Benchmarking Robustness Measures

Clara Pereira
clara.gomes.pereira@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisbon, Portugal

Joana Correia
joana.d.correia@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisbon, Portugal

Vasco Gameiro
vasco.gameiro@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisbon, Portugal

Sérgio Jesus
sergio.jesus@feedzai.com
Feedzai Supervisor
Oporto, Portugal

Inês Silva
ines.silva@feedzai.com
Feedzai Supervisor
Oporto, Portugal

Abstract

Robustness has been a growing topic of interest in machine learning research. Although numerous techniques and methods have been proposed to enhance model robustness to adversarial attacks and data distribution shifts, there is no literature discussing a single metric to holistically assess model robustness in both scenarios. In this study, we benchmark existing measures of model generalization, uncertainty, and overall performance, examining their correlations with performance decline on out-of-distribution data. We train more than 50 deep neural networks on the CIFAR-10 dataset and apply Kendall's Tau to uncover potential relationships between these metrics and out-of-distribution robustness. We conclude that from the extensive set of metrics analyzed, none is able to explain robustness of the model in the conventional sense. However, using alternative definitions of robustness, we find moderate correlations between some metrics and both out-of-distribution settings.

Keywords

Robustness, Data Drifts, Adversarial Attacks, Deep Learning

1 Introduction

Machine learning systems typically aim to learn the structure of their training data to generalize to new data points sampled from the same distribution. Although assuming consistent performance on both training and test sets is a useful baseline, real-world scenarios often require models that remain robust despite shifts in input distribution.

In this study robustness refers to a model's ability to maintain reliable performance and produce accurate predictions when exposed to adversarial perturbations and changes in data distribution. Adversarial attacks involve carefully engineered inputs designed to deceive a classifier and induce misclassification, whereas data drifts represent the natural evolution of data distributions over time.

Although there are several specialized methods and frameworks for evaluating the generalization and robustness of a model against adversarial attacks and data drifts, there remains a significant gap in the literature. Current approaches do not provide a unified framework capable of measuring model robustness across diverse scenarios. We hypothesize that a metric or a set of metrics can encompass model robustness to both adversarial attacks and data drifts.

In this paper, we address this gap by systematically evaluating and comparing existing approaches under a common set of criteria, offering insights into metrics that could guide the development of more robust models, specifically focusing on image classification. Our main contribution is a comprehensive study on a set of proxy robustness metrics and their effect on actual model robustness to adversarially attacked and shifted distribution datasets.

The key findings from our study are summarized below:

(1) Within the most widely used generalization and classification metrics, none exhibits a strong correlation with standard robustness measures across data drifts and adversarial attacks.
(2) Performance metrics in distribution (E.g. Test Accuracy, F1-Score, etc.) and Weight Norms over margins are moderately correlated with relative robustness for both adversarial attacks and data distribution shifts.
(3) The in-distribution Margin and entropy are moderately correlated with uncertainty robustness.
(4) Sharpness metrics are generally moderately correlated with relative and absolute robustness for adversarial attacks.

The code for our implementation and experiments can be found at https://github.com/vascomgameiro/Robustness-Metrics.

2 Related Work

Accurate estimation of model performance on out-of-distribution (OOD) data remains a significant challenge in current machine learning research. Previous research has proven that models with superior generalization capabilities are inherently more robust to data shifts and adversarial attacks [5, 24]. Motivated by these findings, we explore theoretically grounded metrics that estimate generalization capabilities and study their effectiveness to measure robustness holistically.

2.1 Generalization Metrics

A model is said to generalize well if it maintains its performance on unseen, in-distribution data. To shed light on why some models generalize better than others, researchers have sought metrics that capture the model's specific behaviors, such as complexity and uncertainty.

Metrics computed directly from the network's output, including margin, entropy, and cross-entropy, have been foundational in evaluating model generalization. Jiang et al. [11] demonstrated that larger margins, lower cross-entropy, and higher entropy are often associated with better generalization. These findings support the intuitive notion that models confident in their predictions (i.e., having high margins and entropy) tend to perform better on data similar to the training set.

However, subsequent studies have highlighted limitations of these output-based metrics in more complex settings. Carlini et al. [25] observed that large margins become less reliable predictors of generalization in highly parameterized models. Additionally, while lower cross-entropy and higher entropy indicate generalization in both in-distribution and adversarial scenarios, these metrics often fail to account for robustness under distributional shifts [4, 24].

An alternative to output-based metrics are norm-based metrics, which focus on weight magnitudes, network architecture interactions, and the distance between untrained and trained models. Neyshabur et al. [19] linked theoretical bounds for generalization in deep learning to network complexity and optimization procedures. However, their effectiveness has been critiqued. The success of norm-based metrics has been attributed to inherent model complexity rather than a direct causal relationship with generalization [11, 19].

Among the most promising generalization metrics are sharpness-based measures, such as the original sharpness measure, magnitude-aware sharpness (MagSharpness), PAC-Bayes sharpness, and magnitude-aware PAC-Bayes sharpness (MagPAC). These metrics assess the sensitivity of a model's loss landscape and have consistently demonstrated strong empirical correlations with generalization [4, 11]. By evaluating how flat or sharp the minima are in the loss landscape, sharpness-based metrics offer a more nuanced understanding of model robustness.

2.2 Robustness to Data Shifts

Despite ongoing research on out-of-distribution robustness, there remains a lack of universally applicable metrics that can accurately estimate robustness to natural data shifts. Many existing studies focus on specific domains, thereby limiting the broader applicability of their findings. Nevertheless, a consistent trend observed in the literature is the strong linear relationship between in-distribution (IID) accuracy and accuracy on naturally shifted data [5].

In addition to performance-based metrics, new perspectives advise assessing uncertainty in natural data shifts scenarios [16]. In complex, real-world settings, evaluating accurate predictions alone is insufficient. The reliability of these predictions may deteriorate due to increased uncertainty resulting from changes in the underlying data distribution.

2.3 Adversarial Robustness

While data drifts represent natural shifts in input distributions, adversarial attacks involve intentionally perturbed inputs designed to mislead classifiers. These attacks exploit vulnerabilities in a model's decision boundary, making an understanding of the underlying optimization problem critical for improving robustness.

Insights into the loss landscape have proven fundamental in comprehending model resilience to adversarial attacks. Stutz et al. [24] demonstrated that flatter average and worst-case flatness during model training enhanced adversarial robustness by reducing a model's sensitivity to perturbations. Although techniques like Adversarial Training with Weight Perturbations (AT-AWP) are beyond the scope of our research, these findings underscore the potential of generalization-focused metrics to explain adversarial resilience.

3 Methodology

Evaluating model robustness comprehensively is a challenging task that demands both rigor in experimental design and attention to practical transferability. In this study, we strategically select datasets, models, and out-of-distribution scenarios to ensure that the results are both representative and meaningful. Each methodological decision, ranging from model architectures to statistical evaluation methods, is grounded on previous studies. By systematically evaluating three-classes of metrics and their relationship with model robustness, we contribute to the broader literature and provide insights that can inform future studies.

3.1 Choosing a Dataset

Given the aim of this study, the training dataset choice is an important one. Models should be trained on a widely benchmarked dataset, that is diverse and non-trivial, while respecting computational constraints. With all these requirements in mind, a good first candidate was Tiny ImageNet, a lower-resolution, downsized version of ImageNet [3].

For reliability and correctness of results, only classes common to all test sets should be seen during training. Due to this fact a considerable amount of training examples was discarded, in order to induce compatibility with an OOD test set.

