Weekly Wrap-up

Progress Highlights and Insights

Contents

Last Week Recap

Question 1: How are Concepts included in the training?

Question 2: How is the loss function calculated?

The Loss Function

- Regularizartion where the explanation loss is the penalty that reduces overfitting.
- "Guiding Attention for Self-Supervised Learning with Transformers" (Deshpande et al., 2020)
  - $L = L_{cls} + \lambda L_{expl}$
  - $L_{expl} = ||A - H||^2_F$

Practical part

- Intro to Pytorch course
- Hugging Face and pretrained models

Intro to Pytorch course

- Finished the Intro to Pytorch course
  - Learned how to train neural networks , RNNs and LSTMs

How to use pretrained models and tokenizers

- Implemented a sentiment analysis model from scratch including
  - Scraping the data, cleaning and preprocessing
  - Using Bert base model and a feed forward network as a classifier.

```
output = model(input_ids, attention_mask)
_, prediction = torch.max(output, dim=1)

print(f'Review text: {review_text}')
print(f'Sentiment  : {class_names[prediction]}')
```

```
Review text: I love completing my todos! Best app ever!!!
Sentiment  : positive
```

Concept Transformer Code

- Progress Understanding the code and how concepts work within the model training.
- Building the architecture
  - Pre-trained layers
  - Extracted Modules
  - Current Progress

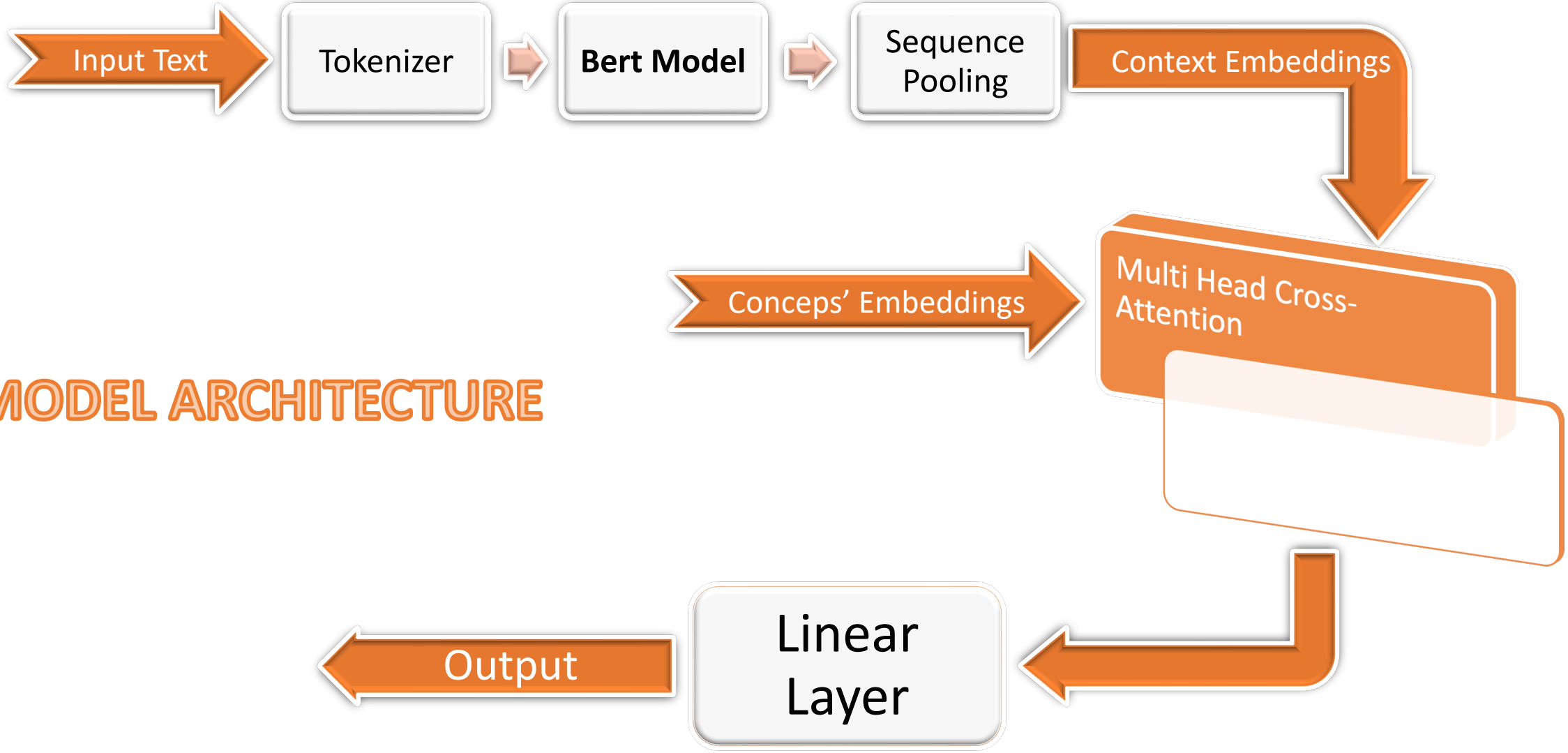
How concepts work within the model training

- Concepts are represented as a tensor of learnable parameters initialized to zeros.
- Get updated during training to result in final embeddings.

Building the model architecture

- The model is built to pass inputs through
  - Tokenizer
  - Pre-trained model
  - Sequence Pooling
  - Cross-attention with concepts
  - Linear Layer
- Already built model but not trained yet https://github.com/OMAR-AHMED-SAAD/Concept-Transformer

MODEL ARCHITECTURE



Current challenges

- Finding a dataset to start training

Target Dataset

Text	Enterpreneur	Traits
Tweets	1	Extroversion, agreeableness
Posts	0	Openness, neuroticism

Available Datasets

Text	Traits
Tweets	Big 5 or MBTI

Potential solutions

- Can we train a model to add traits to collected data with the entrepreneurs labelling
- Explore other Concepts that may have available datasets
- Suggestions?

References

- ATTENTION-BASED INTERPRETABILITY WITH CONCEPT TRANSFORMERS
- Guiding Attention for Self-Supervised Learning with Transformers

Thank You