Week 4 - Transformers

Adding the Encoder

Week 4 Objectives

- Utilizing one of the datasets mentioned in the references
- Clean file structure into Data, Model, and Train sections
- Adding the Encoder to last week’s Transformer (Decoder only)
- Scaling your model in size deliberately and tracking metrics
- Train on GPU
- Monitoring progress using Weights & Biases

Text-to-SQL

Datasets:

- Textbooks are all you need
  - Prompt
  - Many fields
  - response
- Text_to_sql
  - Question
  - Answer
  - Context

How to combine? → (Prompt, Response) + (Question, Answer)

What to do with text_to_sql context?

Outstanding Questions

- HuggingFace uses Apache Arrow under the hood
  - Better than DataFrames? Need to look into
- End SentencePiece input = 26 mil rows → How to use SP at scale?
- Use of tiny_stories for text_to_sql? Use of Textbooks non-SQL results?
  - How to get an intuition for transfer learning?

Clean file structure

All on github

- tiny_piece.model
- multi_head_with_pos_encod_weights_0_100000.pt
- bash_gpt
- bash_gpt_evaluation
- bash_gpt_inference
- bleu_metrics
- dataset_bes
- parameters_per_layer
- server
- tokenizer_bes
- train_bash_gpt


```python
class attention(torch.nn.Module):
    def __init__(self, is_causal = False):
        super().__init__()
        #MHA
        self.num_heads = num_heads
        self.head_dim = input_embedding_dimensions // self
        self.is_causal = is_causal

        self.register_buffer('mask', torch.tril(torch.ones
        self.mha_final_linear_layer = nn.Linear(input_embe

    def forward(self,positional_embeddings, q,k,v):
```

```python
class positionalEmbeddings(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.input_embedding_dimensions = input_embedding_dimensions
        self.input_embeddings = nn.Embedding(vocab_size, input_embedding_dimensions)
        self.positional_encoding = nn.Embedding(max_sequence_length, input_embedding_dimensions)

    def forward(self, inputs):
        input_embeddings = self.input_embeddings(inputs)
        positional_embeddings = self.positional_encoding(torch.arange(inputs.size(1)).to(device))
```

```python
class feedForward(nn.Module):
    def __init__(self):
        super().__init__()
        self.ff_first_linear = nn.Linear(input_embedding_dimensions,ff_dimension)
        self.ff_relu = nn.ReLU()
        self.ff_second_linear = nn.Linear(ff_dimension,input_embedding_dimensions)
        self.drop   = torch.nn.Dropout(drop_rate)

    def forward(self, mha_norm_output):
        ff_first_linear_output = self.ff_first_linear(mha_norm_output)
        ff_relu_output = self.ff_relu(ff_first_linear_output)
```