However, and despite this significant reduction in dataset size, computational challenges arose. The trade-off between simplicity of the model and its overall performance posed as an impediment to insightful analysis: only more complex models were able to achieve acceptable results, and these required unreasonable training times given our resources.

As a result, we selected what we believe to be the next best candidate for our research: CIFAR-10 [13]. Although it does not offer as wide an array of out-of-distribution alternatives as the original dataset, CIFAR-10's more modest size, both in terms of examples and classes, makes it better suited to our computational constraints.

To further accommodate these constraints, we reduced the dataset size from 60,000 examples to 30,000. For model development, we used an $80\% - 20\%$ train-test split, followed by a $90\% - 10\%$ split of the training set into training and validation subsets.

3.2 Model Training

To evaluate the robustness on a heterogeneous set of models, we conducted experiments using custom convolutional neural networks (CNNs) designed with varying architectural specifications.

We chose 5 different hyperparameters related to optimization and architecture design (depth and width of the network, dropout, learning rate, optimizer), with 2 to 3 choices for each parameter. With this configuration $2^4 \times 3 = 48$ models were generated and trained on the CIFAR-10 dataset.

Additionally, a custom architecture inspired by VGG models [23] was implemented and trained with different optimization hyperparameters (with the same possible choices as the aforementioned models), resulting in 6 more models.

The models were trained with resort to early stopping: a model was trained for a maximum of 100 epochs, with a patience of 10 epochs. For early stopping criterion we chose validation set's accuracy, since we are interested in training the model for the best generalization capability possible, within reasonable training time. This training approach not only avoids overfitting, but also accelerates the training process, which is crucial given the computational resources available. In total, our pool of models that will be used to assess robustness on distributional shifts consists of 48 + 6 deep-learning models.

For more details on architecture and hyperparameter choices, please refer to Appendix A.

These models achieved a broad range of in-distribution accuracy values (Figure 1), indicating different model predicting capabilities.

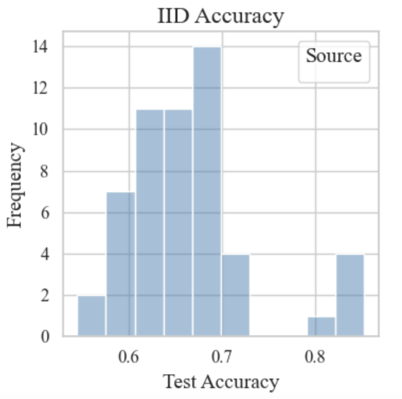

Figure 1: Test Accuracy distribution

3.3 Robustness Scenarios

In previous literature, the most common approach to quantify robustness against adversarial and distributional shifts is to measure the *accuracy gap*. Let T_{IID} denote an in-distribution test set drawn i.i.d. from the training distribution $\mathcal{D}$, and let T_{OOD} denote a test set sampled from a potentially different distribution $\mathcal{D}_{OOD}$. Consider a model M trained on data sampled from $\mathcal{D}$ (that is independent from T_{IID}). We measure the accuracy of M on any dataset T via $\text{Acc}(M, T)$. The accuracy gap, $\delta_{\text{Acc}}(M)$, is then defined as

$$\delta_{\text{Acc}}(M) \;=\; \text{Acc}(M, T_{\text{IID}}) \;-\; \text{Acc}(M, T_{\text{OOD}}) \tag{1}$$

Although δ_{Acc} captures the absolute drop in accuracy between in-distribution and out-of-distribution sets, it does not account for the *proportional* drop. For instance, a 25% gap in accuracy affects a model differently if $\text{Acc}(M, T_{\text{IID}})$ is 95% rather than 65%. Therefore, we also propose measuring the *relative* or *proportional* accuracy drop,

$$\delta_{\text{Rel}}(M) \;=\; \frac{\text{Acc}(M, T_{\text{IID}}) \;-\; \text{Acc}(M, T_{\text{OOD}})}{\text{Acc}(M, T_{\text{IID}})} \tag{2}$$

This formulation ensures that the same absolute gap is contextualized by the model's baseline performance on T_{IID}.

Finally, to obtain a more comprehensive perspective, we extend both δ_{Acc} and δ_{Rel} to analogous Log-loss gaps, offering a fuller view of how each model certainty degrades under adversarial and distributional shifts.

In this study, we examine three out-of-distribution scenarios - Adversarially Perturbed Sets, Natural Shifts, and Synthetically Corrupted Sets (a controlled form of distributional shift).

3.3.1 Adversarial Attacks. To evaluate each model's susceptibility and robustness under targeted perturbations, we subject them to five white-box adversarial attacks - FGSM [6], PGD [15], CW [2], DeepFool [17], and BIM [14].

These popular attacks, implemented via the `torchattacks` library, have full access to model parameters and apply gradient-based manipulations to induce misclassification. A demonstration of fast adversarial example generation (FGSM) applied to ImageNet is shown in Figure 2. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can induce miss classification [6].

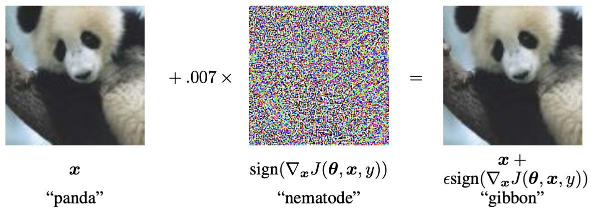
+ .007 ×
=
$\boldsymbol{x}$
"panda"
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"

Figure 2: Visual representation of a FGSM attack - from [6]

For a fair assessment of accuracy gaps, we apply these perturbations to the original IID test set, generating five new test sets. Because most of our models were not adversarially trained, default attack settings resulted in near-zero accuracy. Consequently, we fine-tuned key hyperparameters (e.g., perturbation magnitude, iteration count) to ensure a heterogeneous range of accuracy drops. For additional details on the attacks, please refer to Appendix B.

3.3.2 Natural Shifts. CIFAR-10.1 [22] offers a natural distributional shift relative to the standard CIFAR-10 dataset. It consists of 2,000 images sampled from the same underlying distribution several years after CIFAR-10's creation, and was designed to serve as an independent benchmark for generalization. Empirical evidence reports significant accuracy drops across multiple models, which was attributed to a "minute" distribution shift [22].

3.3.3 Synthetically Corrupted. CIFAR-C [7] is an artificially generated extension of CIFAR-10 designed to evaluate model performance under common real-world corruptions (e.g., Gaussian noise, blur, weather effects, digital distortions). The dataset contains 19 corruption types, each with five severity levels, increasing from mild to severe. Although the corruptions are synthetic, they simulate real-world degradation patterns such as camera noise and environmental conditions, therefore serving as a controlled approximation of distributional shifts. Five new datasets - one for each level of severity - were created based on CIFAR-C's corruptions. These contain the same images as the IID test set, similarly to what was done with the adversarial attacks.

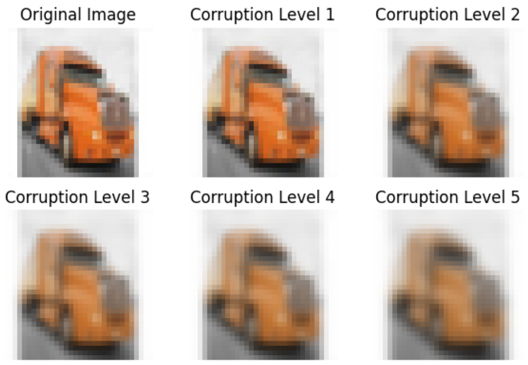

Figure 3: Different levels of “Gaussian Blur" corruption for image with label “Truck"

In the analysis we will refer to these corrupted datasets as Test_10ci , with i being the respective corruption level of each set.

