On the reversed bias-variance tradeoff in deep ensembles

Author(s):
Kobayashi, Seijin; von Oswald, Johannes; Grewe, Benjamin

Publication date:
2021-07-23

Permanent link:
https://doi.org/10.3929/ethz-b-000501624

On the reversed bias-variance tradeoff in deep ensembles

Seijin Kobayashi [*1] **Johannes von Oswald** [*1] **Benjamin Grewe** [1]

Abstract

Deep ensembles aggregate predictions of diverse neural networks to improve generalisation and quantify uncertainty. Here, we investigate their behavior when increasing the ensemble members' parameter size - a practice typically associated with better performance for single models. We show that under practical assumptions in the overparametrized regime far into the double descent curve, not only the ensemble test loss degrades, but common out-of-distribution detection and calibration metrics suffer as well. Reminiscent to deep double descent, we observe this phenomenon not only when increasing the single member's capacity but also as we increase the training budget, suggesting deep ensembles can benefit from early stopping. This sheds light on the success and failure modes of deep ensembles and suggests that averaging finite width models perform better than the neural tangent kernel limit for these metrics.

1. Introduction

Deep neural network ensembles (Lakshminarayanan et al., 2017) are a scalable and conceptually simple way to improve test set generalisation and obtain reliable uncertainty quantification. This led to the development of numerous algorithms (Srivastava et al., 2014; Lee et al., 2015; Gal & Ghahramani, 2016; Huang et al., 2017; Garipov et al., 2018; Wen et al., 2020; von Oswald et al., 2021; Rame & Cord, 2021) which share the common goal of finding an ensemble of models that are functionally diverse but all fit the training data well. While connections and extensions to Bayesian ensembling exist (Pearce et al., 2020; He et al., 2020; Wilson & Izmailov, 2020), uncertainty estimates of deep ensembles can still be misleading and insufficient (Zhou et al., 2002; Fort et al., 2020; Rahaman & Thiery, 2020; Ashukha et al., 2020; Nixon et al., 2020; Wen et al., 2021).

[*]Equal contribution [1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland. Correspondence to: Seijin Kobayashi <seijink@ethz.ch>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

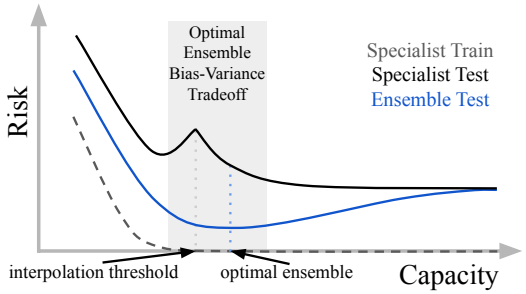

Figure 1. An illustration of the bias-variance tradeoff observed for deep ensembles in commonly used setups. Single models undergo a double descent with improved generalisation when capacity increases. Deep ensembles observe a turning point after the interpolations threshold when members interpolate the data but variance between members remains high.

Despite these drawbacks, deep ensembles remain attractive for out-of-distribution detection and uncertainty quantification due to their implementation simplicity and scalability to modern deep learning applications (Ovadia et al., 2019; Leibig et al., 2017).

Inspired by (Geiger et al., 2020a), in this work we analyse the performance of deep ensembles empirically when increasing the model capacity of each of its ensemble members. This is motivated by the recently observed *double-descent* phenomenon (Belkin et al., 2018; Advani & Saxe, 2017; Opper & Kinzel, 1996): when increasing the model size beyond capacity needed to interpolate the training data, test loss of neural networks trained with (stochastic) gradient descent decreases again, often surpassing the optimal performance reached in the under-parametrized regime (c.f. Figure 1). This phenomenon can be theoretically studied and partly explained under the light of the neural tangent kernel (Jacot et al., 2018; Geiger et al., 2020a) and supports the large set of evidence as well as the common belief in support of improved performance as a function of model size (Nakkiran et al., 2020; Krizhevsky et al., 2012; Brown et al., 2020).

