Research Proposal: A Path Towards More Robust and Statistically Efficient Machine Learning.

Learning robust representations with high statistical efficiency is important for reducing the cost of building safe AI systems. Current training methods often require large datasets to learn robust models. In contrast, humans can learn to solve the same problems with a relatively small amount of data. Closing this gap in robustness and statistical efficiency could help us build more trustworthy models, reduce the economic and social cost of building large datasets, and improve our understanding of how general behaviour emerges during model training.

While the formulation of a classical ML problem as minimising a loss function over a distribution is well-founded [13], applying SGD over a training set introduces a variety of opportunities for improvement. Specifically, when applying SGD, we risk **(1)** getting stuck in a local minimum of the loss function, **(2)** learning spurious correlations in the training set, and **(3)** limiting the amount of information extracted from a single batch.

This research proposal explores how we can take a step towards addressing these shortcomings to improve the robustness and statistical efficiency of model training. This will help reduce the cost of training more safe and trustworthy models, impacting a wide range of problems, for example, in computer-aided drug design - where it is expensive to gather large volumes of data - or training autonomous vehicles - where robustness to a wide range of uncommon circumstances is critical. Each of the three opportunities can be considered independently, which introduces some redundancy if any of the proposed methods are unsuccessful.

The first opportunity identified above for improving SGD is to **reduce the chance of converging to a local minimum with low generalisation error**. Although there has been significant previous work in this area [7, 5, 2, 1], new characterisations of the loss surface for non-linear models [such as 20] could yield improvements to the problem.

One approach could consider how to better initialise models towards better attractors in the loss function. Warm-up has been shown to improve robustness and convergence speed [11], yet there is relatively little work considering alternative methods for traversing the weight space before executing the traditional SGD procedure. For example, Li et al. [10] show models that generalise better often follow an indirect path over plateaus in the loss function before descending to a local minimum [see also 6]. We might imagine that these models find a better minimum because they 'search' more of the weight space before selecting the 'best' minimum point to descent to. This motivates an investigation into work that explicitly encourages this behaviour, for example through an additional warm-up routine that encourages traversing the weight space parallel to the contours of the loss function.

The second problem identified above is that **SGD limits how much we can learn from each data point**. Consider the problem of learning from a batch of new data in a continual learning setting. To obtain high statistical efficiency, we should maximise the mutual information between the update to our model and our batch of data, while obtaining robustness requires us to limit the complexity of the resulting model [17].[1]

Previous work on improving statistical efficiency focuses primarily on meta-learning single-step update rules [4]. However, comparatively little work considers the underlying problem of how to balance the conflicting objectives of statistical efficiency and robustness. For example, we could extract more information from a single batch of data during continual learning by performing

[1] Here 'complexity' refers to the information content of a black-box model, which does not necessarily correspond to its parameter count. This is important in the context of recent work that has shown that increasing model size often improves generalisation [16, 21].

multiple gradient steps on the same batch. In practice, this would lead to overfitting, so we can use a method such as LDIFS [15] to constrain the update to behaviour that lies within a fixed-radius ball centred at the original model. By adaptively varying the ball's radius, we can control this efficiency-robustness trade-off. We could also investigate how to extend methods to constrain the size of model updates, for example, by extending LDIFS to non-Euclidean distance metrics.

Even if we are able to efficiently learn the empirical distribution of our dataset, **sampling our training set from the true distribution we want to model can introduce spurious correlations into our model** (point **(2)** above) [19]. I propose a two part approach to this problem: first, we should use active learning to more tightly control the training set, and, second, we should build models that are inherently less prone to learn specific 'shortcuts' that do not generalise well outside of the training set.

Active learning and curriculum learning can be used to control the model's training distribution to encourage better generalisation and robustness for each newly labelled data point [14]. For example, Korakakis et al. [8] show that, by focusing training on areas of the model's representation space with relatively little data, we can improve both in-distribution, and out-of-distribution generalisation. However, we still do not have a good understanding about what features of our training set will lead to certain model behaviours. We should investigate how we can use active learning or curriculum learning to answer this question. For example, we could consider encoding invariances into the representation space by identifying the representations of three points, say r_0, r_1, and r_2, and then training on a new point that has a representation close to $r_0 + (r_1 - r_2)$.

