Do Deeper Convolutional Networks Perform Better?

Eshaan Nichani [1, *] Adityanarayanan Radhakrishnan [1, *] Caroline Uhler [1]

October 20, 2020

Abstract

Over-parameterization is a recent topic of much interest in the machine learning community. While over-parameterized neural networks are capable of perfectly fitting (interpolating) training data, these networks often perform well on test data, thereby contradicting classical learning theory. Recent work provided an explanation for this phenomenon by introducing the double descent curve, showing that increasing model capacity past the interpolation threshold can lead to a decrease in test error. In line with this, it was recently shown empirically and theoretically that increasing neural network capacity through width leads to double descent. In this work, we analyze the effect of increasing depth on test performance. In contrast to what is observed for increasing width, we demonstrate through a variety of classification experiments on CIFAR10 and ImageNet32 using ResNets and fully-convolutional networks that test performance worsens beyond a critical depth. We posit an explanation for this phenomenon by drawing intuition from the principle of minimum norm solutions in linear networks.

1 Introduction

Traditional statistical learning theory argues that over-parameterized models will overfit training data and thus generalize poorly to unseen data [8]. This is explained through the bias-variance tradeoff; as model complexity increases, so will variance, and thus more complex models will generalize poorly. Modern deep learning models, however, have been able to achieve state-of-the-art test accuracy by using an increasing number of parameters [13, 21, 9]. In fact, while over-parameterized neural networks have enough capacity to interpolate randomly labeled training data [27], in practice training often leads to interpolating solutions that generalize well.

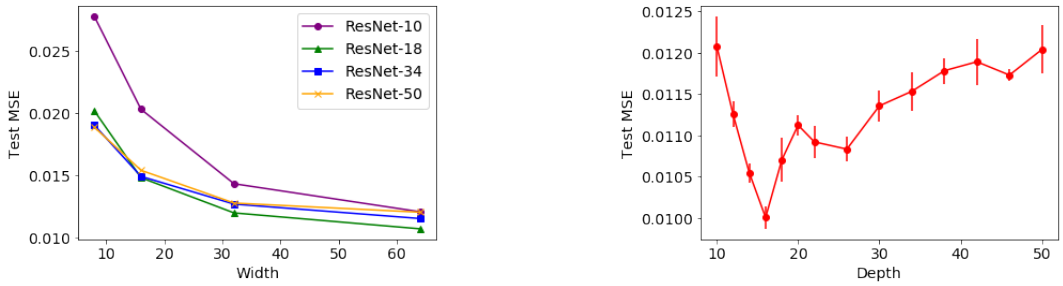
To reconcile this apparent conflict, the double descent risk curve was proposed in [2], where beyond the interpolation threshhold, the risk decreases as model complexity increases. In neural networks, model complexity has thus far mainly been analyzed by varying network width. Indeed, in line with double descent, the authors in [25, 17, 2] demonstrated that increasing width beyond the interpolation threshhold while holding depth constant can decrease test loss.

However, model complexity in neural networks can also be increased through depth. In this work, we study the effect of depth on test performance while holding network width constant. In particular, we focus on analyzing the effect of increasing depth in convolutional networks. These networks form the core of state-of-the-art models used for image classification and serve as a prime example of a network with layer constraints. In this paper we answer the following question: *What is the role of depth in convolutional networks?*

In contrast to what has been shown for increasing model complexity through width, we demonstrate that the test performance of convolutional networks worsens when increasing network depth beyond a critical point, suggesting that double descent does not happen through depth. Figure 1 demonstrates the difference between increasing width and depth in ResNets [9] trained on CIFAR10. In particular, Figure 1a shows

[1]Laboratory for Information & Decision Systems, and Institute for Data, Systems, and Society, Massachusetts Institute of Technology

[*] Equal Contribution



(a) ResNet-18 and ResNet-34 of increasing width on CIFAR10.

(b) ResNet of width 64 on CIFAR10.

