EMPIRICAL STUDY OF THE DECISION REGION AND ROBUSTNESS IN DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

In general, the Deep Neural Networks (DNNs) is evaluated by the generalization performance measured on the unseen data excluded from the training phase. Along with the development of DNNs, the generalization performance converges to the state-of-the-art and it becomes difficult to evaluate DNNs solely based on the generalization performance. The robustness against the adversarial attack has been used as an additional metric to evaluate DNNs by measuring the vulnerability of them. However, few researches have been performed to analyze the adversarial robustness in terms of the geometry in DNNs. In this work, we perform empirical study to analyze the internal properties of DNNs which affect model robustness under adversarial attacks. Especially, we propose the novel concept *Populated Region Set (PRS)* where train samples populated more frequently to represent the internal properties of DNNs in the practical setting. From the systematic experiments with the proposed concept, we provide empirical evidences to validate that the low PRS ratio has strong relationship with the adversarial robustness of DNNs.

1 INTRODUCTION

With the steep improvement of the performance of Deep Neural Networks (DNNs), they are expanding their applications to the real world, such as autonomous driving or healthcare (LeCun et al., 2015; Miotto et al., 2018; Huang & Chen, 2020). For the real world application, one may need to choose the best model among the candidates. Traditionally, the generalization performance is used to evaluate the models, which measures the objective score on the test dataset excluded in the training phase (Bishop, 2006). However, it is non-trivial to evaluate DNNs based on the single metric. For example, if two networks with the same structure have the similar test accuracy, it is ambiguous to select the better one.

Robustness against the adversarial attacks that measures the vulnerability can be an alternative to evaluate DNNs (Szegedy et al., 2013; Goodfellow et al., 2014; Gu & Rigazio, 2014; Huang et al., 2015; Jakubovitz & Giryes, 2018; Yuan et al., 2019; Zhong et al., 2021). Adversarial attacks aim to induce the model misprediction by perturbing the input with small magnitude. Most of previous work focus on the way to find the adversarial samples by utilizing the model properties such as gradients with respect to the loss function. In a sense that the adversarial attack seeks to find the perturbation path on the model prediction surface over the input space, robustness can be expressed in terms of the geometry of the model. However, few researches has been performed to interpret the robustness with concept of the geometric properties of DNNs.

In geometric viewpoint, the internal properties of DNNs is represented by the boundaries and the regions (Baughman & Liu, 2014). It is shown that the DNNs with piece-wise linear activation layers composes a lot of piece-wise linear regions, and the maximal number of these regions is mathematically related to the expressivity of DNNs (Montúfar et al., 2014; Xiong et al., 2020). As these approaches only provide the upper bound for the expressivity with the same structured model, it does not explain how much information that the model actually expresses.

In this work, we investigate the relationship between the internal properties of DNNs and the robustness. Especially, our approach analyzes the internal characteristics in the perspective of the decision boundary (DB) and the decision region (DR), referred as the basic components of DNNs (Fawzi et al., 2017). To avoid insensitivity of the maximal number of linear regions in the same structure

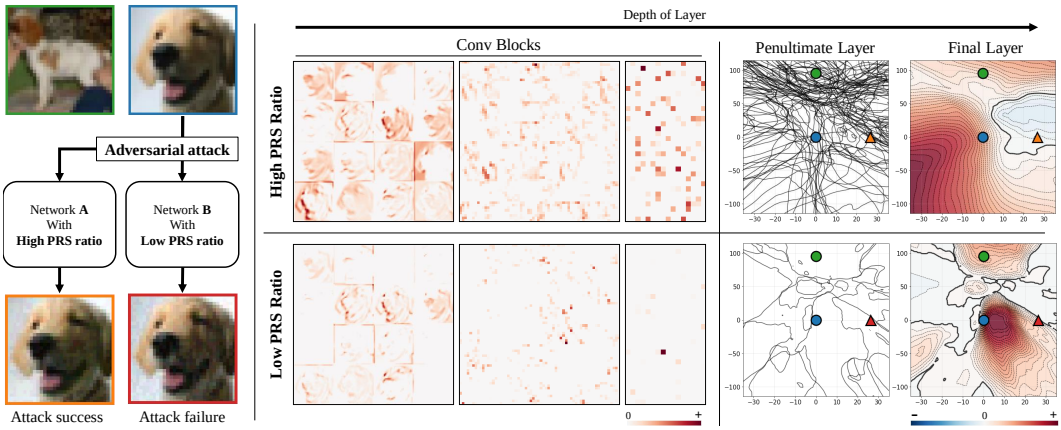


Figure 1: An illustrative comparison of two networks which have the different PRS ratio (High and Low) with similar test accuracy. (Left column) Each colored box image corresponds to the colored dot in the right column. Orange/Red box image is perturbed image in each network with adversarial attack respectively. (Middle column) The network with low PRS ratio learns more sparse feature representation than high PRS ratio. (Right column) Decision boundaries and regions in the section of input space for the target layer. Contour indicates the logit value for the dog class.

