BoolQ with Transformer

Yelin Zhang

Preprocessing

- Lower case
- Special character removal
- Non ascii word removal
- No truncation
- Concatenation of question and passage
  - Special token separator <sep>
- Padding only to max length of every batch

Input/Output

- Input: 64 x 622
- Input Transformer: 64 x 622 x 512
- Hidden: 64 x 512 x 512
- Output: 64 x 1

- Question + Passage max length: 622
- Embedding dimension: 512
- Batch size: 64

Model

- Embedding layer (622, 512)
- Positional encoding layer
- Transformer layers
- Linear layer (512, 512)
- ReLu
- Linear layer (512, 1)
- Sigmoid

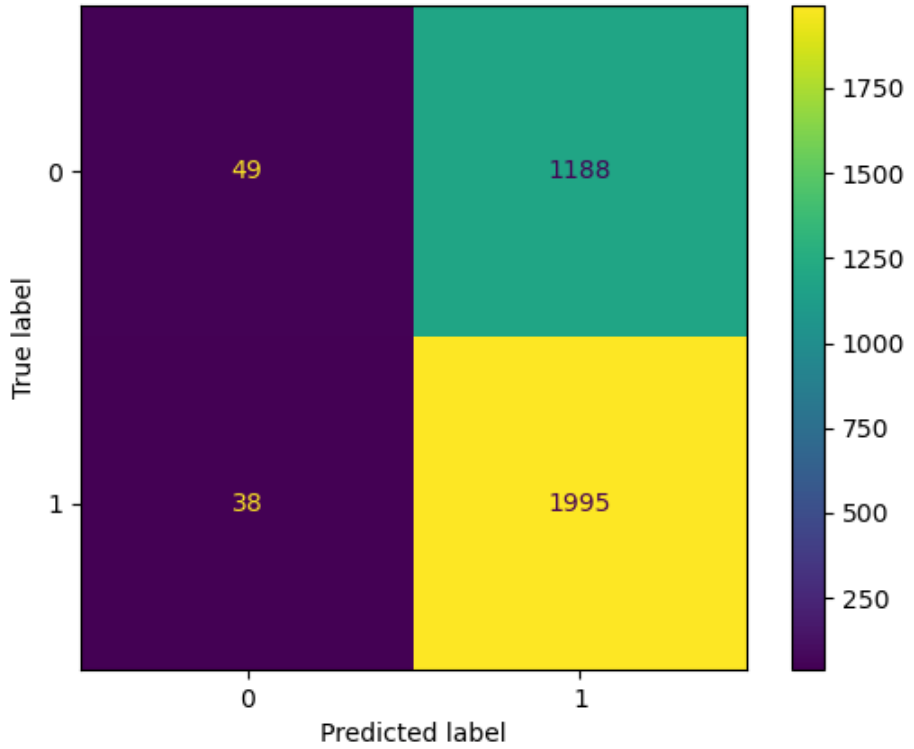
Experiments

- Positional encoding type (Sinusoidal, Embedding)
- Attention heads (8, 16)
- Learning rate (0.001, 0.0001)

= 2*2*2 = 8 Experiments

Result

	Train	Valid	Test
Loss	0.6517	0.6745	0.6640
Accuracy	0.5581	0.5951	0.6287
Recall	-	-	0
Specifcity	-	-	0



Final Parameters: Positional Embedding, Attention heads: 8, Learning rate 1-e4

Comparison test metrics

	Word embeddings	RNN	Transformer
Accuracy	0.6165	0.6397	0.6287

Interpretation

- Expectation 70% as an improvement over RNN
- More parameters need more training data
- More epochs needed
  - Model needs to learn Words Embeddings and Positional Embeddings
  - Training and Validation loss was flat
- RNN was selected by accuracy not loss
  - probably selected an overfit model

Learnings

- Embeddings take quite a few epochs to learn
- Lot of customization options for transformers
- Learning rate is still the dominant hyperparameter