Understanding Deep Learning Requires Rethinking Sharpness

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

The geometric flatness of neural network minima has long been associated with desirable generalisation properties. In this paper, we extensively explore the hypothesis that robust, calibrated and functionally similar models sit at flatter minima, inline with prevailing understandings of the relationship between flatness and generalisation. Contrary to common assertions in the literature, we find a relationship between increased sharpness, generalisaton, calibration and robustness in neural networks across architectures when using Sharpness Aware Minimisation, augmentation and weight decay as regulariser controls. Our findings suggest that the role of increased sharpness should be considered independently for individual models when reasoning about the geometric properties of neural networks. We show that sharpness can be related to generalisation and safety-relevant properties against the flatter minima found without the use of our regularisation controls. Understanding these properties calls for a re-thinking of the role of sharpness in geometric landscapes.

1. Introduction

Neural network architectures with different implicit biases have been observed to have different geometric properties around the minima, with flatness being ascribed to performance improvements [18]. Literature has stated the desirability of flat minima, due to having wide error margins [25]. Empirical studies have sought to support this idea [5, 12, 24]. However this has been argued against with work showing that models can have arbitrarily sharpened minima [3] and a growing body of literature has questioned the requirement of flatness for generalisation generally [1, 22, 28]. Following notions of sharp minima from [3] geometric sharpness measure have been redefined to satisfy reparameterization invariance criteria. Metrics which satisfy the reparameterization invariance condition have been introduced such as Fisher-Roa-Norm [19] and Relative-Flatness [24] which provide improved minima landscape evaluation and reaffirmed the empirical observation of negative correlation between increased sharpness of a loss landscape and generalization [24].

The properties of minima flatness are typically attributed to desirable properties of neural networks such as improvements in generalisation provided by Sharpness Aware Minimization (SAM) [5], improved transferability of the learnt representations [20] and the improvements residual connections yield [17]. However outside of generalisation, other desirable properties of neural network minima exist such as robustness to adversarial perturbations [9], calibration of a models [7] and functional diversity [27], which we consider safety critical evaluations. The relationship between flatness and these safety critical evaluations are lesser known. We extensively explore the hypothesis that robust, calibrated and functionally similar models sit at flatter minima, inline with existing understandings of flatness and generalisation.

We use the CIFAR [13] and TinyImageNet [16] datasets to train the ResNet [8] VGG [26], and ViT [4] architectures. We apply the regularisers weight decay [14], data augmentation and Sharpness Aware Minimization (SAM) while recording the impact of these controls on the geometric properties of the network, alongside our safety critical evaluations.

Our findings can be summarised as follows:

1. Contrary to existing literature, we find that neural networks, across architectures have a relationship between increased sharpness, accuracy, calibration, robustness and functional similarity over our baseline condition.
2. Notions of flatness need to be reconsidered as we find that the geometric properties of a neural network are highly dependent on architecture and dataset.
3. In some cases increased sharpness can serve as a proxy for assessing safety critical properties of neural networks beyond accuracy.

2. Sharpness, Generalization and Safety Critical Evaluations

Sharpness Metrics: We employ three measures of sharpness from literature, namely Fisher-Rao Norm [19], Relative Flatness [24], and SAM-Sharpness [5]. We formalise these metrics in Appendix Section A.

Calibration Evaluation: Calibrated neural networks are important as they enable an understanding of the true error of a model against its predictive confidence. Neural networks such as ResNets have been show empirically to be severely overconfident [7] which results in reduced trustworthiness. To measure calibration we use Expected Calibration Error (ECE) [7], a low ECE indicates a well-calibrated model.

Functional Diversity Evaluation: Functional diversity is important for understanding how similar neural networks are in their representation space [27]. Functional analysis has been used to explain how ensembles work [6] with increased diversity of networks being argued to enable better ensemble performance [21]. However, other literature has shown that representation convergence [27] enables improved ensemble performance. Close functional similarity indicates stability of the learnt representations against the stochastic nature training which we would argue is better for models. To measure functional similarity we use Prediction Disagreement, we consider a low disagreement desirable.