3.4 Evaluation Criteria

To evaluate the quality of a robustness measure μ, it is essential to consider both accuracy and log-loss gaps in out-of-distribution scenarios. While accuracy gaps are a standard metric for assessing model robustness, they primarily reflect the model’s ability to predict the correct class. However, they fail to capture the confidence and calibration of predictions - a critical aspect in high-stakes fields such as healthcare and finance. Log-loss, on the other hand, provides a more comprehensive perspective by quantifying the certainty of model predictions. Furthermore, evaluating proportional gaps allows us to understand how robustness metrics scale relatively to their baseline performance, enabling fair comparisons across models with varying initial accuracies.

To quantify the concordance between the robustness measure μ and the observed robustness gaps δ, we use Kendall’s rank correlation coefficient τ. Kendall’s τ is particularly well-suited for our study as it is more interpretable and robust to outliers compared to alternatives like Spearman’s ρ. It also offers better handling of smaller datasets, making it appropriate for our scenarios, where the number of models and experimental conditions is limited. Unlike Pearson’s correlation, which assumes linear relationships and is sensitive to outliers, Kendall’s τ focuses on ordinal associations and is non-parametric. Additionally, Kendall’s τ has been widely used in similar studies, reinforcing its validity as a standard for comparing robustness metrics [11, 26].

4 Exploratory Data Analysis

Before conducting our correlation analysis, we performed an exploratory data analysis (EDA) of our models’ performance across the chosen out-of-distribution test sets. This initial analysis encompassed both univariate and bivariate approaches to comprehensively understand the distribution and relationships of key performance metrics.

For the univariate analysis, we calculated descriptive statistics and constructed histograms for each OOD test set’s accuracy, and log-loss, as well the respective absolute gap, and proportional gap. The analysis yielded two significant observations: the performance metrics did not exhibit normal-like distributions across the test sets, reinforcing the use of a non-parametric correlation coefficient (see Figure 13); the DeepFool adversarial attack resulted in consistently low and nearly homogeneous accuracies, ranging between 2% and 4%, making it a candidate set to be excluded from the analysis.

To explore the relationships between in-distribution performance and OOD performance, we plotted a grid of scatter plots comparing in-distribution accuracy with OOD accuracies across different test sets (Figure 4). The scatter plots revealed a linear relationship between in-distribution and various OOD accuracies, corroborating previous studies that identified similar linear correlations [5]. Notably, the DeepFool attack emerged as a clear outlier, displaying minimal dependence on in-distribution performance. Additionally, VGG models consistently outperformed other architectures across all OOD test sets.

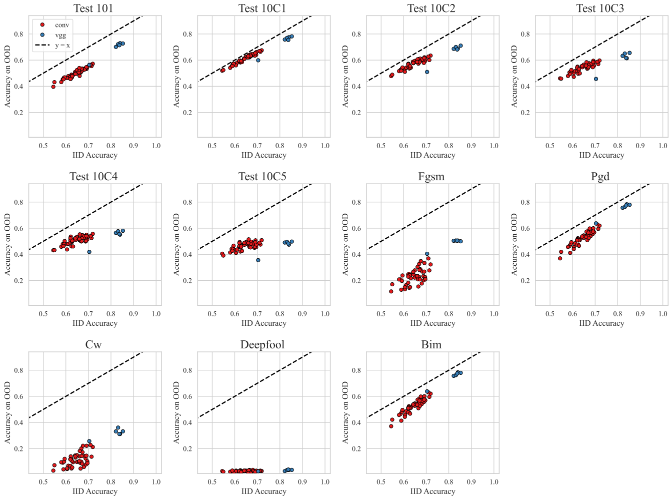

Figure 4: Scatter Plot: Out-of-Distribution vs In-Distribution Test Accuracies

With the objective of determining whether robustness to data drifts, corruptions, and adversarial attacks displayed similar patterns, we computed correlation matrices to assess the relationships between the different gaps across different OOD test sets (See Figure 5).

The correlation analysis revealed that robustness to synthetic corruptions does not generalize to natural distribution shifts. Performance on CIFAR-10.1 was entirely uncorrelated with corrupted test sets, contradicting our initial assumption that corrupted data would exhibit similar behavior to real-world shifts. As including these sets would introduce extraneous noise and given that robustness to corruptions lies outside our primary research scope — we excluded them from further analysis.

Concerning the adversarial sets, unexpectedly, there was only a weak correlation between robustness to data shifts and adversarial attacks, excluding the DeepFool attack (See 6). Among the adversarial attacks, Projected Gradient Descent (PGD) and Basic Iterative

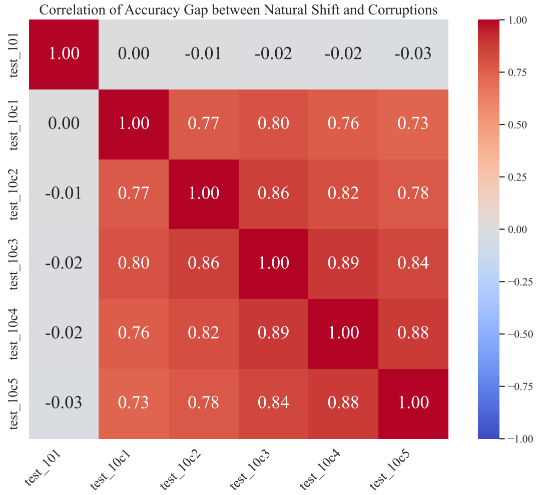

Figure 5: Correlation matrix: Accuracy-Gaps between Natural and Corrupted Sets

Method (BIM) displayed an expected almost perfect correlation, due to their similar iterative approach.

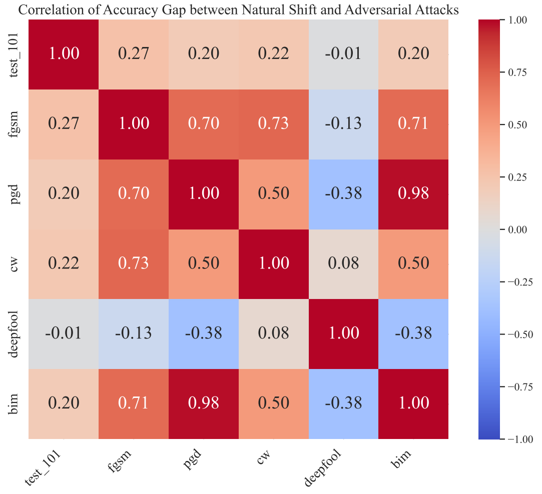

Figure 6: Correlation matrix: Accuracy-Gaps between Natural Shift and Adversarial Attacks

Additionally, the DeepFool attack's extreme severity resulted in nearly uniform and extremely low accuracy scores, which undermined its utility for further analysis by reducing variability in performance metrics. Consequently, we decided to exclude the DeepFool attack from subsequent analyses. This exclusion ensures that our analysis remains relevant and avoids skewed results caused by overly aggressive adversarial perturbations. The DeepFool results further validate our choice to tune adversarial attacks to be less severe, as we aim to achieve a more comprehensive representation of robustness that aligns with the natural variations observed in real-world scenarios.

5 Performance of Robustness Metrics

5.1 Baseline Metrics

Before exploring more sophisticated metrics, we first examined baseline metrics derived directly from the models' outputs. We investigated two main categories: standard classification (e.g., Accuracy, F1-Score, Precision, Recall) and uncertainty-related metrics (e.g., ROC AUC, Entropy, Log-Loss). Additionally, we computed Conditional Value at Risk (CVaR), a widely researched metric in robustness literature. To visually assess the correlation between out-of-distribution performance across different scenarios and these baselines, we generated correlation heatmaps.