assumption, we propose the novel concept *Populated Region Set (PRS)* which is a set of DRs containing at least one sample included in the train dataset. Since the PRS can be considered as the feasible complexity of the model, we hypothesize that the size of PRS is related to the robustness of network. To validate our hypothesis, we perform systematic experiments with various structure of DNNs and dataset. Our observations are summarized as follows:

- The models with same structure can have different size of PRS, although the models have similar generalization performance. In experiments, we observe that this difference causes different robustness of the network.
- We empirically show that the size of PRS is related to the robustness against the adversarial attack. The model with the small size of PRS tends to show higher robustness compared to the large one. We further observe that the model with the small size of PRS has the relatively parallel parameters compared to the large one in the final layer.
- We verify that the the size of intersection of PRS from train/test dataset is related to the robustness of model. The model with high PRS inclusion ratio of test samples has higher robustness than low PRS inclusion ratio.
- We identify that the model with small size of PRS learns the sparse feature representation. In quantification, we observe the proportional relation between the size of PRS and sparsity of feature representation.

2 Related Work

Adversarial robustness For the real-world application of DNNs, the adversarial attack referred as the vulnerability of DNNs (Goodfellow et al., 2014) is mainly used to validate the reliability of the trained network. As an early stage for adversarial attacks, the fast gradient sign method (FGSM) (Goodfellow et al., 2014) based on the gradient with respect to the loss function and the multi-step iterative method (Kurakin et al., 2016) are proposed to create the adversarial examples to change the model prediction with the small perturbation. Recently, many researches for effective attack in the various settings (e.g., white-box or black-box) are performed to understand the undesirable decision of the networks (Madry et al., 2017; Shaham et al., 2018; Madry et al., 2018; Chen et al., 2020). In terms of factors affecting robustness, Yao et al. (2018) provide the evidence to argue that training with large batch size can degrade the robustness of the model against the adversarial attack in the perspective of the hessian spectrum. In contrast, Kamath et al. (2019) propose that the model with

a constant ratio between learning rate and batch size does not degrade model robustness even with large batch size as it converges to the flatter minima.

Geometric Analysis inside Deep Neural Networks With the increasing interest in the expressive power of DNNs, there are several attempts to analyze DNNs in a geometric perspective (Dauphin et al., 2014; Choromanska et al., 2015). Among them, the characteristics of decision boundary or regions formulated by the DNNs are mainly discussed. Montúfar et al. (2014) show that the cascade of the linear layer and the nonlinear activation organizes the numerous piece-wise linear regions. They show that the complexity of the decision boundary is related to the maximal number of these linear regions, which is determined by the depth and the width of the model. Xiong et al. (2020) extend the notion of the linear region to the convolutional layers and show the better geometric efficiency of the convolutional layers. Fawzi et al. (2018) reveal that classification regions in DNNs are topologically connected and the decision boundary of natural images is flat in most directions. It has also been shown that the manifolds learned by DNNs and distributions over them are highly related to the representation capability of a network (Lei et al., 2018). While these studies highlight that the benefits of increasing expressivity of DNNs as the number of regions increases, interpreting the vulnerability of DNNs with the geometry is another important topic. Yang et al. (2020) have shown that the model with the thick decision boundaries induce the robustness. Moosavi-Dezfooli et al. (2019) have shown that the decision boundary with small curvature acquires the high robustness of the model. These approaches focus on the decision boundaries, while this paper suggests to focus on the decision regions, which are composed by the surrounding decision boundaries.

3 Internal Property of DNNs

This section describes the internal properties of DNNs in the perspective of decision boundaries (DBs) and regions (DRs). The DBs of DNNs classifier is mainly defined as the borderline between DRs for classification, where the prediction probability of the class i and the neighboring class j are same (Fawzi et al., 2018). To expand the notion of DBs and DRs to the internal feature-level, we re-define the DBs in the classifier which generalizes the existing definition of DBs. Then, we propose the novel concept *Populated Region (PR)* that describes the specific DRs used from the network for train samples. PR is used to analyze the relation between the trained parameters and the characteristics of networks.