Surprisingly this picture can change when shifting focus towards ensembles of deep networks, as reported in (Geiger et al., 2020a) and reproduced by (Lee et al., 2020). Indeed, when pushing the capacity of each ensemble member

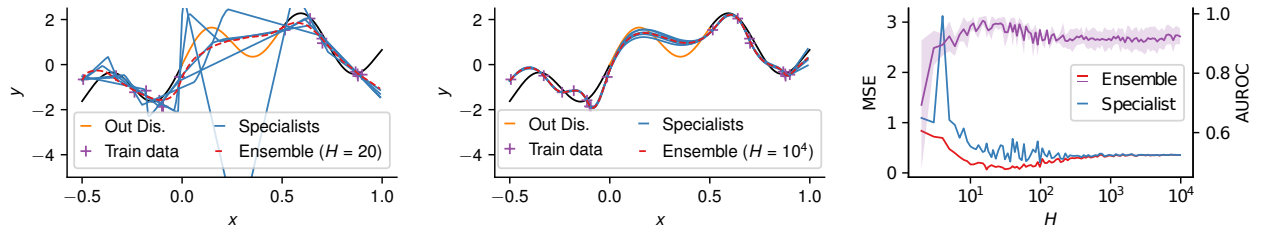

Figure 2. Specialist and ensemble over varying hidden layer size H on a low dimensional regression problem. After a certain capacity $H_E \approx 20$ ensemble performance starts to decline. *Left:* High variance of the specialist predictions outside the training is observed for limited capacity. *Center:* Specialists coincide within training distribution and show small variance outside. *Right:* Increased OOD detection (purple) performance measured by the prediction variance as well as improved MSE for an ensemble (red) of limited capacity ($H = 20$). For $H \to \infty$ the specialist (blue) and ensemble performance converge.

far into the double descent curve, while the ensemble test loss decreases at first it increases again shortly after the interpolation threshold in some cases. See Figure 5 for a reproduction of this finding. Analysed by (Geiger et al., 2020b), the authors argue for the equivalence of ensembles and single models when increasing the network layers width $H \to \infty$ for different initialisation schemes leading to *lazy training* (Chizat et al., 2019) and the *mean-field* setting. In practice, it remains unclear which of the regimes is superior but evidence exist showing a dependence on the considered problem and network architecture (Arora et al., 2019; Geiger et al., 2020b). Here, we focus on deep neural networks which weights are initialised with variance $1/H$, essential to avoid vanishing and exploding gradients (He et al., 2015; Glorot & Bengio, 2010). In this regime when $H \to \infty$, optimisation with gradient descent can be described by a deterministic neural tangent kernel (NTK), which implicitly biases models at convergence towards a single functional solution, independent from the initialization.

We argue for a *reversed* bias-variance tradeoff for deep ensembles in this practical setup. In the infinite width limit, the variance vanishes and the models converge to the NTK, which empirically performs worse than the high variance finite-width models. This suggests to avoid the following two regimes for deep ensembles: the classic underparametrised high bias and the overparametrised *modern interpolation* regime when $H \to \infty$ (Belkin et al., 2018).

Our contributions are as follows:

1. We extend the findings of (Geiger et al., 2020a) and show that the benefit of deep ensembles disappears beyond a certain capacity H_E of the ensemble members, not only in terms of test loss but also of standard out-of-distribution (OOD) and calibration metrics, converging to the performance of a single member when H becomes very large.

2. We follow (Nakkiran et al., 2020) and study test accuracy and calibration of deep ensembles with varying ensemble member capacity during training. Intriguingly, we observe that also late in training, deep ensemble accuracy as well as calibration degrades.

We define a neural network function with L layers $f = \sigma_L \circ \sigma_{L-1} \circ \dots \sigma_1$ as a chain of transformations $\sigma_l(h) = \sigma(W_l h + b_l)$ with W_l the weight matrix, b_l the bias vector of layer l followed by $\sigma : \mathbb{R} \to \mathbb{R}$, a non-linearity applied elementwise. Throughout our experiments we use $\sigma(x) = max(0, x)$ i.e. the rectified linear unit (ReLU). We denote the width of the layer l by H_l, H_0 as the input dimension and H_L the output dimension. We use models with the same width $H_l = H$ across all hidden layers in the following. We call the ensemble members *specialists* and denote the number of specialists by K. Note that to obtain a deep ensemble throughout this study, we rely solely on different random weight initialisation via Kaiming (He et al., 2015), as well the variability introduced by minibatch gradient descent in the CIFAR-10 experiments.