5.1.1 Results. The analysis between the baseline metrics and **accuracy gap** δ_{Acc} across the different scenarios resulted in predominantly non-significant correlations (see Figure 14). Specifically, no substantial correlations were observed between the baseline metrics and distribution shifts or most adversarial attacks. There were exceptions, in the context of adversarial robustness, where most metrics exhibited moderate correlations with PGD and BIM accuracy gaps, with correlation coefficients ranging from -0.37 to 0.39.

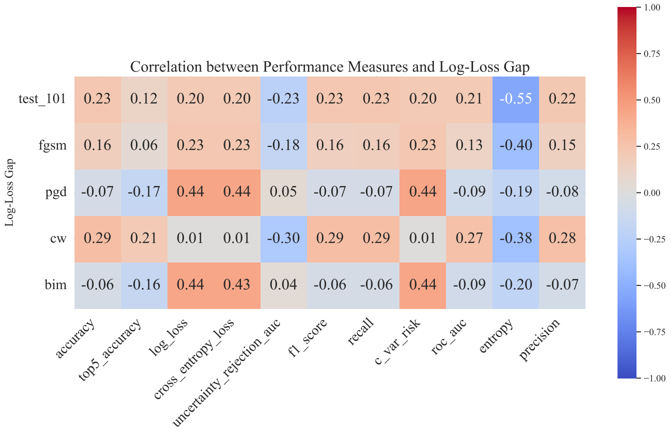

Figure 7: Correlation Matrix: Performance Metrics vs Log-Loss Gap

In contrast, the evaluation of **log-loss gaps** revealed more noteworthy relationships (see Figure 7). Notably, entropy demonstrated moderate to high correlations with most scenarios, achieving the strongest correlation with the shift gap at a coefficient of -0.55. In adversarial settings, correlations were less consistent, with approximately half of the attacks showing moderate correlations around -0.40 and the remaining attacks exhibiting weaker correlations closer to -0.20. The negative coefficients indicate that higher entropy values are associated with smaller log-loss gaps and thus a higher robustness. This finding aligns with the existing literature [11, 16].

For **proportional accuracy gaps**, the results were more consistent (see Figure 8). Uncertainty measures exhibited moderate positive correlations, whereas classification metrics showed negative correlations across all scenarios. These findings suggest that models with better in-distribution performance experience smaller proportional drops in both data shift and adversarial attack scenarios. This conclusion reinforces our earlier observation that superior in-distribution performance correlates with enhanced out-of-distribution robustness.

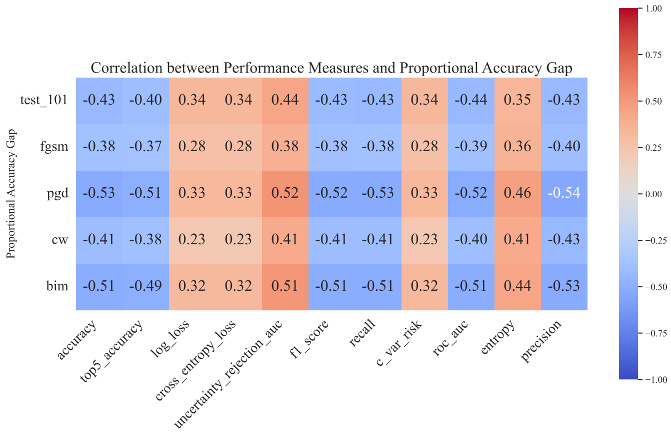

Figure 8: Correlation matrix: Performance Metrics vs Proportional Accuracy Gap

The **proportional log-loss gap** displayed behaviour similar to both the standard log-loss gap and the proportional accuracy gap, with most performance metrics showing moderate correlations across all scenarios (see Figure 17). Entropy remained the most consistently correlated metric, while both BIM and PGD demonstrated only weaker correlations.

Overall, traditional baseline performance metrics exhibit limited correlation with robustness. However, certain uncertainty-related metrics, particularly entropy, show significant correlations with log-loss gaps, suggesting their potential utility in predicting robustness in specific contexts. Additionally, our findings support the strategy that developing models with strong in-distribution generalization is the most effective approach for achieving robust performance under out-of-distribution scenarios.

5.2 Norm and Margin-Based Metrics

Norm-based metrics on a network's parameters are frequently proposed as indicators of model complexity [11], which is believed to be associated with generalization performance [1, 20]. In this study, we examine the following norms and their variants leveraging the margin γ : Spectral Norm, Frobenius Norm, Trace Norm, and Path Norms.

The margin γ of a model is calculated using its output logits [10]. Specifically, for an input X with target class y, the margin of the output $M(X)$ is defined as:

$$\gamma(X) = M(X)[y] - \max_{i \neq y} M(X)[i] \tag{3}$$

where $M(X)[y]$ is the logit corresponding to the target class, and $\max_{i\neq y} M(X)[i]$ is the highest logit among the non-target classes. The margin γ is then taken as the 10th percentile of these margins across all outputs. Previous work by [10] has indicated that margin is a predictor of the generalization gap in in-distribution settings.

In the work of [11] all norm and margin measures are divided by squared margin. We divide only by the margin in most of the implementations, due to specific constraints of this study: given the heterogeneity of our pool of models, some obtained less than 90% in-distribution accuracy, in which case negative values for γ appear. In the negative domain, a larger absolute value indicates the model has high confidence in the wrong prediction. If we square the margin, we lose perception of this fact: a large γ^2 is an indicative of a very certain model, however it is not explicit in which "direction" it is certain, i.e., if it outputs correct predictions with high certainty or miss-classifies examples with high certainty. Note that in the work of [11] all of the models studied achieved more than 99% accuracy, so the discussion on negative margins was not relevant. Nevertheless, some metrics were left with squared margin for coherence with previous work.
For more details on norm and margin-based metrics, please refer to Appendix C.2.

5.2.1 Results. The **accuracy gap**, our standard robustness metric, exhibited weak to moderate correlations with norm-based measures (see Figure 16). None of the metrics demonstrated a strong relationship with robustness to data shifts, with most correlations being negligible. Notably, metrics that combine norms with margins, such as the *Spectral Norm over Margin*, showed moderate correlations with PGD and BIM accuracy gaps, with correlation coefficients ranging from −0.33 to −0.42. This suggests that these metrics can partially capture adversarial robustness.

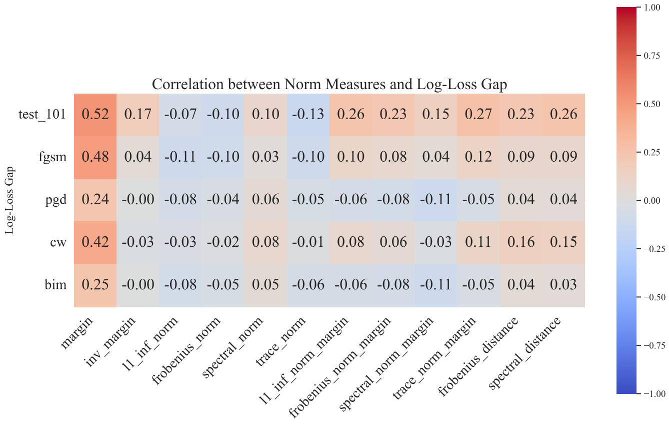

Figure 9: Correlation Matrix: Norm Metrics vs Log-Loss Gap

The analysis of the **log-loss gap** revealed more varied results (see Figure 9). The *margin* itself exhibited the strongest correlation with the shifted dataset, achieving a coefficient of 0.52. In adversarial settings, correlations were less consistent, with approximately half of the attacks showing moderate correlations around 0.45, while the remaining attacks exhibited weaker correlations closer to 0.25. The positive coefficients indicate that higher values of margin are associated with worse log-loss robustness. Which does not necessarily contradict the notion that larger margins contribute to better generalization [11, 16], as those conclusions were drawn in ID settings.

