Reproducibility report formatting instructions for ML Reproducibility Challenge 2020

Anonymous Author(s)
Affiliation
Address
e-mail

Reproducibility Summary

Scope of Reproducibility
The paper proposes Contextual Decomposition Explanation Penalization (CDEP) to make a model more reliable and robust to spurious dataset correlations and dataset shifts. This is done by penalising the model so that the model’s predictions align with the provided explanations. This work aims to experimentally verify the CDEP algorithm on the ISIC skin cancer, Decoy MNIST and Stanford Sentiment Tree classification datasets, and show that a model trained with CDEP is more robust than a normally trained network.

Methodology
We reimplemented the PyTorch codebase provided by the authors in tf.keras and reproduced experiments across the three aforementioned tasks. For each task we used the same architectures/models as described in the paper, and the same hyperparameters wherever specified. We attempted to retain the data preprocessing steps in the original codebase, but made modifications where necessary. Additionally, we verified the outputs of both codebases using unit tests. All experiments were carried out on Google Colaboratory using T4 Tesla GPUs for a total of around 140 hours.

Results
We reproduced the results on DecoyMnist within 1% margin of the reported results. For the SST dataset the difference between penalised and unpenalised model was within 3% of the difference reported in the paper for the gender variant. The unpenalised model did better for the bias and random variants. For ISIC all models were reproduced but could not be run due to computational constraints. We did not insentence subtrees for training due to computational constraints.

What was easy
Both the paper and the codebase were written in a clear and concise manner, and it was easy to understand the major concepts and ideas. For the most part we were able to use the same dataset preprocessing steps that were described in the repository. The simplicity and accessibility of the code and datasets allowed us to focus on the novel aspects that the paper proposes.

What was difficult
Dealing with PyTorch and tf.keras discrepancies was the most difficult and time-consuming. Several functions needed to be reimplemented from scratch, and this was especially cumbersome for the MaxUnpool2D method and the code needed to be modified to deal with the difference in the pretrained models provided by each library. Some dataset downloading code needed to multithreaded to download it in a practical time frame. CDEP required more overhead than normal training, and the ISIC experiment could not be run with the original dataset size.

Communication with original authors
We reached out for guidance in selecting hyperparameters as well as in understanding how evaluation was done on the Decoy MNIST task. We are grateful for their prompt replies providing the same.

Submitted to ML Reproducibility Challenge 2020. Do not distribute.