Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith

Presenter: Liyan Tang

Reference: https://arxiv.org/pdf/2004.10964.pdf

Motivation

Representations learned by large pretrained models achieve strong performance across many tasks with datasets of varying sizes drawn from a variety of sources.

Question: Do the large pretrained models work universally or is it still helpful to build separate pretrained models for specific domains?

Setting

Model: RoBERTa (pre-trained on corpus derived from multiple sources).

We consider four domains:

- Biomedical (BIOMED) papers
- Computer science publications
- Newstext from REALNEWS
- AMAZON reviews

and eight classification tasks (two in each domain).

Domain Similarity

	PT	News	Reviews	BioMed	CS
PT	100.0	54.1	34.5	27.3	19.2
News	54.1	100.0	40.0	24.9	17.3
Reviews	34.5	40.0	100.0	18.3	12.7
BioMed	27.3	24.9	18.3	100.0	21.4
CS	19.2	17.3	12.7	21.4	100.0

Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to RoBERTa's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Eight classification tasks

Domain	Task	Label Type	Train (Lab.)	Train (Unl.)	Dev.	Test	Classes
BioMed	ChemProt	relation classification	4169	-	2427	3469	13
	†RCT	abstract sent. roles	18040	-	30212	30135	5
CS	ACL-ARC	citation intent	1688	-	114	139	6
	SciERC	relation classification	3219	-	455	974	7
News	HyperPartisan	partisanship	515	5000	65	65	2
	†AGNews	topic	115000	-	5000	7600	4
Reviews	†Helpfulness	review helpfulness	115251	-	5000	25000	2
	†IMDB	review sentiment	20000	50000	5000	25000	2

Our tasks represent both high- and low-resource (≤ 5K labeled training examples)

Domain-Adaptive Pretraining (DAPT)

Definition: Domain-Adaptive Pretraining (DAPT) refers to continue pretraining LM (in this paper, RoBERTa) on a large corpus of unlabeled domain-specific text.

Expection: the more dissimilar the domain, the higher the potential for DAPT.

We use an off-the-shelf RoBERTa-base model and perform supervised fine-tuning of its parameters for each classification task. Before that, we pre-train ROBERTA on each domain for 12.5K steps. This phase of pretraining results in four domain-adapted LMs, one for each domain.

Results

Dom.	Task	RoBa.	DAPT	¬DAPT
BM	ChemProt	$81.9_{1.0}$	**$84.2_{0.2}$**	$79.4_{1.3}$
	†RCT	$87.2_{0.1}$	**$87.6_{0.1}$**	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	**$75.4_{2.5}$**	$66.4_{4.1}$
	SciERC	$77.3_{1.9}$	**$80.8_{1.5}$**	$79.2_{0.9}$
News	HyP.	$86.6_{0.9}$	**$88.2_{5.9}$**	$76.4_{4.9}$
	†AGNews	**$93.9_{0.2}$**	**$93.9_{0.2}$**	$93.5_{0.2}$
Rev.	†Helpful.	$65.1_{3.4}$	**$66.5_{1.4}$**	$65.1_{2.8}$
	†IMDB	$95.0_{0.2}$	**$95.4_{0.2}$**	$94.1_{0.4}$

We observe that DAPT improves over RoBERTa in all domains, demonstrating:

1. the benefit of DAPT when the target domain is more distant from ROBERTA’s source domain.
2. DAPT may be useful even for tasks that align more closely with ROBERTA’s source domain.

Sanity Check

Is the improvements over RoBERTa attributed simply to exposure to more data, regardless of the domain?

In this setting, for NEWS, we use a CS LM; for REVIEWS, a BIOMED LM; for CS, a NEWS LM; for BIOMED, a REVIEWS LM. We use the vocabulary overlap statistics to guide these choices (least vocab overlap).

Results

Dom.	Task	RoBa.	DAPT	¬DAPT
BM	ChemProt	$81.9_{1.0}$	$\mathbf{84.2}_{0.2}$	$79.4_{1.3}$
	†RCT	$87.2_{0.1}$	$\mathbf{87.6}_{0.1}$	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$\mathbf{75.4}_{2.5}$	$66.4_{4.1}$
	SciERC	$77.3_{1.9}$	$\mathbf{80.8}_{1.5}$	$79.2_{0.9}$
News	HyP.	$86.6_{0.9}$	$\mathbf{88.2}_{5.9}$	$76.4_{4.9}$
	†AGNews	$\mathbf{93.9}_{0.2}$	$\mathbf{93.9}_{0.2}$	$93.5_{0.2}$
Rev.	†Helpful.	$65.1_{3.4}$	$\mathbf{66.5}_{1.4}$	$65.1_{2.8}$
	†IMDB	$95.0_{0.2}$	$\mathbf{95.4}_{0.2}$	$94.1_{0.4}$

- DAPT significantly outperforms adapting to an irrelevant domain (¬DAPT), suggesting the importance of pretraining on domain-relevant data.
- ¬DAPT results in worse performance than even RoBERTa on end-tasks.
- In some cases, continued pre-training on any additional data is useful

Task-Adaptive Pretraining (TAPT)

Definition: Task-adaptive pretraining (TAPT) consists of a second phase of pretraining RoBERTa, but only on the available task-specific unlabeled training data (a cheaper adaptation technique compare to DAPT).

