On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines

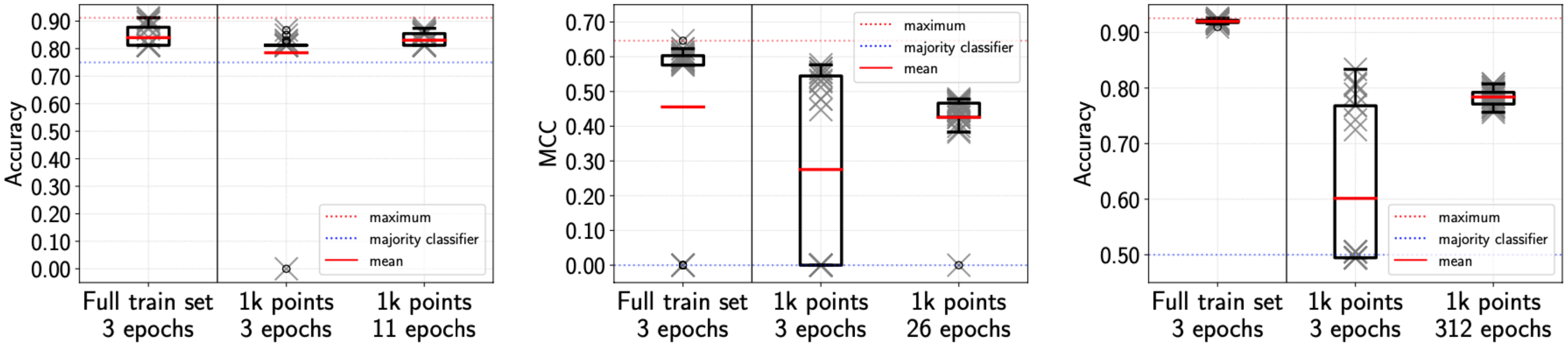
Marius Mosbach, Maksym Andriushchenko, Dietrich Klakow

Motivation

- Fine-tuned pre-trained language models are *everywhere*.
- Previous works observed **instabilities** during fine-tuning (Devlin et al. (2019), Phang et al. (2018), Dodge et al. (2020)):
    - Changing only the *random seed* leads to large differences in down-stream task performance (e.g. accuracy, F1, MCC).
    - Some fine-tuning runs fail entirely, leading to chance performance.
- Why is fine-tuning prone to failures and how can we improve it's stability?

Devlin et al. (2019) - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Phang et al. (2018) - Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks
Dodge et al. (2020) - Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping

