FREE LUNCH FOR FEW-SHOT LEARNING: DISTRIBUTION CALIBRATION

COMP6248: DEEP LEARNING REPRODUCIBILITY REPORT

Claudiu Ionut Dan
University of Southampton
cid1u17@soton.ac.uk
29314127

Jingran Ding
University of Southampton
jrd1g20@soton.ac.uk
32407556

Mihnea Ionut Onet
University of Southampton
mio1g17@soton.ac.uk
29556732

1 INTRODUCTION AND TARGET QUESTION

Few shot learning represents a class of machine learning problems that have to work and learn using datasets that are limited in size, dimensionality or diversity. This approach is most encountered in the field of computer vision and natural language processing where algorithms are expected to gain a high level understanding of the data from only a few number of feature-dense examples.

The work (Yang et al. (2021)) addressed in this report proposes a simple method of recalibration of the skewed training data distribution by transferring information from the dense regions of the data space to the sparse ones, with the only assumption being that the distribution of these transferable features draws close to a Gaussian distribution (Salakhutdinov et al. (2012)).

Additionally, it is claimed that this approach outperforms the previous state of the art by about 5% in terms of accuracy performance and that it can be applied to any classifier while also being agnostic to the algorithm used to extract the training features. This paper will challenge these claims by attempting to reproduce the original experiments while also exploring the performance of the distribution recalibration on additional environments, including different machine learning algorithms, datasets or combinations of training parameters. All the results will be reported in the following sections beginning with the experimental methodology and details about the implementation.

2 IMPLEMENTATION DETAILS AND EXPERIMENTAL METHODOLOGY

The main setting of the paper is a few-shot classification problem on the CUB (Welinder et al. (2010)), miniImageNet (Ravi & Larochelle (2017)) and tiredImageNet (Ren et al. (2018)) datasets. The experiments use an N-way-K-shot methodology where N classes are sampled from the set of underrepresented data and only K labeled examples are provided for each class. These are then split into a support set, used for training, and a query set, composed of q test cases, used for testing. It is important to note that the classification is done over the feature space constructed from the penultimate layer of the WideResNet and, according to the original paper, this model can be swapped with any other suitable classification architecture.

Since one of the main assumptions of this approach is that the feature space can be approximated using a Gaussian distribution, the authors use Tukey's Ladder of Powers Transformations in order to make the real distribution more Gaussian-like For the calibration of the feature distribution, the mean μ and covarinace Σ are computed based on the extracted feature vectors and are then used to transfer some of the statistical information to the underrepresented classes. This transfer is done using the Euclidean distance to the closest k classes to a specific feature sample $\bar{x}$ from the support set while also taking α into consideration which is the degree of dispersion of the features sampled from the adjusted distribution (see Yang et al. (2021) for the comprehensive mathematical description of this process).

3 ANALYSIS OF RESULTS AND FINDINGS

To reproduce the results, we have reimplemented the algorithm described in the original paper which can be found on GitHub ("Free Lunch For Few Shot Learning" repository: https://github.com/COMP6248-Reproducability-Challenge/Free-Lunch). However, for other parts which do not depend on what is described in the paper, such as preparing the data, we have used the already existing code of the authors, since that is not directly linked to what is shown in the paper

and is more of a pre-processing task to arrange the data in a convenient form to run the algorithm on. On top of that, we have added our own code to reproduce the experiments and the visualisations. It should be noted that our rewritten code is two times faster than the original code.

3.1 TUKEY TRANSFORMATION

In the paper, it is stated that Tukey's Ladder of Powers transformation is used to reduce skewness and make the distributions more Gaussian-like. Furthermore, it is stated the algorithm is agnostic to any feature extractor. However, if features contain zero values, then Tukey's transformation will not work for some of the parameters shown in their plots.

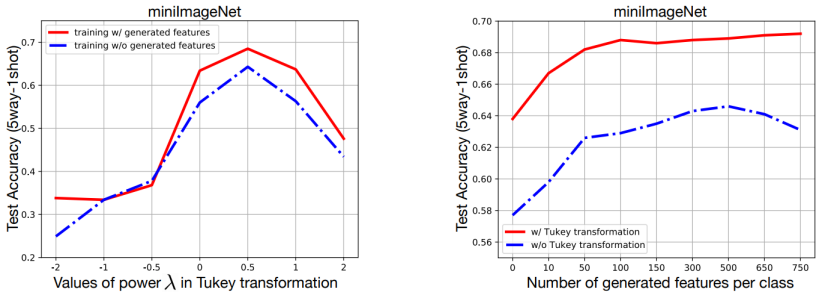

Figure 1: Original experiments for various lambdas and number of generated features

For $\lambda \neq 0$, Tukey's transformation is $x_{new} = x^{\lambda}$. However, if x contains zero values and λ is -2 (as shown in Figure 1), then $x_{new} = \infty$. This problem also appears when $\lambda = 0$ and x contains zero values because then, according to Tukey's transformation, $x_{new} = log(x_i) = log(0) = -\infty$. The data used in the paper contains zero values and the previous code-breaking scenarios occur when trying to reproduce the plot (Figure 1 left). Therefore, we were not able to reproduce the plot for the variation of λ, since the paper does not mention how to handle features with zero values.