Figure 1: (a) As explained by double descent, increasing width in ResNets trained on CIFAR10 results in a decrease in test error. (b) In contrast, increasing the depth of ResNets trained on CIFAR10 results in an increase in test loss (results are averaged across 3 random seeds).

that increasing width leads to a decrease in test error even when training accuracy is 100%. This effect is captured by the double descent curve. On the other hand, Figure 1b demonstrates that training ResNets of increasing depth but fixed width leads to an increase in test error. Since network depth is a form of model complexity, this behavior contradicts what is expected based on double descent. It is therefore critical to carefully analyze and understand this phenomenon.

The main contributions of our work are as follows:

1. We conduct a range of experiments in the classification setting on CIFAR10 and ImageNet32 using ResNets and fully-convolutional networks, and consistently demonstrate that test performance worsens beyond a critical depth (Section 3). In particular, we observe that the test accuracy of convolutional networks approaches that of fully connected networks as depth increases.

2. To gain intuition for this phenomenon we analyze linear neural networks. We demonstrate that increasing depth in linear neural networks with layer constraints (e.g. convolutional networks or Toeplitz networks) leads to a decrease in the Frobenius norm and stable rank of the resulting linear operator. This implies that increasing depth leads to poor generalization, when solutions of lower Frobenius norm (e.g. solutions learned by linear fully connected networks) do not generalize (Section 4).

3. Against conventional wisdom, our findings indicate that when models are near or past the interpolation threshold (e.g. achieving 100% training accuracy), practitioners should *decrease* the depth of these networks to improve their performance. Our results also provide evidence that the driving force behind the success of deep learning is not the depth of the models, but rather their width.

2 Related Work

We begin with a discussion of recent works analyzing the role of depth in convolutional networks (CNNs). The authors in [25] studied the bias-variance decomposition of deep CNNs and showed that as depth increases, bias decreases and variance increases. In particular, they observed that generally the magnitude of bias is greater than that of variance, and thus overall risk decreases. While the focus of their analysis on depth was not on the interpolating regime, they suggested that it may be possible for deeper networks to have increased risk. We here extend their experimental methodology for training ResNets and demonstrate that, indeed, deeper networks have increased risk.

The authors in [18] studied the role of convolutions, but focused on the benefit of sparsity in weight sharing. Their work analyzed the effect of depth on fully-convolutional networks, but only considered models of two depths. The authors in [22] analyzed the role of depth in student-teacher CNNs, which differs from our analysis of training CNNs of varying depth from scratch on CIFAR10.

Other works have aimed to understand the role of depth in CNNs by characterizing implicit regularization in over-parameterized deep CNNs. The authors in [20] characterized the inductive bias of over-parameterized autoencoders and demonstrated that with sufficient depth, these networks become locally contractive around

training examples. The role of depth in autoencoders in the more restrictive setting of a single training example was studied in [28]. Finally, the authors in [19] studied optimization in deep CNNs and showed that increasing depth increases representational power, while increasing width smooths the optimization landscape. While each of these works identified forms of implicit regularization which occur with depth in CNNs, they did not provide an explicit connection to generalization in CNNs used for classification, which is the focus of our work.

On the other hand, previous works studying generalization via double descent have primarily focused on over-parameterization through increasing width. In particular, the authors in [2] and [17] demonstrated that double descent occurs when increasing the width of neural networks trained on MNIST [14] and CIFAR10 respectively. Several works demonstrated double descent theoretically [7, 3, 15, 16, 4, 1], but analyzed linear or shallow non-linear models with an increasing number of features. Our work performs a similar empirical analysis to [17], but on the impact of depth instead of width in CNNs, thereby identifying contrasting behaviors between the two different ways of increasing model complexity.

3 Empirical Evidence in Non-linear Classifiers

We now present our main set of experiments demonstrating that the test accuracy of convolutional networks decreases when increasing depth past a critical threshold. We begin with a demonstration of this phenomenon for fully-convolutional networks applied to CIFAR10 and ImageNet32. We then demonstrate that this phenomenon holds also for ResNets applied to CIFAR10. Our training methodology is outlined in Appendix C.