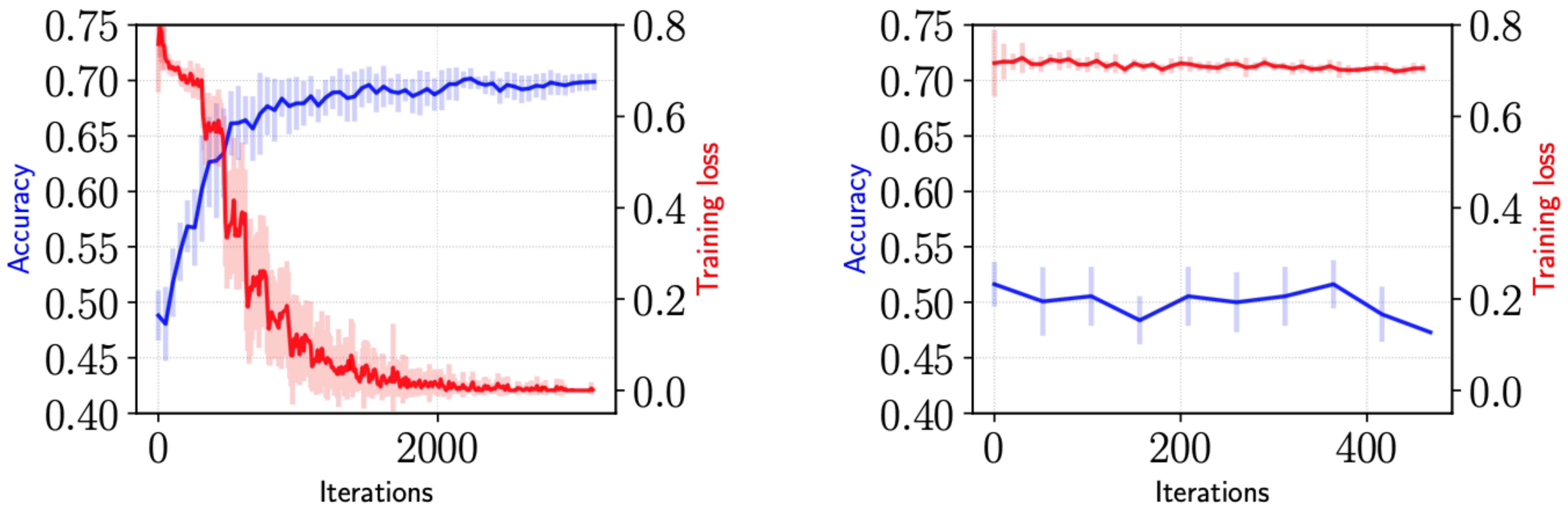
Hypothesis 1: Small training datasets



(a) MRPC (b) CoLA (c) QNLI

- Small datasets are **not** causing fine-tuning instability.
- Fixing the number of epochs is sub-optimal.
- **Number of iterations** is crucial to get back original fine-tuning stability.

