LAST LAYER RE-TRAINING: A STUDY ON REPRODUCABILITY

Ajayesh Saini, Damian Smith, Rory Wilson, Quan Yuan [*]
School of Electronics & Computer Science
Southampton University, UK
{as11n23, ds5n23, rw7g20, qy1g18}@soton.ac.uk

ABSTRACT

Reproducability is vital to scientific research. We attempt to reproduce several claims regarding Last Layer Re-Training by working from an ICLR 2023 paper, and by using published code. While similar improvements are recreated, errors and imprecision in the paper mean that we are unable to confirm the specific results reported.

1 INTRODUCTION

Our team attempted to reproduce the following claims of LAST LAYER RE-TRAINING IS SUFFICIENT FOR ROBUSTNESS TO SPURIOUS CORRELATIONS by Kirichenko et al. (2022) (hereafter: THE PAPER).

- Although neural network classifiers can largely rely on spurious features, they ”still often learn core features associated with the desired attributes of the data”. THE PAPER Sec 4.1
- ”Simple last layer retraining can match or outperform state-of-the-art approaches on spurious correlation benchmarks, but with profoundly lower complexity and computational expenses”. THE PAPER Sec 6

Half of our team used THE PAPER ’s provided code: D. Smith - Sec 6, Q. Yuan - Sec 4.1. The other half attempted to write new code to reproduce the results using only the details in THE PAPER: A. Saini - Sec 6, R. Wilson - Sec 4.1

The principle claim of THE PAPER is to increase the worst group accuracy through DFR.

2 EXPERIMENTATION WITH SUPPLIED CODE

2.1 SEC 4.1: FEATURE LEARNING ON WATERBIRDS DATA

We regenerated the Waterbirds datasets used to generate Table 1 in THE PAPER. The supplied code was then used to attempt to reproduce the baseline results reported there.

Our results (Table 1) showed a similar pattern to THE PAPER, but all our accuracies were slightly lower than theirs. We ran a single 50 Epoch training for each combination. In Table 3, our estimate for baseline standard deviation was 5%, so our results are compatible with THE PAPER’s observation that ”ERM classifiers trained on Waterbirds with Original and FG-Only images achieve similar FG-Only accuracy.”

However, THE PAPER was insufficiently detailed for us to recreate the Balanced dataset (see Sec 3.1).

2.2 SEC 6: FEATURE REWEIGHTING IMPROVES ROBUSTNESS

We tried reproducing the Baseline and Deep Feature Reweighting (DFR) results reported in Table 2 of THE PAPER for the Waterbirds dataset. The authors’ code was downloaded

[*] Alphabetical order by last name

Table 1: ERM Classifiers trained on regenerated datasets (THE PAPER, Table 1) & our results

	TEST DATA (WORST ACC)			
	(PAPER)		**(OURS)**	
TRAIN DATA	**ORIGINAL**	**FG ONLY**	**ORIGINAL**	**FG ONLY**
Original (95%)	73.8%	93.7%	69%	92%
Original (100%)	38.4%	94%	33%	90%
FG-Only	75.2%	95.5%	70%	94%
Avg Difference (Paper-Ours)	4.2%	2.5%		

from `github.com/PolinaKirichenko/deep_feature_reweighting` and branched to allow our team to make minor changes at `https://github.com/DHLSmith/DPBIA_deep_feature_reweighting`. Following instructions in the README.md, the tarball of images was downloaded from `https://nlp.stanford.edu/data/dro/waterbird_complete95_forest2water2.tar.gz` The README.md file was reasonably detailed and included command line arguments to run the tests described.

2.2.1 MAIN OBSERVATIONS

Reviewing the code in train_classifier.py and the parameters from README.md, the supplied code mostly matches Appendix C of THE PAPER.

We observed that some of the Waterbirds images are small enough that augmentation may sample outside the image area. The majority groups contain relatively more (Table 2) ”too small” images than the minority groups. This imbalance may cause a spurious correlation.

Their code scales images by 256/224 (linear) and then centre crops to 224x224 pixels in evaluation. This operation is not noted in THE PAPER. See Sec 3.1 re. normalisation discrepancies.

The code labels groups 0-3 where THE PAPER uses G_{1-4} in Fig 6, but the meanings and sizes of the groups match. In Sec 2, THE PAPER defines the groups differently, which is problematic.

Table 2: Small images

GROUP	**COUNT**	**TOO SMALL**	**%**
0 0	3498	116	3.3
0 1	184	4	2.2
1 0	56	1	1.8
1 1	1057	37	3.5

The criterion used for training in train_classifier.py is torch.nn.CrossEntropyLoss(). This is not stated in THE PAPER.

Running train_classifier.py five times with different SEED values (Table 3) did not match Table 2 of THE PAPER. (We believe the values were for the test group labelled ”wb” rather than ”wb_val”, though the two were very similar.) Mean accuracy is far from that stated, and std is twice as big for worst group. Their Baseline value (98.1) is much better than the DFR mean accuracy, and we suspect it was a typographical error.

DFR_evaluate_spurious.py was run seven times on a single model (trained with SEED 747) as shown in Table 3. Our accuracies were slightly lower than in THE PAPER but showed improvements over the Baseline values, especially for the Worst groups. The std value was an order of magnitude higher than that reported in THE PAPER.

