Models Trained By Hyperparameter Testing

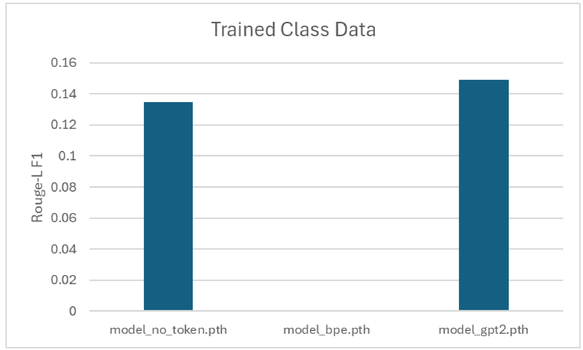






Interpretation: Models Trained on Tokenizers showed greater results rather than models trained without, more layers = better results

Results on Model Trained from Class Data



Interpretation: Tokenizers showed better performance, however showed difficulty with tokenizers such as BPE and BBPE (hallucinations/decoding errors).