Robustness Evaluation: Robustness of a neural network can provide insights into the strength of the features a network has learned; offering guarantees for safety-critical real-world settings [9]. To quantify robustness we evaluate on adversarial corruptions dataset CIFAR10-C and CIFAR100-C which include, but are not limited to, Impulse Noise, JPEG and Contrast corruptions [9], to analyse this we record the mean corruption accuracy, a high corruption accuracy shows high robustness.

Each of the evaluations described above are crucial for developing safe real-world AI systems and extend beyond considerations of accuracy alone. Therefore, we see an important direction of research opening in observing these properties of neural networks within existing frameworks for analysing generalisation. We formally describe and further elaborate on each of the evaluation metrics in Appendix Section B.

2.1. Experimental Setup

In this paper we want to explore the relationship between geometric properties of neural networks, generalisation and safety relevant measures. In this setting we use 10 initialisations, seeds 0-9. These initialisations are used across all model conditions and we retain the data order for training on each specific seed, such that all models on the same seed start at the same point in the geometric landscape and could hypothetically reach the same minima. The training controls we use are Baseline, Baseline + SAM, Augmentation, Augmentation + SAM, Weight Decay and Weight Decay + SAM. We apply these separately as it allows us to isolate the effect of each condition. In a controlled setup such as this, it is possible to have effective comparisons that are like-for-like and account only for the impact of the specific control being used.

The controls are described as follows:

Baseline A model trained in a vanilla setting that has no extra regularisation terms applied – for each architecture and dataset we define how the baseline model is created in Appendix Section C. For each seed the baseline provides an insight into the standard geometric, generalisation and safety relevant evaluations that should be expected in a vanilla setting for each architecture and dataset.

Weight Decay, Augmentation and SAM: We use Weight decay (at a rate of $5e-4$) and Augmentation (random rotation and crop) as independently applied explicit regularisers to understand their effects on the network. SAM sharpness is an extra optimization process that leverages second-order information and is hypothesised to reduce sharpness of a resulting model, empirically it has seen large performance benefits over traditional optimization [5], though some literature argues that SAM does not seek flatter minima [28]. If SAM sharpness does find flatter minima then one would expect that application of SAM would cause the models in every condition to be flatter. We record how the application of Weight Decay, Augmentation and SAM impact the geometric properties of the network and its safety critical evaluations.

We are interested in the relation between safety critical features and geometric properties of neural networks, as existing flatness literature would suggest these properties would be found at flatter minima. Inspired by the literature, we state the two potential outcomes of the experiment as follows:

- o_1: Neural networks trained from the same initialization using the same data order under the application of different regularisers result in models that perform better on safety evaluations and have **relatively flatter geometric properties compared to the baseline**.
- o_2: Neural networks trained from the same initialization using the same data order under the application of different regularisers result in models that perform better on safety evaluations and have **relatively sharper geometric properties compared to the baseline**.

3. Results

In the main body of the paper we present the results for the ResNet18 trained on CIFAR10, CIFAR100 and TinyImageNet datasets, in Appendix Section C we detail the training settings and sharpness metric settings for each dataset. The results for the VGG and ViT architectures can be found in Appendix Sections D and E, respectively, however it is important to note that the analysis given in the main body largely describes what is observed across architectures. The following tables present the impact of each of the explicit regularisers on the ResNet-18 with respect to accuracy, ECE, corruption accuracy and functional similarity against sharpness metrics. The TinyImageNet results excludes corruption accuracy and Relative Flatness. We report the Mean and ± 1 SEM [2] over 10 models for each table.

Table 1: Results for ResNet-18 Trained on CIFAR10. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness	Relative Flatness
Baseline	0.720 ±0.002	0.186 ±0.001	58.614 ±0.201	0.282 ±0.001	0.009 ±0.000	4.052E-06 ±2.173$E-07$	34.607 ±0.757
Baseline + SAM	0.794 ±0.001	0.108 ±0.001	66.342 ±0.164	0.168 ±0.000	0.029 ±0.002	3.072E-05 ±1.686$E-05$	75.093 ±1.693
Augmentation	0.886 ±0.001	0.077 ±0.001	68.755 ±0.219	0.121 ±0.001	1.101 ±0.060	1.693E-02 ±1.417$E-03$	2903.220 ±89.243
Augmentation + SAM	**0.908** ±0.000	**0.014** ±0.001	**71.419** ±0.283	**0.069** ±0.000	1.529 ±0.009	1.291E-02 ±1.913$E-03$	4970.972 ±30.139
Weight Decay	0.721 ±0.002	0.174 ±0.002	58.562 ±0.227	0.281 ±0.001	0.018 ±0.001	8.493E-06 ±6.908$E-07$	59.767 ±3.009
Weight Decay + SAM	0.802 ±0.001	0.096 ±0.001	67.079 ±0.117	0.162 ±0.001	0.035 ±0.002	3.051E-05 ±1.409$E-05$	88.807 ±2.336