[1] On our Windows 10 PC containing an Nvidia RTX 2070 with 8GB of dedicated GPU RAM, we were unable to run train_classifier.py with batches of 128 (per THE PAPER), using 32 instead - this would be expected to give sub-optimal results. Training took $\tilde{2}0$ hours, so we only ran it once, hence no std value

Table 3: Baseline and DFR comparison for Waterbirds (THE PAPER, Table 2) & our results

	WATERBIRDS		CELEBA	
METHOD	**WORST(%)**	**MEAN(%)**	**WORST(%)**	**MEAN(%)**
Baseline (Paper)	74.9 ± 2.4	98.1 ± 0.1	46.9 ± 2.8	95.3 ± 0
Baseline (Ours)	72.7 ± 5.0	91.1 ± 0.3	37.4[1]	95.1
$\text{DFR}_{\text{Tr}}^{\text{Val}}$ (Paper)	92.9 ± 0.2	94.2 ± 0.4	88.3 ± 1.1	91.3 ± 0.3
$\text{DFR}_{\text{Tr}}^{\text{Val}}$ (Ours)	91.1 ± 2.1	93.1 ± 1.1	88.0 ± 0.8	91.1 ± 0.1

CelebA All images in CelebA are 178x218 pixels, so all groups will be subject to equivalent levels of padding when augmented/cropped. Our results (Table 3) are compatible with those in THE PAPER (taking into account the smaller batch size). The DFR was run five times on the single-trained model (Seed 16).

3 CODE CREATION FROM PUBLISHED DESCRIPTION

Our code is committed at `https://github.com/COMP6258-Reproducibility-Challenge/Dont-Prop-Back-in-Anger`

3.1 SEC 4.1: FEATURE LEARNING ON WATERBIRDS DATA

This section describes an attempt to reproduce the results in Table 1 of THE PAPER, using only implementation details from the THE PAPER along with some reasonable assumptions.

This reproduction does not include the classifier fine-tuned on the ’Balanced (50%)’ training dataset, as we could not locate it to download. THE PAPER describes how it was created, but the method of replacement used is unspecified, as are the replacement backgrounds, meaning that an accurate reconstruction of this dataset was not possible. See Appendix B.1 of THE PAPER for details.

Hyperparameters are used exactly as specified where they are provided. The transformation applied to the testing data is not specified, but we assumed an unscaled centre crop to the correct input shape. We also assumed the unstated loss function to be Cross Entropy.

We note that the supplied code uses transforms.Normalize() during training and testing, but this is not documented in THE PAPER, so it was not implemented in our code. We used a seed value of 3141592654.

The description in THE PAPER of how the groups are used was hard to interpret, and it was only with reference to the supplied code that it became clear that training is done using only labels (waterbird=1/landbird=0), whereas the testing is against all four groups. This perhaps reflects our lack of domain knowledge.

Table 4: ERM classifiers on Waterbirds (THE PAPER, Table 1) & our results

	TEST DATA (WORST ACC)	
CLASSIFIER TRAINING DATA	**ORIGINAL (95%)**	**FG-ONLY**
Original - 95% (Paper)	73.8%	93.7%
Original - 95% (Ours)	80.7%	90.5%
Original - 100% (Paper)	34.8%	94%
Original - 100% (Ours)	49.7%	86.3%
FG-Only - 0% (Paper)	75.2%	95.5%
FG-Only - 0% (Ours)	81.3%	92.4%

Our results (Table 4) correlate with the results from THE PAPER in that in each case, FG-Only performs better than Original-95%. However, their ”Original - 100%” classifier performs much worse than ours on the ”Original” test data and much better than ours on the ”FG-Only” test data, which is surprising. Their classifier performed better than ours on the FG-Only data, which implies it learned more core features and relied less on spurious correlations, so we would have expected it to perform better than ours on the data with spurious correlations as well.

3.2 Sec 6: Feature Reweighting Improve Robustness

This section describes an attempt to reproduce an experiment done in Sec 6 of The Paper based on its description. Specifically, using multiNLI datasets with the BERT model.

NLI involves determining whether a given premise sentence logically entails, contradicts, or is neutral with respect to a given hypothesis sentence (Williams et al., 2018). The Paper uses a 50-20-30 split of train, validation and test data using metadata from the GroupDRO paper (Sagawa et al., 2020), but did not state this; we had to infer it.

The Paper does not clarify what exact negation words are used, but we infer they are as (Sagawa et al., 2020) indicates: ’nobody’, ’no’, ’never’ and ’nothing’.

There is a discrepancy between the dataset and the metadata from (Sagawa et al., 2020). The group labelling differs between those two sources. This confusion has led to The Paper mislabelling their results. Taking this into account, we reproduced exactly their group sizes.

In our code, data was tokenised using BertTokenizer, with truncation set to ‘true’ and padding to ‘max length’. The max length used was tuned to 128; this did lead to some tokens overflowing, but it only affected 0.5% of data. These values were not provided in The Paper. Tokenisation takes approximately 6 minutes. The model is trained on the BERT model using the specification given by The Paper. Training takes approximately 3 hours and 45 minutes. This was done for five epochs in accordance with The Paper. The worst group accuracy on test data was similar to that reported in The Paper.

Last layer retraining (DFR) was implemented as described in The Paper, increasing the accuracy of the worst group compared with Baseline values, but only by 5% instead of the 9% reported in The Paper.

4 Conclusion

For our chosen experiments in The Paper, we were able to confirm that Last Layer Retraining does improve worst group accuracy compared to the baseline performance.

We were not able to reproduce the exact results reported in The Paper. Many aspects of The Paper lacked detail, or contained minor errors, which may explain the discrepancies, and which make it harder to extend work in the field.

Acknowledgments

Thanks to Jeff McCann (ECS Systems Support Technician) for configuring our PC for this project. We are also grateful to Dr A Marcu for background and insight.

References

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts, April 2020. URL `http://arxiv.org/abs/1911.08731`. arXiv:1911.08731 [cs, stat].

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, February 2018. URL `http://arxiv.org/abs/1704.05426`. arXiv:1704.05426 [cs].