3.2 DISTRIBUTION ESTIMATION VISUALISATION

The paper uses t-SNE to reduce the sampled features to 2D. However, t-SNE varies greatly depending on the initialisation, and the parameters used have a high impact on the result. The paper does not state the parameters of t-SNE used which means that we were not able to reproduce some parts of the visualisations.

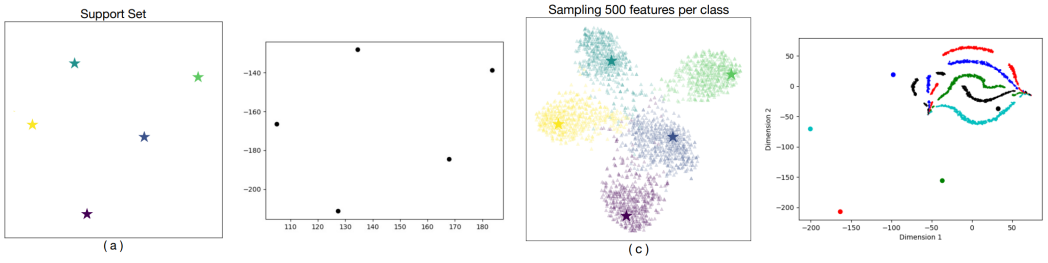

Figure 2: Original support set vs reproduced support set (first two), and original generated features vs reproduced generated features (last two)

Figure 2 shows that our reproduced support set features are similar to the original. We used the standard parameters for scikit-learn's t-SNE. However, when trying to visualise the sampled features, the results are not similar anymore. Figure 2 shows how the samples look in 2D using the standard t-SNE parameters. Obviously, that is very different from what is shown in the paper and that is most likely because the authors ran t-SNE under certain parameters which are not mentioned.

3.3 EXPLORING TRAINING PARAMETERS AND THE NUMBER OF GENERATED FEATURES

In order to fully validate the claims of the original paper, we also tried to test the performance of the data recalibration for a wide range of values for each major parameter in the training pipeline. We explored the evolution of accuracy for various α, λ and k values and compared them with the results presented by the authors. Our results can be seen in Figure 3. The general behaviour that we observed was similar to the expected one, however there were slight differences in terms of the top accuracy values achieved. One important aspect to consider is the granularity of range of values explored for parameters, in particular for λ where the authors decided that the most optimal value for the 5way-1shot scenario would be 0.5 given the evolution presented in their paper. However, for our experiments, upon increasing the granularity of the range of values, we found that the most optimal value for λ would instead be 0.6.

