Fine-Tuning is not (Always) Overfitting Artifacts*

Jeremie Bogaert[1], Emmanuel Jean[2],
Cyril de Bodt[1] & François-Xavier Standaert[1]

[1] ICTEAM/UCLouvain - Belgium. [2] Multitel - Belgium

Abstract. Since their release, transformers, and in particular fine-tuned transformers are widely used for text-related classification tasks. However, only a few studies try to understand how fine-tuning actually works and existing alternatives, such as feature-based transformers, are often overlooked. In this work, we study a French transformer model, CamemBERT, to compare the fine-tuned and feature-based approaches in terms of their performances, interpretability and embedding space. We observe that while fine-tuning has a limited impact on performances in our case study, it significantly affects the intepretability (by better isolating words that are intuitively connected to the classification task) and embedding space (by summarizing the majority of the relevant information into a fewer dimensions) of the results. We conclude by highlighting open questions regarding the generalization potential of fine-tuned embeddings.

1 Introduction

Since their release a few years ago, transformers [11] have been of interest for the whole Natural Language Processing (NLP) community. Models such as BERT [6] and its variants (RoBERTa [12], AlBERT [13], ...) achieve state-of-the-art results on numerous classification tasks and datasets. These models are typically first pre-trained on big amounts of data in order to learn how to transform texts into embeddings, and then adapted to an end task. To do so, [6] presented two different approaches: the fine-tuning approach, *where a simple classification layer is added to the pre-trained model and all parameters are jointly fine-tuned on a down stream task*, and the feature-based approach, often referred to as the frozen one, *where fixed features are extracted from the pre-trained model and fed to another model*. While the former is usually preferred, and may marginally improve the model performances [8], both models reach similar results on many classification tasks [5, 9]. However, as many other complex models, neural networks are prone to learn statistical quirks in data, which can results in adopting shallow heuristics that succeed for most training examples, instead of learning underlying generalizations that they are intended to capture [1, 4], which we denote as overfitting artifacts. It raises the question of the impact of fine-tuning in this respect, which this paper tackles by assessing different properties of fine-tuned vs. frozen models based on two French text classification tasks using CamemBERT [14]. We first observe quantitatively that, for both tasks, both approaches reach similar performances and none of them overfits

* Work supported by the Service Public de Wallonie Recherche, grant n°2010235-ARIAC by DIGITALWALLONIA4.AI. FXS is Senior Research Associate of the F.R.S.-FNRS.

data. We then use recent explanation methods [7] to identify the most important words for both frozen and fine-tuned models. We find that fine-tuned models rely on more "intuitively-relevant" words for the considered classification tasks than the frozen ones, leading to more interpretable results. We also visualize the impact of fine-tuning on the data representation through nonlinear dimension reduction methods [3], observing that the embedding structure of the fine-tuned models leads to much more separated clusters and concentrated information across fewer dimensions. As the fine-tuned approach outperforms the frozen one in all regards for our case studies, we conclude by discussing the possible loss of generality that fine-tuned embeddings can lead to (compared to frozen ones). For this purpose, we evaluate the performance loss in the context of a task A when using a fine-tuned embedding obtained for a different task B. We observe that the loss in performances is negligible whereas the impact on the embedding space is more significant, raising the challenge of better understanding such a context from the interpretability viewpoint as an interesting open problem.

2 Tasks, datasets, and experiment design

We use two datasets to conduct our experiments. The first one is a home-made dataset, called the InfOpinion dataset (`https://github.com/jebogaert/Fine-tuning_analysis`). It contains 10,000 news extracted from the "Radio-Télévision Belge de la communauté Française" (RTBF) corpus and it was built to train and evaluate a classification model distinguishing between texts from the journalistic *opinion* genre (such as editorials, commentaries, reviews), which are considered to be subjective, and texts belonging to the *information* genre (press agency dispatches, news articles), meant to be more objective. This binary categorization relies solely on the articles' annotation by the RTBF as either *opinion* or *information*. The second one is the Allocine dataset (`https://huggingface.co/datasets/allocine`). It is composed of 200,000 movie reviews that can be positive (50.02%) or negative (49.98%). Both datasets are split in 3 parts: a training set (80%), a validation set (10%) and a test set (10%). For both datasets, the task is to predict the binary category of a given text.

We start our experiments by training both the fine-tuned and frozen models. In the first case, we fine-tune the model on the training set during 2 epochs, as presented in [6]. In the second case, we use the base version of CamemBERT [14] as our transformer model. We first probe the internal representation of each text from the training set at the last layer of the pre-trained model, just before the RoBERTa classification head. As in [5], we then use the first token's representation as a text embedding and train a RoBERTa classification head during 20 epochs on top of these embeddings. The model accuracy is evaluated at each epoch on the validation set and, at the end of the training process, on the test set. Once both models are trained, we use the Layer-wise Relevance Propagation (LRP) method [7] to collect word-level explanations for every text. We compute the mean relevance across all texts for each word and compare the top words for our two models. We finally study the latent

representations of both the frozen and the fine-tuned models. We use multi-scale t-SNE [3] to visualize the text embeddings (which are in 798-D) in 2-D, while also analyzing the concentration of information through Principal Component Analysis (PCA). We conclude by initiating a discussion on the loss of generality induced by fine-tuning, by using text embeddings fine-tuned on one classification task for another task. The code of our experiments is available from `https://github.com/jebogaert/Fine-tuning_analysis`.