An alternative approach to learning models that avoid overfitting to our training set is to introduce inductive bias into our model. Veličković and Blundell [18] suggest a direct approach could be to align our model's architecture with the task we are attempting to solve. While this and other similar approaches have been shown to be promising directions, they do not necessarily eliminate all shortcuts a model may learn. For example, one observed reason for poor robustness is due to *superposition*, where a model represents multiple concepts with a single activation [3]. These different, superposed concepts may not overlap in the training set, however, at inference-time, superposition could lead to poor generalisation due to concepts that share activations interfering with each other. A solution may be to force the model to avoid superposition, for example by adding a new loss term that encourages feature activations to be either high or low, but not somewhere in between. Such a method could also draw parallels with (synchronised) spiking neural nets [12], and may improve model interpretability [3, 9].

References

[1] Animashree Anandkumar and Rong Ge. "Efficient approaches for escaping higher order saddle points in non-convex optimization". In: *29th Annual Conference on Learning Theory*. Ed. by Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 81–102. URL: https://proceedings.mlr.press/v49/anandkumar16.html.

[2] Pratik Chaudhari et al. "Entropy-SGD: biasing gradient descent into wide valleys". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019 (2016). URL: https://api.semanticscholar.org/CorpusID:13807351.

[3] N. Elhage et al. *Toy Models of Superposition*. 2022.

[4] 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*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 1126–1135. URL: https://proceedings.mlr.press/v70/finn17a.html.

[5] Pavel Izmailov et al. "Averaging weights leads to wider optima and better generalization". In: *arXiv preprint arXiv:1803.05407* (2018).

[6] Yiding Jiang et al. *Fantastic Generalization Measures and Where to Find Them*. 2019. arXiv: 1912.02178 [cs.LG]. URL: https://arxiv.org/abs/1912.02178.

[7] Nitish Shirish Keskar et al. "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=H1oyRlYgg.

[8] Michalis Korakakis, Andreas Vlachos, and Adrian Weller. "ALVIN: Active Learning Via INterpolation". In: *arXiv preprint arXiv:2410.08972* (2024).

[9] Michael Lan et al. "Sparse autoencoders reveal universal feature spaces across large language models". In: *arXiv preprint arXiv:2410.06981* (2024).

[10] Hao Li et al. "Visualizing the Loss Landscape of Neural Nets". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

[11] Liyuan Liu et al. "On the Variance of the Adaptive Learning Rate and Beyond". In: *ArXiv* abs/1908.03265 (2019). URL: https://api.semanticscholar.org/CorpusID:199528271.

[12] Wolfgang Maass. "Networks of spiking neurons: The third generation of neural network models". In: *Neural Networks* 10.9 (1997), pp. 1659–1671. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(97)00011-7. URL: https://www.sciencedirect.com/science/article/pii/S0893608097000117.

[13] David JC MacKay. "Information Theory, Inference, and Learning Algorithms". In: (2003).

[14] Sören Mindermann et al. "Prioritized training on points that are learnable, worth learning, and not yet learnt". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 15630–15649.

[15] Jishnu Mukhoti et al. "Fine-tuning can cripple your foundation model; preserving features may be the solution". In: *arXiv preprint arXiv:2308.13320* (2023).

[16] Vaishnavh Nagarajan and J Zico Kolter. "Uniform convergence may be unable to explain generalization in deep learning". In: *Advances in Neural Information Processing Systems* 32 (2019).

[17] Behnam Neyshabur. "Implicit Regularization in Deep Learning". In: *arXiv e-prints* (2017), arXiv–1709.

[18] Petar Veličković and Charles Blundell. "Neural algorithmic reasoning". In: *Patterns* 2.7 (2021).

[19] Tan Wang et al. *Causal Attention for Unbiased Visual Recognition*. 2021. arXiv: 2108.08782 [cs.CV]. URL: https://arxiv.org/abs/2108.08782.

[20] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. "A Critical View of Global Optimality in Deep Learning". In: *ArXiv* abs/1802.03487 (2018). URL: https://api.semanticscholar.org/CorpusID:125214941.

[21] Chiyuan Zhang et al. "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3 (2021), pp. 107–115.