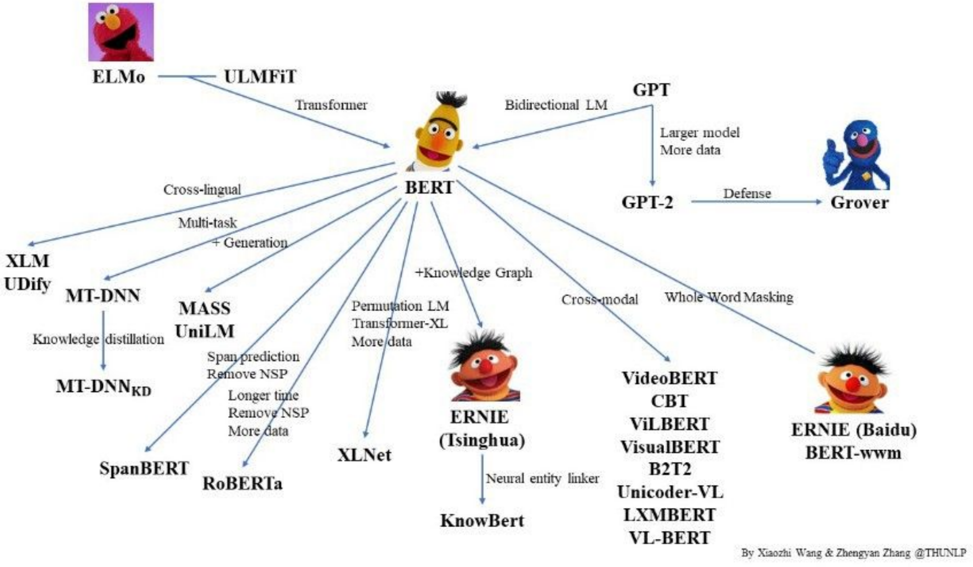
Domain Adversarial Fine-Tuning as an Effective Regularizer

Giorgos Vernikos, Katerina Margatina, Alexandra Chronopoulou, Ion Androutsopoulos

Transfer Learning in NLP



scarcity of labeled data for NLP tasks
➜ implicit data augmentation

overfitting to small datasets
➜ transfering from unsupervised task improves sample complexity and overall performance (Dai & Le, Yogotama et al.)

Transfer Learning in NLP

Two-stage process:

Pretraining
Training on unlabeled generic data in a variant of language modelling

Fine-Tuning
Initialize with the pretrained Language Model (LM) &
Train on the labeled task-specific data

Challenges: Overfitting, Catastrophic Forgetting

Transfer Learning in NLP

How to improve fine-tuning?

- ❑ Additional/Multitask training on labeled data or language modelling (Howard & Ruder, Liu et al., Phang et al., Gururangan et al.)
- ❑ Optimization stability (parameter freezing, lower learning rates, more iterations) (Howard & Ruder, Chronopoulou et al., Mosbach et al.)
- ❑ Penalize deviations from the parameters of the pretrained model (Kirkpatrick et al., Wiese et al., Lee et al.)
- ❑ Enforce constraints on the high-level representations of the model (Zhu et al., Cao et al., Jiang et al., Aghajanyan et al.)

Proposed approach:

domain Adversarial Fine-Tuning as an Effective Regularizer (**AFTER**)

Loss of **general-domain** representations as a form of catastrophic forgetting.

Adversarial loss that enforces invariance of text representations across different domains during fine-tuning.

The adversarial term acts as a **regularizer** that preserves most of the general-domain knowledge captured during pretraining.

