Memorisation versus Generalisation in Pre-trained Language Models

Michael Tanzer
m.tanzer@imperial.ac.uk

Sebastian Ruder
ruder@google.com

Marek Rei
marek.rei@imperial.ac.uk

Introduction

State-of-the-art pre-trained language models have been shown to memorise facts and perform well with limited amounts of training data. To gain a better understanding of how these models learn, we study their generalisation and memorisation capabilities in noisy and low-resource scenarios. We also propose an extension based on prototypical networks that improves performance in low-resource named entity recognition tasks.

Experimental setting

Datasets Two scenarios: noisy and low-resource settings. The noise is simulated by randomly permuting some of the labels in the training set. In order to investigate memorisation we train the models on datasets that contain only a small number of examples for a particular class. We focus on the Named Entity Recognition (NER) task and employ the CoNLL03 (Sang and De Meulder, 2003), JNLPBA (Collier and Kim, 2004), and WNUT17 (Derczynski et al. 2017) datasets.

Language Models We use BERT-base (Devlin et al., 2019) as the main language model for our experiments. We compare BERT's behaviour with that of other pre-trained transformers such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020). We also report performance for a bi-LSTM-CRF (Lample at al., 2016) model with combined character-level and word-level representation.

Acknowledgements

Michael is funded by the UKRI CDT in AI for Healthcare (Grant No. P/S023283/1).

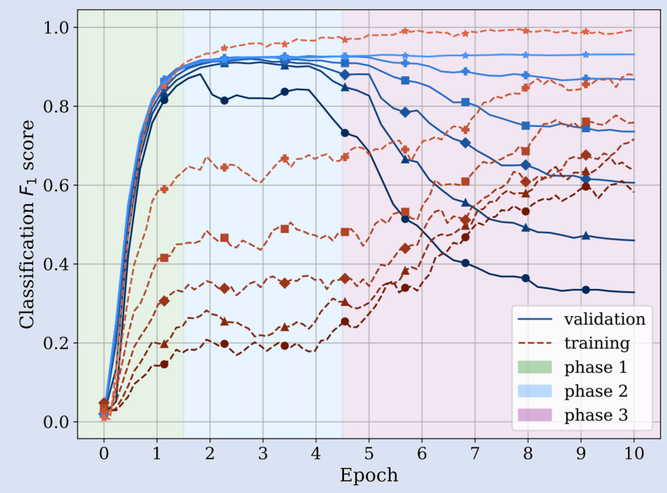
Imperial College London

Generalisation in noisy settings

We train BERT on 6 versions of datasets with noise from 0% to 50% in steps of 10% and report the results.

We find three phases of the training:

- **Fitting**: the generalisation is learned. Improvement for training and validation performance.
- **Settling**; the performance plateaus, neither the training and validation performance changes considerably. The duration of this phase is inversely proportional to the amount of noise added.
- **Memorisation**: the model memorises the noise. Training performance improve while validation performance is degraded.



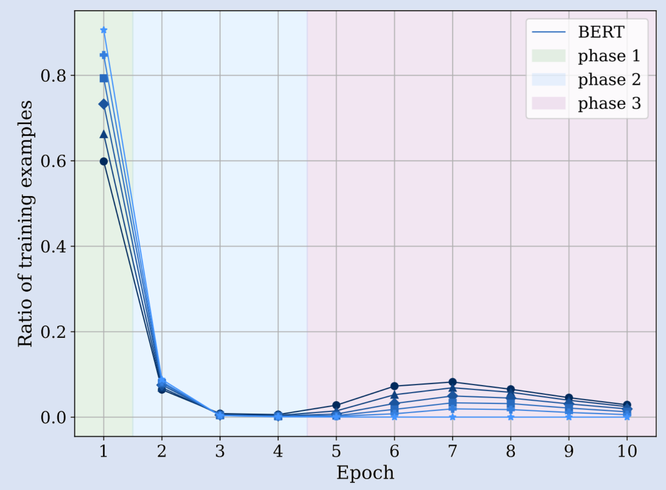
Key insight about the three phases of the training:

- The settling phase is longer in BERT compared to large pre-trained models for other modalities. In some cases the second phase is not observed at all.
- The top validation performance is high for all datasets despite the high noise. This suggests a strong robustness to noise.