3.1 Decision Boundary and Region

Let the classifier with L number of layers be $F(x) = f_L(\sigma(f_{L-1}\sigma(\cdots\sigma(f_1(x)))) = f_{L:1}(x)$ where x is the sample in the input space $\mathcal{X} \subset \mathbb{R}^{D_x}$ and $\sigma(\cdot)$ denotes the non-linear activation function[1]. For the l-th layer, $f_l(\cdot)$ denotes the linear operation and $f^i_{l:1}(\cdot)$ denotes the value of i-th element of the feature vector $f_{l:1}(x) \in \mathbb{R}^{D_l}$. We define the DB for the i-th neuron of the l-th layer.

Definition 1 (Decision Boundary (DB)) *The i-th decision boundary at the l-th layer is defined as*

$$B^i_l = \{x | f^i_{l:1}(x) = 0, \quad \forall x \in \mathcal{X}\}.$$

We note that the internal DB B^i_l ($l < L$) divides the input space $\mathcal{X}$ based on the hidden representation of the l-th layer (i.e., existence of feature and the amount of feature activation). There are total D_l boundaries and the configuration of the DBs are arranged by the training. As input samples in the same classification region are considered as the same class, the input samples placed on the same side of the internal DB B^i_l share the similar feature representation. In this sense, we define the internal DR, which is surrounded by internal DBs.

Definition 2 (Decision Region (DR)) *Let $\mathbf{V}_l \in \{-1, +1\}^{D_l}$ be the indicator vector to determine which side of decision boundaries of l-th layer. Then the decision region $DR_{\mathbf{V}_l}$, which shares the sign of feature representation, is defined as*

$$DR_{\mathbf{V}_l} = \{x | sign(f_{l:1}(x)) = \mathbf{V}_l, \quad \forall x \in \mathcal{X}\}.$$

[1]Although there are various activation functions. We only consider ReLU activation for this paper.

Figure 1 indicates the internal properties for two networks trained on CIFAR-10 with similar test accuracy. The right column in Figure 1 depicts the internal DBs and DRs in the network with high/low PRS ratio (top and bottom). We randomly select 2 test images (blue and green box) and generate adversarial images for blue box (orange and purple box) in each network respectively. We make hyperplane with these images to visualize the DBs and DRs in the 2D space. We identify that the configuration of DBs and DRs seems to be different, although the two networks have same structure and similar test accuracy.

3.2 Populated Region Set

It is well-studied that the number of DRs is related to the representation power of DNNs (Montúfar et al., 2014; Xiong et al., 2020). In particular, the expressivity of DNNs with partial linear activation function is quantified by the maximal number of the linear regions and this number is related to the width and depth of the structure. We believe that although the maximal number can be one measure in the function aspect, the trained DNNs with finite train data[2] cannot handle the entire regions to solve the task. To only consider DRs which the network uses in the training process, we devise the train-related regions where train samples populated more frequently. We define the Populated Region Set (PRS) which is a set of DRs containing at least one sample included in the train dataset. PRS will be used to analyze the relation between the geometrical property and robustness of DNNs in the practical aspect.

Definition 3 (Populated Region Set (PRS)) *From the set of every DRs, the Populated Region Set of the model f and the given dataset $\mathbf{X}$ is defined as*

$$PRS(\mathbf{X}, f, l) = \{DR_{\mathbf{V}_l} | x \in DR_{\mathbf{V}_l} \; \exists \, x \in \mathbf{X}, \; \forall \, \mathbf{V}_l \in \{-1, 1\}^{D_l}\}.$$

Then we can define Populated Region as a union of decision regions in PRS as

$$PR(\mathbf{X}, f, l) = \cup_{DR \in PRS(\mathbf{X}, f, l)} DR.$$

We note that the size of PRS is bounded to the size of given dataset $\mathbf{X}$. When $|PRS(\mathbf{X}, f, l)| = |\mathbf{X}|$, each sample in train dataset is assigned to each distinct DR in the l-th layer. To compare the PRS of networks, we define the PRS ratio, $\frac{|PRS(f, \mathcal{X}, l)|}{|\mathcal{X}|}$, which measures the ratio between the size of PRS and the given dataset. Figure 2 indicates that the training results for convolutional neural network with 6 convolution blocks (CNN-6) on the CIFAR-10 with different PRS ratio.

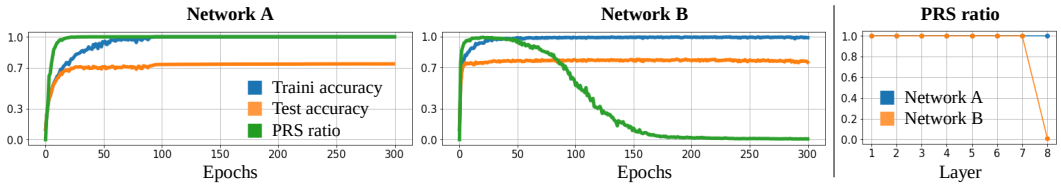


Figure 2: (First column) Train/Test accuracy and the ratio of PRS on CNN-6 with high PRS ratio. (Second column) Train/Test accuracy and the ratio of PRS on CNN-6 with low PRS ratio. (Third column) The number of PRS for the depth of each layers. We call these two CNN-6 as network A and B respectively throughout the paper.

