Within-layer Diversity Reduces Generalization Gap

Firas Laakom [1] **Jenni Raitoharju** [2] **Alexandros Iosifidis** [3] **Moncef Gabbouj** [1]

Abstract

Neural networks are composed of multiple layers arranged in a hierarchical structure jointly trained with a gradient-based optimization. At each optimization step, neurons at a given layer receive feedback from neurons belonging to higher layers of the hierarchy. In this paper, we propose to complement this traditional 'between-layer' feedback with additional 'within-layer' feedback to encourage diversity of the activations within the same layer. To this end, we measure the pairwise similarity between the outputs of the neurons and use it to model the layer's overall diversity. By penalizing similarities and promoting diversity, we encourage each neuron to learn a distinctive representation and, thus, to enrich the data representation learned within the layer and to increase the total capacity of the model. We theoretically and empirically study how the within-layer activation diversity affects the generalization performance of a neural network and prove that increasing the diversity of hidden activations reduces the generalization gap.

1. Introduction

Neural networks are a powerful class of non-linear function approximators that have been successfully used to tackle a wide range of problems. They have enabled breakthroughs in many tasks, such as image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012a), and anomaly detection (Golan & El-Yaniv, 2018). However, neural networks are often over-parameterized, i.e., have more parameters than data. As a result, they tend to overfit to the training samples and not generalize well on unseen

[1]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland [2]Programme for Environmental Information, Finnish Environment Institute, Jyväskylä, Finland [3]Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark. Correspondence to: Firas Laakom <firas.laakom@tuni.fi>.

examples (Goodfellow et al., 2016). While research on double descent (Belkin et al., 2019; Advani et al., 2020; Nakkiran et al., 2020) shows that over-parameterization does not necessarily lead to overfitting, avoiding overfitting has been extensively studied (Neyshabur et al., 2018; Nagarajan & Kolter, 2019; Poggio et al., 2017) and various approaches and strategies have been proposed, such as data augmentation (Goodfellow et al., 2016; Zhang et al., 2018), regularization (Kukačka et al., 2017; Bietti et al., 2019; Arora et al., 2019), and dropout (Hinton et al., 2012b; Wang et al., 2019; Lee et al., 2019; Li et al., 2016), to close the gap between the empirical loss and the expected loss.

Diversity of learners is widely known to be important in ensemble learning (Li et al., 2012; Yu et al., 2011) and, particularly in deep learning context, diversity of information extracted by the network neurons has been recognized as a viable way to improve generalization (Xie et al., 2017a; 2015). In most cases, these efforts have focused on making the set of weights more diverse (Yang et al.; Malkin & Bilmes, 2009). However, diversity of the activations has not received much attention.

To the best of our knowledge, (Cogswell et al., 2016) is the only work in neural network context which considers diversity of the activations directly. They propose an additional loss term using cross-covariance of hidden activations, which encourages the neurons to learn diverse or non-redundant representations. The proposed approach, known as DeCov, has empirically been proven to alleviate overfitting and to improve the generalization ability of neural network, yet a theoretical analysis to prove this has so far been lacking. Moreover, modeling diversity as the sum of the pair-wise cross-covariance, it can capture only the pairwise diversity between components and is unable to capture the higher-order "diversity".

In this work, we start by theoretically showing that the within-layer activation diversity boosts the generalization performance of neural networks and reduces overfitting. Moreover, we propose a novel approach to encourage activation diversity within a layer. We propose complementing the 'between-layer' feedback with additional 'within-layer' feedback to penalize similarities between neurons on the same layer. Thus, we encourage each neuron to learn a distinctive representation and to enrich the data representation

learned within each layer. We propose three variants for our approach that are based on different global diversity definitions.

Our contributions in this paper are as follows:

- Theoretically, we analyse the effect of the within-layer activation diversity on the generalization error bound of neural network. As shown in Theorems 1-6, we express the upper-bound of the estimation error as a function of the diversity factor. Thus, we provide theoretical evidence that the within-layer activation diversity can help reduce the generalization error.

- Methodologically, we propose a new approach to encourage the 'diversification' of the layers' output feature maps in neural networks. The main intuition is that by promoting the within-layer activation diversity, neurons within a layer learn distinct patterns and, thus, increase the overall capacity of the model.

- Empirically, we show that the proposed within-layer activation diversification boosts the performance of neural networks.

2. Generalization error analysis

In this section, we analyze how the within-layer activation diversity affects the generalization error of a neural network. Generalization theory (Zhang et al., 2017; Kawaguchi et al., 2017) focuses on the relation between the empirical loss and the expected risk defined as follows:

$$L(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{Q}}[l(f(\boldsymbol{x}),y)], \tag{1}$$

where $\mathcal{Q}$ is the underlying distribution of the dataset. Let $f^* = \arg\min_f L(f)$ be the expected risk minimizer and $\hat{f} = \arg\min_f \hat{L}(f)$ be the empirical risk minimizer. We are interested in the estimation error, i.e., $L(f^*) - L(\hat{f})$, defined as the gap in the loss between both minimizers (Barron, 1994). The estimation error represents how well an algorithm can learn. It usually depends on the complexity of the hypothesis class and the number of training samples (Barron, 1993; Zhai & Wang, 2018).

In this work, we are interested in the effect of the within-layer activation diversity on the estimation error. In order to study this effect, we assume that with a high probability τ, the distance between the output of each pair of neurons, $(\phi_n(\boldsymbol{x}) - \phi_m(\boldsymbol{x}))^2$, is lower bounded by d_{min}^2 for any input $\boldsymbol{x}$. Intuitively, if two neurons n and m have similar outputs for many samples, their corresponding similarity d_{min} will be small. Otherwise, their similarity d_{min} is small and they are considered "diverse". By studying how d_{min} affects the generalization of the model, we can theoretically understand how diversity affects the performance of neural networks.

To this end, we derive generalization bounds for neural networks using d_{min}.

