Coefficient-based fine-tuning of language models on distribution shift

Suma Kasa, Antoine Bigeard and Samuel Barry – CS 330, Fall 2022

Problem statement

Pre-trained models have been highly successful in a wide range of NLP tasks. However, the models are very brittle to the distribution shifts, and techniques like transfer learning or fine-tuning are required to use these models on a new distribution.

The most common fine-tuning approaches are to fine-tune all the parameters or the last few layers of the model. In this project, **we explore the effects of fine-tuning different subsets of layers on two text classification tasks:** hate speech detection and sentiment analysis.

For each task, we combine different datasets with various ratios to create distribution shifts (e.g. $p_{A_{shift}} = 0.9 * p_A + 0.1 * p_B$) and compare the classical fine-tuning approaches with a new approach we introduce, coefficient-based variable fine tuning.

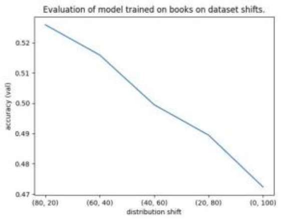
Motivation and prior work

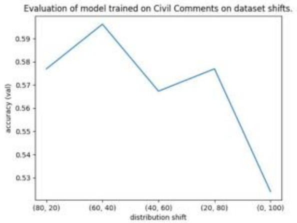
Large language models have huge number of parameters and **fine-tuning all the layers cause the model to overfit and not to perform optimally** when evaluated on data from a different distribution. Miller et al. and Lazadirou et al. show that fine-tuning all model parameters do not generalize well for data from distributional shifts, in spite of increasing the model size.

Surgical fine-tuning paper shows that for CNN, fine-tuning just the early layers and freezing the other layers of the model result in an improvement in performance in spite of distributional shifts for certain settings. Following this idea, we propose Coefficient-Based Variable Fine-Tuning, **combining the parameters of a fully fine-tuned model and a model with frozen layers in the ratio α and 1-α.**

I – Performance deterioration of distribution shift

We load a base BERT model and tune all the layers for a specific task A for an extensive period. We then create **datasets** with **various level of distribution shifts** : $p_{A_\delta} = (1-\delta) * p_A + \delta * p_B, \delta \in [0,1]$), and **evaluate how the (fixed) trained model performs** for increasing value of δ:

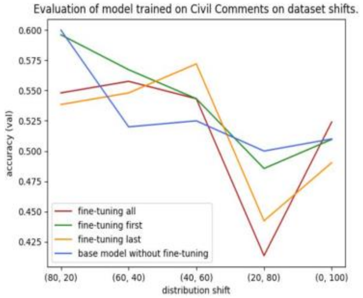


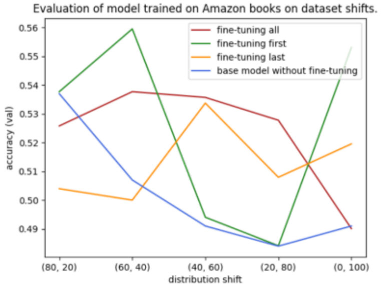


II – Classical surgical fine tuning approaches

We then experiment with classical fine-tuning methods proposed in the CNN surgical fine-tuning paper. We conduct the experiments by fine-tuning the first two, middle two and the last two layers of BERT encoder.

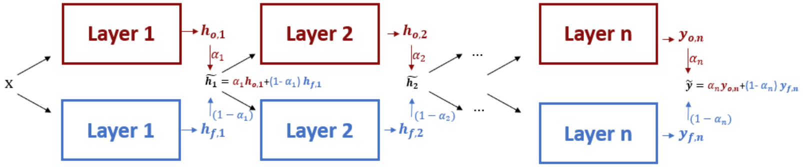
We compare the results for each shift and each fine-tuning method to the baselines of both hate speech detection and sentiment analysis tasks (part I.). On those two tasks, **fine-tuning all and first layers seems to perform better**. Logically, **fine-tuning outperforms the base model** which was not updated to account for distribution shift.



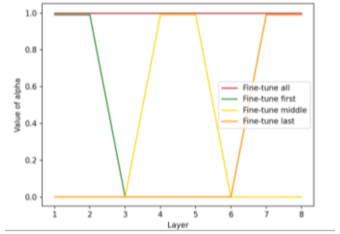


III.A – Introducing Coefficient-Based Variable Fine Tuning (CBV FT)

The sections I and II focused on fine-tuning predefined layers. To go further, **we introduce a concept** never explored to our knowledge: **Coefficient-Based Variable Fine Tuning**.



This represents an **extension of classical surgical fine tuning**.
For instance, first-layers fine tuning corresponds to $\alpha_1 = \alpha_2 = 1$ and $\alpha_3 = \ldots = \alpha_n = 0$:



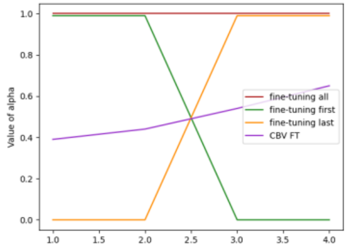
However here, the **model** has the possibility to be **way more flexible** in its choice of layers to fine tune.

III.B – Results of Coefficient-Based Variable Fine Tuning (CBV FT)

We compare our fine-tuning method and the classical surgical fine tuning approaches.

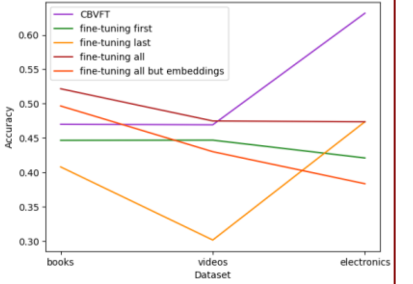
To do so, we fine-tune and evaluate on 3 datasets:

- Amazon books
- Amazon videos
- Amazon electronics



Our method is able to:

- **Outperform classical surgical fine tuning** approaches (incl. the all layers fine-tuning ones) on average
- **Yield consistent results** on the 3 datasets
- However, it **fails to outperform all layers** fine-tuning **on our biggest dataset, books**



Some room for improvement remains as:

- The process of **optimizing the alphas is particularly slow** for usual learning rate values
- **We don't fine-tune the embeddings layer** as the algorithm otherwise only focus on these part of the network

Conclusion and future work

We propose Coefficient-Based Variable Fine-Tuning method, a linear combination of a model with fine-tuned parameters and a model with frozen parameters. **The model performs better than other classical surgical fine-tuning methods** like fine-tuning first, middle and last layers of the model. **In the future**, we would like to further research which layer the algorithm prefers by **constraining α** and conduct more experiments to **confirm the robustness of our method** in different environments.

On a longer time horizon, it would be interesting to investigate the following ideas:

- Introduce an optimization-based approach to determine the optimal values of **α**
- Look into the interpretability behind the layers chosen by the algorithm
- Quantify the robustness of CBVFT by investigating even larger distribution shifts