2. Experiments

2.1. Low dimensional regression problem

We begin with an illustrative example of a 1-dimensional regression problem to give an intuition on the reduced variance outside the training data of deep ensembles. We follow (Gal, 2016) and create a training dataset by sampling uniformly $N = 20$ input data x from $[-0.5, 0] \cup [0.5, 1]$ while

$$y = x + \sin(4x) + \sin(13x) \tag{1}$$

with $x + \epsilon$ and $\epsilon \sim \mathcal{N}(0, 0.3^2)$. The model is a one-hidden layer neural network $f(x) = \sigma_2 \circ \sigma_1$ with frozen weight of the first layer (Belkin et al., 2018). We initialise the weights and bias parameter $W_1 \sim \mathcal{N}(0, 1)$ and $b_1 \sim \mathcal{N}(0, 1)$ but scale the variance of the second weight matrix with the number of hidden units H_1 i.e. $W_2 \sim \mathcal{N}(0, 2/H_1)$ and set $b_2 = 0$. We train the network for 10^5 steps by minimising the mean-squared error loss (MSE) with ADAM (Kingma & Ba, 2015) which we found to work the best for this problem. We also scaled the learning rate with the hidden layer size $\gamma = \frac{0.1}{H_1}$. Note that the trainable parameters are $W =$

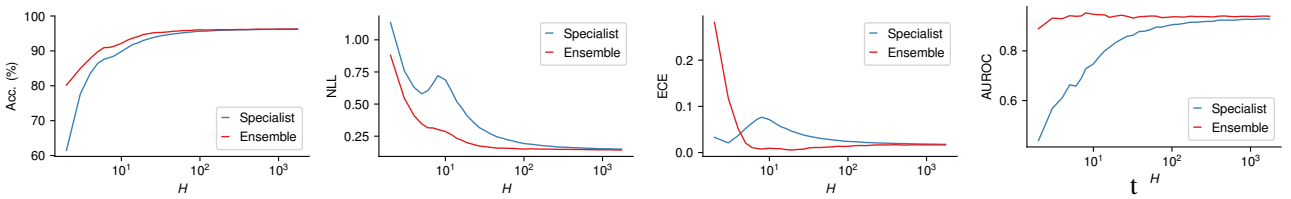

Figure 3. Ensemble performance on a 2 hidden layer fully-connected neural network with different width H trained with full gradient decent on a subset of MNIST with the cross-entropy loss. *Left & center left*: The single specialist as well as the ensemble observe double descent behaviour in accuracy and test loss (NLL) around $H = 10$. *Right & center right*: Expected calibration error (ECE) for the deep ensemble is lowest roughly around $H = 20$ but increases consistently for $H \to \infty$. Similar behaviour for OOD detection (AUROC) is obsereved which peaks around $H \approx 10$ but degrades for growing capacity to the single models performance.

$\{W_2, b_2\}$. We found our results dependant on the variability of the data creation and therefore averaged all results over 3 different draws of training data. K is set to 50.

Our results are depicted in Figure 2. In this setup we do not observe a clear double descent test MSE for the specialist when varying H. Nevertheless test loss decreases steadily for $H \to \infty$ as predicted by double descent.

A different behaviour for the ensemble when varying H is observed. While improving upon the specialist performance over all capacity values H after $H_E \approx 15$ we see decreasing performance for growing specialist capacity $H \gg H_E$. Note that the ensemble and specialist show strong resemblance in MSE after $H > 1000$ as described by (Geiger et al., 2020a). Intriguingly, when using the variance of the specialist's predictions for out-of-distribution data, 1000 data points sampled from $[0, 0.5]$, and 1000 test data sampled from the training distribution, a similar trend occurs. Ensembles with $H \approx 20$ improve upon all other capacity models notably even when $H \to \infty$.The functional diversity of the ensemble with $H = 20$ compared to $H = 10^4$ is shown in the left and right panel of Figure 2 resp.