Several techniques have been used to quantify the estimation error, such as PAC learning (Hanneke, 2016; Arora et al., 2018), VC dimension (Sontag, 1998; Harvey et al., 2017; Bartlett et al., 2019), and the Rademacher complexity (Xie et al., 2015; Zhai & Wang, 2018; Tang et al., 2020). The Rademacher complexity has been widely used as it usually leads to a tighter generalization error bound (Sokolic et al., 2016; Neyshabur et al., 2018; Golowich et al., 2018). In this work, we also rely on the Rademacher complexity to study diversity. We seek a tighter upper bound of the estimation error and show how the within-layer diversity, expressed with d_{min}, affects the bound. We start by deriving such an upper-bound for a simple network with one hidden layer trained for a regression task and then we extend it for a general multi-layer network and for different losses. The proofs are provided as supplementary material.

2.1. Single hidden-layer network

Here, we consider a simple neural network with one hidden-layer with M neurons and one-dimensional output trained for a regression task. The full theoretical characterization of the setup can be summarized in the assumptions presented in Appendix 6.1

Our main goal is to analyze the estimation error bound of the neural network and to see how its upper-bound is linked to the diversity, expressed by d_{min}, of the different neurons. The main result is presented in Theorem 1.

Theorem 1. *Under Assumptions 1, there exist a constant A, such that with probability at least $\tau^Q(1-\delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq \left(\sqrt{\mathcal{J}} + C_2\right)A$$
$$+ \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2\sqrt{\frac{2\log(2/\delta)}{N}}, \tag{2}$$

where $\mathcal{J} = C_4^2\big(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2)\big)$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Theorem 1 provides an upper-bound for the estimation error. We note that it is a decreasing function of d_{min}. Thus, we say that a higher d_{min}, i.e., more diverse activations, yields a lower estimation error bound. In other words, by promoting the within-layer diversity, we can reduce the generalization error of neural networks.

2.2. Binary classification

We now extend our analysis of the effect of the within-layer diversity on the generalization error in the case of a binary classification task, i.e., $y \in \{-1, 1\}$. The extensions of Theorem 1 in the case of a hinge loss and a logistic loss are presented in Theorems 2 and 3, respectively.

Theorem 2. *Using the hinge loss, there exist a constant A, such that with probability at least $\tau^Q(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq A + (1 + \sqrt{\mathcal{J}})\sqrt{\frac{2\log(2/\delta)}{N}}, \quad (3)$$

where $\mathcal{J} = C_4^2(MC_5^2 + M(M - 1)(C_5^2 - d_{min}^2/2))$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Theorem 3. *Using the logistic loss $l(f(x), y) = \log(1 + e^{-yf(x)})$, there exist a constant A such that, with probability at least $\tau^Q(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq \frac{A}{1 + e^{\sqrt{-\mathcal{J}}}}$$
$$+ \log(1 + e^{\sqrt{\mathcal{J}}})\sqrt{\frac{2\log(2/\delta)}{N}}, \quad (4)$$

where $\mathcal{J} = C_4^2(MC_5^2 + M(M - 1)(C_5^2 - d_{min}^2/2))$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

As we can see, also for the binary classification task, the error bounds of the estimation error for the hinge and logistic losses are decreasing with respect to d_{min}. Thus, employing a diversity strategy can improve the generalization also for the binary classification task.

2.3. Multi-layer networks

Here, we extend our result for networks with P (> 1) hidden layers. We assume that the pair-wise distances between the activations within layer p are lower-bounded by d_{min}^p with a probability τ^p. In this case, the main theorem is extended as follows:

Theorem 4. *There exist a constant A such that, with probability of at least $\prod_{p=0}^{P-1}(\tau^p)^{Q^p}(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq (\sqrt{\mathcal{J}^P} + C_2)A$$
$$+ \frac{1}{2}\left(\sqrt{\mathcal{J}^P} + C_2\right)^2\sqrt{\frac{2\log(2/\delta)}{N}}, \quad (5)$$

where Q^p is the number of neuron pairs in the pth layer, defined as $Q^p = \frac{M^p(M^p - 1)}{2}$, and $\mathcal{J}^P$ is defined recursively using the following identities: $\mathcal{J}^0 = C_3^0 C_1$ and $\mathcal{J}^p = M^p C^{p2}\big(M^{p2}(L_\phi C_3^{p-1}\mathcal{J}^{p-1} + \phi(0))^2 - M(M - 1)\frac{d_{min}^{p}{}^2}{2}\big)$, for $p = 1, \ldots, P$.

In Theorem 4, we see that $\mathcal{J}^P$ is decreasing with respect to d_{min}^p. Thus, we see that maximizing the within-layer diversity, we can reduce the estimation error of a multi-layer neural network.

2.4. Multiple outputs

Finally, we consider the case of a neural network with a multi-dimensional output, i.e., $\boldsymbol{y} \in R^D$. This yields the following two theorems:

Theorem 5. *For a multivariate regression trained with the squared error, there exist a constant A such that, with probability at least $\tau^Q(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq (\sqrt{\mathcal{J}} + C_2)A$$
$$+ \frac{D}{2}(\sqrt{\mathcal{J}} + C_2)^2\sqrt{\frac{2\log(2/\delta)}{N}}, \quad (6)$$

where $\mathcal{J} = C_4^2(MC_5^2 + M(M - 1)(C_5^2 - d_{min}^2/2))$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Theorem 6. *For a multi-class classification task using the cross-entropy loss, there exist a constant A such that, with probability at least $\tau^Q(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq \frac{A}{D - 1 + e^{-2\sqrt{\mathcal{J}}}}$$
$$+ \log\left(1 + (D - 1)e^{2\sqrt{\mathcal{J}}}\right)\sqrt{\frac{2\log(2/\delta)}{N}}, \quad (7)$$

where $\mathcal{J} = C_4^2(MC_5^2 + M(M - 1)(C_5^2 - d_{min}^2/2))$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Theorems 5 and 6 extend our result for the multi-dimensional regression and classification tasks, respectively. Both bounds are inversely proportional to the diversity factor d_{min}. We note that for the classification task, the upper-bound is exponentially decreasing with respect to d_{min}. This shows that increasing diversity within the layer yields a tighter generalization gap and, thus, theoretically guarantees a stronger generalization performance.