Table 2: Results for ResNet-18 Trained on CIFAR100. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness	Relative Flatness
Baseline	0.530 ±0.002	0.220 ±0.001	38.760 ±0.085	0.452 ±0.000	0.080 ±0.008	1.762E-03 ±1.521$E-03$	32.085 ±0.313
Baseline + SAM	0.556 ±0.002	0.191 ±0.002	41.888 ±0.098	0.410 ±0.000	0.109 ±0.004	1.031E-03 ±6.142$E-04$	123.791 ±4.185
Augmentation	0.697 ±0.002	0.185 ±0.001	44.613 ±0.169	0.288 ±0.001	0.981 ±0.043	1.451E-01 ±1.779$E-02$	2766.925 ±178.669
Augmentation + SAM	**0.705** ±0.001	0.145 ±0.001	**45.428** ±0.217	**0.269** ±0.000	1.140 ±0.010	1.022E-01 ±8.144$E-03$	4196.832 ±52.606
Weight Decay	0.521 ±0.003	**0.099** ±0.005	37.868 ±0.265	0.474 ±0.001	0.235 ±0.032	2.015E-03 ±1.509$E-03$	136.969 ±7.484
Weight Decay +SAM	0.543 ±0.001	0.106 ±0.002	40.604 ±0.222	0.444 ±0.001	0.488 ±0.019	1.882E-03 ±5.944$E-04$	360.271 ±16.190

Table 3: Results for ResNet-18 Trained on TinyImageNet. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Condition	Test Accuracy	Test ECE	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness
Baseline	0.604 ± 0.001	0.303 ± 0.001	0.238 ± 0.000	0.101 ± 0.035	4.928E-04 $\pm 1.036E-04$
Baseline + SAM	0.638 ± 0.000	0.199 ± 0.001	0.186 ± 0.000	0.126 ± 0.029	3.377E-04 $\pm 2.849E-05$
Augmentation	0.578 ± 0.001	0.119 ± 0.001	0.473 ± 0.000	6.056 ± 0.026	1.893E+00 $\pm 7.702E-02$
Augmentation + SAM	0.594 ± 0.000	**0.056** ± 0.002	0.440 ± 0.000	5.796 ± 0.010	1.665E+00 $\pm 5.139E-02$
Weight Decay	0.604 ± 0.001	0.265 ± 0.000	0.222 ± 0.000	0.062 ± 0.008	2.708E-04 $\pm 1.086E-05$
Weight Decay + SAM	**0.641** ± 0.001	0.180 ± 0.001	**0.185** ± 0.000	0.103 ± 0.005	3.056E-04 $\pm 4.175E-06$

Regularisers Can Make Geometric Landscapes Sharper: For every architecture on the CIFAR datasets we find that the Baseline control always records the lowest values for the sharpness metrics Fisher Rao Norm, SAM Sharpness and Relative Flatness. Surprisingly, this coincides with this control having the worst performing results on test accuracy and safety evaluations, as seen in Tables 1 and 2. The models that perform best across our experiments always have higher sharpness values than this baseline condition. As a result, having lower sharpness values does not correlate with increased generalisation or safety properties which is contrary to popular belief would suggest. Our findings posit that sharpness can be a desirable property.

SAM Does Not Only Promote Flatness: Literature has stated that SAM finds flatter points in the loss landscape thereby corresponding to improved generalisation [5]. Our results in Tables 1, 2 and 3 in the main body and in Appendix Sections D and E challenge this belief. In CIFAR10 in Table 1 we see that the application of SAM as a control increases the sharpness values for all sharpness metrics in all conditions with the only exception being Augmentation + SAM for SAM Sharpness. We also see that Augmentation + SAM has the best performance across evaluations and can be described as the sharpest model. The findings hold for the sharpness metrics apart from SAM Sharpness for CIFAR100 in Table 2. It is important to note that there are instances when the application of SAM makes a sharp landscape flatter but typically this is inconsistent, as seen in Table 3.