2.2. MNIST

We next study the common classification problem MNIST (LeCun et al., 2010) to investigate the robustness of the previous finding. Following (Belkin et al., 2018), we decrease the training set size to $N = 10000$ but use the cross-entropy loss with help of the common softmax operation after the output layer. K is set to 10.

The fully-connected 2-layer neural network is again optimised with full gradient descent with a learning rate of 0.1 and 10000 epochs for all H. The weights $W_2, W_3 \sim \mathcal{N}(0, 2/H)$ are initialised with kaiming initialisation. On top of the negative log likelihood (NLL), we measure the commonly used *expected calibration error* (ECE) (Naeini et al., 2015). Finally, we use the predictive entropy, which should be low for in-distribution and high for out-distribution data, for out-of-distrobution detection on FashionMNIST (Xiao et al., 2017) and permuted MNIST. Performance is summarized through the area under the receiver operating characteristics curve (AUROC).

We observe a clear and prominent double descent for the specialist as well as the ensemble, see Figure 3. The NLL of the ensemble decreases monotonically and stays constant after $H \approx 100$ but we do not observe an advantage for ensembles with reduced capacity H. We observe a different picture for the ECE and OOD. In both cases, the ensemble performance peaks shortly after the second descent curve before degrading again. For $H \to \infty$ we observe the ensemble converging and closely resembling the specialists performance on all four considered metrics.

2.3. CIFAR-10 and CIFAR-100

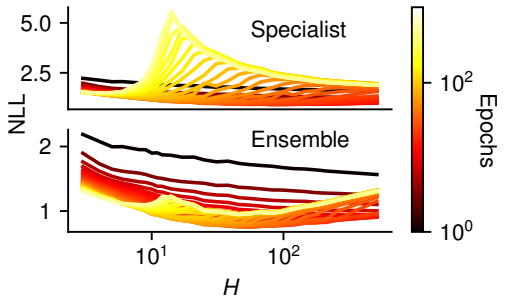

Figure 4. NLL when training a convolutional neural network with on CIFAR-10 for varying hidden layer width H. NLL for the specialist and deep ensemble observe a clear double descent while the ensemble NLL increases again after $H \approx 100$. This trend is preserved over training (epochs). Advantages of early stopping for the deep ensemble and specialist are clearly visible for all capacities after the interpolation threshold.

We now focus on the more difficult CIFAR-10 and CIFAR-100 image classification problem.

For CIFAR-10 the model consists of 3-convolutional layers followed by one fully-connected layer and is now trained with stochastic gradient descent with a batch size of 250, learning rate 0.01 and weight-decay (0.0005) for 1000 epochs for all H. Note that this setup strongly resembles common neural network training setups. Weight decay stabilizes training around the interpolation threshold but shows

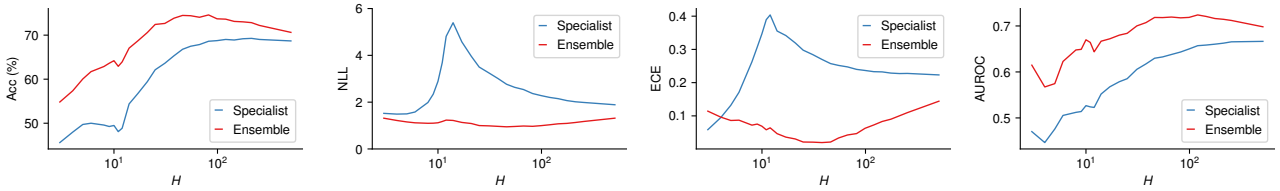

Figure 5. Specialist and ensemble performance on a 3-convolutional neural network with different width H trained with stochastic gradient decent on CIFAR-10 with the cross-entropy loss and weight decay (0.0005). *Left & Center left*: A clear double descent for ensemble and specialist occurs for NLL and test set accuracy. After an optimal ensemble capacity around $H_E = 90$ the ensemble performance decreases. *Right & Center right*: Expected calibration error (ECE) for the ensemble is lowest roughly around H_E but increases consistently for $H \to \infty$. This trend is reoccuring for OOD detection which peaks around H_E and declines afterwards.