Forgetting of learned information

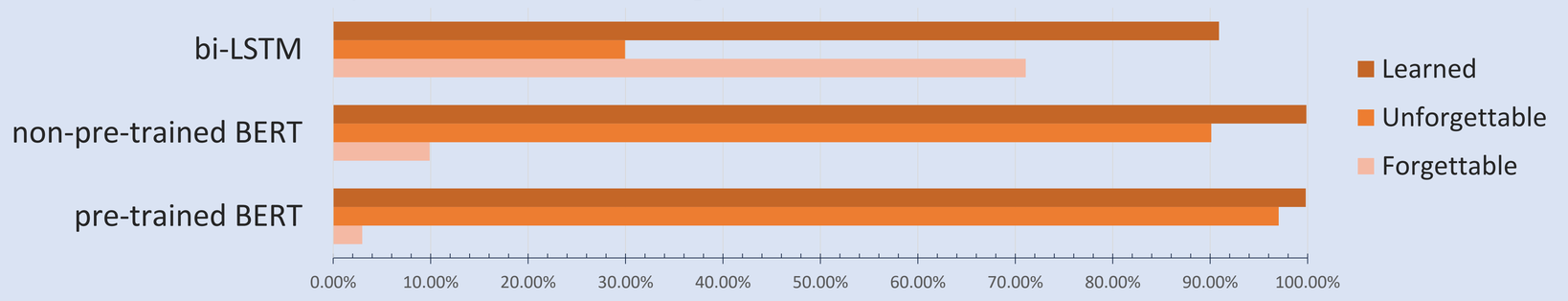
To study memorisation, we use the following concepts:

- **Learned** example: every example that was classified correctly at least once during the training.
- **Forgettable** example: those points that were learned and then forgotten during the trained.
- **Unforgettable** example: all the examples that were learned and never forgotten until the end of the training.



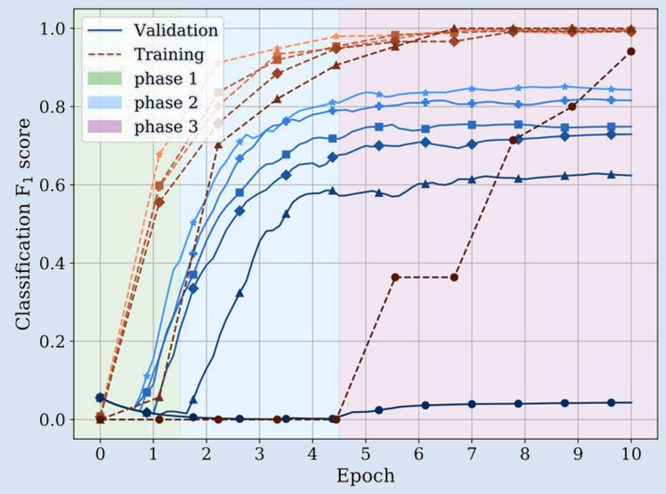
Studying learning and memorisation we can see how:

- BERT forgets much less compared to a bi-LSTM model.
- The pre training process plays a considerable role in retaining information during the training.
- BERT learns most of the examples during the Fitting phase. In the settling phase it simply stops learning any new examples and finally in the memorization phase it starts memorizing the noise.



BERT in low-resource scenarios

We will now examine if the same behaviour applies in low-resource scenarios where a minority class is only observed very few times. To simulate a low-resource scenario, we remove from the CoNLL03 training set all sentences containing tokens with the minority labels MISC and LOC except for a predetermined number of such sentences.



We show the classification performance for the training and validation datasets on 6 different versions of CoNLL03 where we left 5 to 105 sentences with the LOC class. Darker colours represent fewer sentences. We only report the F1 score on the LOC class.

- For fewer available sentences the model's ability to generalise is greatly reduced.
- When only 5 sentences are available, the LOC tokens are treated as noise and learned during the memorisation phase.

ProtoBERT

To address BERT's poor performance on the few-shot learning datasets, we propose a new model: ProtoBERT, that is a combination of BERT, with all its pre-trained knowledge, combined with the few-shot capabilities of prototypical networks (Snell et al., 2017).

ProtoBERT vastly outperforms BERT when fewer sentences are available. As the number of sentences increases the two are roughly equivalent, and in some cases ProtoBERT even outperforms BERT in non-low-resource scenarios tasks.