Transferability and Stability of Learning With Limited Labelled Data in Multilingual Text Domain

Branislav Pecher[1,2*]

[1]Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[2]Kempelen Institute of Intelligent Technologies, Mlynske nivy 5, Bratislava, Slovakia
branislav.pecher@kinit.sk

Abstract

Using the learning with limited labelled data approaches to improve performance in multilingual domains, where small amount of labels are spread spread across languages and tasks, requires knowing the transferability of these approaches to new datasets and tasks. However, the lower data availability makes the learning with limited labelled data unstable, resulting in randomness invalidating the investigation, when it is not taken into consideration. Nevertheless, previous studies that perform benchmarking and investigation of such approaches mostly ignore the effects of randomness. In our work, we want to remedy this by investigating the stability and transferability, for effective use in the multilingual domains with specific characteristics.

1 Introduction

Learning with limited labelled data approaches, such as meta-learning, transfer learning or weakly supervised learning, are used to achieve high performance in case of lack of labels. Applying these approaches to domains, characterized by a lack of labels spread across various languages and tasks can be challenging. To effectively do this, we need to investigate how affected these approaches are by the changes in datasets and tasks and how they can perform in this domain, in comparison to each other. In other words, we need to know the transferability of these approaches to previously unseen features and tasks.

To effectively investigate the transferability of these approaches, we need to be aware of their training stability and take into account the various sources of randomness that influence it. The lower data availability, coupled with lower data quality, makes the training process more unstable and prone to the effects of uncontrolled randomness [Ye *et al.*, 2021; Lu *et al.*, 2021; Zheng *et al.*, 2021; Bragg *et al.*, 2021]. Even small changes in data and parameters can lead to massive changes in performance, such as changing order of samples leading from state-of-the-art results to simple random predictions [Lu *et al.*, 2021]. Without taking the stability into consideration, the replicability of the approaches suffer leading to biased comparisons. Investigating other properties will lead to results that are influenced by the randomness, making them biased and unusable. Despite this, previous works that compare the approaches on benchmark datasets or investigate other properties of these approaches still mostly ignore the training stability and its influence.

We want to address this in our work and investigate the transferability and stability of learning with limited labelled data, when applied to the domain of multilingual text data, with the available labels spread across languages and tasks.

2 Focus of Our Study

In our work, we first focus on the *stability* of the training process of the learning with limited labelled data approaches. We define stability as property that indicates what influence the *small scale, random* changes in the input *data* and *parameters* have on the outputs of the model. The small scale changes are an effect of inherent randomness of the training process, such as different splits of data, or random initialisation, which are limited in their effect. Therefore they cannot be easily controlled for, but must be observed and taken into consideration. This results in stability affecting all the other properties. For this reason, we focus on it as the first property we investigate, to make the investigation of other properties more unbiased.

To investigate the stability, we first identify the points of the training process, where some kind of randomness is introduced. We call them *randomness points* and differentiate between two groups of them, the ones related to randomness in data (such as data splits, order of data or other noise), and the ones related to model-specific parameters (such as random initialisation and minor changes to specific hyperparameters). By observing the behaviour of the approaches around these randomness points, we can identify which effects of randomness have the most influence on them. This will allow us to better address them, improving the replicability of the approaches.

Besides improving the replicability and making the investigation of other properties, and the comparisons between approaches, more unbiased, knowing the source of instability leads to more efficient use of the approaches in practice. We can identify the parts of training process on which to focus during the optimisation, while also identifying the parts that

*Supervised by: **Maria Bielikova** and **Ivan Srba**, Kempelen Institute of Intelligent Technologies

have almost no impact on the variability of results, allowing us to disregard them without taking methodologically incorrect shortcuts.

After investigating the training stability of the learning with limited labelled data, we will focus on their effective use in the multilingual text domain, with labels spread across languages and tasks. In order to improve their effective use, we will investigate the *transferability* of these approaches to new, previously unseen datasets and tasks. This leads us to our definition of transferability, which indicates how the outputs of the model change, when different, but related *dataset* and *task definition* is used instead. As the dissimilarity between datasets and tasks, resulting in negative transfer, is still an open problem for the approaches [Hospedales *et al.*, 2020], even though many of the approaches are designed to use these datasets and tasks as a source of additional knowledge, investigating the transferability capabilities is important for the effective use of the approaches.

