Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim,
Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, Samuel R. Bowman

We compare 20 tasks and task combinations for pretraining NLP models. Shockingly, language modeling is the most effective.

Motivation

- Language modeling pretraining is very effective, but many other tasks have been previously proposed to pretrain NLP models
- Research questions:
  - In a controlled setting, which task is best for pretraining models?
  - Which tasks can be productively *combined* with pretrained language models?

Task	|Train|	Task Type
WNLI	634	coreference resolution (NLI)
RTE	2.5k	NLI
MRPC	3.7k	paraphrase detection
STS	7.0k	sentence similarity
CoLA	8.5k	acceptability
SST	67k	sentiment
QNLI	105k	QA (NLI)
DisSent (WikiText-103)	311k	discourse marker prediction
QQP	364k	paraphrase detection
MNLI	393k	NLI
MT En-Ru	3.2M	translation
MT En-De	3.4M	translation
SkipThought (WikiText-103)	4M	next sentence prediction
LM (WikiText-103)	4M	language modeling
Reddit response prediction	18M	response prediction
LM (Billion Word Benchmark)	30M	language modeling

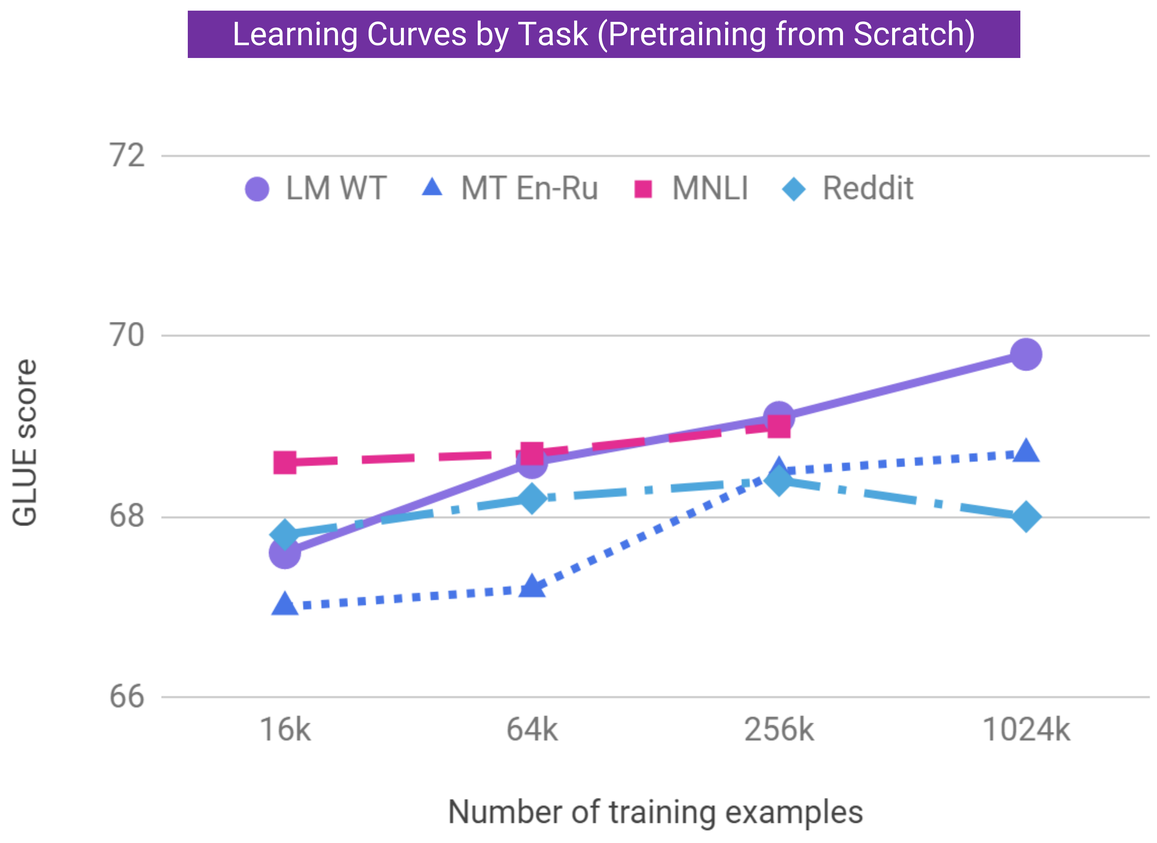
Methodology

- Collect a lot of standard NLP tasks (see table)
- Pretrain a model from scratch or finetune an existing model (ELMo, BERT) on each task