Important Safety Properties Can Exist at Sharper Minima: Alongside our finding that models in the Baseline control are always flatter for the CIFAR datasets, we also find that the lowest ECE, highest corruption accuracy and lowest prediction disagreement always belong to a control that is sharper than the Baseline control. Our results indicate that sharpness can be an important property for safety evaluations – we posit that this could be due to tighter decision boundaries that have been suggested to exist at sharper minima [10]. With this perspective we understand that all learning problems do not require wide decision boundaries with some tasks necessitating tight decision boundaries enabled by sharpness.

There is No One Geometric Goldilocks Zone for Sharpness: While for the CIFAR datasets we often see a positive relationship between incresed sharpness, generalisation and safety evaluations it is not consistent that the sharpest model across conditions provides the best performance for generalisation and safety measures. However, the model that does perform the best in this regard is typically sharper than the Baseline condition. As result, we argue that neither extreme flatness or sharpness is ideal for learning tasks but that a learning task does require a level of sharpness above that provided by the implicit regaulairstaion of an architecture itself to perform well across these metrics. Furthermore, given that sharpness values and evaluations are specific to each architecture on each dataset we argue that generally the correct sharpness value is dependant on these factors and does not exist as a universal constraint.

4. Conclusion

Our paper seeks to understand the dynamics of neural network training and geometric properties – by connecting geometric properties to other evaluations of safety metrics such as expected calibration error, corruption accuracy and functional similarity moving beyond traditional accuracy evaluations. We find across numerous architectures and datasets, that when Weight Decay, Augmentation and SAM sharpness are used as controls that neural networks access sharper minima, pointing towards a requirement for sharpness to gain the best performance across generalisation and safety relevant measures. Through this, we posit that given the relationship between loss landscape geometry and decision boundaries that this relationship can be explained as a requirement of tighter decision boundaries for different learning tasks. Moreover, our work calls for a deeper exploration of geometric properties of neural networks and argues that understanding deep learning requires rethinking commonly held beliefs regarding flat and sharp minima.

References

[1] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

[2] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005.

[3] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[5] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

[6] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. URL https://arxiv.org/abs/1912.02757.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[10] W. Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K. Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. In Jessica Zosa Forde, Francisco Ruiz, Melanie F. Pradier, and Aaron Schein, editors, *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pages 87–97. PMLR, 12 Dec 2020. URL https://proceedings.mlr.press/v70/dinh17b.

[11] Cheongjae Jang, Sungyoon Lee, Frank C. Park, and Yung-Kyun Noh. A reparametrization-invariant sharpness measure based on information geometry. In Alice H. Oh, Alekh Agarwal,

Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=AVh_HTC76u.

[12] Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/69b5534586d6c035a96b49c86dbeece8-Abstract-Conference.html.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[14] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, page 950–957, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558602224. URL https://dl.acm.org/doi/10.5555/2986916.2987033.

[15] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

[16] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

[17] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

[18] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

[19] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 888–896. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/liang19a.html.

[20] Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23ao.html.

[21] Haiquan Lu, Xiaotian Liu, Yefan Zhou, Qunli Li, Kurt Keutzer, Michael W Mahoney, Yujun Yan, Huanrui Yang, and Yaoqing Yang. Sharpness-diversity tradeoff: improving flat ensembles with sharpbalance. *arXiv preprint arXiv:2407.12996*, 2024.

[22] Israel Mason-Williams, Fredrik Ekholm, and Ferenc Huszar. Explicit regularisation, sharpness and calibration. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=ZQTiGcykl6.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[24] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34: 18420–18432, 2021. URL https://openreview.net/forum?id=sygvo7ctb_.

[25] Jurgen Schmidhuber and Sepp Hochreiter. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994. URL https://proceedings.neurips.cc/paper/1994/file/01882513d5fa7c329e940dda99b12147-Paper.pdf.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. URL https://arxiv.org/pdf/1409.1556.pdf.

[27] Yipei Wang, Jeffrey Siskind, and Xiaoqian Wang. Great minds think alike: The universal convergence trend of input salience. *Advances in Neural Information Processing Systems*, 37: 71672–71704, 2024.