negligible qualitative difference otherwise, see Figure 10 in the SM. We also show successful optimisation to virtually zero train loss for $H > 10$ in the SM Figure 7. In order to quantify OOD detection of the models trained on CIFAR-10, we measure the average AUROC of discriminating based on the predictive entropy between the CIFAR-10 test data and SVHN, Netzer et al. (2011), LSUN, Yu et al. (2015), Tiny ImageNet as well as on CIFAR-100 subset. K is set to 8.

As already observed by (Geiger et al., 2020a) a clear picture arises as shown in Figure 5: deep ensemble NLL reaches a minimum around $H_e \approx 60$ and increases afterwards. We extend this finding to the ECE, OOD AUROC and robustness to data corruption (Figure 13 in the SM), and see this trend reoccurring. Intriguingly, the optimal capacity for these metrics roughly coincide and also translates into decreasing test set accuracy after H_e, see Figure 5. Reminiscent to observations of (Nakkiran et al., 2020) we also see the ensemble performance degrading after a certain training epoch $T_E \approx 100$. Figure 4 and Figure 6 display NLL and ODD over the course of training visualised with brighter colors corresponding to later stages of training (in log scale). NLL as well as OOD improve during training, peak around $H_T \approx 100$ and then degrade. This indicates a clear advantage of early stopping for the ensemble. See SM for similar behaviour of the ECE (Figure 12). Note that the optimal capacity over the training duration shifts but performance always degrades with growing H after a certain threshold.

Results on CIFAR-100 show similar trends and are described in detail in the SM.

3. Discussion

We empirically studied deep ensembles under the light of the double descent phenomenon observed when increasing the ensemble members capacity $H \to \infty$ while using standard initialization schemes. Our results indicate a clear trend emerging: after a certain threshold H_E, test risk as well as common calibration, OOD detection and robustness metrics degrade. After this turning point, we observed deep ensembles converging to the performance of single models leading to a negligible effect of model averaging. We therefore support suggestions by (Geiger et al., 2020a) to choose the network capacity of ensemble members wisely. This not only leads to decreased computational overhead and memory savings but also to performance improvements on in- and out-of-distribution data. Similar suggestions have been discussed for large scale neural networks ensembles under a fixed memory budget (Lobacheva et al., 2020; Wang et al., 2021; Littwin et al., 2020). Further, deep ensemble improvements on CIFAR-10 were obtained by early stopping.

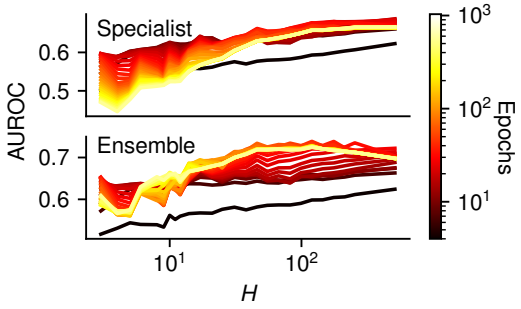

Figure 6. Out-of-distribution detection (AUROC) when training a convolutional neural network on CIFAR-10 with varying hidden layer width H. OOD for the specialist and deep ensemble observe increase with growing H before a peak around $H \approx 100$ for the deep ensemble is observed. This trend is preserved over training (epochs) and advantages of early stopping for the deep ensemble and specialist are clearly visible for almost all model.

The aforementioned findings suggest that recent interest in economical ensembles is justified. Common techniques involve parameter sharing and hence reduce the capacity of single ensemble members. This would result in performance and calibration improvements if the ensemble is pushed into a beneficial scarce capacity regime. Further investigations of deep ensemble uncertainty and calibration performance in the *mean-field* regime are left for further study.

We identified success modes of deep ensembling and conclude that counter-intuitively increasing the ensemble members size can hurt performance of deep ensembles. This goes against common beliefs in deep learning where increasing model and data set size is typically associated with performance improvements.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks, 2017.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 33*. 2019.