From first and second columns in Figure 2, we can identify that although two networks have similar train/test accuracy, the PRS ratio is largely different. Third column in Figure 2 indicates that the PRS ratio for the depth of layers in each model with 300th epoch. The network B only has the diminished PRS ratio at the layer $l = 8$ (the penultimate layer). From the above observation and the fact that the penultimate layers are widely used as feature extraction, we only consider the PRS ratio at the penultimate layer in the remained paper.

Experimental setups For the systematic experiments, we select three different structures of DNNs to analyze; (1) convolutional neural network with 6 convolution blocks (CNN-6) (2) VGG-16 (Simonyan & Zisserman, 2014) and (3) ResNet-18 (He et al., 2016). We train each network with cross-entropy loss over three datasets; MNIST (LeCun & Cortes, 2010), F-MNIST (Xiao et al., 2017)

[2]In general, the number of train data is smaller than the maximal number of the linear region.

and CIFAR-10 (Krizhevsky et al., 2009). To extensively analyze the relation between PRS ratio and robustness, we consider multiple epochs of each model with similar test accuracy as independent models. The details for the used networks are described in the Appendix.

4 Robustness for Adversarial Attack

In this section, we perform the experiments to analyze the relationship between the PRS ratio and the robustness. We evaluate the robustness of the network using fast gradient sign method (FGSM) (Goodfellow et al., 2014) and iterative method (Kurakin et al., 2016) which are mainly used as the adversarial attacks. The train/test adversarial attacks using entire samples are performed for the various perturbation ϵ and the two types of attacks (untargeted/targeted) are considered. We further observe that the PRS ratio is also related to the degree of parallelism for the parameters in the final layer.

4.1 PRS and Robustness for Models

At first, we compare two models with similar test accuracy but different PRS ratio (Network A and B in Figure 2). Figure 3 indicates the results of robustness under various adversarial attacks with various epsilons. We identify that the network B (low PRS ratio) is more robust than the network A (high PRS ratio) in entire adversarial attacks.

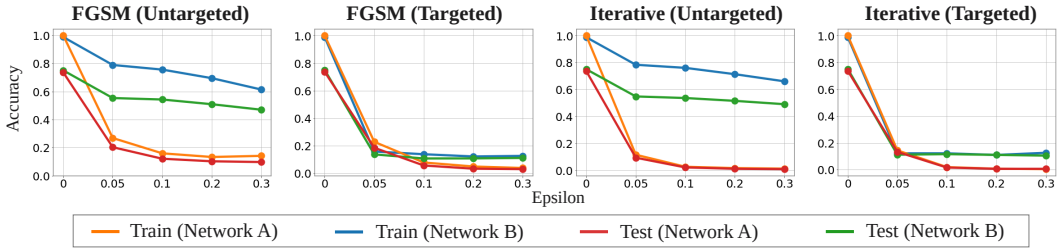


Figure 3: Train/Test accuracy under various adversarial attacks on network A and B. The x-axis indicates perturbation ϵ and y-axis indicates train/test accuracy after attack.

As the untargeted attacks show the remarkable difference between two models, we focus on the analysis of the untargeted adversarial attacks. Among the untargeted attacks, because the FGSM and iterative method show similar trend, we only consider the untargeted FGSM to quantify the robustness of DNNs considering the computation efficiency in the remained paper. To reduce the result of adversarial attacks with various ϵ into the single score, we refer to the robustness score (RS) as the area of under curve for the changed accuracy. To verify the relationship between PRS ratio and robustness, we measure the PRS ratio and RS in the entire models. Figure 4 indicates the experiment results grouped by the model structure.

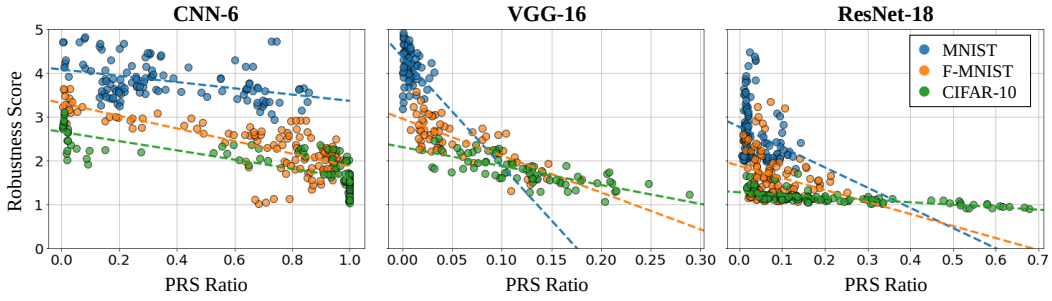


Figure 4: The relationship between PRS ratio and RS in various models and dataset. The colored dot means that the independent models. The colored dashed line indicates the trend for each dataset.

