An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models

Lifu Tu[1,4] Garima Lalwani[2] Spandana Gella[2] He He[3,4]

[1]Toyota Technological Institute at Chicago

[2]Amazon AI

[3]New York University

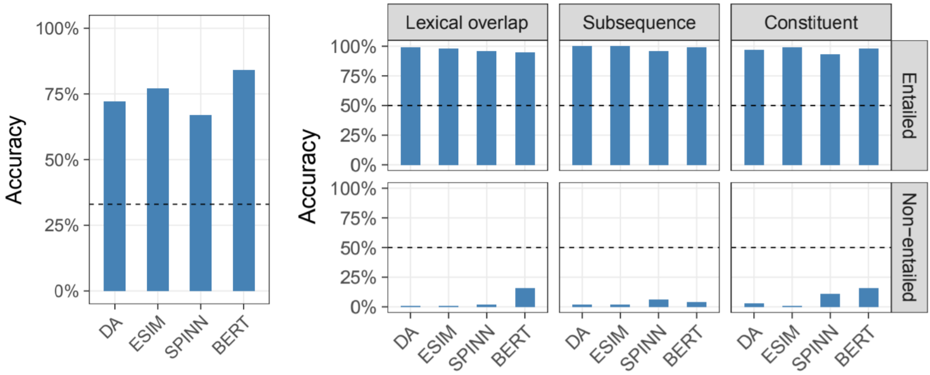
[4]Most work was done during working at Amazon

EMNLP 2020

Table of Contents

Motivation

Models achieve high accuracy on benchmarks, however, perform poorly on the challenging datasets [McCoy et al., 2019] .



- Spurious correlations is learned.
- How to improve robustness to spurious correlations?

NLI

Representative example from MNLI [Williams et al., 2017]
P: The doctor mentioned the manager who ran.
H: The doctor mentioned the manager
entailment

Representative example from HANS [McCoy et al., 2019]
P: The actors who advised The manager saw the tourists.
H: The manager saw the tourists
non-entailment!

PI

Representative example from QQP [Iyer et al., 2017] :
S1: Bangkok vs Shanghai?
S2: Shanghai vs Bangkok?
paraphrase

Representative example from PAWS$_{QQP}$ [Zhang et al., 2017] :
S1: Are all dogs smart or can some be dumb?
S2: Are all dogs dumb or can some be smart?
non-paraphrase!

Word overlap-based heuristic that works for training examples **fails** on the test data

Table of Contents

Pre-training Improve Robust Accuracy

Recently, people find pre-training improve robustness. [Hendrycks et al. (2019, 2020); Li et al. (2019)]

However, could we answer the following questions?

- What role does longer fine-tuning play?
  - Minority examples require longer fine-tuning.
- How do pre-trained models generalize to out-of-distribution data?
  - Minority patterns in the training set
- When do they generalize well given the inconsistent improvements?
  - Different minority patterns may require varying amounts of training data

What Role does Longer Fine-tuning Play?

We observe longer fine-tuning:

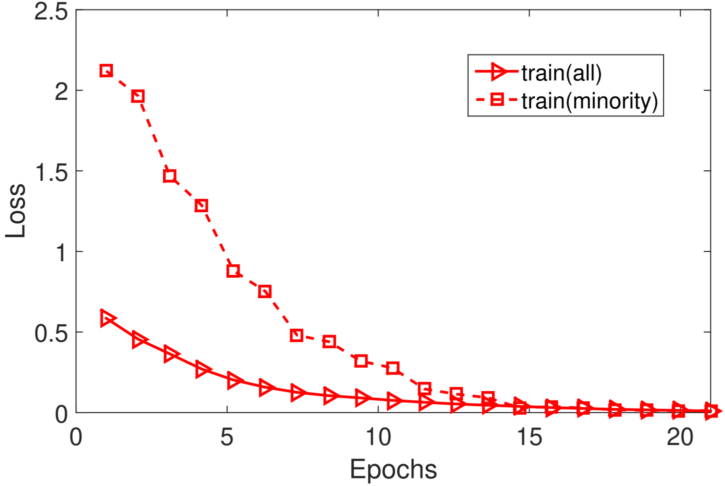
- in-distribution accuracy saturates quickly
- improves accuracy on challenging examples

Hypothesis: minority examples require longer fine-tuning.