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

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*, 2018.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems 33*. 2019.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective, 2020.

Gal, Y. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems 31*. 2018.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d' Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020a.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, Nov 2020b. ISSN 1742-5468. doi: 10.1088/1742-5468/abc4de. URL http://dx.doi.org/10.1088/1742-5468/abc4de.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*. PMLR, 2010.

He, B., Lakshminarayanan, B., and Teh, Y. W. Bayesian deep ensembles via the neural tangent kernel. In *Advances in Neural Information Processing Systems 34*. 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. *arXiv preprint arXiv:1502.01852*, February 2015.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and surface variations, 2019.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 32*. 2018.

Kingma, D. P. and Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*. 2017.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *ATT Labs. Available: http://yann.lecun.com/exdb/mnist*, 2010.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study, 2020.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. Why M heads are better than one: training a diverse ensemble of deep networks. *arXiv preprint: arXiv:1511.06314*, November 2015.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7, 2017.

Littwin, E., Myara, B., Sabah, S., Susskind, J., Zhai, S., and Golan, O. Collegial ensembles. In *Advances in Neural Information Processing Systems 34*. 2020.

Lobacheva, E., Chirkova, N., Kodryan, M., and Vetrov, D. On power laws in deep ensembles. In *Advances in Neural Information Processing Systems 33*. 2020.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pp. 2901–2907, 2015.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Nixon, J., Tran, D., and Lakshminarayanan, B. Why aren't bootstrapped neural networks better? In *NeurIPS Workshop: I Can't Believe It's Not Better Workshop*, 2020.

Opper, M. and Kinzel, W. *Statistical Mechanics of Generalization*. Springer New York, 1996.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 33*, 2019.

Pearce, T., Leibfried, F., and Brintrup, A. Uncertainty in neural networks: Approximately bayesian ensembling. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 2020.

Rahaman, R. and Thiery, A. H. Uncertainty quantification and deep ensembles, 2020.

Rame, A. and Cord, M. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *International Conference on Learning Representations*, 2021.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

von Oswald, J., Kobayashi, S., Sacramento, J., Meulemans, A., Henning, C., and F. Grewe, B. Neural networks with late-phase weights. In *International Conference on Learning Representations*, 2021.

Wang, X., Kondratyuk, D., Christiansen, E., Kitani, K. M., Movshovitz-Attias, Y., and Eban, E. On the surprising efficiency of committee-based models, 2021.

Wen, Y., Tran, D., and Ba, J. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

Wen, Y., Jerfel, G., Muller, R., Dusenberry, M. W., Snoek, J., Lakshminarayanan, B., and Tran, D. Combining ensembles and data augmentation can harm your calibration, 2021.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhou, Z.-H., Wu, J., and Tang, W. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 2002.

A. Additional visualisations & results

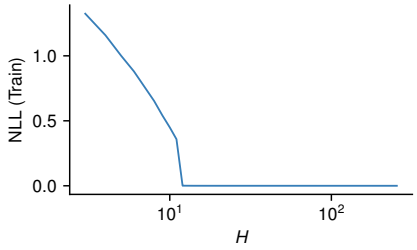

Figure 7. Final train loss after 1000 epochs when minimizing the cross entropy loss in the CIFAR-10 setup. The interpolation threshold can be observed after $H = 10$ after which the loss becomes virtually 0.

Mean Corruption Error experiment In order to investigate the robustness of deep ensembles as a function of the ensemble members' complexity, we used the corruptions dataset proposed by Hendrycks & Dietterich (2019), freely available at `https://github.com/hendrycks/robustness`. The authors propose 15 noise sources such as random Gaussian noise, spatter or contrast changes to deform the input data and report the model test set accuracy on the corrupted dataset under 5 severity levels (noise strengths).