3 Results and discussion

3.1 Model performances

Table 1 reports the accuracies of both models on both datasets. For the Allocine dataset, the fine-tuned model is slightly (but statistically significantly, $p \ll 0.01$) better, confirming previous works like [5, 9]. For the Infopinion dataset, the differences are not even statistically significant ($p = 0.34$). We also note that the performances on the test set are similar to the one on the training set (given in parentheses), suggesting that no overfitting of the data occurs.

	Accuracy fine-tuned	Accuracy frozen
Allocine	97.07 (97.49)	94.31 (93.77)
InfOpinion	95.60 (96.36)	96.20 (97.10)

Table 1: Classification accuracy on the Allocine and the InfOpinion datasets.

3.2 Discriminant words

We continue our study by applying the LRP explanation method [7] to obtain word-level explanations for each text in the dataset. We then compute the mean relevance across all appearances of each word to get a list of the most discriminant words according to the models (independent of the context surrounding these words). In order to limit noise, we only consider words appearing at least 100 times in the datasets. We finally highlight the difference between the two models by ranking the words according to difference in mean relevance between the frozen and fine-tuned models. Below are the most different words for the Allocine dataset (movie reviews)[1], translated into English:

- **More important for fine-tuned**: bouleversant (upsetting), régal (threat), adoré (liked), émouvant (touching), plat (flat), bijou (jewel), magnifique (magnificent), merveille (wonder), poignant (touching), éviter (to avoid).
- **More important for frozen**: regardez (watch), vieilli (aged), revu (already seen), penser (to think), Besson, sauce, proche (close), mourir (die), visuellement (visually), d' (of), dessin (drawing), animation.

[1] Similar trends hold with the InfOpinion dataset, not displayed due to space limits.

Interestingly, and despite model interpretation is admittedly harder to evaluate on quantitative bases, we observe that the improved performances of the fine-tuned model are not based on data artifacts but that, instead, its top words well reflect the considered task (i.e., they are connected to the appreciation of the movies). This is in contrast with the frozen approach, where the list of most relevant words rather relate to the movie style (which, without additional context, is not directly related to its appreciation). This suggests that despite similar performances, the frozen model is not based on the same easily interpretable features as the fine-tuned one. We next try to give a complementary viewpoint to this observation by analyzing the embedding spaces of the two models.

3.3 Text embeddings visualisation

The next figures show projections of the text internal embeddings for the frozen and fine-tuned models on both datasets, computed with multi-scale t-SNE [3].

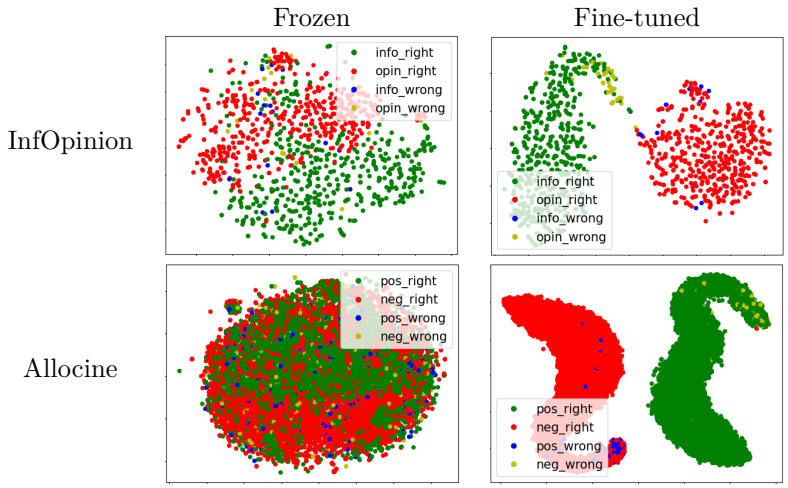


Table 2: t-SNE visualisations of the model's embedding space.

These results suggest that the decision boundaries (in two dimensions) are clearer and that clusters with distinct labels are more separated using the fine-tuned internal text representations than the frozen ones, supporting the previous results of [8]. In addition to pushing clusters with different labels far away from each other, the fine-tuning also concentrates the information across fewer dimensions. Through a PCA of the internal text embeddings, we additionally computed the % of retained variance with respect to the number of Principal Components (PC). It turns out that the 1^{st} PC captures more than 95% of the variance for the fine-tuned internal text embeddings, whereas more than 200 PCs are necessary to preserve the same amount of variance in the frozen case.

