BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch?

Joel Niklaus[†], **Daniele Giofre**[†]

[†]*Thomson Reuters Labs*, Zug, Switzerland, {joel.niklaus}/{daniele.giofre}@thomsonreuters.com

Key Contributions

Train Longformer base on legal data using ~ 24 V100 GPU days

Performs well on summarization datasets BillSum and PubMed

Why Should One Care?

Pretraining is very expensive

No Legal Longformer available yet

Method

Apply Replaced Token Detection (RTD) Task to Longformer

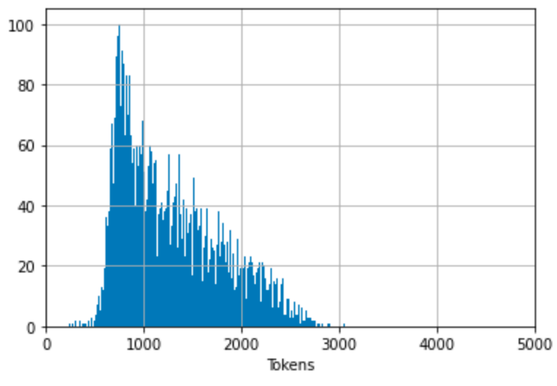
Couple with randomly initialized decoder for summarization

Pretraining Data

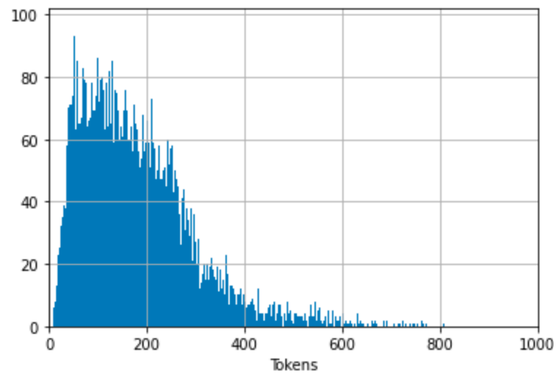
PileOfLaw Subset	Dataset Size	# Words	# Documents
caselaw			
CL Opinions	**59.29GB**	**7.65B**	**3.39M**
diverse			
Total	**73.04GB**	**8.91B**	**2.1M**
CL Opinions	8.74GB	1.13B	500K
CL Docket Entries and Court Filings	17.49GB	1.80B	500K
U.S. State Codes	6.77GB	829.62M	157
U.S. Code	0.27GB	30.54M	43
EUR-Lex	1.31GB	191.65M	106K
Edgar Contracts	7.26GB	0.97B	500K
Atticus Contracts	31.2GB	3.96B	488K

Table 1: The datasets used for pretraining our models. Abbreviations: CL: Court Listener

Evaluation Data

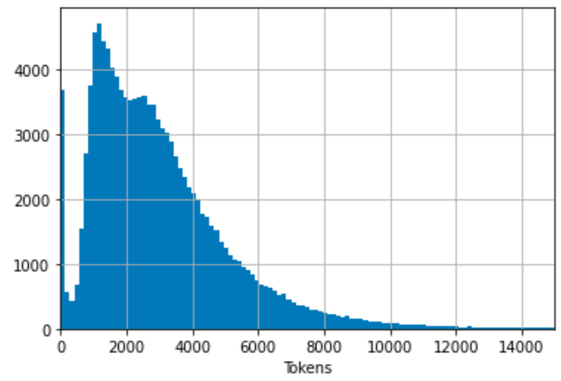


(a) Input Text
Mean: 1289, Median: 1166
75-Quant: 1644, 95-Quant: 2290, Max: 3055

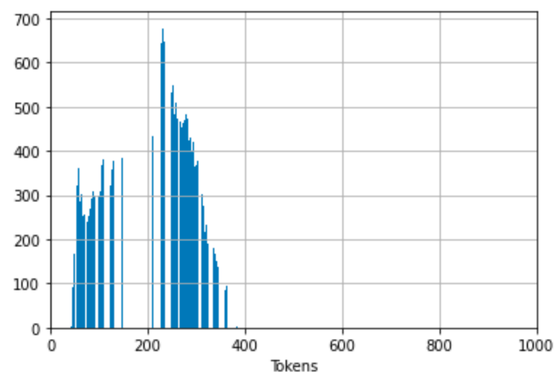


(b) Summary
Mean: 179, Median: 157
75-Quant: 240, 95-Quant: 398, Max: 808

Figure 6: Histograms for the BillSum training set (18949 samples).