[28] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36:1024–1035, 2023.

Appendix A. Sharpness Metrics

This section describes the sharpness metrics Fisher-Rao norm, SAM-Sharpness and Relative Flatness. Information Geometric Sharpness (IGS) [11] is also a suitable sharpness metric candidate, however we omitted it from this study as the calculation of this metric exceeds feasible computation for large-networks and dataset sizes.

Fisher-Rao Fisher-Rao Norm [19] uses information Geometry for norm-based complexity measurement. It provides a reparametrisation invariant measure for loss landscape sharpness measuring, as verified by [24].

SAM-Sharpness We define SAM-sharpness as the average difference across 100 different locations of 0.005 rho away the original model and calculate the SAM sharpness from these models as defined by [22] and [5].

Relative Flatness [24] define the sharpness measure Relative Flatness– their results show that it has the strongest correlational between flatness and a low generalisation gap. Relative Flatness sharpness is calculated between the feature extraction layer and the classification of the neural network and represents a highly expensive measure due to its calculation of the trace of the hessian of these output matrices.

Appendix B. Safety Critical Metrics

Expected Calibration Error Calibration is the deviation of predicted confidence of a neural network and the true probabilities observed in the data, [7] explored how ResNets are poorly calibrated and are often over confident. To calculate Expected Calibration Error (ECE) we use the Lighting AI Pytorch Metrics implementation of Multiclass Calibration Error[1] Implemented from [15].

Functional Diversity To provide an intuitive understanding of functional diversity we are interested the deviations between models top-1 predictions, the metric we focus on for this is:

- **Prediction Disagreement:** The disagreement between the top-1 predictions of two models on the test dataset. A lower prediction disagreement results in a models that agree more on top-1 predictions.

Robustness Evaluations We employ the CIFAR10-C and CIFAR100-C datasets provided by [9] to observe how geometric properties interact with the robustness of a neural network. An example of the perturbations used is presented in Figure 1, the corruptions have 5 levels of severity per perturbation.

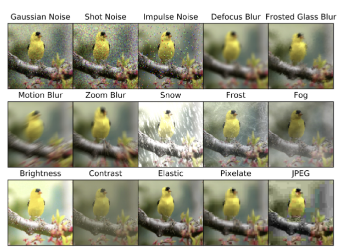

Figure 1: Examples of Adversarial Corruptions on ImageNet dataset examples from [9]

Corruption Accuracy (cACC) The metric we used for this robustness analysis is corruption accuracy. It represents the accuracy of a classifier (f) on the perturbed test dataset ($\mathcal{D}_{corruption}$).

$$\mathbf{cACC^f_c} = \left(\sum_{s=1}^{5} E^f_{s,c}\right) / (5 * c) \tag{1}$$

Appendix C. Experimental Settings

All models are trained using NVIDIA A100 GPU's and each sharpness metric is calculated using the same GPU setup - as models output layer becomes larger for transitions between CIFAR10,

1. Calibration Error documentation from Lighting AI:https://lightning.ai/docs/torchmetrics/stable/classification/calibration_error.html#

CIFAR100 and TinyImageNet the calculation of sharpness metrics increases by an order of magnitude. It should be noted that while Fisher Rao Norm is computationally inexpensive to calculate, SAM sharpness takes a factor of time longer and Relative Flatness is the most computationally expensive measure from a time and memory perspective.

CIFAR10 Training: To train the **baseline** architectures on the CIFAR10 dataset we use the following settings: We use SGD with the momentum hyperparameter at 0.9 to minimize cross entropy loss for 100 epochs, using a batch size of 256 a learning rate of 0.001. For all architectures in the **SAM condition** we use the same settings as above but with SAM an extra optimization step occurs. We use SAM with the hyperarameter rho at the standard value of 0.05. For the **Augmentation condition** we use the baseline conditions with the augmentations Random Crop with a padding of 4 and a fill of 128 alongside a Random Horizontal Flip with a probability of 0.5. Finally for the **Weight Decay condition** we use the same setup as the baseline condition but with the addition of the weight decay value set at $5e - 4$.