Proposed approach: AFTER

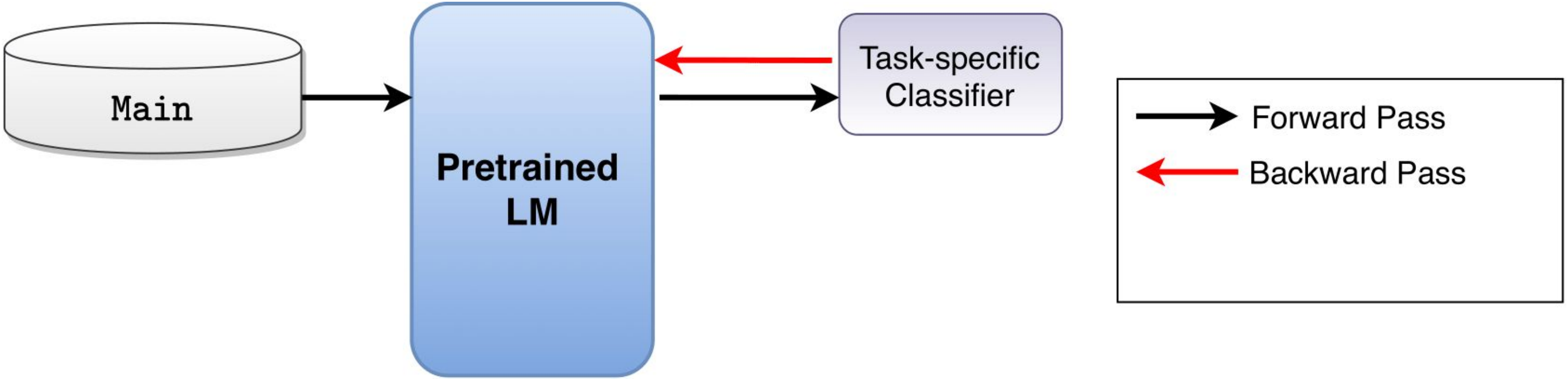
Regularize the extent to which the pretrained parameters are allowed to adapt to the target task domain.

Objective:

$$\mathcal{L}_{\text{AFTER}} = L_{Main} - \lambda L_{Domain}$$

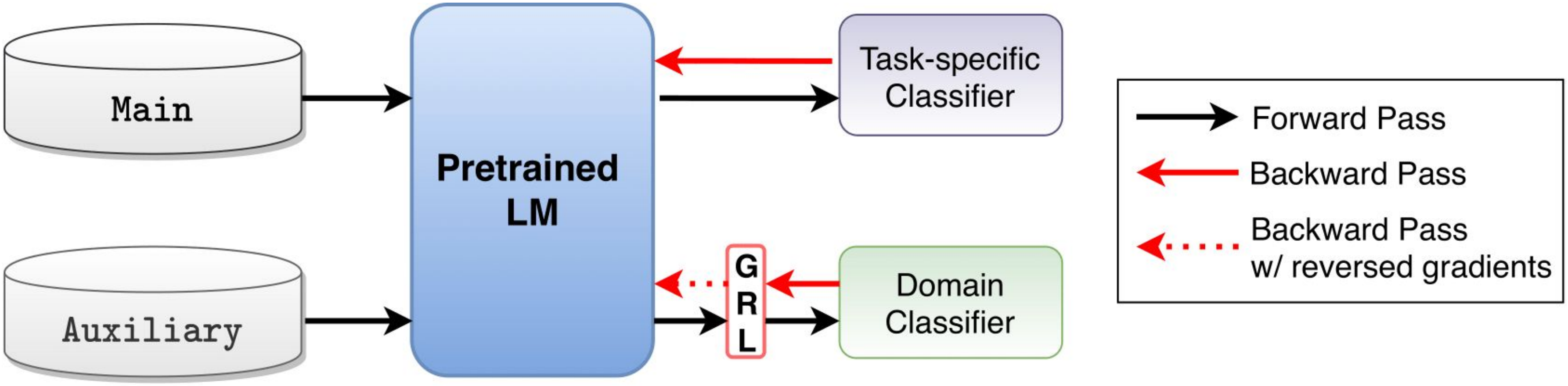
L_{Main} is the task-specific loss,
L_{Domain} refers to the auxiliary task of discriminating between in-domain and out-of-domain samples,
λ controls the importance of the second term

Model Architecture



Standard Fine-Tuning (SFT)

Model Architecture



AFTER

Datasets & Tasks

DATASET	DOMAIN	N_{train}
Main		
CoLA	Miscellaneous	8.5K
SST-2	Movie Reviews	67K
MRPC	News	3.7K
RTE	News, Wikipedia	2.5K
Auxiliary		
AG NEWS	Agricultural News (NEWS)	120K
EUROPARL	Legal Documents (LEGAL)	120K
AMAZON	Electronics Reviews (REVIEWS)	120K
PUBMED	Medical Papers (MEDICAL)	120K
MATHEMATICS	Mathematics Questions (MATH)	120K