From the Figure 4, we identify that the PRS ratio has inverse proportional relation to the RS in almost cases. The complicated models (VGG-16 and ResNet-18) have lower PRS ratios compared to the CNN-6, and show even lower PRS ratios on the simple dataset (i.e., MNIST and F-MNIST).

PRS and Robustness for Training We also analyze how the PRS ratio and RS change during the training. We measure the PRS ratio and the RS for three different networks trained on CIFAR-10. Figure 5 indicates the test accuracy, PRS ratio and RS for each training epoch. Although the test accuracy of each model converges after 50th epoch, the PRS ratio keeps decreasing after 50th epoch and the RS keeps increasing.

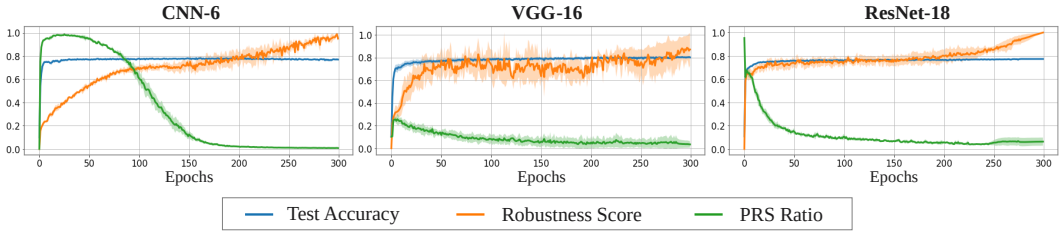


Figure 5: PRS ratio and RS for training epochs in three different networks with CIFAR-10. The blue line indicates test accuracy, the orange line indicates the RS, and the green line indicates the PRS ratio. The shaded lines depict the standard deviation for 5 fixed random seeds.

4.2 PRS AND FINAL LAYER

From the above observations, we empirically confirm that the PRS ratio is related to the robustness against adversarial attack. In order to investigate the evidence that the low PRS ratio causes robustness for the gradient-based attack, we perform the additional analysis on the failed attack samples. In the gradient-based attack, as the magnitude of gradient is crucial component to success, we first count the ratio of the zero gradient samples in the failed attack samples. The left column in Figure 6 indicates the ratio of zero gradient samples (red bar) in the entire failed attack samples. We identify that the ratio of the zero-gradient samples is smaller than non zero-gradient samples.

We note that the non-zero gradient in failed attacks mean that the perturbed samples (attacked samples) maintain the index of the largest logit as the true class. To analyze the reason of failure, we examine the change of logits after attack. This change is shown in the the second column of Figure 6. We can identify that the attacked logit vector is almost parallel with the true logit vector. Although the untargeted FGSM succeeds to reduce the logit value for the corresponding true class, other logit values decrease at the same time, resulting in the maintenance of the prediction.

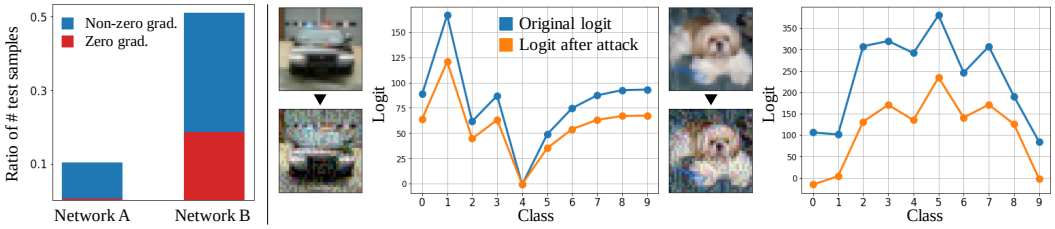


Figure 6: (First column) The comparison of the ratio of zero gradient in the failure attack for the test untargeted FGSM with $\epsilon = 0.3$ (network A and B). The red bar means the ratio of zero gradient samples. (Second column) The illustrative examples for the failed samples with non-zero gradient and the corresponding logits. After attack, the logits are changed in parallel with the original logits.