For **proportional accuracy gaps**, norm-based metrics over margins provided more consistent insights (see Figure 10). The metrics demonstrated moderate negative correlations across all datasets, including both adversarial attacks and the shifted dataset. Specifically, the *Norms over Margin* achieved correlations of approximately −0.4 with proportional accuracy gaps across the five scenarios.

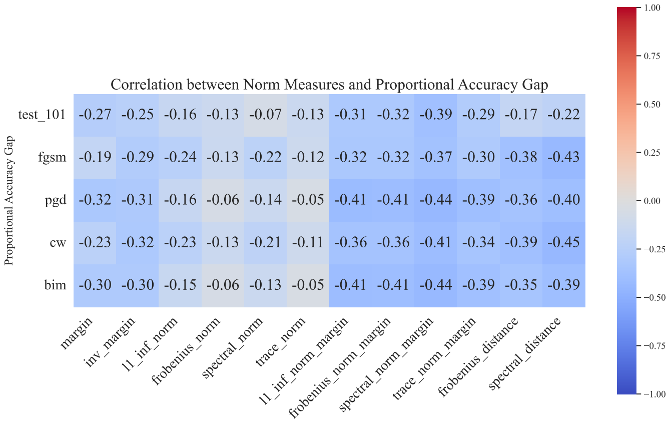

Figure 10: Correlation Matrix: Norm Metrics vs Proportional Accuracy Gap

The **proportional log-loss gap** exhibited a similar trend to the standard log-loss gap (see Figure 17). The *margin* were strongly positively correlated with $\delta_{log-loss}$, with coefficients ranging from 0.29, in the CW attacked set, to 0.63, in the naturally shifted set.

Although no norm-based measure in this study successfully correlated with the standard robustness measure, some metrics show potential for specific use cases. The *margin* demonstrated promise in estimating both absolute and relative robustness, particularly in scenarios where uncertainty is critical. Additionally, all the *norm over margin* measures yielded promising results, showing moderate positive correlations with relative accuracy robustness.

5.3 Sharpness Metrics

The notion of sharpness is closely related to the loss function landscape. Training a model can be reduced to the search of a minimum of the loss function in weight space. Intuitively, a very steep function will be more sensitive to perturbations of its input than a flatter one. Hence, when training a deep network, if one finds a sharp minimum, small perturbations to its weights will significantly (and most likely negatively) affect the model's predictions and consequently its accuracy. Flatter minima in the loss function has been associated with both IID generalization [8, 12], and adversarial robustness [24].

One way to measure this sharpness is through PAC-Bayesian methods, which have been showing promising insights in studying deep network generalization [11, 18]. Following the work of [11], we evaluate sharpness by finding σ_{PAC_Bayes} and $\sigma_{Sharpness}$. Generally, σ is the largest number which will cause a deviation of at most 10% in the model's accuracy:

$$\mathrm{E}_{u \sim N(\mu, \sigma^2 I)}\left[Acc(M_{w+u})\right] \leq 0.1$$

The current implementations of the Binary Search used to estimate σ_{PAC_Bayes} consist of the following key steps:

(1) **Applying Gaussian Noise:** Gaussian noise is applied to the model's parameters based on the current value of σ at a given step of the binary search process.
(2) **Estimating Perturbed Accuracy:** The accuracy of the model with perturbed weights is computed to assess the impact of the applied noise on the model's performance.
(3) **Adjusting Search Bounds:** Depending on the observed accuracy gap between the perturbed and original models, the upper and lower bounds for σ are updated to refine the search range.

The $\sigma_{Sharpness}$ search builds upon this procedure by introducing gradient ascent steps on the perturbed model in an attempt to assess worst-case scenario deviations. For more details and notation, please refer to Appendix C.3

A limitation of standard sigma search methods is that perturbing weights without considering their magnitudes can result in unrealistic parameter changes, such as sign flips, which can drastically alter the loss landscape. To mitigate this, we adopt magnitude-aware perturbations, which scale perturbations relative to the magnitude of each parameter, and therefore maintain the stability of the parameter signs [11].

5.3.1 *Implementation Challenges.*

Implementing sharpness-based metrics presented several challenges due to the absence of robust and well-documented implementations in popular frameworks. Existing implementations in TensorFlow and Skorch were not only poorly documented but also difficult to integrate into our PyTorch-based codebase. Additionally, discrepancies between the pseudocode provided in [11] and available implementations introduced further obstacles.

Another significant limitation in the literature is the lack of guidance for tuning the numerous hyperparameters involved in sharpness calculations. To address this, we meticulously monitored the sigma search process for a subset of randomly selected models and fine-tuned hyperparameters to ensure convergence while minimizing computational overhead. For instance, during the sharpness sigma search, we found that, for our models, the learning rate for gradient ascent needed to be reduced by a factor of 100 for models without schedulers and by 1,000 for models with schedulers. Using the training learning rate resulted in convergence to upper or lower bounds, hindering effective sigma optimization.

Furthermore, we identified a limitation in the standard sharpness sigma implementation, where the norm of the difference between original and perturbed weights was normalized to match sigma after each gradient ascent step. This normalization often caused sigmas to converge to lower bounds or attain values too small for effective gradient steps. To overcome this, we incorporated weight clipping as utilized in MAG Sigma Search. The primary difference between the two methods lies in the noise generation process: one approach is agnostic to the weights being perturbed, while the other accounts for them, as in the PAC-Bayesian search.

5.3.2 Results. Sharpness-based metrics exhibited mixed correlations with **accuracy robustness** to data shifts and adversarial attacks. Most metrics, including PAC-Bayes and standard sharpness measures, showed non-significant correlations with robustness to data shifts 11. However, in adversarial settings, moderate correlations were observed. Notably, *Sharpness_MAG_Sigma* consistently achieved a correlation coefficient of approximately −0.40, indicating a moderate inverse relationship between sharpness and adversarial robustness.

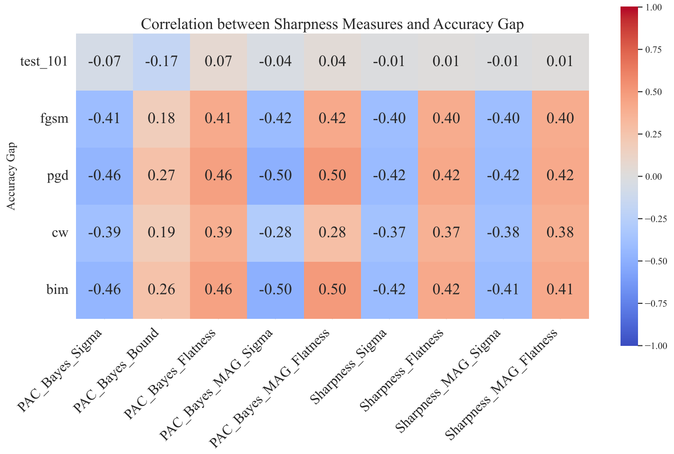

Figure 11: Correlation Matrix - Sharpness Measures vs Accuracy Gaps

The correlations between sharpness-based metrics and **log-loss gaps** were generally weak (see Figure 18. For shifted datasets, correlations ranged from low to moderate, with absolute coefficients between 0.18 and 0.39. The PAC-Bayes Bound achieved a correlation of −0.39, suggesting a relationship between log-loss robustness and sharpness. In adversarial settings, sharpness-based metrics exhibited no significant correlations with log-loss gaps, underscoring their limited ability to capture uncertainty-based robustness.

