IM1102-232433M - Deep Neural Engingeering assignment 1

Review of "On the Connection Between Pre-training Data Diversity and Fine-tuning Robustness"

Alexander Van Hecke (852631385)

March 16, 2024

Summary

In Ramanujan et al's article "On the Connection Between Pre-training Data Diversity and Fine-tuning Robustness" (1), the authors examine the influence of pre-training data on downstream robustness against distribution shift. Previous research on the influence of pre-training data focused either on downstream performance instead of robustness (2–4), or on architecture and pre-training algorithm variations when examining model robustness (5–7).

The study describes a number of experiments comparing robustness of models obtained using transfer learning against models trained from scratch. In a first series of experiments, the effects on downstream robustness of **(i)** pre-training set size, **(ii)** label granularity (coarse vs fine grained labels), **(iii)** label semantics (alignment with downstream task) and **(iv)** image diversity (number of categories and per-class image diversity) within a data distribution are examined. A second experiment examines fine-tuning robustness behaviours across different pre-training data sources, including different types of synthetic datasets. Both completely synthetic datasets (8) and thematically relevant generated datasets using stable diffusion (9) are considered. Finally, the use of self-supervised pre-training is considered.

The research shows that pre-training set size and label granularity influence robustness against distribution shift most. Furthermore, it is shown that even though the use of the synthetic FractalDB-1K dataset (8) improves downstream robustness over training from scrach, it is noticeably less effective than using natural image data. The use of a synthetic dataset generated using stable diffusion however, yields similar improvements in downstream robustness compared to natural image datasets, at least up to a certain pre-training dataset size. Using a larger stable diffusion dataset shows that the robustness benefits begin to saturate and do not keep up with the benefits obtained from using natural image datasets. Finally, the use of self-supervised pre-training datasets is shown to have similar downstream robustness benefits as the use of natural image datasets.

Associated theory and related work

The authors have given a thorough overview of current literature. An overview of why transfer learning is used is given as general context. In particular, in computer vision tasks, the use of models pre-trained on the ImageNet dataset is found to be a de-facto standard for transfer learning in several domains.

According to the authors, current research on the use of pre-trained models mostly focuses on the effects on downstream performance, in particular accuracy. However, the authors of (1) do not focus on performance, but rather on robustness against (natural) distribution shift. Much work has been done on robustness, and the authors have looked at approaches using variations of architecture and pre-training algorithms. An architecture dynamically deciding whether to pass a target image through the fine-tuned layers or the pre-trained layers is discussed in (7). Adversarial robustness is introduced in (6), and a mapping between source and target labels is learned in (5).

The effect of the pre-training dataset on robustness was already examined in earlier work. A wide range of natural distribution shifts was examined in (10), where it was determined that a model pre-trained on ImageNet yields the best robustness. Contrastively trained language-image models such as CLIP are examined in (11). The authors of this paper examine why models such as CLIP result in robustness gains,

and conclude that the more diverse training distribution is the main cause. However, in (12) it is shown that simply gathering huge datasets is not sufficient for obtaining better robustness.

The authors of the reviewed paper (1) focus on natural distribution shifts, meaning the shifts are induced by natural processes (such as varying weather conditions). They examine the effect of variations of features (dataset size, label granularity, label semantics, image diversity and dataset diversity) of dataset on robustness. Both supervised and self-supervised pre-trained models are considered.

We have found some papers not mentioned in (1) discussing the robustness of pre-trained models against distribution shift. The accuracy and robustness of supervised, self-supervised and auto-encoder based models when presented with out-of-distribution data are compared in (13). The authors of (13) conclude that self-supervised models are consistently more robust against distribution shifts than supervised and auto-encoder models. Yet, the authors of the reviewed paper (1) have primarily investigated supervised models and found that self-supervised models yielded no better robustness than supervised models, even when varying pre-training set size.

An empirical study on distribution shift robustness is done in (14). One of the things they examine is pre-training dataset size. They conclude there is a positive correlation between pre-training dataset size and robustness against distribution shift. Furthermore, the positive correlation between pre-training and model robustness was already examined in (15).