3. Within-layer activation diversity

As shown in the previous section, promoting diversity of activations within a layer can lead to tighter generalization bound and can theoretically decrease the gap between the empirical and the true risks. In this section, we propose a novel diversification strategy, where we encourage neurons within a layer to activate in a mutually different manner, i.e., to capture different patterns. To this end, we propose an additional within-layer loss which penalizes the neurons that activate similarly. The standard loss function $\hat{L}(f)$ is augmented as follows: $\hat{L}_{aug}(f) = \hat{L}(f) + \lambda \sum_{i=1}^P J^i$, where J^i expresses the overall similarity of the neurons within the ith layer and λ is the penalty coefficient for the diversity loss. Our proposed diversity loss can be applied to a single layer or multiple layers in a network. For simplicity, let us focus on a single layer.

Let $\phi_n^i(\boldsymbol{x}_j)$ and $\phi_m^i(\boldsymbol{x}_j)$ be the outputs of the nth and m^{th} neurons in the ith layer for the same input sample $\boldsymbol{x}_j$. The similarity s_{nm} between the the nth and m^{th} neurons can be obtained as the average similarity measure of their outputs for N input samples. We use the radial basis function to

express the similarity:

$$s_{nm} = \frac{1}{N} \sum_{j=1}^{N} \exp\big(-\gamma||\phi_n^i(\boldsymbol{x}_j) - \phi_m^i(\boldsymbol{x}_j)||^2 \big), \quad (8)$$

where γ is a hyper-parameter. The similarity s_{nm} can be computed over the whole dataset or batch-wise. Intuitively, if two neurons n and m have similar outputs for many samples, their corresponding similarity s_{nm} will be high. Otherwise, their similarity s_{mn} is small and they are considered "diverse". Based on these pair-wise similarities, we propose three variants for the overall similarity J^i:

- **Direct:** $J^i = \sum_{n \neq m} s_{nm}$. In this variant, we model the global layer similarity directly as the sum of the pairwise similarities between the neurons.

- **Det:** $J^i = -\det(\mathbf{S})$, where $\boldsymbol{S}$ is defined as $\boldsymbol{S}_{nm} = s_{nm}$. This variant is inspired by the Determinantal Point Process (DPP) (Kulesza & Taskar, 2010; 2012).

- **Logdet:** $J^i = -\text{logdet}(\mathbf{S})$[1]. This variant has the same motivation as the second one. We use logdet instead of det as logdet is a convex function over the positive definite matrix space.

It should be noted here that the first proposed variant, i.e., direct, similar to Decov (Cogswell et al., 2016), captures only the pairwise diversity between components and is unable to capture the higher-order "diversity", whereas the other two variants consider the global similarity and are able to measure diversity in a more global manner.

4. Experiments

To demonstrate the effectiveness of our approach and its ability to reduce the generalization gap in neural networks, we conduct image classification experiments on the ImageNet-2012 dataset (Russakovsky et al., 2015) using the ResNet50 model (He et al., 2016). The diversity term is applied on the last intermediate layer, i.e., the global average pooling layer. The training protocol is presented in the appendix 6.3.

We analyse the effect of the two parameters: γ, which is the RBF parameter used to measure the pair-wise similarity between two units, and λ, which controls the contribution of the global diversity term to the global loss, on both the final performance of the models and its generalization ability, i.e., generalization gap. The analysis is presented in Figure 1. As it can be seen, promoting the within-layer diversity consistently reduces overfitting and decreases the generalization gap for most of the hyperparameters values.

Moreover, we note that global modeling of diversity, i.e., det and logdet variants, yield tighter generalization gaps between the train and test errors compared to the non-global direct approach. In fact, while direct variant decreases the generalization gap compared to the standard approach, it decreases it only by 0.5% for most hyperparameter values, whereas, for the more global approaches, i.e., det and logdet, the generalization gap is less than 1.1% in multiple cases compared to the gaps 2.87% and 2.50% achieved by the standard approach and the direct variant, respectively.

For the direct variant (the curves in blue), we note that the performance of the method is not sensitive the hyperparameters, and the method achieves its best performance for low values of λ and γ. For the det variant (the curves in orange), we note that it significantly improves the generalization ability of the model. However, there is a trade-off between the generalization gap and the final error. In fact, emphasizing diversity and using a high weight for the diversity term significantly decreases the generalization gap. This damages the performance of the model compared to the standard approach. For example, with $\lambda = 0.01$ and $\gamma = 10$, the generalization gap of the model is 0.9% compared to 2.87% of the standard. However, the test error for this model gets up to 24.42% compared to 23.87% for the standard. For lower values of λ, the model is able to significantly outperform the standard approach on both the test error and the generalization gap. For the logdet variant (green curves), we note that, in terms of generalization gap, the approach consistently outperforms the standard approach. Using a small value for λ, the model yields lower error rates than the standard approach. For high values of λ, the error rates become similar to the standard approach but with a lower generalization gap. This variant is not sensitive to the hyperparameter γ. Additional empirical results are presented in appendix 6.4.