To explain the change of the logit vector in parallel, we hypothesize that the DBs corresponding each class node have similar configuration in the input space. However, it is intractable to measure the similarity between DBs in the entire network, because of the highly non-linear structure and high dimensional input space. To simplify our hypothesis, we only measure the cosine similarity between parameters in the final layer.

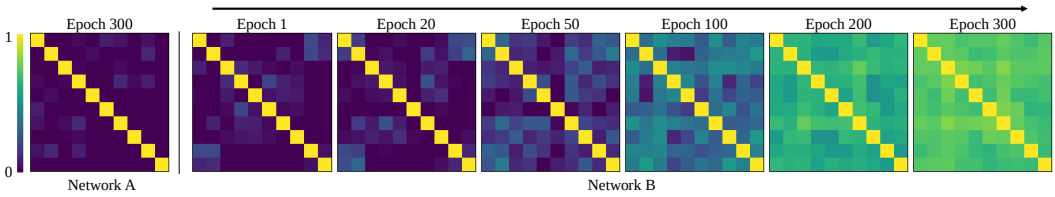


Figure 7: (First column) Cosine similarity matrix for the final layer on the network A. (Second column) The similarity matrix for epochs on the network B. As the epoch increases, the cosine similarity for each parameter increases.

Figure 7 indicates that the similarity matrix for network A and B. When we compare the matrix between two models at 300th epoch, we identify that the network B (low PRS ratio) has greater cosine similarity between each parameter in the final layer. We note that the cosine similarity between each parameter in the final layer can be considered as the degree of the parallelism for the normal vectors in the linear classifier. We also confirm that the decrease of PRS ratio is aligned with the increase of the similarity of parameters in the second column when we consider the graph in Figure 2. To verify the relationship between PRS ratio and the cosine similarity between the parameters in the final layer, we measure the PRS ratio and the cosine similarity between each parameter in the entire models. Figure 8 indicates the experiment results grouped by the model structure. We identify that the PRS ratio has the inverse proportional relation to the cosine similarity between each parameter in the final layer.

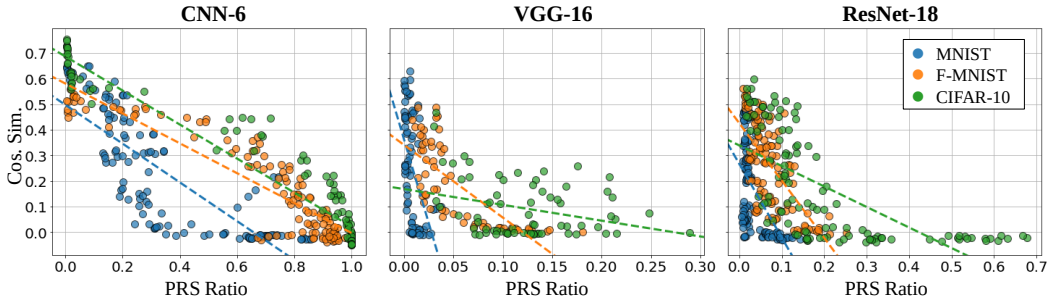


Figure 8: The trend between PRS ratio and cosine similarity in various models and dataset. The colored dot means that the independent models. The colored dashed line indicates the trend for each dataset.

5 PRS AND TEST DATASET

Regarding the model as a mapping function from the input space to the feature space, handling the unseen data to the known feature domain is significant for the generalization performance of the model. In that sense, if the majority of samples from the test dataset are assigned to the train PR, the model can be considered to learn the informative and general concept of feature mapping. For example, if the arbitrary test sample is mapped to the train PR, we expect that the similar decision can appear. However, it is non-trivial to guess which type of decision appears when the test sample is mapped to out of the train PR.

To investigate the difference between the test samples which are included/excluded in the train PR, we perform adversarial attack for each group. For the controlled comparison, we divide inclusion and exclusion group each with 1k correctly predicted test samples. Figure 9 shows the results of targeted/untargeted FGSM and iterative method on CNN-6 trained with CIFAR-10 for 5 fixed random seeds respectively. Although the accuracy of each test group decreases as the epsilon becomes larger, we observe that the inclusion group is more robust for all types of attacks than the exclusion group.