When investigating the transferability, we do not expect that one approach will be the best for each task and domain. In addition, due to sheer amount of possibilities, we do not expect to be able to find any pattern for determining which approach can better deal with the dissimilarity, softening the requirements on the similarity between tasks and datasets (although we expect some approaches to perform better overall than others). Therefore, we plan to investigate the transferability similarly to the stability. We will first try to identify points that affect the transferability (which we will call transfer points). Afterwards, we will investigate the behaviour of the approaches around these points. This will allow us to determine what points to focus on and how to control them in order to achieve better transferability of the approaches, making them better suited for the specifics of our domain.

We plan to investigate these properties on a selection of approaches for learning with limited labelled data. We focus on the approaches that are used today and perform sufficiently well when combined with deep learning. The focus will be both on the approaches that utilize an existing larger set of unlabelled data, as well as on the ones that make use of related datasets. As their characteristic may make them more usable in different domains and on other tasks, focusing on both groups is important. In the first phases of our work, the focus will be on meta-learning, from the second group, and the semi-supervised learning, from the first group. After sufficiently having explored these two sets of approaches, we will expand to weakly supervised learning and transfer learning. Due to there being large number of approaches in each group, we will first select representative approaches for the most important philosophies and investigate the property on them (such as using representative approaches for optimisation and metric based meta-learning).

As there may be interactions between the different points and their values, we will perform the investigation in a hierarchical manner, changing the values for all of them at the same time and not using fixed values for any of the investigated parameters. This poses a challenge of how to select the different randomness and transfer points to investigate. Due to the number of possible combinations being in tens of thousand, it is not feasible to investigate all of them at the same time. Furthermore, the points that are not currently investigated will need to have their values fixed to prevent any noise from them. Finally, by including the interactions into the experiments, we will need to take them into account when evaluating them. Therefore, we will make use of more advanced tests and models during the evaluation, such as Linear Mixed Models, in a similar fashion to [Boquet *et al.*, 2019].

2.1 Preliminary Results

We have conducted a preliminary investigation of the stability for the optimisation based meta-learning approaches on a simple text sentiment analysis task. First, we have identified a set of randomness points. From this set, we have investigated the effect of different *data splits*, by training and evaluating the selected approaches on different train-test splits, and the effect of the random *model initialisation*, *task sampling* and *sample order* in one, by running the training multiple times without fixing the random seeds. We have investigated three basic meta-learning models: Model Agnostic Meta-Learning (MAML) [Finn *et al.*, 2017], its first order version (FOMAML) [Finn *et al.*, 2017] and Reptile [Nichol *et al.*, 2018]. In addition, we also varied the amount of labels available during training, to determine how the models behave with decreasing number of labels. This study has already shown us differences in the individual models, with Reptile being susceptible to the setup of hyperparameters, but showing almost no variance contributed by the repeated runs, when the number of labels is low and the hyperparameters are sufficiently optimised. On the other hand, MAML shows constant good stability overall, only being outperformed by Reptile in a single case.

We have also applied the learning with limited labelled approaches to the domain we are dealing with. We have used a simple transfer learning approach, where a model is first trained on dataset with larger amount of labels and then fine-tuned on dataset with low number of labels, and the Reptile meta-learning approach on a stance detection task. The results of this experiment already show promising results, as the performance on both datasets was significantly improved. However, we have noticed a significant instability in the results, in both the transfer and meta-learning approach. This gives us further incentive to investigate the training stability before focusing more in-depth on the transferability property.

References

[Boquet *et al.*, 2019] Thomas Boquet, Laure Delisle, Denis Kochetkov, Nathan Schucher, Boris N Oreshkin, and Julien Cornebise. Reproducibility and stability analysis in metric-based few-shot learning. *RML@ ICLR*, 3, 2019.

[Bragg *et al.*, 2021] Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. FLEX: Unifying evaluation for few-shot NLP. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International*

Conference on Machine Learning - Volume 70, ICML'17, page 1126–1135. JMLR.org, 2017.

[Hospedales *et al.*, 2020] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[Lu *et al.*, 2021] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[Nichol *et al.*, 2018] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018.

[Ye *et al.*, 2021] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *CoRR*, abs/2104.08835, 2021.

[Zheng *et al.*, 2021] Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Jian Li, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. Fewnlu: Benchmarking state-of-the-art methods for few-shot natural language understanding. *arXiv preprint arXiv:2109.12742*, 2021.