4 Conclusion and further works

This works highlights the specificities of the fine-tuned vs. frozen models on two different data sets and classification tasks. While both methods can reach similar accuracies and none of the methods overfits the training data in our case studies, fine-tuned models appear to be more interpretable and to concentrate the information of their text embeddings in fewer dimensions, despite being more complex in terms of number of updated parameters compared to frozen ones.

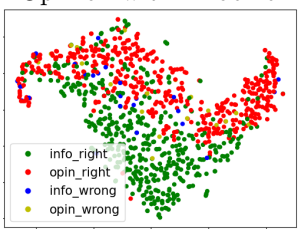
These promising results for the fine-tuned approach naturally raise the question of the possible loss of generality they could come up with. As an opening experiment towards discussing this question, we finally performed a cross fine-tuning experiment, where a model is first fine-tuned on a task, and its text embeddings are then used to perform another task. A similar experiment was presented in [8] but focused on different classification tasks with the same dataset. We extend it to different tasks based on different datasets. The table below shows the accuracy obtained through this process:

	F-t Allocine	F-t InfOpinion	Frozen
Allocine	97.07	94.32	94.31
InfOpinion	95.50	95.60	96.20

Table 3: Classification accuracy on the cross-task finetuning experiment.

It is noteworthy that, even when using a different dataset, fine-tuning reaches very competitive accuracies, suggesting that it can preserve most of the relevant information of the frozen internal text embeddings. The figures below then depict the internal text embeddings of each dataset using cross-task fine-tuning:

InfOpinion with Allocine f-t



Allocine with InfOpinion f-t

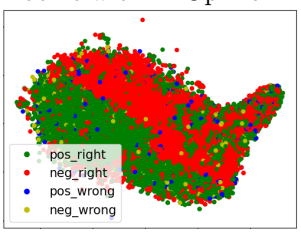


Table 4: t-SNE of the model's embedding space with cross-task finetuning.

This last figure is more contrasted since it suggests that only a part of the "information concentration" observed in Section 3.3 is maintained in a cross-task setting. This is also confirmed by the number of PC needed to capture 95% of the variance, which is worth 37 for the left figure (InfOpinion with Allocine fine-tuning) and 89 for the right one (Allocine with InfOpinion fine-tuning).

How much this last observation is due to the good/bad generalization potential of fine-tuned models or to the differences/similarities of the datasets and

tasks considered in our experiments is an interesting open question. As a first step towards answering it, it would be interesting to quantify whether performance drops accumulate after multiple cross-trainings, and whether the intermediate "information concentration" of these last figures translates into reduced interpretability, for example using the techniques used in Section 3.2.

Besides, our case studies rely on the Camembert model for the French language. How our conclusions can be reproduced for other models and languages is therefore another interesting direction for further investigations.

References

[1] R. Thomas McCoy et al., Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, in CoRR, abs/1902.01007, 2019, arXiv

[2] A. Merchant et al., What Happens To BERT Embeddings During Fine-tuning?, in Proceedings of the Third BlackboxNLP Workshop (EMNLP), p 33–44, November 2020, published by ACL

[3] C. de Bodt et al., Fast Multiscale Neighbor Embedding, in IEEE Trans. Neural Netw. Learn. Syst., p 1-15, 2020

[4] O. Kovaleva et al., Revealing the Dark Secrets of BERT, in Proceedings of the 2019 EMNLP-IJCNLP, p 4364–4373, November 2019, published by ACL

[5] M. E. Peters et al., To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks, in Proceedings of the 4th Workshop on Representation Learning for NLP, p 7–14, August 2019, published by ACL

[6] J. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proceedings of the 2019 NAACL-HLT , Volume 1 p 4171–4186, June 2019, published by ACL

[7] H. Chefer et al., Transformer Interpretability Beyond Attention Visualization, in CVPR 2021, p 782–791, published by Computer Vision Foundation / IEEE

[8] Y. Zhou and V. Srikumar, A Closer Look at How Fine-tuning Changes BERT, in Proceedings of the 60th Annual Meeting of the ACL, Volume 1 p 1046–1061, May 2022, published by ACL

[9] N. F. Liu et al., Linguistic Knowledge and Transferability of Contextual Representations, in Proceedings of the 2019 NAACL-HLT, Volume 1 p 1073–1094, 2019, published by ACL

[10] S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, in Proceedings of the 2018 NAACL-HLT, Volume 2 p 107–112, June 2018, published by ACL

[11] A. Vaswani et al., Attention is All you Need, in NEURIPS 2017, p 5998–6008, 2017

[12] Y.Liu et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, in CoRR 2019, arXiv

[13] Z. Lan et al., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, in ICLR 2020, April 2020, published by OpenReview.net

[14] L. Martin et al., CamemBERT: a Tasty French Language Model, in Proceedings of the 58th Annual Meeting of the ACL, p 7203–7219, July 2020, published by ACL