Scaling in Size

Total Trainable Params: **78,468,160**

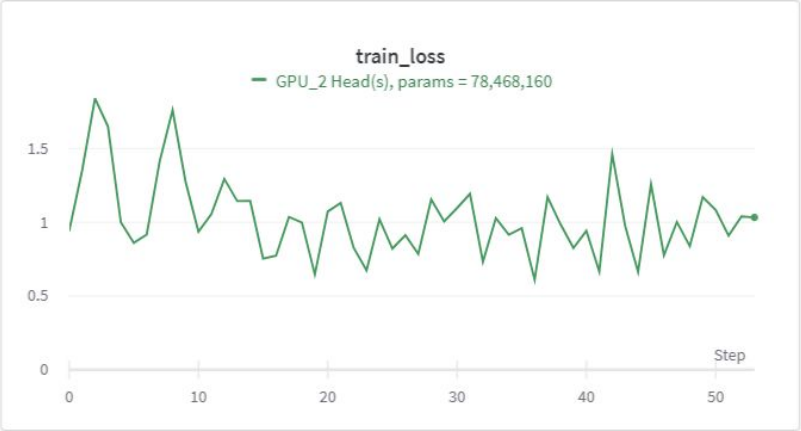
20 minutes per epoch on MLX GPU

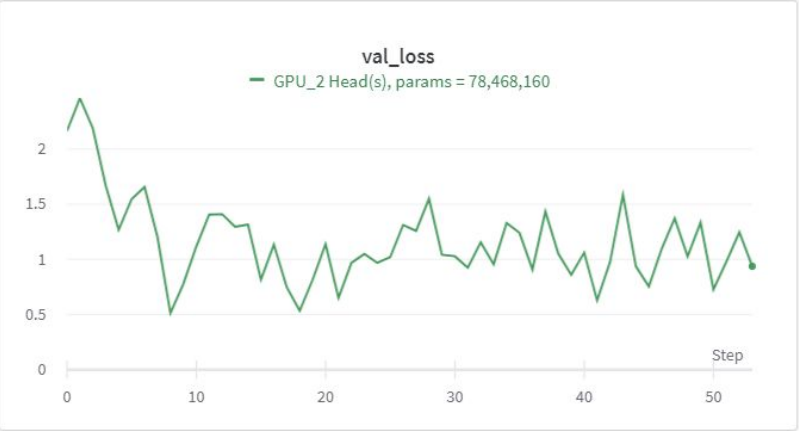
Laptop GPU 10x faster than CPU

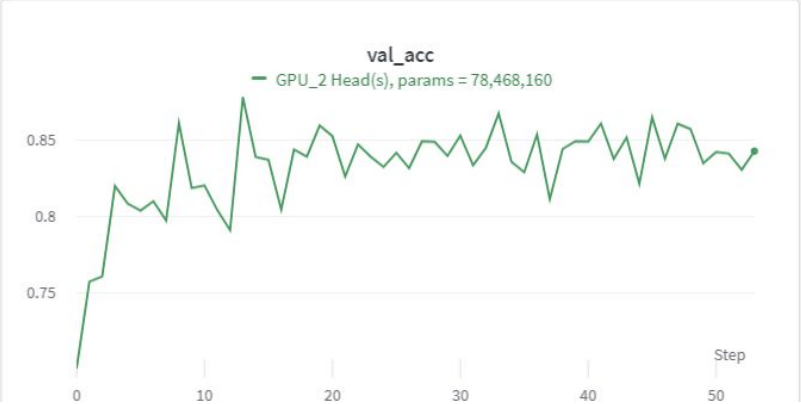
MLX GPU 4x faster than Laptop GPU

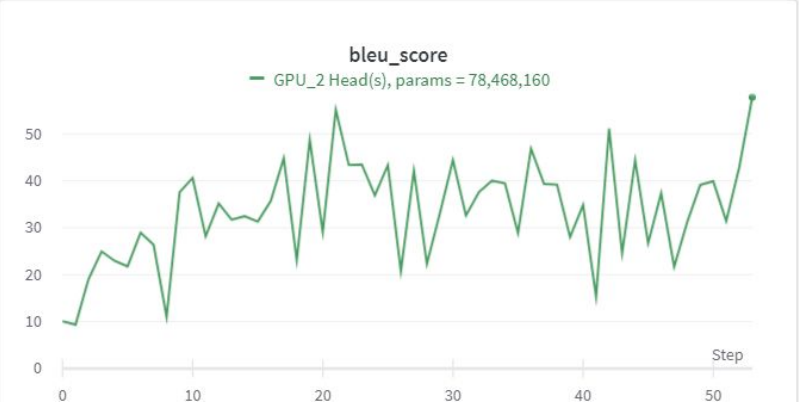
Modules	Parameters
enc_positional_embeddings.input_embeddings.weight	20480000
enc_positional_embeddings.positional_encoding.weight	611328
dec_positional_embeddings.input_embeddings.weight	20480000
dec_positional_embeddings.positional_encoding.weight	611328
encoder.0.attention.mha_final_linear_layer.weight	262144
encoder.0.attention.mha_final_linear_layer.bias	512
encoder.0.mha_norm_layer.weight	512
encoder.0.mha_norm_layer.bias	512
encoder.0.feedForward.ff_first_linear.weight	1048576
encoder.0.feedForward.ff_first_linear.bias	2048
encoder.0.feedForward.ff_second_linear.weight	1048576
encoder.0.feedForward.ff_second_linear.bias	512
encoder.0.ff_norm_layer.weight	512
encoder.0.ff_norm_layer.bias	512
encoder.1.attention.mha_final_linear_layer.weight	262144
encoder.1.attention.mha_final_linear_layer.bias	512
encoder.1.mha_norm_layer.weight	512
encoder.1.mha_norm_layer.bias	512
encoder.1.feedForward.ff_first_linear.weight	1048576
encoder.1.feedForward.ff_first_linear.bias	2048
encoder.1.feedForward.ff_second_linear.weight	1048576
encoder.1.feedForward.ff_second_linear.bias	512
encoder.1.ff_norm_layer.weight	512
encoder.1.ff_norm_layer.bias	512
encoder.2.attention.mha_final_linear_layer.weight	262144
encoder.2.attention.mha_final_linear_layer.bias	512
encoder.2.mha_norm_layer.weight	512
encoder.2.mha_norm_layer.bias	512
encoder.2.feedForward.ff_first_linear.weight	1048576
encoder.2.feedForward.ff_first_linear.bias	2048
encoder.2.feedForward.ff_second_linear.weight	1048576
encoder.2.feedForward.ff_second_linear.bias	512
encoder.2.ff_norm_layer.weight	512
encoder.2.ff_norm_layer.bias	512
decoder.0.first_attention.mha_final_linear_layer.weight	262144
decoder.0.first_attention.mha_final_linear_layer.bias	512
decoder.0.mha_first_norm_layer.weight	512
decoder.0.mha_first_norm_layer.bias	512
decoder.0.second_attention.mha_final_linear_layer.weight	262144
decoder.0.second_attention.mha_final_linear_layer.bias	512
decoder.0.mha_second_norm_layer.weight	512
decoder.0.mha_second_norm_layer.bias	512
decoder.0.feedForward.ff_first_linear.weight	1048576
decoder.0.feedForward.ff_first_linear.bias	2048
decoder.0.feedForward.ff_second_linear.weight	1048576