Experimental Details

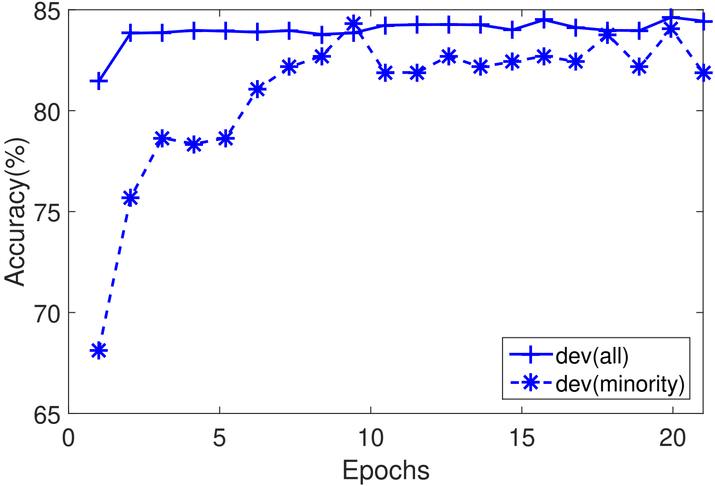
Tasks: NLI
Setting: fine-tuning pre-trained models
Metric: training loss and dev set accuracy

What Role does Longer Fine-tuning Play?



Training loss of minority examples decreases more slowly!

What Role does Longer Fine-tuning Play?



minority examples: epoch 10; all examples: epoch 5.

How do pre-trained models generalize to out-of-distribution data?

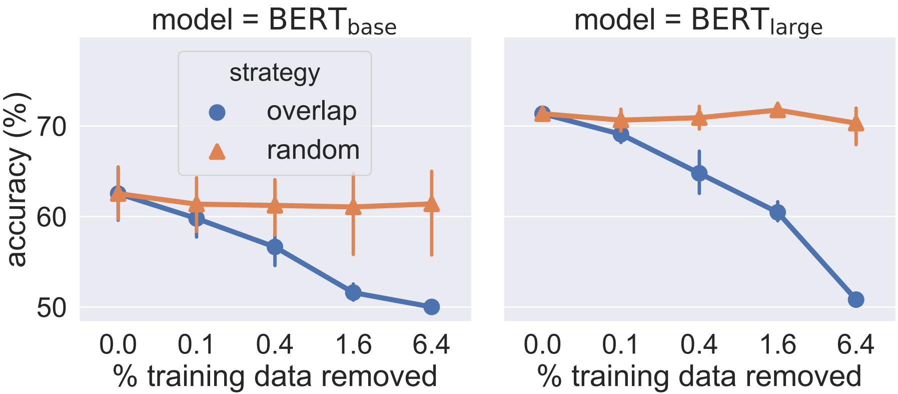
Do pre-trained model enable extrapolation to unseen patterns? no

Hypothesis: pre-trained models generalize better from minority patterns in the training set.

Representative minority example:
"fly from Chicago to New York" vs. "fly from New York to Chicago"

Experimental Details

Task: MNLI
Setting: remove minority (727) only vs. randomly in MNLI training set
Metric: accuracy on the challenging dataset (HANS)



Removing high overlap examples have significantly worse accuracy

When do They Generalize Well Given the Inconsistent Improvements?

Previously we find fine-tuning makes the different improvement on two tasks: NLI and PI.

Why?