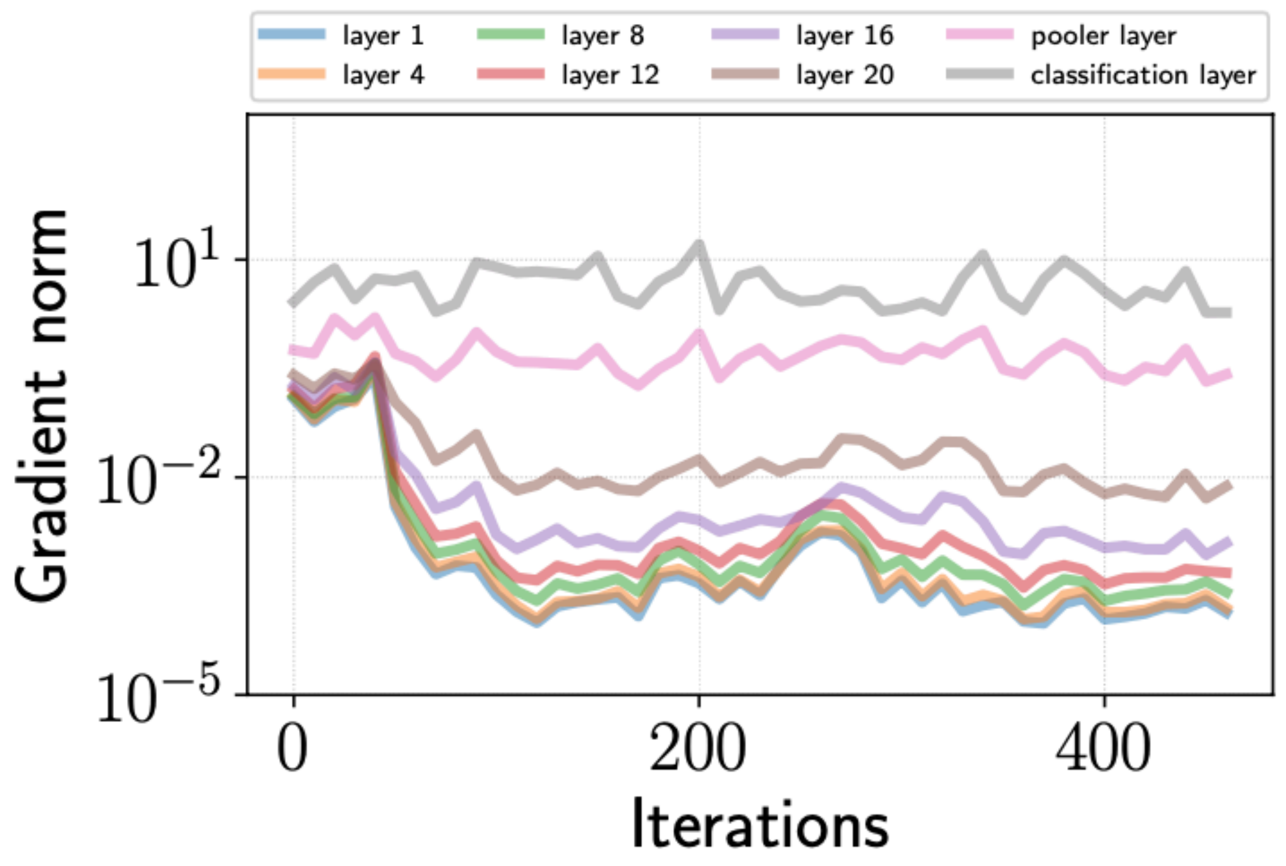
Hypothesis 2: Catastrophic forgetting



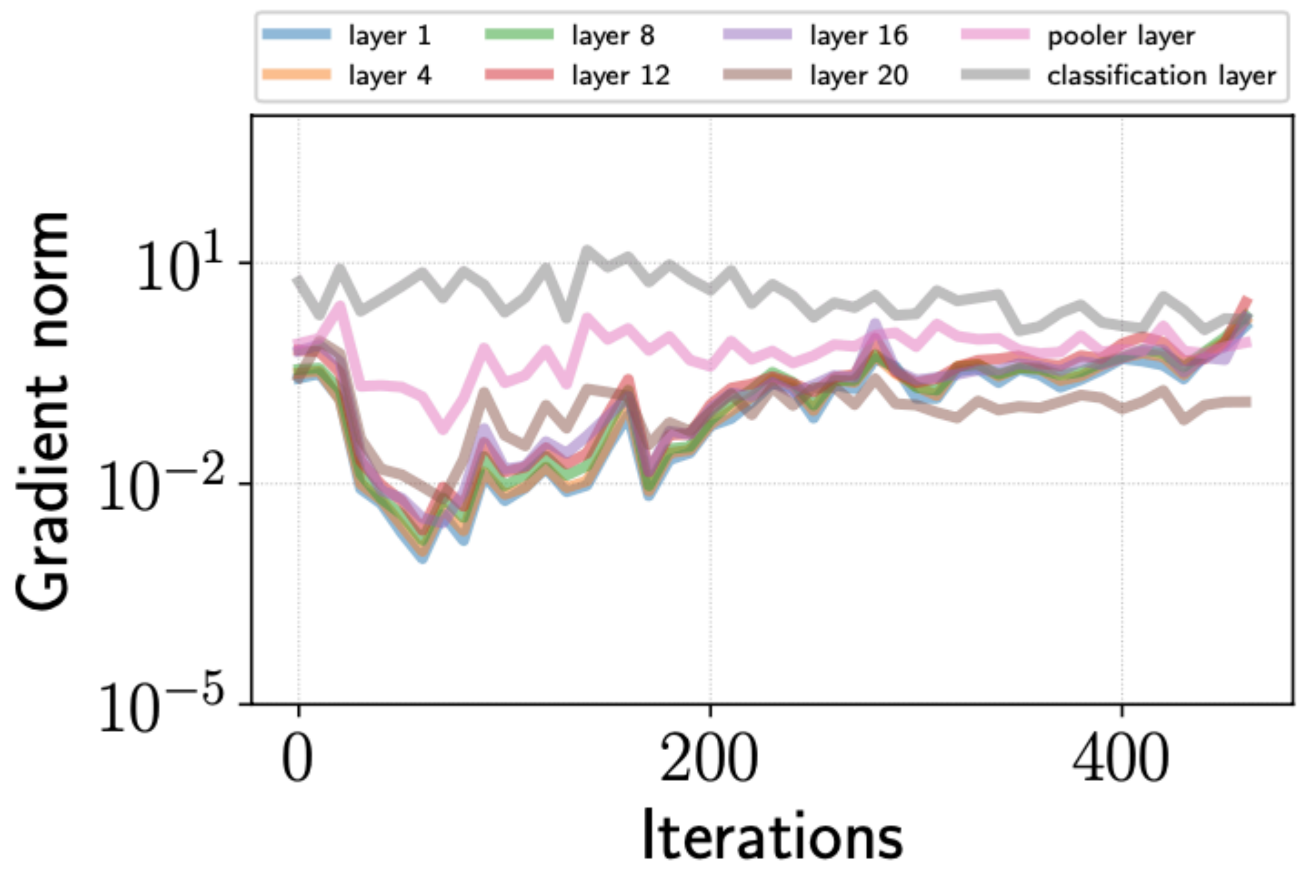
(a) Successful runs (b) Failed runs

- *Catastrophic Forgetting* is **not** causing fine-tuning instability.
- Failed fine-tuning runs **don't learn anything at all**.