Most innovative aspects and strengths

In the reviewed paper (1), the authors have intervened in different ways in the pre-training dataset to examine the effects of these interventions on robustness against distribution shift in the resulting fine-tuned models. Their results can be summarized as follows for supervised learning:

- using pre-training yields better robustness than training from scratch
- pre-training with more data helps to increase robustness. Even using moderate pre-training dataset sizes increases robustness. This parameter influences robustness the most.
- decreasing label granularity decreases robustness. This effect is smaller than intervening in dataset size.
- intervening in label semantics or image diversity does not influence robustness significantly.
- intervening in the data source (generic ImageNet data source versus domain specic iNaturalist data source) does not influence robustness significantly.
- purely synthetic dataset such as FractalDB-1K (8) decreases robustness compared to natural image data. However, this still yields a significant improvement in robustness compared to training from scratch.
- synthetic datasets generated using stable diffusion (9) yield the same robustness compared to natural image data, or at least up to a certain dataset size (150K images). When using larger dataset sizes (1M images), the benefits of pre-training with these kinds of synthetic datasets begin to decrease compared to natural image data.

Their results can be summarized as follows for self-supervised learning :

- self-supervised image-text pre-training data increases robustness
- it does not outperform supervised approaches, even when varying dataset sizes (up to 500M images).

The goal of the authors is to establish guidelines on the creation and use of pre-training datasets. Of note here is that robustness is used as the metric, not accuracy. The trained models are more robust when distribution shift occurs (ideally the accuracy remains the same in in-distribution and out-of-distribution conditions), their accuracy is not necessarily better in in-distribution conditions compared to models trained from scratch.

The most innovative aspects and strenghts of (1) can be summarized as :

- pre-training dataset sizes can be reduced down to 25K images without affecting robustness too much. Smaller dataset sizes than 25K yield worse robustness. This means moderate pre-training dataset sizes suffice, saving compute time and resources.
- very coarse grained labels yield worse robustness, but there is no need to assign very fine grained labels, as from 17 labels (and up) robustness is not affected much. This means labelling efforts, often done by hand, can be simplified.

- the use of domain specific images in pre-training does not influence robustness. In other words, any readily available dataset can be used for pre-training.
- the use of diverse images in pre-training does not influence robustness. In other words, over- or underrepresentation of certain classes is not necessarily a problem for robustness.
- the use of domain specific datasets (datasets containing only images relevant to the domain) does not influence robustness. Again, this means any readily available dataset can be used for pre-training.
- the use of "realistic" synthetic pre-training dataset (i.e. using stable diffusion (9)), is as good as natural image data in increasing robustness, up to a certain dataset size. In other words, when no dataset is availabe, you can generate one without adversily impacting robustness.

Experiment design critique

The following setup is used in each experiment in (1), each time intervening in some way in the pre-training datasets. The ImageNet (wide range of images) and iNaturalist (animal images) datasets are used as pre-training datasets (in one experiment additional synthetic pre-training datasets are considered). A number of neural network architectures (ResNet, ReNext, DenseNet, AlexNet and MobileNet-V3) are pre-trained and then fine-tuned to the iWildCam (wildlife images) dataset, going through a range of number of epochs for each experiment. The fine-tuned models are then evaluated against both in-distribution images and out-of-distribution images available in the iWildCam-WILDS dataset. Macro F1 scores are calculated for both in-distribution and out-of-distribution images and plotted in a scatter plot. Finally, linear trends between in-distribution and out-of-distribution macro F1 scores are calculated and visualised on the plots, along with their 95% bootstrap confidence intervals. Ideally, the in-distribution macro F1 scores are equal to the out-of-distribution macro F1 scores, as this would indicate the model always performs the same, regardless of the presence of any distribution shift. A baseline linear trend of a model trained from scratch is taken from another study, and is plotted as well.

