BoolQ with Pretrained Transformers

Yelin Zhang

Preprocessing

- Concat question and passage with [SEP]
- DeBERTA-v2 Tokenizer
  - SentencePiece subword tokenizer
  - SentencePiece embeddings
  - Compatible with DeBERTA-v3
- No other processing done, because SentencePiece expects raw text
- Padded by max length in every batch with [PAD]

Network architecture

Pretrained Transfomer
- DeBERTA-v3 layers
  - Max input 512, Out 768

- ContextPooler
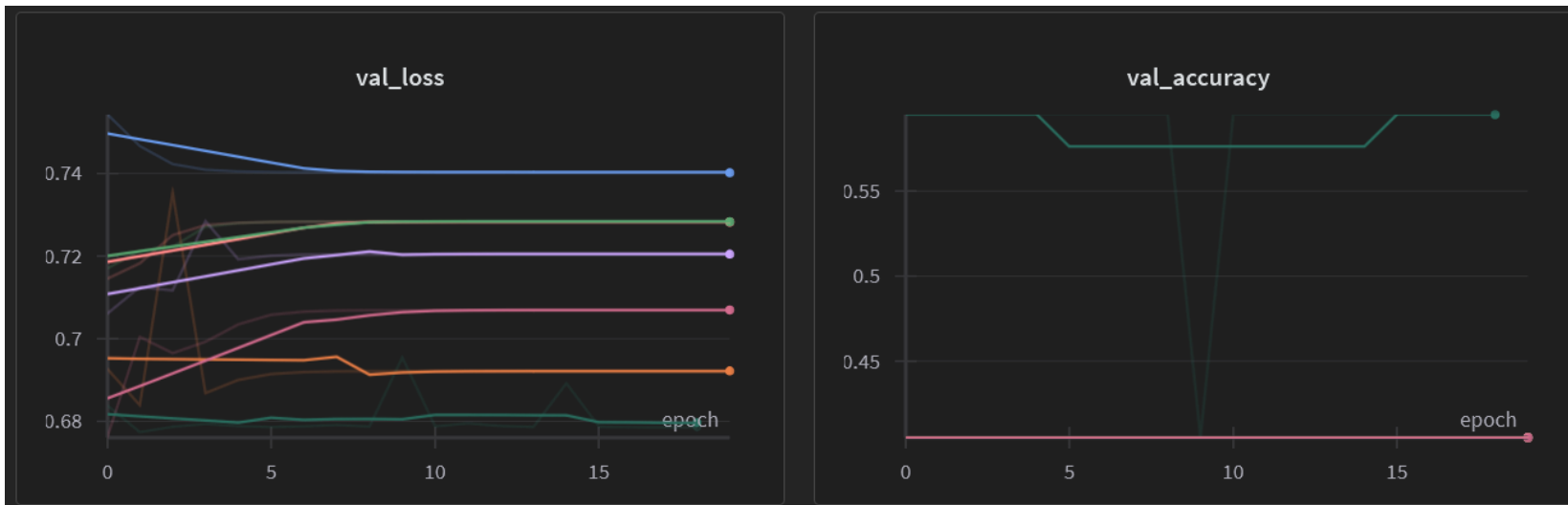 - takes first hidden state [CLS]
  - In 768, Out 768

Classification Head
- Linear
  - In 768, Out 256
  - ReLu

- Linear
  - In 256, Out 1

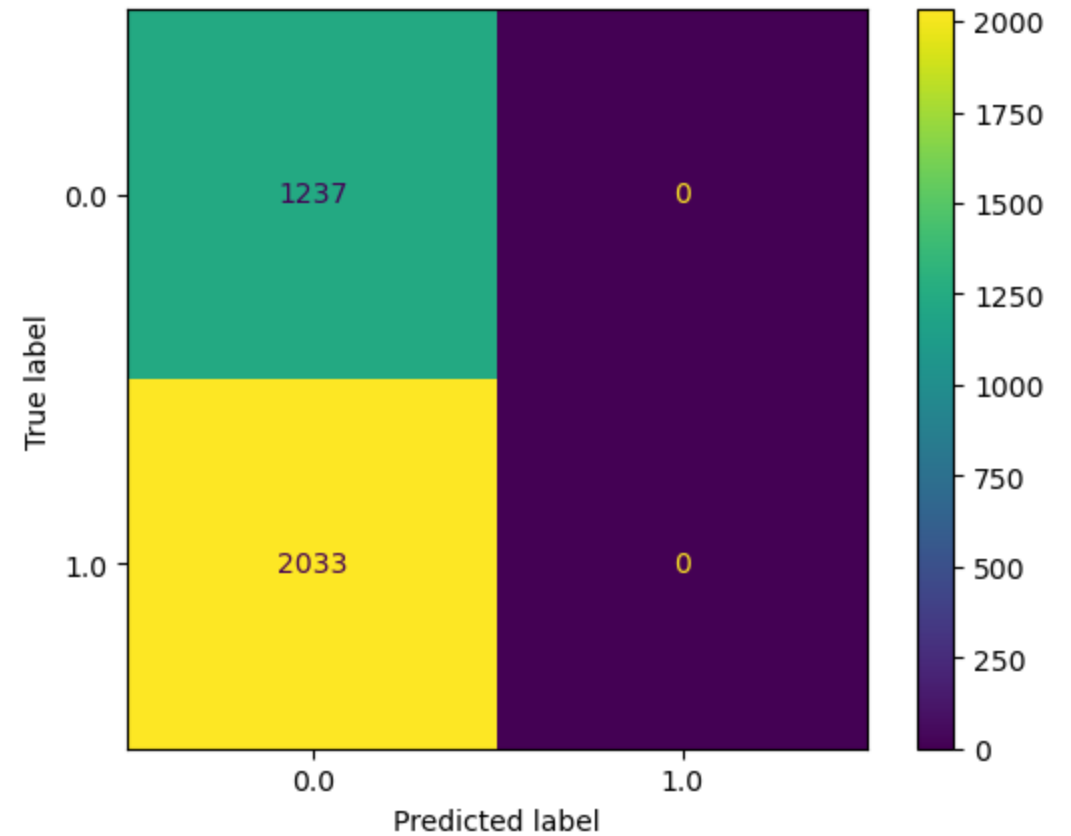Optimizer: AdamW

Loss: BCE with Logits

Experiments

- Learning rate in continuous range 1 – 1e-10
  - potentially unlimited runs
- Completed 7 runs
  - due to long training times

Results

	Train	Valid	Test
Loss	0.6275	0.6839	0.6809
Accuracy	0.25	0.5950	0.6218

Best parameter: Learning rate 0.12

Comparison test metrics

	Majority class	Word embeddings	RNN	Transformer	Pre trained Transformer	
Accuracy	0.6218	0.6165	0.6397	0.6287	0.6218	

Interpretation

- LR over ~0.3: pure minority class classifiers

- LR under ~0.3: pure majority class classifiers

- Reasonable LR (probably at most ~0.01): mix of classifications

- Future: Sweep more restrictive range of learning rates