What happens with failed runs?



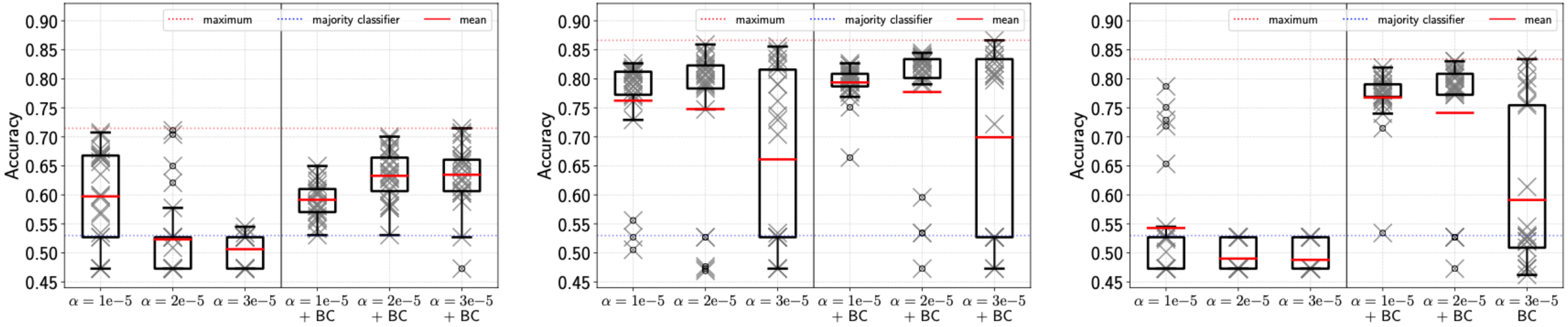
(a) Failed run



(b) Successful run

- **Vanishing gradients** problem occurs early in training.
- This suggests *optimization issues*.