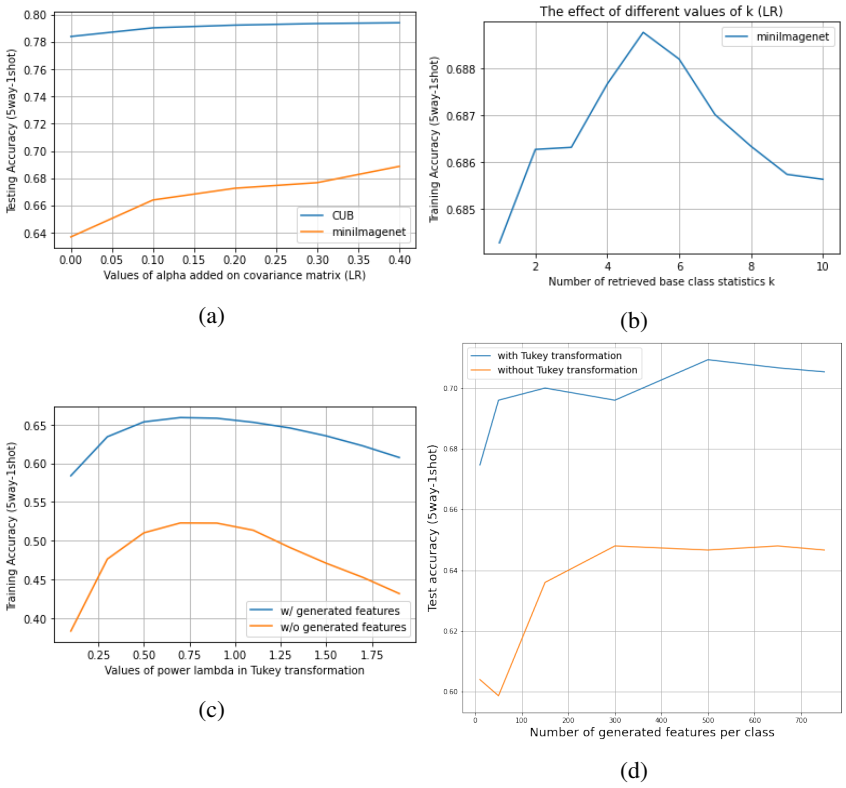

Figure 3: Reproduced plots for varying α (a), k (b), λ (c), and number of generated features (d)

3.4 DIFFERENT CLASSIFIERS

In their experiments, the authors claim that the recalibration can be paired with any classifier without any special tuning needed. Given that the original experiments used a simple logistic classifier, in our experiments we also tried three additional classifiers, namely KNN classifier and Random Forest Classifier and SVC, the last one also being mentioned in the paper, but without showing any explicit results for it. The implementation for the algorithms can be used directly from scikit-learn with the training parameters being the default values in order to remove any tuning bias. For the other variables, we chose the 5ways-1shot setting on the miniImagenet dataset. It is worth noting that the original experiments averaged the accuracy results over 10 thousand tasks for the miniImagenet dataset and 100 thousand tasks for the CUB dataset but due to time constraints we changed the number to 1000 in order to be able to run a feasible amount of experiments. The results of these experiments can be seen in Table 1.

The results we got for our experiments draw close the the ones claimed by the authors. Moreover the accuracy values achieved for the other classifiers that were not mentioned are fairly high too.

Classifier	1 shot (mini)	5 shot (mini)	1 shot (CUB)	5 shot (CUB)
Logistic Regression	69.12 ± 0.8	82.22 ± 0.6	80.12 ± 0.5	90.02 ± 0.4
SVC	69.42 ± 0.4	81.02 ± 0.4	79.42 ± 0.2	80.33 ± 0.7
KNN	66.27 ± 0.6	75.01 ± 0.4	75.01 ± 0.3	74.24 ± 0.6
Random Forest	63.78 ± 0.9	77.33 ± 0.7	77.91 ± 0.4	9.63 ± 0.5

Table 1: Testing accuracy values for different classifiers trained on the miniImagenet dataset for 5way-1shot experiments. All the values are gathered for 5way-Kshots with the k being changed for each run.

Additionally, the same clear difference between the 5way-1shot and 5way-5shot experiments was observed in our case too since the number of available examples per class is always important in a few-shot learning scenario. The scale of this difference also remains similar which is a positive aspect for the reproducibility of the paper.

3.5 General Observations

As for the general observations about the overall quality of the paper, there are a number of negative aspects that could have been improved. One of them is the wrong scaling present in the plots for the training parameters (more specifically the plots for the evolution of λ, α and number of generated features). Even though the accuracy values are represented correctly, their evolution between the sampled points is skewed due to the varying interval on the X-axis. On top of that, the claim that this distribution recalibration approach can be paired with any classifier could be supported by more evidence (having a wider range of classifiers with different complexities would have helped). Moreover, there is another popular dataset for the few-shot classification problem,namely Meta-Dataset, which could have also been used to test the performance of the approach.

Finally, in terms of the time and computational resources needed to replicate the experiments presented by the authors, the scale of the data and number of iterations per run were fairly large which translated into large running periods. However, we followed and recreated the majority of the tasks at a smaller scale and managed to achieve a similar performance and evolution for both 5-way-1shot and 5way-5shot scenarios.

4 Conclusion

In conclusion, the results we got for the "Free Lunch for Few-shot Learning: Distribution Calibration" paper draw close to the ones claimed by the original authors in terms of the accuracy performance and the majority of the experimental parameters. Our analysis shows a stable and significant improvement for the proposed few-shot classification setting given the assumptions of a Gaussian-like feature distribution holds. On the other hand, there were quite a few aspects in the paper that were defined poorly or not at all with the most important ones being the λ domain in the Tukey transform and the t-SNE parameter values. However, even with these negative aspects, the reproducibility level and the overall quality of the paper remain really high and the works described in it represent a real contribution to the general problem of few-shot classification.

References

S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification, 2018.

Ruslan Salakhutdinov, Joshua Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver (eds.), *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pp. 195–206, Bellevue, Washington, USA, 02 Jul 2012. PMLR.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations (ICLR)*, 2021.