Inspired by the relative mean Corruption Error (mCE) introduced by the authors, we propose a slight variant of the metric which intuitively captures the relative robustness of the deep ensemble against single ensemble members when encountering data corruptions. Formally, given an ensemble trained on the standard CIFAR-10 dataset, the relative mCE applied to our setting is computed as

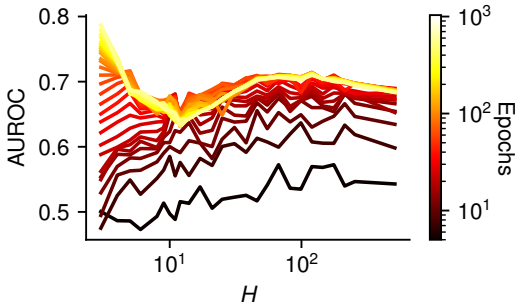

Figure 8. Out-of-distribution detection (AUROC) when training a convolutional neural network on CIFAR-10 with varying hidden layer width H. OOD for the deep ensemble is computed by comparing the variance over the softmax values of the ensemble members on in- and out-of-distribution data. We observe a performance increase with growing H before a peak around $H \approx 100$ for the deep ensemble throughout all training epochs.

$$\text{relative mCE} = \frac{1}{15}\sum_{\epsilon} \text{relative CE}_{\epsilon} \tag{2}$$

where

$$\text{relative CE}_{\epsilon} = \frac{\sum_{s=1}^{5} E_{s,\epsilon}^{Ensemble} - E_{Clean}^{Ensemble}}{\sum_{s=1}^{5} E_{s,\epsilon}^{Specialist} - E_{Clean}^{Specialist}} \tag{3}$$

with $E_{Clean}^{Ensemble}, E_{Clean}^{Specialist}$ resp. the error of the ensemble and single specialist on the uncorrupted CIFAR-10 test set, and $E_{s,\epsilon}^{Ensemble}, E_{s,\epsilon}^{Specialist}$ the error on the same test set corrupted by noise source ϵ with severity s.

A score of 1 would indicate that the rate of performance degradation against data corruption is similar between the ensemble and the specialist. See Figure 13 for the relative mCE as a function of model capacity, computed on the same ensemble trained in and discussed in section 2.3.

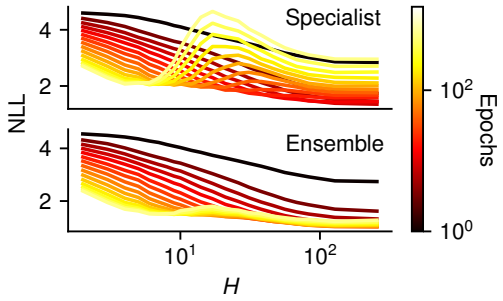

Figure 9. NLL when training ResNet18 with on CIFAR-100 for varying hidden layer width H. NLL for the specialist and deep ensemble observe a clear double descent. Advantages of early stopping for the deep ensemble and specialist are clearly visible for all capacities after the interpolation threshold.

CIFAR100 We investigate the phenomenon on the CIFAR-100 image classification problem using modern deep learning techniques. As the classification mode, we use a ResNet18 with varying width as described in (Nakkiran et al., 2020). The model is now trained with ADAM (Kingma & Ba, 2015) on the cross-entropy loss, with a batch size of 128, learning rate 0.0001 and without weight-decay for 1500 epochs for all H. Furthermore, the data is augmented using the standard pytorch `RandomCrop(32, padding=4)` and `RandomHorizontalFlip` transformations.

With an ensemble of $K = 8$, we measure the test-set accuracy, test-set NLL, ECE as well as the AUROC of discriminating based on the predictive entropy between the CIFAR-100 test data and SVHN, Netzer et al. (2011), LSUN, Yu et al. (2015), Tiny ImageNet as well as on CIFAR-10 subset. Results for the ensemble as well as specialists are shown in Fig. 11 and Fig. 9.