Proportional Gaps: Similar trends were observed for proportional gaps (see Figure 12 and 19). Proportional accuracy gaps showed moderate correlations with adversarial robustness but weak correlations with robustness to data shifts. Proportional log-loss gaps demonstrated generally weak correlations across both adversarial and shifted datasets. These findings suggest that while flatter loss landscapes may marginally correlate with adversarial robustness, their association with relative performance robustness under data shifts is limited.

Overall, our results indicate that sharpness-based metrics are not well-suited for comprehensively measuring model robustness. Although they align with prior studies in capturing adversarial robustness, their performance concerning standard and proportional log-loss gaps is negligible. A potential explanation for these results lies in the sigma search methodology, which optimizes sigma to induce a specific accuracy gap (10%) through perturbations. Consequently, the sigma is tailored to accuracy-based robustness, which may not effectively capture log-loss-based robustness.

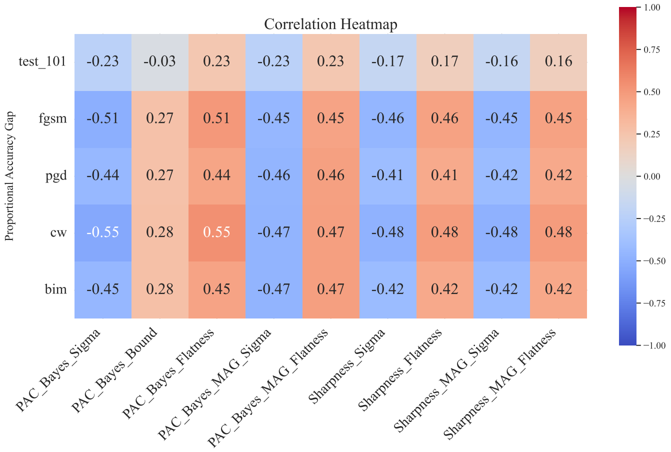

Figure 12: Correlation Matrix - Sharpness Measures vs Proportional Accuracy Gap

6 Conclusions

In this study, we investigated whether a single metric could reliably predict both adversarial and data drifts robustness. To this end, we evaluated a broad range of generalization, uncertainty, and classification metrics on models tested against adversarial attacks and a naturally shifted dataset. Our findings suggest that no single measure fully captures a model's robustness across these distinct domains. Despite the absence of a universal metric, our analysis resulted in several noteworthy insights. We proposed extending the conventional accuracy gap (δ_{Acc}) to both relative accuracy gap(δ_{Rel}) and log-loss gaps, consequently capturing the proportional impact of performance loss and quantifying the increase in uncertainty. Our results indicate that, in scenarios where predictive certainty is critical, models displaying lower in-distribution margins and higher entropy exhibit increased robustness to uncertainty, in both adversarial and distribution-shift contexts. Meanwhile, in terms of proportional performance, models with stronger in-distribution performance and higher weight norms over margin tend to maintain stability relative to their baseline accuracy. Finally, our findings extend previous observations on sharpness correlating well with adversarial robustness, to also being predictive of relative robustness.

7 Limitations and Future Work

Our study lays the groundwork for exploring holistic metrics that unify the evaluation of robustness against adversarial attacks and data shifts. While our findings provide valuable insights, they also highlight significant gaps and open questions that merit further investigation, as we believe there is considerable potential for future research in this area.

A key limitation of our study was the computational power constraints faced when training models, forcing us to rely on simpler CNN configurations. Future work should consider a wider range of model architectures and training regimes to ensure more heterogeneous results, as well as potentially uncover new behaviors. Ideally, this would include experimenting with pre-trained models, adversarially trained networks, and cost functions specifically designed to enhance model robustness.

Another major challenge in this study is the lack of naturally shifted datasets for CIFAR-10. This limitation constrains the transferability of our insights, as the absence of additional naturally shifted data does not allow us to confirm or contest our findings with further evidence. More comprehensive studies should replicate our experiments on larger and more varied image datasets - such as ImageNet, SVHN, or MNIST - to evaluate the generality of our conclusions. Given the limited availability of naturally shifted versions for many datasets, future research could also focus on improved synthetic generation techniques that better mimic real-world distributional changes. While our work centres on image classification, extending it to natural language or tabular tasks could enhance our understanding of how robustness metrics perform in different domains.

In the context of adversarial attacks, only white-box attacks were analysed in this study, as they are among the most commonly researched approaches in the field. However, expanding this analysis to include other attack types, such as black-box methods, could offer additional insights into model robustness under varying levels of adversarial knowledge. Black-box attacks, which lack direct access to model parameters or gradients, may represent a more realistic scenario in many practical applications. This study did not, of course, test all available generalization metrics. Given the extensive range of metrics in the literature, it remains an open question whether untested measures could yield better results. Furthermore, a more in-depth investigation into appropriate hyperparameter settings for these metrics would be beneficial, as our choices, particularly for sharpness-based metrics, were shaped primarily by computational feasibility.

Our notions of robustness were also limited. Although we expanded the traditional perspective by examining relative and uncertainty-based robustness, we still relied on comparing performance gaps between IID and OOD data. More sophisticated approaches could consider noisy gaps, subgroup and worst-group robustness, or baseline models that provide more meaningful points of reference. Lastly, the experimental rigor of this study was constrained by available computational resources. Future research should involve multiple random seeds during model training, repeated trials for sigma search to reduce sampling noise, and repeated metric evaluations. These steps would support the calculation of confidence intervals and enable more robust statistical testing, ultimately enhancing the reliability of the conclusions drawn.

References

[1] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks, 2017.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[4] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733, 2020.

[5] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, January 1997.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[10] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions, 2019.

[11] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[12] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[14] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[15] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

[16] Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

[17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[18] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning, 2017.

[19] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[20] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks, 2015.

[21] Konstantinos Pitas, Mike Davies, and Pierre Vandergheynst. Pac-bayesian margin bounds for convolutional neural networks, 2018.

[22] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10?, 2018.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[24] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7807–7817, 2021.

[25] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[26] Aleksandar Vakanski and Min Xian. Evaluation of complexity measures for deep learning generalization in medical image analysis. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, October 2021.

A Custom CNN architecture and hyperparameters

The CNN models followed a simple architecture based on stacking blocks of convolution modules, which were followed by fully connected layers.
The blocks were comprised of a single convolution layer with kernel size 3, stride 1, and padding that preserves dimensions, followed by a maxpooling layer of kernel size 2 and stride 2, without padding. After each hidden fully connected layer there was a possible probability of dropout. ReLu activation and batch normalization [9] was always performed after convolution layers and fully connected layers.
The choices for each hyperparameter are specified below:
Optimizers: {Stochastic Gradient Descent (SGD), Adam}
Learning Rate: $\{10^{-2}, 10^{-3},$ "Scheduler"$\}$. If Learning Rate = "Scheduler", then it will start as 10^{-2} and be divided by a factor of 2 if there is no improvement on validation loss for more than 2 epochs.
Depth: $\{2, 4\}$. Refers to both convolution blocks and hidden fully connected layers: for example, if depth = 4, the model will have 4 convolution blocks and 4 fully connected hidden layers.
The following hyperparameters are dependent on the model's depth:
Depth 2
Dropout: $[\{0, 0\}, \{0.5, 0.2\}]$
Width: $\{[8, 16, 160, 80], [32, 64, 320, 160]\}$. As width we consider both the nr of filters in the convolution layer and nr of nodes in the fully connected layers. Thus we present the choices as a tuple [nr filters, nr nodes].
Depth 4
Dropout: $[\{0, 0, 0, 0\}, \{0.5, 0.3, 0.3, 0.2\}]$
Width: $\{[4, 8, 16, 32, 200, 200, 160, 80], [8, 16, 32, 64, 400, 320, 160, 80]\}$

Despite its random appearance, the values chosen for widths are a compromise between having heterogeneity in the pool of models and constructing models that are trainable considering the computational constraints we faced. The previous settings yield the $2^4 \times 3 = 48$ simple models. Additionally, a VGG-inspired model was constructed and trained with the same choices for optimizer and lr as stated before, creating $3 \times 2 = 6$ different models.
Model specifications:
*Conv(kernel = 3, stride =1, filters = 64) + Maxpool(kernel = 2, stride = 2) + Dropout(0,5)

* 3 AveragePool(kernel = 2, stride = 2) 1 fully connected layer, 512 nodes 1 classification layer Batch normalization performed after every convolution.