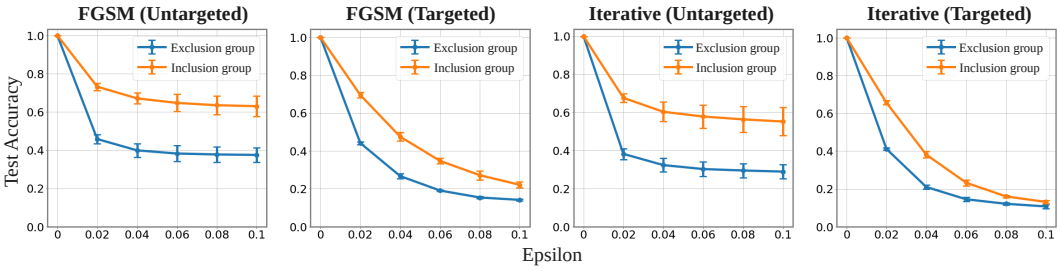


Figure 9: The change of accuracy under adversarial attacks for inclusion/exclusion groups for CNN-6 on CIFAR-10 for 5 fixed random seeds. The blue/orange line shows the exclusion/inclusion group respectively. The exclusion group shows more vulnerable to adversarial attacks.

Figure 10 indicates the PRS ratio and the inclusion ratio of test samples for the train PR. We compute the inclusion ratio as the ratio of the test dataset mapped to the train PR. In Figure 10, we identify that the PRS ratio has inverse proportional relationship to the inclusion ratio. As we already verify that the included test samples show high robustness, we empirically confirm that the low PRS ratio is related to the robustness under adversarial attack.

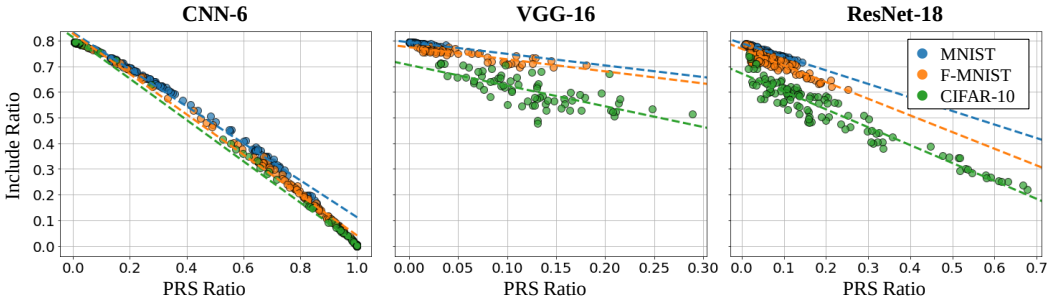


Figure 10: The relation between the PRS ratio and the inclusion ratio for CNN-6, VGG-16 and ResNet-18 on CIFAR-10. The colored dashed line indicates the trend for each dataset.

6 PRS AND TRAINED FEATURES

This section explores the trained features of the models with the different PRS ratios. At first, we visualize the featuremap directly for each depth of layers. Figure 11 indicates the illustrative examples of featuremaps. We identify that the model with the low PRS ratio learns more sparse features compared to the model with the high PRS ratio.

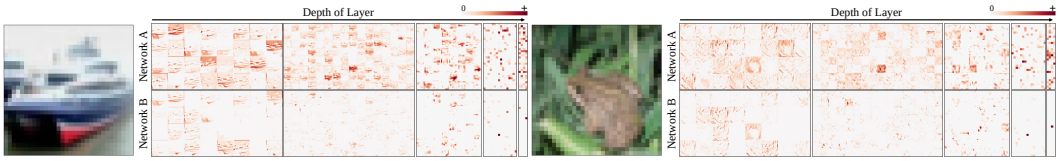


Figure 11: Visualization of feature map for CNN-6 trained on CIFAR-10. (First row) The featuremaps for the network A. (Second row) The featuremaps for the network B.

As the sparse features are considered as the independent and informative representation (Ranzato et al., 2007; Lee et al., 2007), if the PRS ratio can cause the sparse feature representation in the various cases, we conclude that the PRS ratio is related to the informative features. To verify our hypothesis, we measure the trend between the PRS ratio and the average sparsity for each network. The average sparsity of the model is calculated by taking the average of the ratio of zero-valued feature over the train dataset for every activation layer.

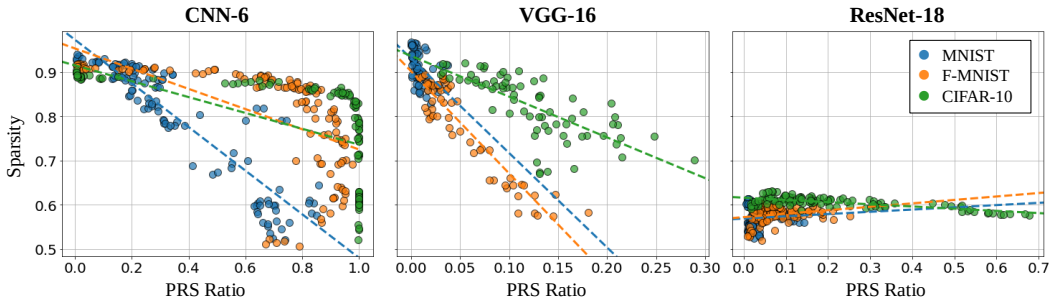
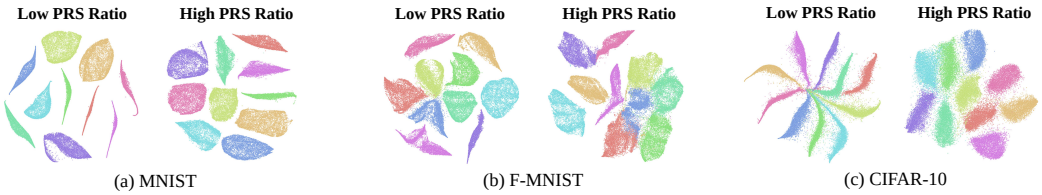