Weights and Biases









BLEU

- < 10: Almost useless
- 10–19: Hard to get the gist
- 20–29 :The gist is clear, but has significant grammatical errors
- 30–40: Understandable to good translations
- 40–50: High quality translations
- 50–60: Very high quality, adequate, and fluent translations
- > 60: Quality often better than human

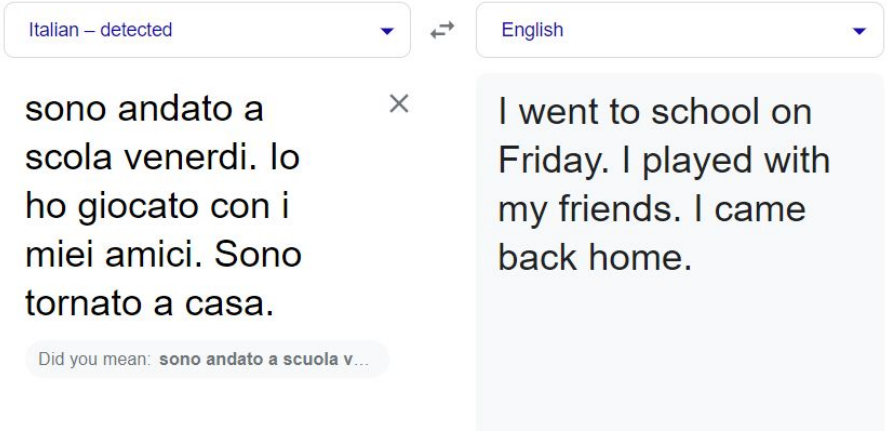
Results

```
training time since start: 0 hours 0 minutes
train_loss: 1.1184, val_loss: 1.1360, val_acc: 0.8392, bleu_score: 36.871708
*****************************************************************************************************

User prompt -> I went to school on Friday. I played with my friends. I came back home in the afternoon.

Ba$H_GPT Translator -> Sono andato a scuola venerdì. Io ho giocato con i miei amici. Sono tornato a casa.

*****************************************************************************************************
```



Results

```
train_loss: 1.1184, val_loss: 1.1360, val_acc: 0.8392, bleu_score: 36.871708
*****************************************************************************************

User prompt -> I went to school on Friday. I played with my friends. I came back home in the afternoon.

Ba$H_GPT Translator -> Sono andato a scuola venerdì. Io ho giocato con i miei amici. Sono tornato a casa.

*****************************************************************************************
```

```
train_loss: 1.0328, val_loss: 0.9377, val_acc: 0.8428, bleu_score: 57.830984
*****************************************************************************************

User prompt -> I went to school on Friday. I played with my friends. I came back home in the afternoon.

Ba$H_GPT Translator -> Sono andata a scuola il venerdì, ho rubato con i miei amici. Sono venuta a casa. Sono tornato a scuola.

*****************************************************************************************
```

Italian – detected

Sono andata a scuola il venerdì, ho rubato con i miei amici. Sono venuta a casa. Sono tornato a scuola.

English

I went to school on Friday, I stole with my friends. I came home. I went back to school.

Lessons Learnt

- I was underestimating the importance of a clean file structure and github.
- Can LLM evaluation be done without human evaluation?