Setting: Task of interest covers only a subset of the text available within the broader domain.

Expectation: In cases where the task data is a narrowly-defined subset of the broader domain, pretraining on the task dataset itself or data relevant to the task may be helpful.

Result

Domain	Task	RoBERTa	Additional Pretraining Phases DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

TAPT consistently improves the RoBERTa baseline for all tasks across domains. Even on the news domain, which was part of RoBERTa pre-training corpus, TAPT improves over RoBERTa, showcasing the advantage of task adaptation; The last column shows the combination of DAPT and TAPT, which yields even better performance.

Cross-Task Transfer

Transfer-TAPT: exploring whether adapting to one task transfers to other tasks in the same domain.

E.x. Further pre-train the LM using the unlabeled data from A, fine-tune it with the labeled data B, where A and B are from the same domain, and observe the effect.

Result

BioMed	RCT	ChemProt
TAPT	$87.7_{0.1}$	$82.6_{0.5}$
Transfer-TAPT	$87.1_{0.4}$ (↓0.6)	$80.4_{0.6}$ (↓2.2)

CS	ACL-ARC	SciERC
TAPT	$67.4_{1.8}$	$79.3_{1.5}$
Transfer-TAPT	$64.1_{2.7}$ (↓3.3)	$79.1_{2.5}$ (↓0.2)

News	HyperPartisan	AGNews
TAPT	$89.9_{9.5}$	$94.5_{0.1}$
Transfer-TAPT	$82.2_{7.7}$ (↓7.7)	$93.9_{0.2}$ (↓0.6)

Reviews	Helpfulness	IMDB
TAPT	$68.5_{1.9}$	$95.7_{0.1}$
Transfer-TAPT	$65.0_{2.6}$ (↓3.5)	$95.0_{0.1}$ (↓0.7)

These results show the differences in task distributions within a domain. Further, this could also explain why adapting only to a broad domain is not sufficient, and why TAPT after DAPT is effective.

Augmenting Training Data for TAPT

Setting: Dataset is often downsampled to collect annotations. The larger unlabeled corpus is thus expected to have a similar distribution to the task's training data (we call this curated data).

Result

simulated low-resource setting

Pretraining	BioMed RCT-500	News HyP.	Reviews IMDB †
TAPT	$79.8_{1.4}$	$90.4_{5.2}$	$95.5_{0.1}$
DAPT + TAPT	$83.0_{0.3}$	$90.0_{6.6}$	$95.6_{0.1}$
Curated-TAPT	$83.4_{0.3}$	$89.9_{9.5}$	$95.7_{0.1}$
DAPT + Curated-TAPT	$\mathbf{83.8}_{0.5}$	$\mathbf{92.1}_{3.6}$	$\mathbf{95.8}_{0.1}$

Curating large amounts of data from the task distribution is extremely beneficial to end-task performance.

Recommendation: release a large pool of unlabeled task data to aid model adaptation through pretraining.

Automated Data Selection for TAPT

Setting: Consider a low-resource scenario without access to large amounts of unlabeled data to adequately benefit from TAPT, as well as absence of computational resources necessary for DAPT.

Use the idea of kNN to select k candidates similar to each task sentence based on embedding space as unlabeled data.

Result

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

1. kNN-TAPT outperforms TAPT for all cases.
2. As we increase k, kNN-TAPT performance steadily increases, and approaches that of DAPT.

Note: curating large in-domain data is expensive.

DAPT vs TAPT

Computational Requirement: TAPT is nearly 60 times faster to train than DAPT on a single v3-8 TPU and storage requirements for DAPT on this task are 5.8M times that of TAPT.

TAPT uses a far smaller pretraining corpus, but one that is much more task-relevant, This makes TAPT much less expensive to run than DAPT

Take Away

1. The more dissimilar the domain (target domain vs. pretraining domain), the higher the potential for DAPT.
2. It's important to do further pretraining on domain-relevant data.
3. Compared to DAPT, TAPT uses a far smaller pretraining corpus, but one that is much more task-relevant.
4. The performance of TAPT is often competitive with that of DAPT.
5. Curating large amounts of data from the task distribution is extremely beneficial to end-task performance.
6. Combined domain- and task-adaptive pre- training achieves the best performance on all tasks.