5. Conclusions

In this paper, we proposed a new approach to encourage 'diversification' of the layer-wise feature map outputs in neural networks. The main motivation is that by promoting within-layer activation diversity, neurons within the same layer learn to capture mutually distinct patterns. We proposed an additional loss term that can be added on top of any layer. This term complements the traditional 'between-layer' feedback with an additional 'within-layer' feedback encouraging diversity of the activations. We theoretically proved that the proposed approach decreases the estimation error bound and, thus, improves the generalization ability of neural networks. This analysis was further supported by experimental results showing that such a strategy can indeed improve the performance of state-of-the-art networks.
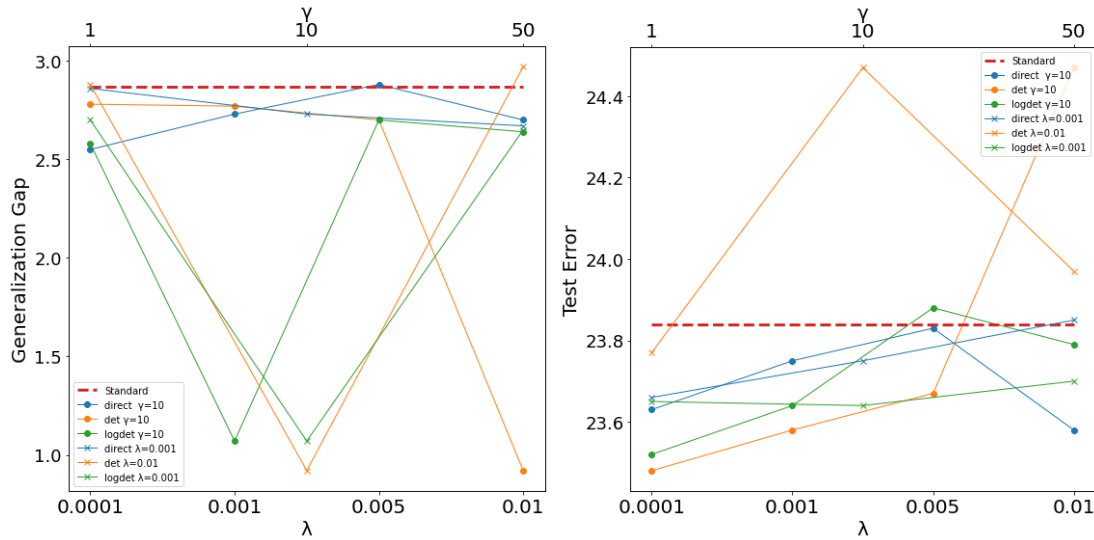
---

[1]This is defined only if $\boldsymbol{S}$ is positive definite. It can be shown that in our case $\boldsymbol{S}$ is positive semi-definite. Thus, in practice we use a regularized version $(\boldsymbol{S} + \epsilon \boldsymbol{I})$ to ensure the positive definiteness.

Figure 1. Sensitivity analysis of λ and γ on both the model accuracy and its generalization ability

Acknowledgments

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. pp. 254–263. Proceedings of Machine Learning Research, 10–15 Jul 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 7413–7424, 2019.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, pp. 930–945, 1993.

Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, pp. 115–133, 1994.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, pp. 463–482, 2002.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, pp. 63–1, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Bietti, A., Mialon, G., Chen, D., and Mairal, J. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning*, pp. 664–674, 2019.

Cogswell, M., Ahmed, F., Girshick, R. B., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations*, 2016.

Golan, I. and El-Yaniv, R. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pp. 9758–9769, 2018.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299, 2018.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*. MIT Press, 2016.

Hanneke, S. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, pp. 1319–1333, 2016.

Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Conference on Learning Theory*, pp. 1064–1068, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal processing magazine*, 29(6):82–97, 2012a.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012b.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Kukačka, J., Golkov, V., and Cremers, D. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.

Kulesza, A. and Taskar, B. Structured determinantal point processes. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2010.

Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

Lee, H. B., Nam, T., Yang, E., and Hwang, S. J. Meta dropout: Learning to perturb latent features for generalization. In *International Conference on Learning Representations*, 2019.

Li, N., Yu, Y., and Zhou, Z.-H. Diversity regularized ensemble pruning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 330–345, 2012.

Li, Z., Gong, B., and Yang, T. Improved dropout for shallow and deep learning. In *Advances in Neural Information Processing Systems*, pp. 2523–2531, 2016.

Malkin, J. and Bilmes, J. Multi-layer ratio semi-definite classifiers. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 4465–4468, 2009.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 11615–11626, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2018.

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J., and Mhaskar, H. Theory of deep learning III: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Lessons from the rademacher complexity for deep learning. 2016.

Sontag, E. D. VC dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, pp. 69–96, 1998.

Tang, Y., Wang, Y., Xu, Y., Shi, B., Xu, C., Xu, C., and Xu, C. Beyond dropout: Feature map distortion to regularize deep neural networks. In *Association for the Advancement of Artificial Intelligence*, pp. 5964–5971, 2020.

Wang, H., Yang, W., Zhao, Z., Luo, T., Wang, J., and Tang, Y. Rademacher dropout: An adaptive dropout for deep neural network via optimizing generalization gap. *Neurocomputing*, pp. 177–187, 2019.

Xie, B., Liang, Y., and Song, L. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pp. 1216–1224, 2017a.

Xie, P., Deng, Y., and Xing, E. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110*, 2015.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017b.

Yang, K., Gkatzelis, V., and Stoyanovich, J. Balanced ranking with diversity constraints. In *International Joint Conference on Artificial Intelligence*, pp. 6035–6042.

Yu, Y., Li, Y.-F., and Zhou, Z.-H. Diversity regularized machine. In *International Joint Conference on Artificial Intelligence*, 2011.

Zhai, K. and Wang, H. Adaptive dropout with rademacher complexity regularization. In *International Conference on Learning Representations*, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations 2018*, 2018.

6. Appendix

6.1. Theoretical assumptions

The full theoretical characterization of the setup can be summarized in the following assumptions:

Assumptions 1.

- *The activation function of the hidden layer, $\phi(t)$, is a positive L_ϕ-Lipschitz continuous function.*

- *The input vector $\boldsymbol{x} \in \mathbb{R}^D$ satisfies $||\boldsymbol{x}||_2 \leq C_1$.*

- *The output scalar $y \in \mathbb{R}$ satisfies $|y| \leq C_2$.*

- *The weight matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_M] \in \mathcal{R}^{D \times M}$ connecting the input to the hidden layer satisfies $||\boldsymbol{w}_m||_2 \leq C_3$.*

- *The weight vector $\boldsymbol{v} \in \mathbb{R}^M$ connecting the hidden-layer to the output neuron satisfies $||\boldsymbol{v}||_2 \leq C_4$.*

- *The hypothesis class is $\mathcal{F} = \{f | f(\boldsymbol{x}) = \sum_{m=1}^{M} v_m \phi_m(\boldsymbol{x}) = \sum_{m=1}^{M} v_m \phi(\boldsymbol{w}_m^T \boldsymbol{x})\}$.*

- *Loss function set is $\mathcal{A} = \{l | l(f(\boldsymbol{x}), y) = \frac{1}{2}|f(\boldsymbol{x}) - y|^2\}$.*