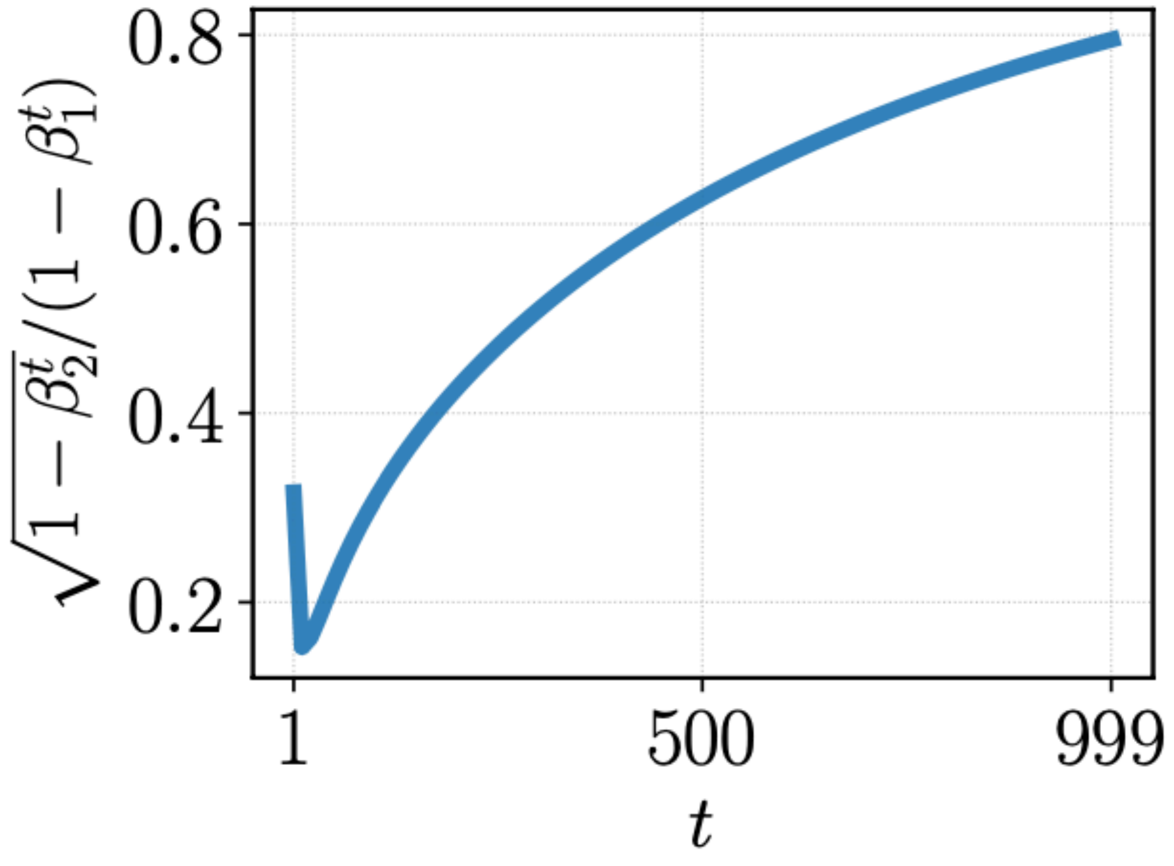
The role of optimization



(a) BERT (b) RoBERTa (c) ALBERT

- **Small step-size** and **bias correction** are crucial for successful fine-tuning.

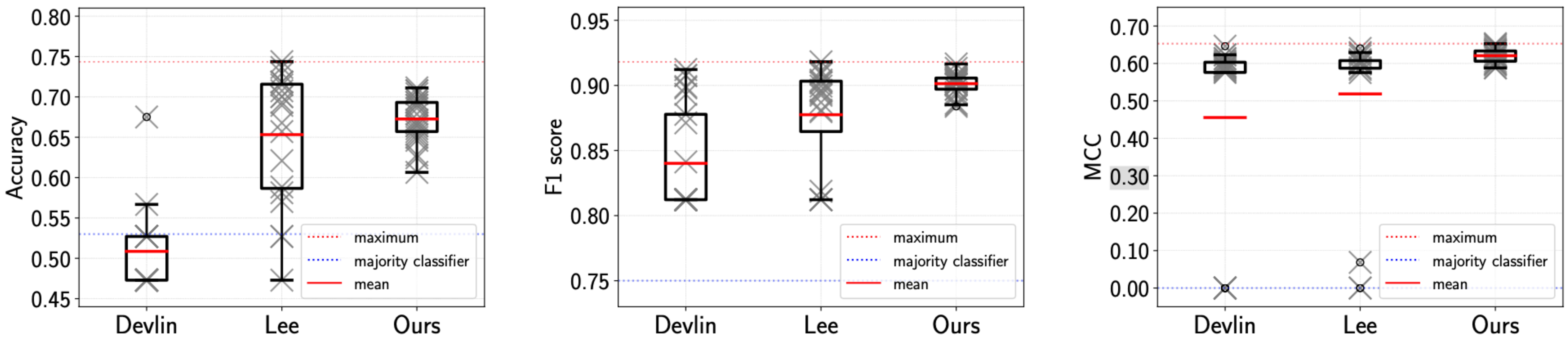
- Failed runs can be avoided, fine-tuning is more stable.



A simple but strong baseline

1. Use **small learning rates** and enable **bias-correction**.

2. **Don't fix number of epochs a priori!** Train for many **iterations.**



(a) RTE (b) MRPC (c) CoLA

Devlin et al. (2019) - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Lee et al. (2020) - Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models

Thanks for listening!

- More experiments and results can be found in the paper.
- Come visit our poster presentation on **May 3rd** @ **09:00 am PDT** **(Poster Session 2)**.
- Link to poster, paper, code.
- Check out concurrent work by Zhang et al. (2021) @ ICLR 2021.

@mariusmosbach

@maksym_andr

@dklakow