4 datasets from the GLUE benchmark as *Main*

5 corpora as *Auxiliary* data from various domains

Results: BERT

	CoLA *Matthews corr.*	SST-2 *Accuracy*	MRPC *Accuracy / F1*	RTE *Accuracy*
BERT SFT	55.5 ± 3.2	92.0 ± 0.5	85.4 ± 1.1 / 89.6 ± 0.6	64.3 ± 3.1
AFTER W/ NEWS	$\mathbf{57.3} \pm 1.5$	$\underline{92.5} \pm 0.4$	$\mathbf{87.5} \pm 1.7$ / $\mathbf{91.1} \pm 1.2$	$\underline{64.7} \pm 1.9$
AFTER W/ REVIEWS	$\underline{57.1} \pm 1.2$	$\underline{92.4} \pm 0.3$	$\underline{86.4} \pm 0.3$ / $\underline{90.1} \pm 0.4$	$\underline{64.6} \pm 0.8$
AFTER W/ LEGAL	55.0 ± 1.5	$\underline{92.4} \pm 0.3$	$\underline{86.6} \pm 0.6$ / $\underline{90.3} \pm 0.5$	$\mathbf{64.8} \pm 1.9$
AFTER W/ MEDICAL	$\underline{55.9} \pm 2.9$	$\mathbf{92.6} \pm 0.3$	$\underline{86.9} \pm 1.3$ / $\underline{90.7} \pm 1.0$	62.6 ± 3.4
AFTER W/ MATH	$\underline{56.1} \pm 2.8$	$\underline{92.3} \pm 0.8$	$\underline{87.3} \pm 0.9$ / $\underline{90.8} \pm 0.7$	62.5 ± 1.3

- AFTER improves performance over SFT on 4 datasets and can reduce variance
- gains are consistent across different *Auxiliary* data (except RTE)

Results: XLNET

	CoLA	SST-2	MRPC	RTE
	Matthews corr.	*Accuracy*	*Accuracy / F1*	*Accuracy*
XLNet SFT	–	93.0 ± 0.7	86.4 ± 0.7 / 90.1 ± 0.5	64.7 ± 4.4
AFTER w/ NEWS	–	$\mathbf{93.9} \pm 0.3$	87.3 ± 0.7 / 91.0 ± 0.5	63.9 ± 2.3
AFTER w/ REVIEWS	–	93.5 ± 0.3	86.9 ± 0.6 / 90.5 ± 0.5	65.1 ± 2.8
AFTER w/ LEGAL	–	93.6 ± 0.5	$\mathbf{87.5} \pm 1.6$ / $\mathbf{90.9} \pm 1.2$	64.8 ± 1.6
AFTER w/ MEDICAL	–	93.3 ± 0.5	87.0 ± 1.1 / 90.5 ± 0.7	64.5 ± 2.1
AFTER w/ MATH	–	$\mathbf{93.9} \pm 0.4$	87.3 ± 1.2 / 90.8 ± 0.9	$\mathbf{66.1} \pm 1.9$

- AFTER improves performance for an even higher-performing LM
- AFTER with BERT outperforms XLNET SFT baseline on two tasks

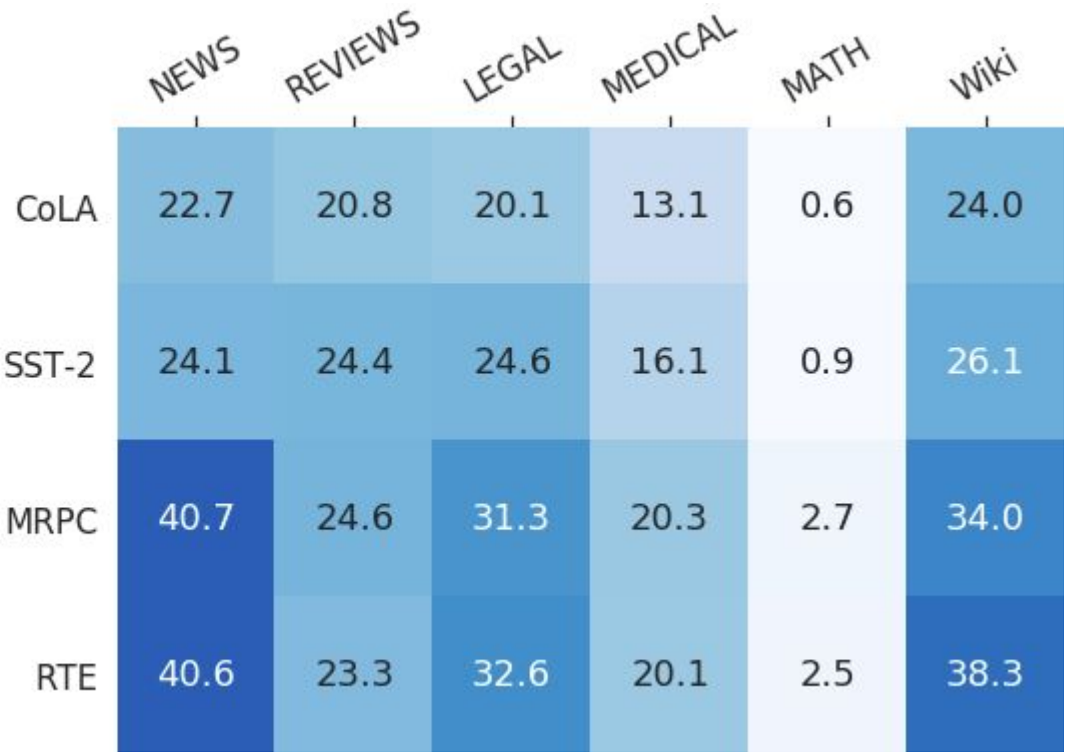
Ablation Study: Domain of the pretraining data