- *With a probability τ, for $n \neq m$, $(\phi_n(\boldsymbol{x}) - \phi_m(\boldsymbol{x}))^2 = (\phi(\boldsymbol{w}_n^T \boldsymbol{x}) - \phi(\boldsymbol{w}_m^T \boldsymbol{x}))^2 \geq d_{min}^2$.*

6.2. Section 2 proofs:

We recall the following two lemmas related to the estimation error and three Rademacher complexity:

Lemma 1. *(Bartlett & Mendelson, 2002) For $\mathcal{F} \in \mathbb{R}^{\mathcal{X}}$, assume that $g : \mathbb{R} \to \mathbb{R}$ is a L_g-Lipschitz continuous function and $\mathcal{A} = \{g \circ f : f \in \mathcal{F}\}$. Then we have*

$$\mathcal{R}_N(\mathcal{A}) \leq L_g \mathcal{R}_N(\mathcal{F}). \tag{9}$$

Lemma 2. *(Xie et al., 2015; Bartlett & Mendelson, 2002) With a probability of at least $1 - \delta$*

$$L(\hat{f}) - L(f^*) \leq 4\mathcal{R}_N(\mathcal{A}) + B\sqrt{\frac{2\log(2/\delta)}{N}} \tag{10}$$

for $B \geq \sup_{\boldsymbol{x},y,f} |l(f(\boldsymbol{x}), y)|$, where $\mathcal{R}_N(\mathcal{A})$ is the Rademacher complexity of the loss set $\mathcal{A}$.

Lemma 3. *(Xie et al., 2015) Under Assumptions 1, the Rademacher complexity $\mathcal{R}_N(\mathcal{F})$ of the hypothesis class $\mathcal{F} = \{f | f(\boldsymbol{x}) = \sum_{m=1}^{M} v_m \phi_m(\boldsymbol{x}) = \sum_{m=1}^{M} v_m \phi(\boldsymbol{w}_m^T \boldsymbol{x})\}$ can be upper-bounded as follows:*

$$\mathcal{R}_N(\mathcal{F}) \leq \frac{2L_\phi C_{134}\sqrt{M}}{\sqrt{N}} + \frac{C_4|\phi(0)|\sqrt{M}}{\sqrt{N}}, \tag{11}$$

where $C_{134} = C_1 C_3 C_4$ and $\phi(0)$ is the output of the activation function at the origin.

Lemma 2 bounds the estimation error using the Rademacher complexity and the supremum of the loss class Lemma 3 provides an upper-bound of the Rademacher complexity for the hypothesis class.

In the following proofs, we use Lipschitz analysis. In particular, a function $f : \mathbb{A} \to \mathbb{R}$, $\mathbb{A} \subset \mathbb{R}^n$, is said to be L-Lipschitz, if there exist a constant $L \geq 0$, such that $|f(\boldsymbol{a}) - f(\boldsymbol{b})| \leq L||\boldsymbol{a} - \boldsymbol{b}||$ for every pair of points $\boldsymbol{a}, \boldsymbol{b} \in A$. Moreover:

- $\sup_{\boldsymbol{x} \in A} f \leq \sup(L||\boldsymbol{x}|| + f(0))$.

- if f is continuous and differentiable, $L = \sup |f'(\boldsymbol{x})|$.

6.2.1. PROOF OF THEOREM 1

In order to find an upper-bound for our estimation error, we start by deriving an upper bound for $\sup_{\boldsymbol{x},f} |f(\boldsymbol{x})|$;

Lemma 4. *Under Assumptions 1, with a probability at least τ^Q, we have*

$$\sup_{\boldsymbol{x},f} |f(\boldsymbol{x})| \leq \sqrt{\mathcal{J}}, \tag{12}$$

where Q is equal to the number of neuron pairs defined by M neurons, i.e. $Q = \frac{M(M-1)}{2}$, and $\mathcal{J} = C_4^2 \big(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2) \big)$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Proof.

$$f^2(\boldsymbol{x}) = \left(\sum_{m=1}^M v_m \phi_m(\boldsymbol{x}) \right)^2 \leq \left(\sum_{m=1}^M ||\boldsymbol{v}||_\infty \phi_m(\boldsymbol{x}) \right)^2$$

$$\leq ||\boldsymbol{v}||_\infty^2 \left(\sum_{m=1}^M \phi_m(\boldsymbol{x}) \right)^2 \leq C_4^2 \left(\sum_{m=1}^M \phi_m(\boldsymbol{x}) \right)^2$$

$$= C_4^2 \left(\sum_{m,n} \phi_m(\boldsymbol{x}) \phi_n(\boldsymbol{x}) \right) = C_4^2 \left(\sum_m \phi_m(\boldsymbol{x})^2 + \sum_{m \neq n} \phi_n(\boldsymbol{x}) \phi_m(\boldsymbol{x}) \right) \tag{13}$$

We have $\sup_m w, \boldsymbol{x}\phi(\boldsymbol{x}) < \sup(L_\phi |\boldsymbol{w}^T \boldsymbol{x}| + \phi(0))$ because ϕ is L_ϕ-Lipschitz. Thus, $||\phi||_\infty < L_\phi C_1 C_3 + \phi(0) = C_5$. For the first term in equation 13, we have $\sum_m \phi_m(\boldsymbol{x})^2 < M(L_\phi C_1 C_3 + \phi(0))^2 = MC_5^2$. The second term, using the identity $\phi_m(\boldsymbol{x})\phi_n(\boldsymbol{x}) = \frac{1}{2}\big(\phi_m(\boldsymbol{x})^2 + \phi_n(\boldsymbol{x})^2 - (\phi_m(\boldsymbol{x}) - \phi_n(\boldsymbol{x}))^2 \big)$, can be rewritten as

$$\sum_{m \neq n} \phi_m(\boldsymbol{x})\phi_n(\boldsymbol{x}) = \frac{1}{2} \sum_{m \neq n} \phi_m(\boldsymbol{x})^2 + \phi_n(\boldsymbol{x})^2 - \big(\phi_m(\boldsymbol{x}) - \phi_n(\boldsymbol{x}) \big)^2. \tag{14}$$