Experiments are done by intervening in the pre-training dataset in different ways :

- limiting the pre-training dataset sizes to (class balanced) sizes of 150K, 100K, 50K, 25K and 5K
- changing label granularity by grouping classes into superclasses, resulting in 5, 17, 37, 85 and 232 classes
- changing label semantics, i.e. focusing on domain relevant images (animals versus objects)
- changing image diversity, using both more data per class and more classes of data
- changing data sources, comparing not only ImageNet and iNaturalist, but also synthetic datasets (8) and (9) and a self-supervised dataset

The experiments are clearly described, however repeatability is not trivial without additional effort :

- the code used does not seem to be publicly available, or at least no link is provided.
- several details on experiments are unclear
  - when changing dataset size, which 1000 classes were chosen?
  - when not changing dataset size, were all 1.2M ImageNet and 600K iNaturalist used?
  - when generating a synthetic dataset using stable diffusion (9), which 80 prompts were used?
  - when changing label semantics, which images were selected exactly?
- it would be useful to summarize all experiments in a table, clearly specifying what was changed in the experiment and also specifying the values of parameters not changed in a specific experiment.

Attention has been given to possible confounding variables in the different experiments. In particular, in the experiment changing label semantics, an additional analysis on a potential confounder because of overlapping labels is done. And in the experiment changing the image diversity, measures are taken to prevent data quantity from being a confounding variable.

The experiments are not statistically validated, only linear trendlines are calculated and visualised. This allows to visually estimate how the different interventions influence robustness, but it does not tell us whether the differences in robustness due to the interventions are statistically relevant.

Presentation evaluation

The article is generally well written, and is sufficiently easy to follow without being simplistic. Sometimes abbreviations are used without introducing them first (e.g. in-distribution ID, out-of-distribution OOD).

The structure mostly follows the different experiments that were performed, and subsections typically describe one of the performed interventions (data quantity, label granularity, label semantics, image diversity, and data sources) in detail. However, when examining interventions in dataset, the authors discuss supervised datasets (section 4.5) and self-supervised datasets (section 5) in different sections of the report. The reason for this separation is not clear nor explained. Since the goal is to examine the effect on robustness when varying datasources, it would perhaps make more sense to eliminate the distinction between the two types of datasets and just examine their effect on downstream robustness in one experiment.

The figures containing examples of pictures from the different datasets (Figures 9, 13 - 15, 17) positively contribute to the understanding of the paper. However, we found the figures detailing robustness (Figures 4 - 8, 10 - 12, 18 - 20, 22) hard to read, as they are quite small and contain a lot of datapoints and information. Full resolution versions of these diagrams in an appendix would be helpful. Additionally, the color palettes used make it hard to distuingish the different variations examined (more so when suffering from colour blindness). Some figures lack axis labels (Figures 21, 22), making it harder to interpret them.

Few tables are used, and in particular one or several tables summarizing the different datasets, quantities and parameters used in the different experiments would be quite helpful. The different parameters can be extracted from the text, but it would be good to make an overview table describing all experiments.

Final decision

We appreciate the main conclusions of the paper :

- when using transfer learning to fine-tune a model for the reasearched problem domain (detection of wildlife in images), the most influential factors for downstream robustness are data quantity and label granularity of the pre-training dataset.
- the use of synthetic natural-looking images generated by tools like stable diffusion have potential to construct pre-training datasets that increase downstream robustness.

Several opportunities for further work are found by the authors :

- the difference in robustness between pre-trained models and models trained from scratch cannot be reproduced across all domains. It is unclear why this is the case.
- the use of different self-supervised (huge) web-crawled corpora appear to significantly improve robustness to distribution shifts. The question of whether such pre-training datasets can be manipulated using the same interventions discussed in the paper, is left for future work.

We would suggest the following minor revisions before this paper can be accepted :

- add a summary table detailing dataset sizes, datasets used, images used, selection criteria, epochs and other (hyper)parameters per intervention
- add axis labels to (Figures 21, 22)
- clarify images detailing robustness (Figures 4 - 8, 10 - 12, 18 - 20, 22). Provide full resolution images in an appendix, and revise color palettes to be colour blind proof and to better distuingish the examined intervention parameters.
- add a link to all code (pre-training and training of models, as well as all code necessary for the interventions) and data used in a public code repository
- always explain or introduce abbreviations
- add additional relevant literature. In particular the results of (13) could be analysed in further detail to try to explain why self-supervised models consistently improve robustness over supervised models in (13), but not in (1).
- add statistical relevance testing for the different experiments

References

1. Ramanujan V, Nguyen T, Oh S, Farhadi A, Schmidt L. On the connection between pre-training data diversity and fine-tuning robustness. Advances in Neural Information Processing Systems [Internet]. 2024 [cited 2024 Mar 9];36. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/hash/d1786f5246c67eefde011599d31b2006-Abstract-Conference.html

2. Huh M, Agrawal P, Efros AA. What makes ImageNet good for transfer learning? [Internet]. arXiv; 2016 [cited 2024 Mar 9]. Available from: http://arxiv.org/abs/1608.08614

3. Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, et al. Big Transfer (BiT): General Visual Representation Learning. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. Cham: Springer International Publishing; 2020. p. 491–507. (Lecture Notes in Computer Science).