CIFAR10 Sharpness: For all sharpness metrics on CIFAR10 we used the entire training dataset to calcualte sharpness across Fisher Rao Norm, SAM Sharpness and Relative Flatness. For the augmentation condition, the training dataset is the augmentations data used to train the model.

CIFAR100 Training: To train the **baseline** architectures on the CIFAR100 dataset we use the following settings: We use SGD with the momentum hyperparameter at 0.9 to minimize cross entropy loss for 100 epochs, using a batch size of 256 a learning rate of 0.01, we also use a Pytorch's [23] Cosine Annealing learning rate scheduler with a Maximum number of iterations of 100. For all architectures in the **SAM condition** we use the same settings as above but with SAM as an extra optimization step occurs and for this we use SAM with the hyperarameter rho at the standard value of 0.05. For the **Augmentation condition** we use the baseline conditions with the augmentations Random Crop with a padding of 4 and a fill of 128 alongside a Random Horizontal Flip with a probability of 0.5. Finally for the **Weight Decay condition** we use the same setup as the baseline condition but with the addition of the weight decay value set at $5e - 4$.

CIFAR100 Sharpness: For both the Fisher Rao Norm and SAM Sharpness metrics on CIFAR100 we used the entire training dataset to calculate sharpness. However, due to the computational burden of calculating Relative Flatness, we only employ 20% of the training dataset to calculate sharpness for this metrics. Once again, for the Augmentation condition, the training dataset is the augmentations data used to train the model.

TinyImageNet Training: On the TinyImagenet dataset we use use pre-trained weights provided for the ResNet18 [2]and VGG19BN [3] by Pytorch - we modify these architectures by removing the existing final layer and replacing it with a final layer with a 200 output classification layer.

To train the **baseline** condition on these architectures using the following settings: We use SGD with the momentum hyperparameter at 0.9 to minimize cross entropy loss for 100 epochs, using a batch size of 256 a learning rate of 0.001. For all architectures in the **SAM condition** we use the same

2. Pytorch ResNet18 ImageNet1K Pretrained Model: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html#resnet18
3. Pytorch VGG19BN ImageNet1K Pretrained Model: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.vgg19_bn.html

settings as above but with SAM as an extra optimization step occurs and for this we use SAM with the hyperarameter rho at the standard value of 0.05. For the **Augmentation condition** we use the baseline conditions with the augmentations Random Resized Crop to the size of 64 and a Random Horizontal Flip with a probability of 0.5. Finally for the **Weight Decay condition** we use the same setup as the baseline condition but with the addition of the weight decay value set at $5e-4$.

TinyImageNet Sharpness: For the Fisher Rao Norm sharpness metric on TinyImageNet we used the entire training dataset to calculate sharpness. However, due to the computational burden of calculating SAM Sharpness, we only employ 20% of the training dataset to calculate sharpness for this metrics. Due to memory constraints on the A100 GPU's we were unable to calculate Relative Flatness for any size of the training dataset on this architecture. Once again, for the Augmentation condition, the training dataset is the augmentations data used to train the model.

Appendix D. VGG

CIFAR10: The Augmentation and SAM condition perform the best for all metrics. It is also the sharpest model with the highest values for Fisher Rao Norm and Relative Flatness and the second highest SAM Sharpness value.

Table 4: Results for VGG-19 Trained on CIFAR10, the mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control	**Test Accuracy**	**Test ECE**	**Corruption Accuracy**	**Prediction Disagreement**	**Fisher Rao Norm**	**SAM Sharpness**	**Relative Flatness**
Baseline	0.782 $\pm$0.001	0.160 $\pm$0.001	64.316 $\pm$0.193	0.204 $\pm$0.000	0.007 $\pm$0.001	1.204E-05 $\pm 6.359E-06$	7.374 $\pm$0.470
Baseline + SAM	0.815 $\pm$0.001	0.108 $\pm$0.001	66.655 $\pm$0.296	0.150 $\pm$0.000	0.296 $\pm$0.011	1.939E-03 $\pm 3.513E-04$	140.164 $\pm$3.149
Augmentation	0.879 $\pm$0.001	0.084 $\pm$0.001	68.497 $\pm$0.199	0.121 $\pm$0.000	1.107 $\pm$0.048	1.780E-01 $\pm 1.932E-02$	688.897 $\pm$26.348
Augmentation + SAM	**0.903** $\pm$0.001	**0.019** $\pm$0.001	**71.268** $\pm$0.196	**0.075** $\pm$0.000	1.342 $\pm$0.008	1.006E-01 $\pm 5.115E-03$	1609.212 $\pm$22.719
Weight Decay	0.782 $\pm$0.001	0.151 $\pm$0.001	64.405 $\pm$0.217	0.202 $\pm$0.000	0.015 $\pm$0.000	1.348E-05 $\pm 1.620E-06$	16.494 $\pm$0.292
Weight Decay + SAM	0.816 $\pm$0.001	0.104 $\pm$0.001	66.827 $\pm$0.286	0.151 $\pm$0.000	0.354 $\pm$0.025	2.565E-03 $\pm 3.723E-04$	157.592 $\pm$5.360