In addition, we have with a probability τ, $||\phi_m(\boldsymbol{x}) - \phi_n(\boldsymbol{x})||_2^2 \geq d_{min}$ for $m \neq n$. Thus, we have with a probability at least τ^Q:

$$\sum_{m \neq n} \phi_m(\boldsymbol{x})\phi_n(\boldsymbol{x}) \leq \frac{1}{2} \sum_{m \neq n} (2C_5^2 - d_{min}^2)$$

$$= M(M-1)(C_5^2 - d_{min}^2/2). \tag{15}$$

By putting everything back to equation 13, we have with a probability τ^Q,

$$f^2(\boldsymbol{x}) \leq C_4^2 \Big(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2) \Big) = \mathcal{J}. \tag{16}$$

Thus, with a probability τ^Q,

$$\sup_{\boldsymbol{x},f} |f(\boldsymbol{x})| \leq \sqrt{\sup_{\boldsymbol{x},f} f(\boldsymbol{x})^2} \leq \sqrt{\mathcal{J}}. \tag{17}$$

$\square$

Note that in Lemma 4, we have expressed the upper-bound of $\sup_{\boldsymbol{x},f} |f(\boldsymbol{x})|$ in terms of d_{min}. Using this bound, we can now find an upper-bound for $\sup_{\boldsymbol{x},f,y} |l(f(\boldsymbol{x}), y)|$ in the following lemma:

Lemma 5. *Under Assumptions 1, with a probability at least τ^Q, we have*

$$\sup_{\boldsymbol{x},y,f} |l(f(\boldsymbol{x}), y)| \leq \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2 \tag{18}$$

Proof. We have $\sup_{\boldsymbol{x},y,f} |f(\boldsymbol{x}) - y| \leq \sup_{\boldsymbol{x},y,f}(|f(\boldsymbol{x})| + |y|) = (\sqrt{\mathcal{J}} + C_2)$. Thus $sup_{x,y,f} |l(f(x), y)| \leq \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2$. $\square$

The main goal is to analyze the estimation error bound of the neural network and to see how its upper-bound is linked to the diversity, expressed by d_{min}, of the different neurons. Now we can prove our main Theorem 1:

Theorem 1 *Under Assumptions 1, there exist a constant A, such that with probability at least $\tau^Q(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq \Big(\sqrt{\mathcal{J}} + C_2 \Big) A + \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2\log(2/\delta)}{N}} \tag{19}$$

where $\mathcal{J} = C_4^2 \big(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2) \big)$, and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Proof. Given that $l(\cdot)$ is K-Lipschitz with a constant $K = sup_{x,y,f}|f(\boldsymbol{x}) - y| \leq (\sqrt{\mathcal{J}} + C_2)$, and using Lemma 1, we can show that $\mathcal{R}_N(\mathcal{A}) \leq K\mathcal{R}_N(\mathcal{F}) \leq (\sqrt{\mathcal{J}} + C_2)\mathcal{R}_N(\mathcal{F})$. For $\mathcal{R}_N(\mathcal{F})$, we use the bound found in Lemma 3. Using Lemmas 2 and 5, we have

$$L(\hat{f}) - L(f^*) \leq 4\Big(\sqrt{\mathcal{J}} + C_2 \Big)\Big(2L_\phi C_{134} + C_4|\phi(0)| \Big)\frac{\sqrt{M}}{\sqrt{N}}$$

$$+ \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2\log(2/\delta)}{N}} \tag{20}$$

where $C_{134} = C_1 C_3 C_4$, $\mathcal{J} = C_4^2 \big(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2) \big)$, and $C_5 = L_\phi C_1 C_3 + \phi(0)$. Thus, taking $A = 4\Big(2L_\phi C_{134} + C_4|\phi(0)| \Big)\frac{\sqrt{M}}{\sqrt{N}}$ completes the proof. $\square$

6.2.2. PROOF OF THEOREMS 2 AND 3

Similar to the proofs of Lemmas 7 and 8 in (Xie et al., 2015), we can show the following two lemmas:

Lemma 6. *Using the hinge loss, we have with probability at least* $\tau^Q(1-\delta)$

$$L(\hat{f}) - L(f^*) \leq 4\Big(2L_\phi C_{134} + C_4|\phi(0)|\Big)\frac{\sqrt{M}}{\sqrt{N}}$$
$$+ (1 + \sqrt{\mathcal{J}})\sqrt{\frac{2\log(2/\delta)}{N}} \quad (21)$$

where $C_{134} = C_1 C_3 C_4$, $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2))$, *and* $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Lemma 7. *Using the logistic loss* $l(f(x), y) = \log(1 + e^{-yf(x)})$, *we have with probability at least* $\tau^Q(1-\delta)$

$$L(\hat{f}) - L(f^*) \leq \frac{4}{1 + e^{\sqrt{-\mathcal{J}}}}\Big(2L_\phi C_{134} + C_4|\phi(0)|\Big)\frac{\sqrt{M}}{\sqrt{N}}$$
$$+ \log(1 + e^{\sqrt{\mathcal{J}}})\sqrt{\frac{2\log(2/\delta)}{N}} \quad (22)$$

where $C_{134} = C_1 C_3 C_4$, $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2))$, *and* $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Taking $A = 4\Big(2L_\phi C_{134} + C_4|\phi(0)|\Big)\frac{\sqrt{M}}{\sqrt{N}}$ in Lemma 6 and Lemma 7 completes the proofs.

6.2.3. PROOF OF THEOREM 4

Theorem 4 There exist a constant A such that, with probability of at least $\prod_{p=0}^{P-1}(\tau^p)^{Q^p}(1-\delta)$, we have

$$L(\hat{f}) - L(f^*) \leq (\sqrt{\mathcal{J}^P} + C_2)A$$
$$+ \frac{1}{2}\Big(\sqrt{\mathcal{J}^P} + C_2\Big)^2\sqrt{\frac{2\log(2/\delta)}{N}} \quad (23)$$

where Q^p is the number of neuron pairs in the pth layer, defined as $Q^p = \frac{M^p(M^p-1)}{2}$, and $\mathcal{J}^P$ is defined recursively using the following identities: $\mathcal{J}^0 = C_3^0 C_1$ and $\mathcal{J}^p = M^p C^{p2}\big(M^{p2}(L_\phi C_3^{p-1}\mathcal{J}^{p-1} + \phi(0))^2 - M(M-1)\frac{d_{min}^{p}{}^2}{2}\big)$, for $p = 1, \ldots, P$.