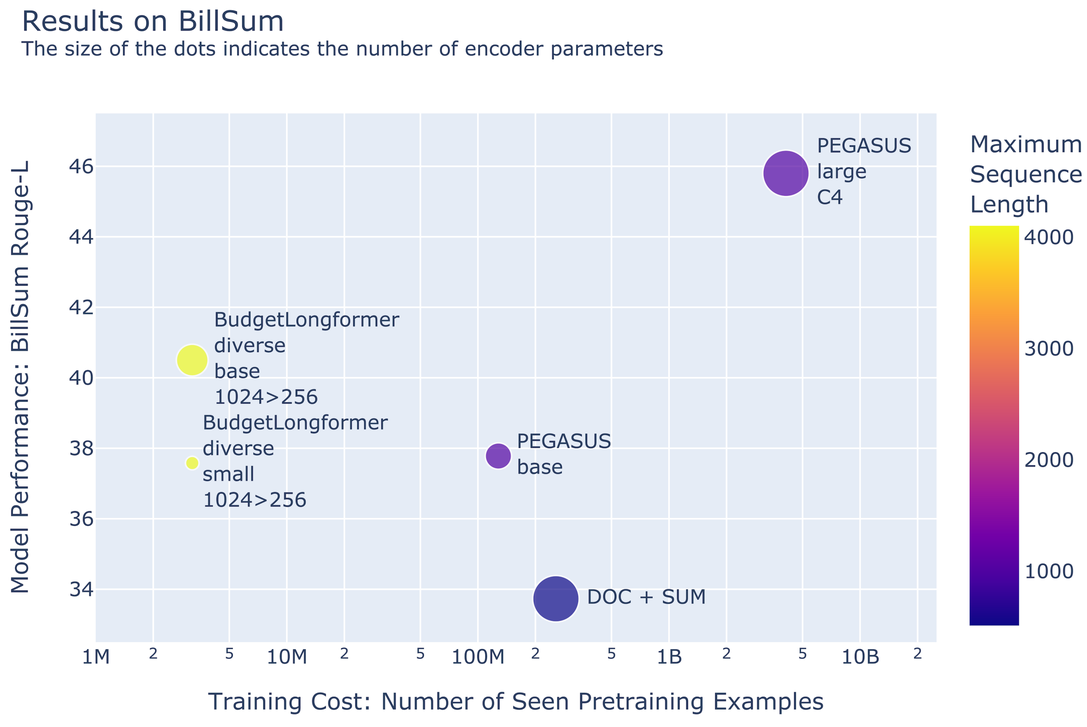
(a) Input Text
Mean: 3044, Median: 2572
75-Quant: 3996, 95-Quant: 7057, Max: 109759

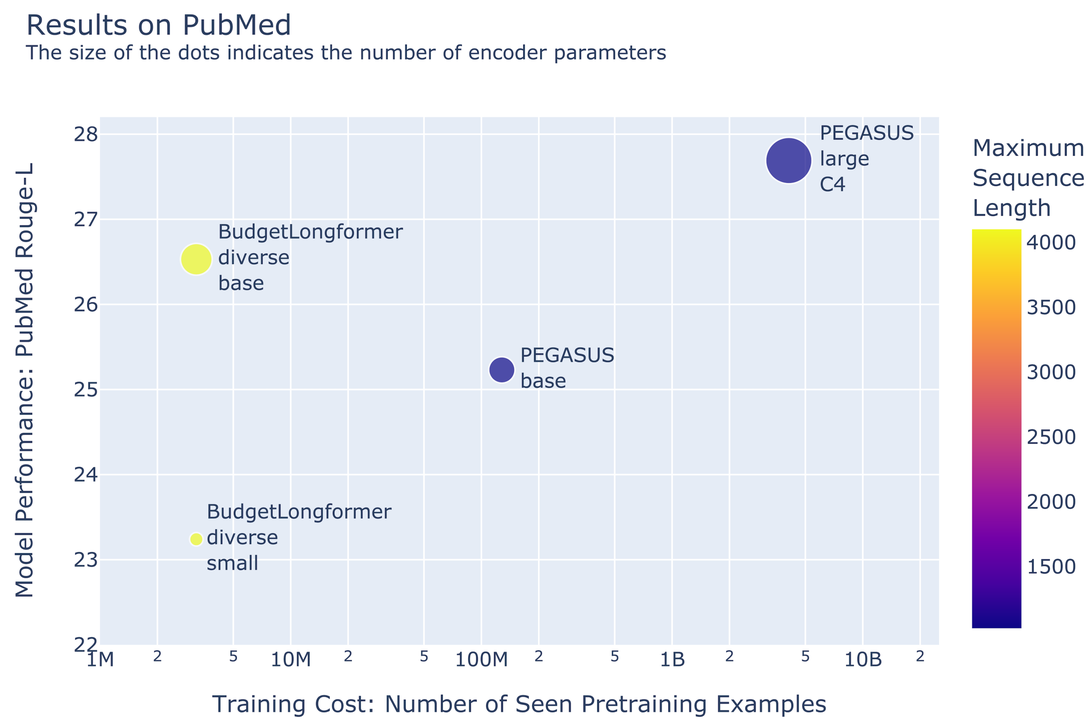


(b) Summary
Mean: 202, Median: 208
75-Quant: 262, 95-Quant: 326, Max: 391

Figure 8: Histograms for the PubMed train set (119924 samples).

Results





Conclusions

Pretraining Longformer with RTD works

It works well on in-domain and out-of-domain summarization (BillSum and PubMed)

Pretraining approach well suited if no large teacher model available

Future Work

More downstream evaluations (e.g. CUAD, MultiLexSum, BigPatent)

More pretraining (larger batch size and more steps)

Warmstart from available checkpoints