B Adversarial Attacks Specification

While the default settings of these attacks often lead to extreme misclassifications, they were adjusted to ensure a more heterogeneous range of accuracy drops across the different models.

B.1 Fast Gradient Sign Method (FGSM)

The FGSM attack is a **one-step gradient-based method** that generates adversarial perturbations by adjusting the input in the direction of the gradient of the loss function with respect to the input image. The attack is designed to maximize the model's loss, causing misclassification with minimal perturbation.

The hyperparameters used in our analysis, taking into account the fact that these attacks are not as aggressive as their default parameters, were:

- Epsilon = 0.005
- Epsilon (ϵ): The magnitude of the perturbation applied to the input image. A higher value results in stronger perturbations and greater misclassification risk.

B.2 Projected Gradient Descent (PGD)

PGD is an **iterative attack** that applies multiple gradient steps, projecting the perturbed image back into a valid data space after each update. It is widely considered **one of the most robust gradient-based attacks**. The hyperparameters used in our analysis, taking into account the fact that these attacks are not as aggressive as their default parameters, were:

- Epsilon = 0.001, alpha = 0.005, steps = 15
- Epsilon (ϵ): The maximum perturbation allowed.
- Alpha (α): The step size at each iteration.
- Number of steps: The number of iterations for generating the adversarial example.

B.3 Carlini-Wagner Attack (CW)

The Carlini-Wagner attack is a powerful attack designed to generate adversarial examples that are visually imperceptible while being highly effective at causing misclassification. **Unlike gradient-based methods, CW optimizes a custom loss function to craft its perturbations**. The hyperparameters used in our analysis, taking into account the fact that these attacks are not as aggressive as their default parameters, were:

- lr = 0.001, steps = 20
- Learning Rate (lr): The learning rate used in the optimization procedure.
- Number of steps: The number of optimization steps to perform.

B.4 DeepFool Attack

DeepFool is an **iterative attack** designed to find the **smallest perturbation needed** to push the input image across the decision boundary of the model. It is a more precise method that computes the minimum adversarial perturbation for each image. The hyperparameters used in our analysis, taking into account the fact that these attacks are not supposed to be as aggressive as their default parameters (in this case they were), were:

- overshoot = 0.02, steps = 20
- Overshoot: A factor to increase the perturbation size for a stronger attack.
- Number of steps: The number of iterations for refinement of the perturbation.

B.5 Basic Iterative Method (BIM)

BIM is an iterative variant of FGSM. It takes the same concept as FGSM, where the input is perturbed in the direction of the gradient of the loss function, but applies it iteratively in smaller steps, projecting the result back into the allowed perturbation space after each step. It is similar to PGD but with a slightly different approach to controlling the perturbation strength. Some authors even refer to BIM as a specific case of PGD with fixed parameters as this second one offers more flexibility in terms of step size and projection strategies

- Epsilon = 0.001, alpha = 0.01, steps = 20
- Epsilon (ϵ): The magnitude of the perturbation applied to each image.
- Alpha (α): The step size for each iteration.
- Number of steps: The number of iterations for generating the adversarial example.

C Metrics

C.1 Performance Metrics

The following metrics were computed to evaluate the model's performance. Simpler metrics such as Precision, Recall, Accuracy and F1-Score are omitted from the following mathematical formulation and explanations for each metric:

$$\text{log\_loss} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{i,y_i}), \tag{4}$$

where p_{i,y_i} is the predicted probability for the true class y_i.

$$\text{cross\_entropy\_loss} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{i,y_i} + \epsilon) \tag{5}$$

where ϵ is a small constant added for numerical stability.

$$\text{entropy}_i = -\sum_{j=1}^{C} p_{i,j}\log(p_{i,j}) \tag{6}$$

where $p_{i,j}$ is the predicted probability of class j for sample i.

Uncertainty AUC : The area under the curve (AUC) computed from the relationship between model errors and entropy as an uncertainty measure.

Conditional Value-at-Risk : Measures the expected loss in the worst-case scenarios (e.g., top 5% of losses):

$$\text{CVaR}_\alpha = \mathbb{E}[L \mid L \geq q_\alpha] \tag{7}$$

where q_α is the α-quantile of the loss distribution L, which we set to 5%.

ROC AUC: The area under the receiver operating characteristic curve, computed for multiclass classification using a one-vs-rest (OvR) approach.

These metrics were computed using logits (model outputs) and true labels. The logits were converted into class probabilities using the softmax function:

$$p_{i,j} = \frac{\exp(\text{logits}_{i,j})}{\sum_{k=1}^{C}\exp(\text{logits}_{i,k})}.$$

Errors and uncertainty metrics (entropy and uncertainty AUC) were further computed to quantify the model's robustness. These measures provide a comprehensive evaluation of both accuracy and uncertainty-based performance.

C.2 Norm Metrics

The metrics we present in this section are derived by bounds on loss functon [21], and mostly rely on various norms, sums, products and fractions of those norms Some additional notation is required to understand this section: for matrix A, $||A||_2$ stands for spectral norm, $||A||_\text{F}$ for frobenious norm, $||A||_\text{T}$ for Trace norm and $||A||_{L_{1,\text{inf}}}$ for $1, \text{inf}$ norm. THe latter is proposed by [?] and consists of .... Model M's architecture with weights w is denoted by w. The depth of a network is denoted by d.

Before calculating norm or sharpness measures, the model was reparametrized by removing batch normalization layers and updating the layers that preceded them accordingly, as done in [11].

Starting with the simplest measures, we calculate the product of a model's parameters:

$$\mu_{prod-of-spec}(M_w) = \prod_{i=1}^{d}||w_i||_2 \tag{8}$$

This metric is also calculated for frobenious, trace, and $L_{1,inf}$ norm.