Proof. Lemma 5 in (Xie et al., 2015) provides an upper-bound for the hypothesis class. We denote by v^p denote the outputs of the pth hidden layer before applying the activation function:

$$v^0 = [w_1^{0T}x, \ldots, w_{M^0}^{0T}x] \quad (24)$$

$$v^p = \Big[\sum_{j=1}^{M^{p-1}} w_{j,1}^p\phi(v_j^{p-1}), \ldots, \sum_{j=1}^{M^{p-1}} w_{j,M^p}^p\phi(v_j^{p-1})\Big] \quad (25)$$

$$v^p = [w_1^{pT}\phi^p, \ldots, w_{M^p}^{pT}\phi^p], \quad (26)$$

where $\phi^p = [\phi(v_1^{p-1}), \cdots, \phi(v_{M^{p-1}}^{p-1})]$. We have

$$||v^p||_2^2 = \sum_{m=1}^{M^p}(w_m^{pT}\phi^p)^2 \quad (27)$$

and $w_m^{pT}\phi^p \leq C_3^p\sum_n\phi_n^p$. Thus,

$$||v^p||_2^2 \leq \sum_{m=1}^{M^p}(C_3^p\sum_n\phi_n^p)^2 = M^p C_3^{p2}(\sum_n\phi_n^p)^2$$
$$= M^p C_3^{p2}\sum_{mn}\phi_m^p\phi_n^p. \quad (28)$$

We use the same decomposition trick of $\phi_m^p\phi_n^p$ as in the proof of Lemma 4. We need to bound $\sup_x\phi^p$:

$$\sup_x\phi^p < \sup(L_\phi|w_j^{p-1T}v^{p-1}| + \phi(0))$$
$$< L_\phi||W^{p-1}||_\infty||v^{p-1}||_2^2 + \phi(0). \quad (29)$$

Thus, we have

$$||v^p||_2^2 \leq M^p C^{p2}\big(M^2(L_\phi C_3^{p-1}||v^{p-1}||_2^2 + \phi(0))^2$$
$$- M(M-1)d_{min}^2/2)\big) = \mathcal{J}^P. \quad (30)$$

We found a recursive bound for $||v^p||_2^2$, we note that for $p = 0$, we have $||v^0||_2^2 \leq ||W^0||_\infty C_1 \leq C_3^0 C_1 = \mathcal{J}^0$. Thus,

$$\sup_{x, f^P \in \mathcal{F}^P}|f(x)| = \sup_{x, f^P \in \mathcal{F}^P}|v^P| \leq \sqrt{\mathcal{J}^P}. \quad (31)$$

By replacing the variables in Lemma 2, we have

$$L(\hat{f}) - L(f^*) \leq 4(\sqrt{\mathcal{J}^P} + C_2)\Bigg(\frac{(2L_\phi)^P C_1 C_3^0}{\sqrt{N}}\prod_{p=0}^{P-1}\sqrt{M^p}C_3^p$$
$$+ \frac{|\phi(0)|}{\sqrt{N}}\sum_{p=0}^{P-1}(2L_\phi)^{P-1-p}\prod_{j=p}^{P-1}\sqrt{M^j}C_3^j\Bigg)$$
$$+ \frac{1}{2}\Big(\sqrt{\mathcal{J}^P} + C_2\Big)^2\sqrt{\frac{2\log(2/\delta)}{N}}$$

Taking $A = \big(\frac{(2L_\phi)^P C_1 C_3^0}{\sqrt{N}}\prod_{p=0}^{P-1}\sqrt{M^p}C_3^p + \frac{|\phi(0)|}{\sqrt{N}}\sum_{p=0}^{P-1}(2L_\phi)^{P-1-p}\prod_{j=p}^{P-1}\sqrt{M^j}C_3^j\big)$ completes the proof. $\square$

6.2.4. PROOFS OF THEOREMS 5 AND 6

Theorem 5 *For a multivariate regression trained with the squared error, there exist a constant A such that, with probability at least* $\tau^Q(1-\delta)$, *we have*

$$L(\hat{f}) - L(f^*) \leq (\sqrt{\mathcal{J}} + C_2)A + \frac{D}{2}(\sqrt{\mathcal{J}} + C_2)^2\sqrt{\frac{2\log(2/\delta)}{N}} \quad (32)$$

where $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2))$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$

Proof. The squared loss $||f(\boldsymbol{x}) - \boldsymbol{y}||^2$ can be decomposed into D terms $(f(\boldsymbol{x})_k - y_k)^2$. Using Theorem 1, we can derive the bound for each term and thus, we have:

$$L(\hat{f}) - L(f^*) \leq 4D(\sqrt{\mathcal{J}} + C_2)\left(2L_\phi C_{134} + C_4|\phi(0)|\right)\frac{\sqrt{M}}{\sqrt{N}}$$

$$+ \frac{D}{2}(\sqrt{\mathcal{J}} + C_2)^2\sqrt{\frac{2\log(2/\delta)}{N}}, \quad (33)$$

where $C_{134} = C_1 C_3 C_4$, $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2))$, and $C_5 = L_\phi C_1 C_3 + \phi(0)$. Taking $A = 4D\left(2L_\phi C_{134} + C_4|\phi(0)|\right)\frac{\sqrt{M}}{\sqrt{N}}$ completes the proof.

$\square$

Theorem 6 *For a multi-class classification task using the cross-entropy loss, there exist a constant A such that, with probability at least $\tau^Q(1 - \delta)$, we have*

$$L(\hat{f}) - L(f^*) \leq \frac{A}{D - 1 + e^{-2\sqrt{\mathcal{J}}}}$$

$$+ \log\left(1 + (D-1)e^{2\sqrt{\mathcal{J}}}\right)\sqrt{\frac{2\log(2/\delta)}{N}} \quad (34)$$

where $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2/2))$ and $C_5 = L_\phi C_1 C_3 + \phi(0)$.