4. Kornblith S, Shlens J, Le QV. Do Better ImageNet Models Transfer Better? In 2019 [cited 2024 Mar 9]. p. 2661–71. Available from: https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html

5. You K, Kou Z, Long M, Wang J. Co-Tuning for Transfer Learning. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2024 Mar 9]. p. 17236–46. Available from: https://proceedings.neurips.cc/paper/2020/hash/c8067ad1937f728f51288b3eb986afaa-Abstract.html

6. Salman H, Ilyas A, Engstrom L, Kapoor A, Madry A. Do Adversarially Robust ImageNet Models Transfer Better? In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2024 Mar 9]. p. 3533–45. Available from: https://proceedings.neurips.cc/paper/2020/hash/24357dd085d2c4b1a88a7e0692e60294-Abstract.html

7. Guo Y, Shi H, Kumar A, Grauman K, Rosing T, Feris R. SpotTune: Transfer Learning Through Adaptive Fine-Tuning. In 2019 [cited 2024 Mar 9]. p. 4805–14. Available from: https://openaccess.thecvf.com/content_CVPR_2019/html/Guo_SpotTune_Transfer_Learning_Through_Adaptive_Fine-Tuning_CVPR_2019_paper.html

8. Kataoka H, Okayasu K, Matsumoto A, Yamagata E, Yamada R, Inoue N, et al. Pre-training without Natural Images. In 2020 [cited 2024 Mar 9]. Available from: https://openaccess.thecvf.com/content/ACCV2020/html/Kataoka_Pre-training_without_Natural_Images_ACCV_2020_paper.html

9. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis With Latent Diffusion Models. In 2022 [cited 2024 Mar 9]. p. 10684–95. Available from: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html

10. Miller JP, Taori R, Raghunathan A, Sagawa S, Koh PW, Shankar V, et al. Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization. In: Proceedings of the 38th International Conference on Machine Learning [Internet]. PMLR; 2021 [cited 2024 Mar 9]. p. 7721–35. Available from: https://proceedings.mlr.press/v139/miller21b.html

11. Fang A, Ilharco G, Wortsman M, Wan Y, Shankar V, Dave A, et al. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). In: Proceedings of the 39th International Conference on Machine Learning [Internet]. PMLR; 2022 [cited 2024 Mar 9]. p. 6216–34. Available from: https://proceedings.mlr.press/v162/fang22a.html

12. Nguyen T, Ilharco G, Wortsman M, Oh S, Schmidt L. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP. Advances in Neural Information Processing Systems [Internet]. 2022 Dec [cited 2024 Mar 9];35:21455–69. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/86a8a512b27f49519594ebe89f66d708-Abstract-Conference.html

13. Shi Y, Daunhawer I, Vogt JE, Torr P, Sanyal A. How robust are pre-trained models to distribution shift? In 2022 [cited 2024 Mar 9]. Available from: https://openreview.net/forum?id=zKDcZBVVEWm

14. Liu Z, Xu Y, Xu Y, Qian Q, Li H, Jin R, et al. An Empirical Study on Distribution Shift Robustness From the Perspective of Pre-Training and Data Augmentation [Internet]. arXiv; 2022 [cited 2024 Mar 9]. Available from: http://arxiv.org/abs/2205.12753

15. Hendrycks D, Lee K, Mazeika M. Using Pre-Training Can Improve Model Robustness and Uncertainty. In: Proceedings of the 36th International Conference on Machine Learning [Internet]. PMLR; 2019 [cited 2024 Mar 9]. p. 2712–21. Available from: https://proceedings.mlr.press/v97/hendrycks19a.html