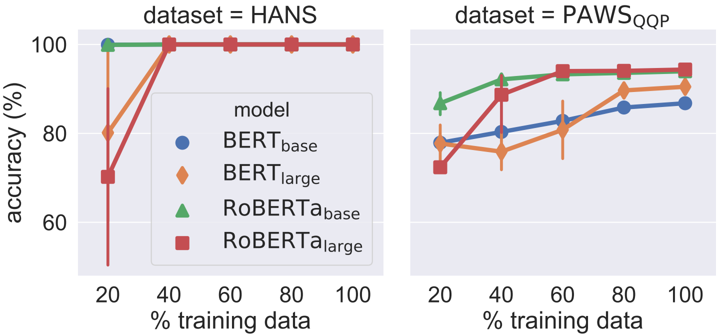
Hypothesis: PAWS have syntactically more complex sentences!

Experimental Details

Tasks: NLI and PI
Setting: fine-tuning pre-trained models on the challenging datasets directly
Metric: accuracy on the challenging dataset

Experimental Details

Fine-tuning pre-trained models on the challenging datasets directly.



PAWS contains longer and syntactically more complex sentences

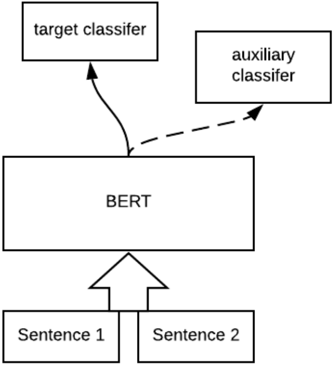
Length: 20.7 (PAWS) VS. 9.2 (HANS)
parse tree height: 11.4 (PAWS) VS. 7.5 (HANS)

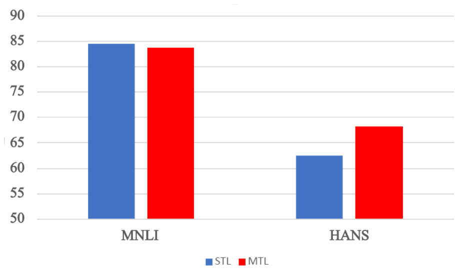
Different minority patterns may require varying amounts of training data

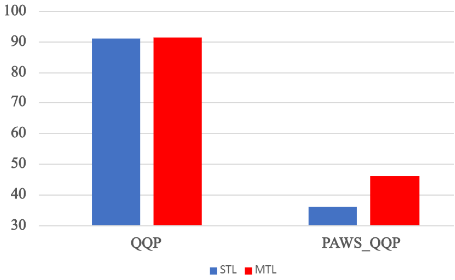
Table of Contents

Increasing the amount of minority examples helps to improve model robustness. How to improve robustness further?

Aggregating generic data from various sources through multi-task learning.







MTL improves robust accuracy and do not hurt in-distribution performance.

How MTL Helps Generalization from Minority Examples?

How to explain the improvement?

- Challenging data in Auxiliary datasets? No
- MTL reduces sample complexity ? Yes

Two Ablation Studies

- Ablation Study 1: removing auxiliary datasets
- Ablation Study 2: remove minority examples from both the auxiliary and the target datasets

Ablation Study1

Setting

Target dataset: QQP
Auxiliary datasets: HANS (challenging dataset) + MNLI + SNLI
remove auxiliary datasets one by one

Removed	**$\mathbf{PAWS_{QQP}}$**	Δ
None	45.9	-
HANS	45.3	-0.6
MNLI	42.3	-3.6
SNLI	44.2	-1.7

The challenging datasets are not much more helpful than benchmark datasets

Ablation Study2

Setting

Target dataset: QQP
Auxiliary dataset: MNLI
Remove minority examples from both the auxiliary and the target datasets

Removed	**PAWS$_{\text{QQP}}$**	Δ
None	45.9	-
random examples		
QQP	44.3	-1.6
MNLI	45.0	-0.9
minority examples		
QQP	38.2	-7.7
MNLI	44.3	-1.6

Improved generalization is from minority examples.

Table of Contents

Conclusions

- Analysis of robustness using pre-trained language models
- Generalization is from a small amount of minority examples.
- More pre-training data, larger model size, and additional auxiliary data can improve robustness

Suggestion to Future Directions

Importance of data diversity

Traditional techniques could still helpful.

Thanks!