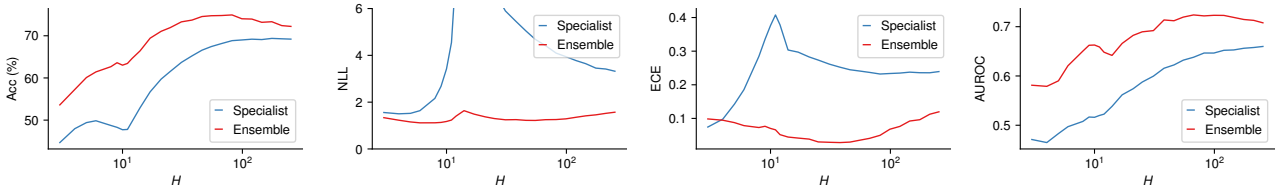

Figure 10. Specialist and ensemble performance on a 3-convolutional and one fully-connected layer neural network with different width H trained with stochastic gradient decent on CIFAR-10 and cross-entropy loss without weight decay. A clear double descent for ensemble and specialist occurs within accuracy (*far left*) and NLL. After an optimal ensemble accuracy and NLL around $H_E = 90$ ensemble performance decreases. *Right plots*: Expected calibration error (ECE) for the ensemble is peaks roughly around 70 but increases consistently for $H \to \infty$. OOD detection (*far right*) increases with growing capacity for the specialist and ensemble while the latter performance peaks around $H \approx 100$ and declines afterwards.

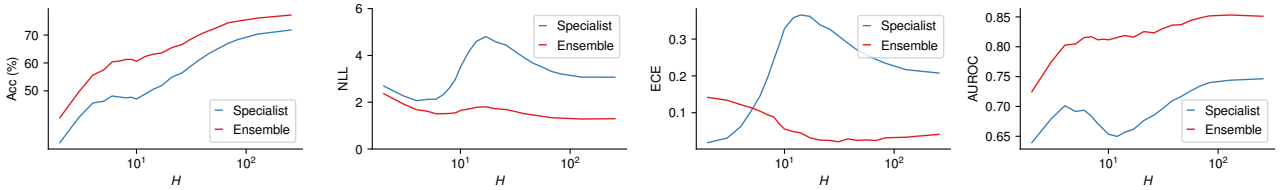

Figure 11. Specialist and ensemble performance on a ResNet18 with different width H trained with ADAM (Kingma & Ba, 2015) on CIFAR-100 and cross-entropy loss without weight decay. A clear double descent for ensemble and specialist occurs within accuracy (*far left*) and NLL. Although the ensemble accuracy and NLL do not degrade after a certain threshold, we observe the discussed negligible effect of ensembling when measuring ECE and OOD after $H \approx 100$ (*right* & *far right*).

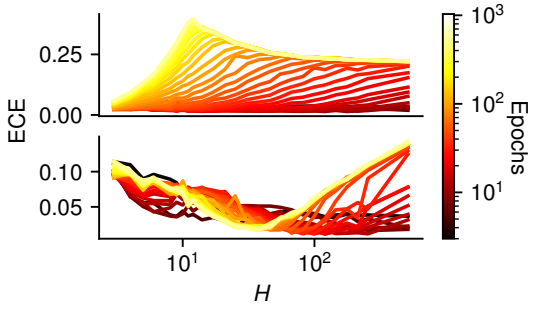

Figure 12. Expected-calibration-error (ECE) over the course of training a convolutional neural network with varying hidden layer width H. ECE consistently decreases for the with growing H before a peak around $H \approx 80$. A clear advantages of early stopping $H \approx 80$ for the deep ensemble is observed.

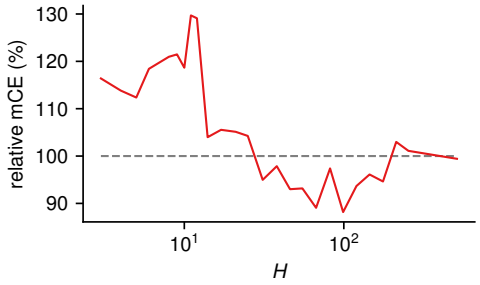

Figure 13. Relative mean corruption error (mCE) in % when training a convolutional neural network on CIFAR-10 with varying hidden layer width H. We observe a performance increase for the deep ensemble wrt. the specialist with growing H before a peak around $H \approx 100$. After this turning point the performance decreases and no advantage between deep ensemble and specialists is observed.