CIFAR100: Augmentation and SAM condition performs the best for test accuracy, corruption accuracy and prediction disagreement. However, for ECE we see that Weight Decay is the best condition. Augmentation and SAM is the second sharpest model for Fisher Roa Norm and SAM sharpness and has the highest value for Relative Flatness. It is important to note that for Weight Decay, with the lowest ECE, that it has higher sharpness values than the baseline condition.

Table 5: Results for VGG-19 Trained on CIFAR100, the Mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness	Relative Flatness
Baseline	0.575 $\pm$0.001	0.253 $\pm$0.000	40.749 $\pm$0.124	0.396 $\pm$0.000	0.050 $\pm$0.005	1.305E-03 $\pm$1.105$E-03$	8.384 $\pm$0.151
Baseline + SAM	0.561 $\pm$0.002	0.232 $\pm$0.002	39.690 $\pm$0.196	0.399 $\pm$0.001	0.167 $\pm$0.005	1.407E-03 $\pm$6.868$E-04$	67.485 $\pm$1.802
Augmentation	0.646 $\pm$0.002	0.222 $\pm$0.002	40.832 $\pm$0.321	0.358 $\pm$0.001	2.287 $\pm$0.091	2.942E-01 $\pm$2.112$E-02$	1430.826 $\pm$53.977
Augmentation + SAM	**0.656** $\pm$0.001	0.157 $\pm$0.001	**41.276** $\pm$0.089	**0.326** $\pm$0.001	1.791 $\pm$0.019	1.983E-01 $\pm$2.004$E-02$	2085.080 $\pm$31.648
Weight Decay	0.584 $\pm$0.001	**0.138** $\pm$0.000	41.266 $\pm$0.112	0.384 $\pm$0.000	0.214 $\pm$0.003	1.380E-03 $\pm$1.047$E-03$	45.728 $\pm$0.073
Weight Decay + SAM	0.553 $\pm$0.002	0.189 $\pm$0.002	38.961 $\pm$0.191	0.429 $\pm$0.001	0.675 $\pm$0.026	4.439E-03 $\pm$1.253$E-03$	153.194 $\pm$6.495

TinyImageNet: The Weight Decay and SAM condition performs best for test accuracy and prediction disagreement. For Weight Decay and SAM condition we see no real difference in the sharpness values. For ECE we see that Augmentation + SAM is the best condition. Augmentation and SAM is the second sharpest model for Fisher Roa Norm and SAM sharpness.

Table 6: Results for VGG-19BN Trained on TinyImageNet, the Mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control	Test Accuracy	Test ECE	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness
Baseline	0.604 ± 0.001	0.303 ± 0.001	0.238 ± 0.000	0.101 ± 0.035	4.928E-04 $\pm 1.036E-04$
Baseline + SAM	0.638 ± 0.000	0.199 ± 0.001	0.186 ± 0.000	0.126 ± 0.029	3.377E-04 $\pm 2.849E-05$
Augmentation	0.578 ± 0.001	0.119 ± 0.001	0.473 ± 0.000	6.056 ± 0.026	1.893E+00 $\pm 7.702E-02$
Augmentation + SAM	0.594 ± 0.000	**0.056** ± 0.002	0.440 ± 0.000	5.796 ± 0.010	1.665E+00 $\pm 5.139E-02$
Weight Decay	0.604 ± 0.001	0.265 ± 0.000	0.222 ± 0.000	0.062 ± 0.008	2.708E-04 $\pm 1.086E-05$
Weight Decay + SAM	**0.641** ± 0.001	0.180 ± 0.001	**0.185** ± 0.000	0.103 ± 0.005	3.056E-04 $\pm 4.175E-06$

Appendix E. Vision Transformer

CIFAR10: We see Augmentation and the Augmentation + SAM conditions perform best and has the highest sharpness values across metrics.