Proof. Using Lemma 9 in (Xie et al., 2015), we have $\sup_{f,\boldsymbol{x},\boldsymbol{y}} l = \log\left(1 + (D-1)e^{2\sqrt{\mathcal{J}}}\right)$ and l is $\frac{D-1}{D-1+e^{-2\sqrt{\mathcal{J}}}}$-Lipschitz. Thus, using the decomposition property of the Rademacher complexity, we have

$$\mathcal{R}_n(\mathcal{A}) \leq \frac{4D(D-1)}{D - 1 + e^{-2\sqrt{\mathcal{J}}}}\left(\frac{2L_\phi C_{134}\sqrt{M}}{\sqrt{N}} + \frac{C_4|\phi(0)|\sqrt{M}}{\sqrt{N}}\right). \quad (35)$$

Taking $A = 4D(D-1)\left(\frac{2L_\phi C_{134}\sqrt{M}}{\sqrt{N}} + \frac{C_4|\phi(0)|\sqrt{M}}{\sqrt{N}}\right)$ completes the proof. $\square$

6.3. Experimental protocol

we conduct image classification experiments on the ImageNet-2012 classification dataset (Russakovsky et al., 2015) using the ResNet50 model (He et al., 2016). The diversity term is applied on the last intermediate layer, i.e., the global average pooling layer for both DeCov and our method. We use the standard augmentation practice for this dataset as in (Zhang et al., 2018; Huang et al., 2017; Cogswell et al., 2016). All the models are trained with a batch size of 256 for 100 epoch using SGD with Nesterov Momentum of 0.9 and a weight decay of 0.0001. The learning rate is initially set to 0.1 and decreases at epochs 30, 60, 90 by a factor of 10.

6.4. Additional experiments

We start by evaluating our proposed diversity approach on two image datasets: CIFAR10 and CIFAR100 (Krizhevsky et al., 2009). They contain 60,000 (50,000 train/10,000 test) 32×32 images grouped into 10 and 100 distinct categories, respectively. We split the original training set (50,000) into two sets: we use the first 40,000 images as the main training set and the last 10,000 as a validation set for hyperparameters optimization. We use our approach on three state-of-the-art CNNs: **ResNext 29-8-16**: we consider the standard ResNext Model (Xie et al., 2017b) with a 29-layer architecture, a cardinality of 8, and a width of 16. **DenseNet-12**: we use DenseNet (Huang et al., 2017) with the 40-layer architecture and a growth rate of 12. **ResNet50**: we consider the standard ResNet model (He et al., 2016) with 50 layers. We compare against the standard networks as well networks trained with DeCov diversity strategy (Cogswell et al., 2016).

All the models are trained using stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 0.0001, and a batch size of 128 for 200 epochs. The initial learning rate is set to 0.1 and is then decreased by a factor of 5 after 60, 120, and 160 epochs, respectively. We also adopt a standard data augmentation scheme that is widely used for these two datasets (He et al., 2016; Huang et al., 2017). For all models, the additional diversity term is applied on top the last intermediate layer. For the hyperparameters: The loss weight is chosen from $\{0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01\}$ for both our approach and Decov and γ in the radial basis function is chosen from $\{0.01, 0.1.1, 10, 50, 100\}$. For each approach, the model with the best validation performance is used in the test phase. Each experiment is repeated three times and we report the average performance over three iterations.

Table 1 reports the average top-1 errors of the different approaches with the three basis networks. We note that, compared to the standard approach, employing a diversity strategy generally boosts the results for all the three models and that our approach consistency outperforms both competing methods (standard and DeCov) in all the experiments. For DenseNet-12, our direct and det variants yield the best performance over CIFAR10 and CIFAR100, respectively. For ResNext-29-08-16, our approach with det term yields on average 0.25% accuracy improvement compared to the standard approach, 0.17% improvement compared to DeCov on CIFAR10. On CIFAR100, the best performance is achieved by the direct variant of our approach. For ResNet50, the three variants of our proposed approach significantly reduce the test errors over both datasets: $0.36\% - 0.54\%$ improvement on CIFAR10 and $1.86\% - 1.94\%$ on CIFAR100.

To further demonstrate the effectiveness of our approach

Table 1. Average classification errors on CIFAR10 and CIFAR100 over three iterations

		Top-1 test Error	
Model	Method	CIFAR10	CIFAR100
DenseNet-12	Standard	07.07	29.25
	DeCov	07.18	29.17
	Ours direct	06.95	29.16
	Ours det	07.04	28.78
	Ours logdet	06.96	29.15
ResNext-29-08-16	Standard	06.93	26.73
	DeCov	06.84	26.70
	Ours direct	06.74	26.54
	Ours det	06.67	26.67
	Ours logdet	06.70	26.67
ResNet50	Standard	08.27	34.06
	DeCov	08.03	32.26
	Ours direct	07.86	32.15
	Ours det	07.73	32.12
	Ours logdet	07.91	32.20

Table 2. Performance of ResNet50 with different diversity strategies on ImageNet dataset

	Top-1 Errors		Generalization
Method	Training	Testing	Gap
Standard	20.97	23.84	2.87
DeCov	20.92	23.62	2.70
Ours direct	20.88	23.58	2.70
Ours det	20.81	23.62	2.77
Ours logdet	22.57	23.64	1.07

and its ability to reduce the generalization gap in neural networks, we conduct additional image classification experiments on the ImageNet-2012 classification dataset and ResNet50. For the hyperparameters, we use the best ones for each approach obtained from CIFAR10 and CIFAR100 experiments.

Table 2 reports the test errors of the different diversity strategies. To study the effect of diversity on the generalization gap, we also report the final training errors and the generalization gap, i.e., training accuracy - test accuracy. As it can be seen, diversity (our approach and DeCov) reduces the test error of the model and yields a better performance. The best performance is achieved by our direct variant. We note that, in accordance with our theoretical findings in Section 2, using diversity indeed reduces overfitting and decreases the empirical generalisation gap of neural networks. In fact, our logdet variant reduces the empirical generalization gap of the model by 1.8% compared to the standard approach.