We then compute the same metric, but divide it by γ:

$$\mu_\text{prod-of-spec}(M_w) = \frac{\prod_{i=1}^{d}||w_i||_2}{\gamma} \tag{9}$$

The following relates the frobenious and the spectral norm of the weights:

$$\mu_\text{fro/spec}(M_w) = \sum_{i=1}^{d}\frac{||W_i||_F}{||W_i||_2} \tag{10}$$

It also makes sense to compute the norm of the **distance** of the trained parameters to the initial parameters:

$$\mu_{frobenius-distance} = \sum_{i=1}^{d}||W_i - W_1^0||_F \tag{11}$$

$$\mu_\text{spec-init-main}(M_w) = \frac{\prod_{i=1}^{d}||W_i||_2 \sum_{j=1}^{d}\frac{||W_j - W_j^0||_F}{||W_j||_2}}{\gamma^2} \tag{12}$$

$$\mu_\text{spec-orig-main} = \frac{\prod_{i=1}^{d}||W_i||_2 \sum_{j=1}^{d}\frac{||W_j||_F}{||W_j||_2}}{\gamma^2} \tag{13}$$

$$\mu_\text{sum-of-fro}(M_w) = d\left(\prod_{i=1}^{d}||W_i||_F\right)^{\frac{1}{d}} \tag{14}$$

$$\mu_\text{sum-of-fro/margin} = d\left(\frac{\prod_{i=1}^{d}||W_i||_F}{\gamma^2}\right)^{\frac{1}{d}} \tag{15}$$

C.3 Sharpness Metrics

As aforementioned, the PAC-Bayesian framework is used to associate sharpness and generalization. The PAC-Bayesian bounds rely on the KL divergence between a prior and a posterior distribution of the model parameters to compute generalization bounds. The prior distribution is taken before observing the training set, and the posterior can be seen as adding gaussian noise to the trained model's parameters. We assume, as done in [18],[11] that these distributions are both Normal, and have the same covariance matrix: $P \sim \mathcal{N}(\mathbf{w}^0, \sigma^2 I)$, $Q \sim \mathcal{N}(\mathbf{w}, \sigma^2 I)$. Under this assumption, the KL divergence becomes simply $\frac{||\mathbf{w}^0 - \mathbf{w}||_2^2}{2\sigma^2}$. The work of [11] takes the KL term in the PAC-Bayes bound and derives multiple metrics.

In practice, σ is the largest number which will cause a deviation of at most 10% in the model's accuracy:

$$\mathrm{E}_{u \sim N(\mu, \sigma^2 I)} [Acc(M_{w+u})] \leq 0.1$$

Metrics:

$$\mu_{flatness} = \frac{1}{\sigma^2} \tag{16}$$

$$\mu_{pac-bayes-init} = \frac{||w - w^0||_2^2}{\sigma^2_{pac-bayes}} + \log\left(\frac{m}{\sigma}\right) + 10 \tag{17}$$

The flatness was computed for each sigma mentioned above - Sharpness sigma, PAC bayes sigma, Sharpness sigma Magnitude aware, PAC bayes sigma Magnitude aware - and the bound (equation 17) was only calculated for the PAC bayes sigma.

D EDA

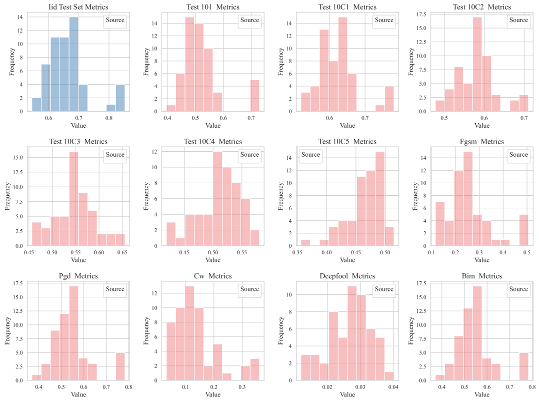

Figure 13: Histograms: Test Set Accuracies

E Correlation Matrices

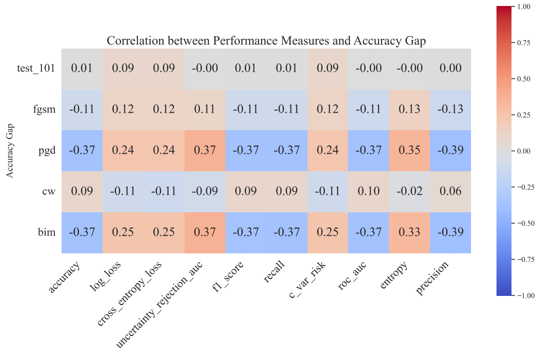

Figure 14: Correlation Matrix: Performance vs Accuracy Gap

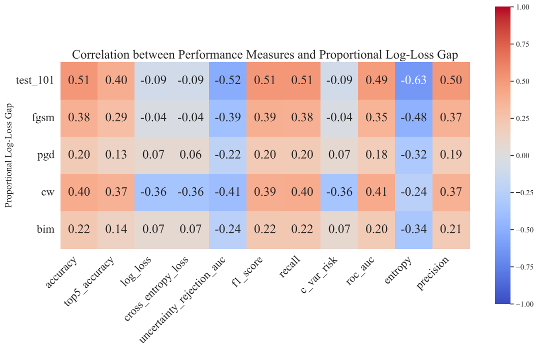

Figure 15: Correlation Matrix: Performance vs Accuracy Gap

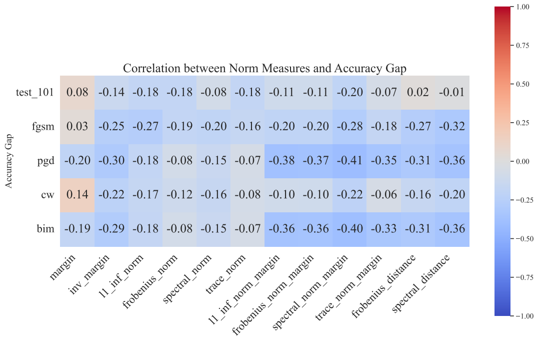

Figure 16: Correlation Matrix: Norm Metrics vs Accuracy Gap

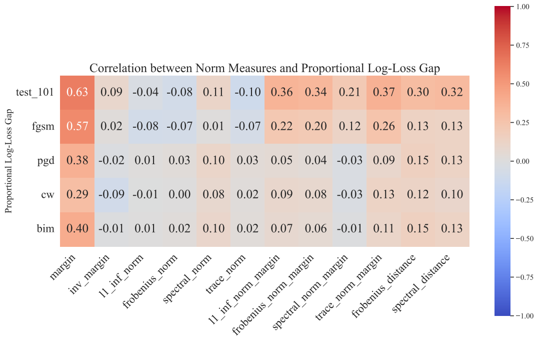

Figure 17: Correlation Matrix: Norm Metrics vs Proportional Log Loss Gap

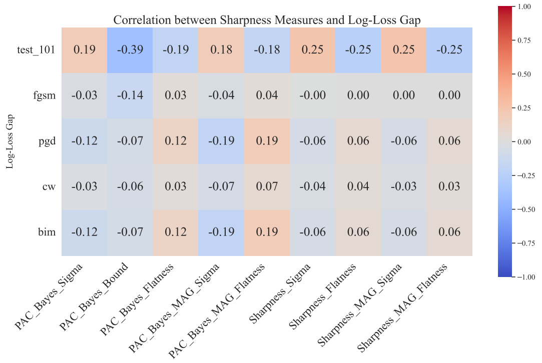

Figure 18: Correlation Matrix: Sharpness vs Log Loss Gap

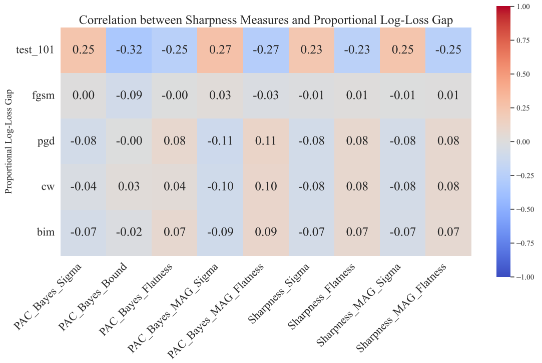

Figure 19: Correlation Matrix: Sharpness vs Proportional Log Loss Gap