Table 7: Results for ViT Trained on CIFAR10, the Mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Control	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness	Relative Flatness
Baseline	0.610 $\pm$0.002	0.308 $\pm$0.002	54.805 $\pm$0.147	0.408 $\pm$0.001	0.049 $\pm$0.001	9.428E-06 $\pm$1.101$E-06$	347.198 $\pm$6.425
Baseline + SAM	0.600 $\pm$0.001	0.276 $\pm$0.001	54.792 $\pm$0.113	0.421 $\pm$0.001	0.350 $\pm$0.018	2.005E-04 $\pm$5.654$E-05$	1459.292 $\pm$82.220
Augmentation	**0.724** $\pm$0.001	**0.019** $\pm$0.001	**64.092** $\pm$0.152	0.217 $\pm$0.001	5.064 $\pm$0.019	5.800E-02 $\pm$6.264$E-03$	38465.647 $\pm$139.905
Augmentation + SAM	0.668 $\pm$0.002	0.030 $\pm$0.001	60.535 $\pm$0.179	**0.201** $\pm$0.001	4.985 $\pm$0.011	8.588E-02 $\pm$4.551$E-02$	18412.664 $\pm$617.822
Weight Decay	0.613 $\pm$0.002	0.301 $\pm$0.002	55.077 $\pm$0.159	0.402 $\pm$0.001	0.073 $\pm$0.001	1.939E-05 $\pm$1.929$E-06$	422.966 $\pm$6.897
Weight Decay + SAM	0.600 $\pm$0.002	0.268 $\pm$0.001	54.797 $\pm$0.125	0.419 $\pm$0.001	0.500 $\pm$0.022	2.708E-04 $\pm$2.804$E-05$	1908.688 $\pm$97.800

CIFAR100: We see Augmentation and the Augmentation + SAM conditions perform best and has the highest sharpness values across metrics.

Table 8: Results for ViT Trained on CIFAR100, the Mean and $\pm$ 1 SEM are recorded over 10 models. Numbers in bold indicate best scores for metrics. For sharpness metrics lower values represent flatter models.

Condition	Test Accuracy	Test ECE	Corruption Accuracy	Prediction Disagreement	Fisher Rao Norm	SAM Sharpness	Relative Flatness
Baseline	0.309 $\pm$0.002	0.402 $\pm$0.002	25.936 $\pm$0.088	0.723 $\pm$0.000	0.144 $\pm$0.013	3.145E-04 $\pm 4.655E-05$	112.185 $\pm$4.246
Baseline + SAM	0.326 $\pm$0.001	0.386 $\pm$0.001	27.628 $\pm$0.097	0.697 $\pm$0.000	0.182 $\pm$0.016	1.622E-03 $\pm 1.249E-03$	124.472 $\pm$30.314
Augmentation	0.508 $\pm$0.001	0.227 $\pm$0.001	38.680 $\pm$0.091	0.483 $\pm$0.001	3.858 $\pm$0.048	5.264E-01 $\pm 5.249E-02$	17401.462 $\pm$143.009
Augmentation + SAM	**0.523** $\pm$0.001	**0.146** $\pm$0.002	**40.275** $\pm$0.097	**0.446** $\pm$0.000	4.364 $\pm$0.029	4.641E-01 $\pm 3.560E-02$	17812.985 $\pm$55.523
Weight Decay	0.325 $\pm$0.001	0.324 $\pm$0.001	27.364 $\pm$0.103	0.700 $\pm$0.000	0.347 $\pm$0.016	2.160E-03 $\pm 1.379E-03$	251.148 $\pm$15.330
Weight Decay + SAM	0.327 $\pm$0.001	0.284 $\pm$0.001	27.739 $\pm$0.069	0.695 $\pm$0.001	1.151 $\pm$0.058	5.322E-03 $\pm 6.859E-04$	1554.595 $\pm$91.649