Does the similarity between the domain of the LMs' pretraining data and the task-specific domain matter?

	RTE	MRPC	CoLA	SST-2
MLM Loss	**2.17**	2.37	2.53	3.39
Overlap with WIKI (%)	**38.3**	34.0	24.0	26.1
AFTER Improvement (%)	**0.8**	2.5	3.2	0.7

general-domain representations created during pretraining ≈ domain-specific representations created during fine-tuning

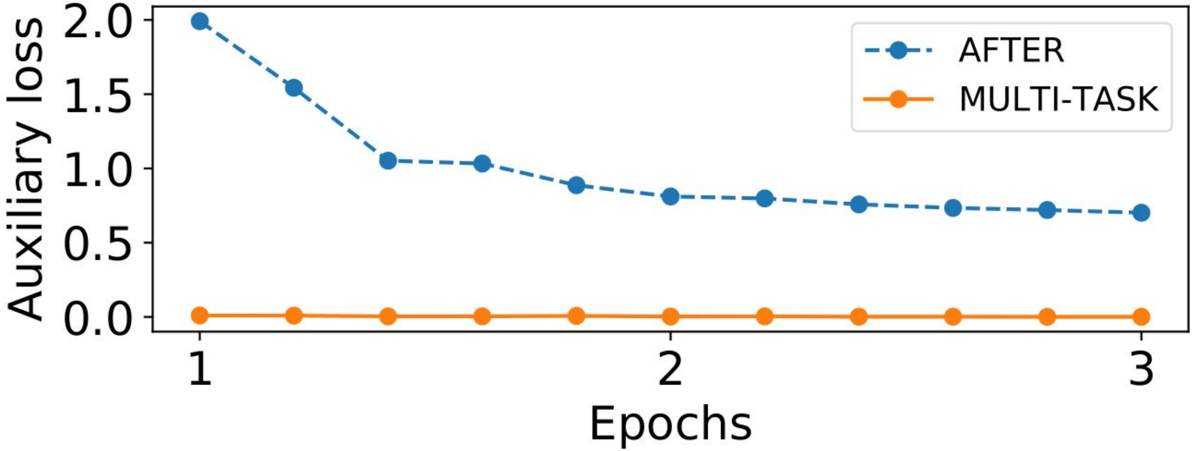
Ablation Study: Domain Distance



We measure the distance between *Main* and *Auxiliary* domains.

No clear pattern emerges, demonstrating, perhaps, the robustness of our approach.

Ablation Study: Domain-invariant vs. Domain-specific



	CoLA	MRPC
AFTER W/ NEWS	**57.3**	**87.5/91.1**
MULTI-TASK W/ NEWS	56.5	86.7/90.5

Conclusions

- We propose AFTER that adds an adversarial domain classification loss to the task-specific loss.
- Our approach does **not require additional labeled data** and is applicable to any transfer learning scenario and model architecture.
- AFTER consistently **outperforms standard fine-tuning**.
- AFTER is more effective when the pretraining and target task data come from different domains and is generally robust to the choice of *Auxiliary* data.

Thank you!

Paper: https://arxiv.org/abs/2009.13366v2

Code: https://github.com/GeorgeVern/AFTERV1.0

Twitter: @gvernikos