Figure 12: The relation between the PRS ratio and the average sparsity for CNN-6, VGG-16 and ResNet-18 on CIFAR-10. The colored dashed line indicates the trend for each dataset.

Figure 12 indicates the relations between the PRS ratio and the average sparsity in entire cases. We identify that CNN-6 and VGG-16 show the inverse proportional relation in the entire dataset, but it is hard to find the apparent relation in ResNet-18. We suspect that (1) the skip connection, or (2) batch normalization can cause this phenomenon.

To further investigate the global feature representations, we visualize the projected feature space using t-SNE (van der Maaten & Hinton, 2008) for the train dataset. In Figure 6, we can identify that the model with the low PRS ratio has more intensive feature representation for each class.



(a) MNIST (b) F-MNIST (c) CIFAR-10

Figure 13: The visualization of the projected feature space with t-SNE for CNN-6 trained with various dataset. The colored code means the class. The model with low PRS ratio tends to have more intensive feature representation compared to the model with high PRS ratio.

7 Conclusion

In this work, we perform empirical study to analyze internal properties of DNNs which affect model robustness under adversarial attacks. We propose the novel concept *populated region set* to derive the relationship between internal properties of DNNs and robustness in the practical setting. From the systematic experiments with the proposed concept, we suggest that the PRS ratio is related to the robustness of DNNs and provide the empirical evidences on this relationship: (1) The network with the low PRS ratio shows high robustness for the gradient-based attack compared to the high PRS ratio. In particular, the model with the low PRS ratio has higher degree of parallelism for the parameters in the final layer which can support the robustness. (2) The network with the low PRS ratio includes more number of test samples in the train PR. We empirically verify that this inclusion ratio is related to the robustness from the observation that the included test samples are more robust than excluded test samples. (3) The network with the low PRS ratio learns sparse feature representation to solve the task. Also its global feature representation shows more intensive than the high PRS ratio.

References

D Richard Baughman and Yih An Liu. *Neural networks in bioprocessing and chemical engineering*. Academic press, 2014.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pp. 1277–1294. IEEE, 2020.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pp. 192–204. PMLR, 2015.

Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *arXiv preprint arXiv:1406.2572*, 2014.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Classification regions of deep neural networks. *arXiv preprint arXiv:1705.09552*, 2017.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

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*, pp. 770–778, 2016.

Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

Yu Huang and Yue Chen. Autonomous driving with deep learning: A survey of state-of-art technologies. *arXiv preprint arXiv:2006.06091*, 2020.

Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 514–529, 2018.

Sandesh Kamath, Amit Despande, and KV Subrahmanyam. On adversarial robustness of small vs large batch training. 2019.

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

Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. *Advances in neural information processing systems*, 20:873–880, 2007.

Na Lei, Zhongxuan Luo, Shing-Tung Yau, and David Xianfeng Gu. Geometric understanding of deep learning. *arXiv preprint arXiv:1805.10451*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *NIPS*, 2014.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9070–9078, 2019.

Marc Ranzato, Christopher Poultney, Sumit Chopra, Yann LeCun, et al. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19:1137, 2007.

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Huan Xiong, Lei Huang, Mengyang Yu, Li Liu, Fan Zhu, and Ling Shao. On the number of linear regions of convolutional neural networks. In *International Conference on Machine Learning*, pp. 10514–10523. PMLR, 2020.

Yaoqing Yang, Rekha Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph Gonzalez, Kannan Ramchandran, and Michael W. Mahoney. Boundary thickness and robustness in learning models. *ArXiv*, abs/2007.05086, 2020.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *arXiv preprint arXiv:1802.08241*, 2018.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

Ziyuan Zhong, Yuchi Tian, and Baishakhi Ray. Understanding local robustness of deep neural networks under natural variations. *Fundamental Approaches to Software Engineering*, 12649: 313, 2021.