- Evaluate models on the target tasks in the GLUE Benchmark (Table 1, **bold**)
- Compare to no pretraining (i.e. only train on target task) and a random encoder (frozen during target task training)

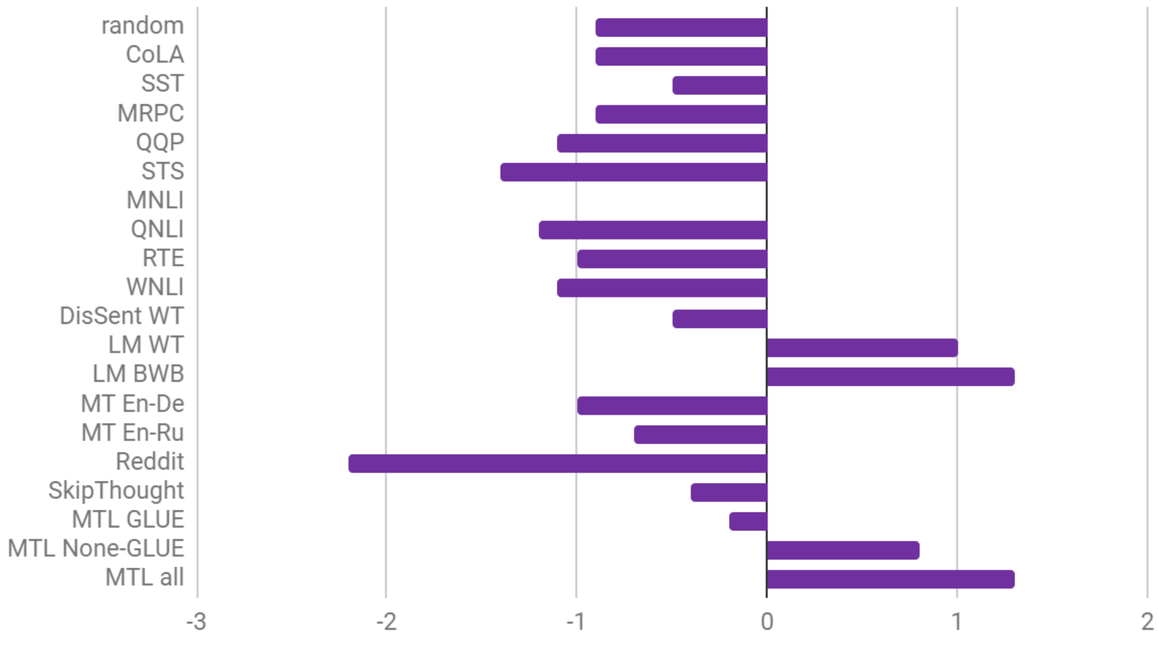
Results and Discussion

- Language modeling is the most effective single pretraining task; most other tasks fail to outperform no pretraining.
- A random encoder is a surprisingly strong baseline.
- Many tasks fail to improve on, and sometimes substantially harm, finetuning BERT for target tasks without intermediate training. We get slightly better results with ELMo (see paper).
- Multitask learning improves over the best single pretraining task, but margins are relatively slim.
- We plot learning curves for all tasks and settings. When pretraining from scratch, language modeling scales most promisingly in number of training examples. In finetuning an existing language model, trends are noisy and unclear (see paper).
- We compute correlations between target tasks and find that many tasks have near-zero or negative correlations.

Learning Curves by Task (Pretraining from Scratch)



GLUE Score by Task (Pretraining from Scratch)



GLUE Score by Task (Intermediate Training of BERT)