3.1 Image Classification with Fully-Convolutional Networks

To understand the role of depth in convolutional networks, we begin with a simplified model of a convolutional network, which we call the *Fully-Conv Net*. The architecture of a Fully-Conv Net of depth d and width w for a classification problem with c classes is depicted in Figure 8 in the Appendix, and consists of the following layers:

- A convolutional layer with stride 1, 3 input filers, and w output filters, followed by batch norm [10] and a LeakyReLU activation [24].
- $d-1$ convolutional layers with stride 1, w input filters, and w output filters, each followed by batch norm and LeakyReLU activation.
- 1 convolutional layer with stride 1, w input filters, and c output filters. This is followed by an average pool of each of the output filters to produce a c-dimensional prediction.

Crucially, this network depends only on convolutional layers, a nonlinear activation, and batch norm; it does not depend on other components commonly found in deep learning architectures such as residual connections, dropout, downsampling, or fully connected layers. We note that this model is not designed to necessarily perform well, but rather to isolate and understand the effect of increasing the number of convolutional layers.

We trained the Fully-Conv Net on 2, 5, and 10 classes from CIFAR10 [12]. All experiments were performed using 5 random seeds to reduce the impact of random initialization. Models were trained using Adam [11] with learning rate 10^{-4} for 2000 epochs, and we selected the model with the best training accuracy over the course of training. We used the Cross Entropy loss, and down-sampled images to 16×16 resolution to reduce the computational burden. See Appendix C for a list of all classes used. The resulting train and test accuracies are shown in Figure 2. As expected, as depth increases, training accuracy becomes 100%. However, beyond a critical depth threshold, the test accuracy begins to degrade sharply. Furthermore, the value of this critical depth appears to increase as the number of training classes increases.

In addition to CIFAR10, we also applied the Fully-Conv Net to subsets of ImageNet32 [5], which is ImageNet downsampled to size 32×32. We again trained on 2, 5, and 10 classes, using the same training procedure as for CIFAR10. Training and test accuracies for ImageNet32 are shown in Figure 3. Again, we observe that as depth increases past a critical value, test performance degrades.

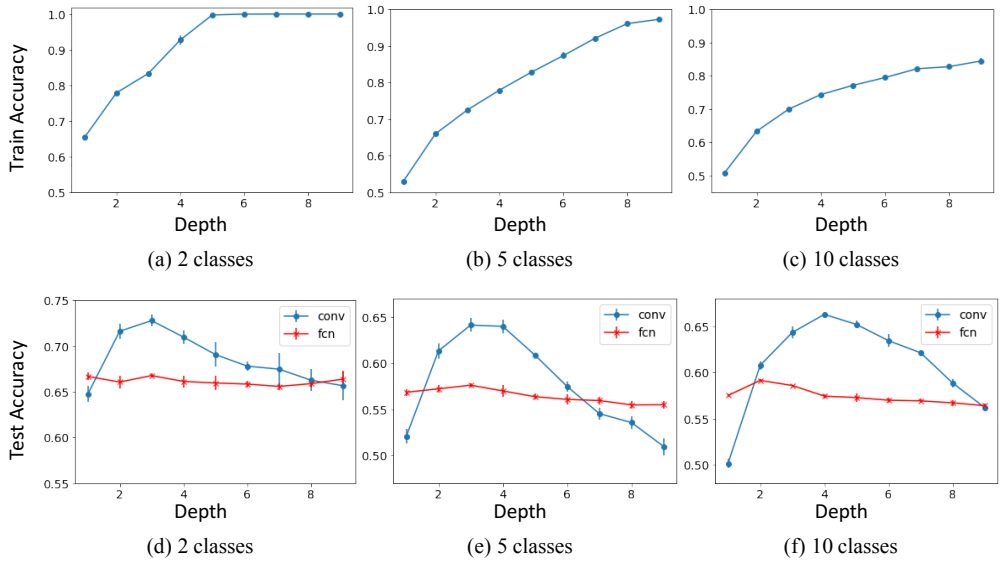


Figure 2: Train and test accuracy of the Fully-Conv Net as a function of depth, for CIFAR10 input images. Increasing depth beyond a critical value leads to a decrease in test accuracy. As depth increases, the performance of the Fully-Conv Net approaches that of a wide fully connected network (shown in red). All experiments are performed across 5 random seeds.

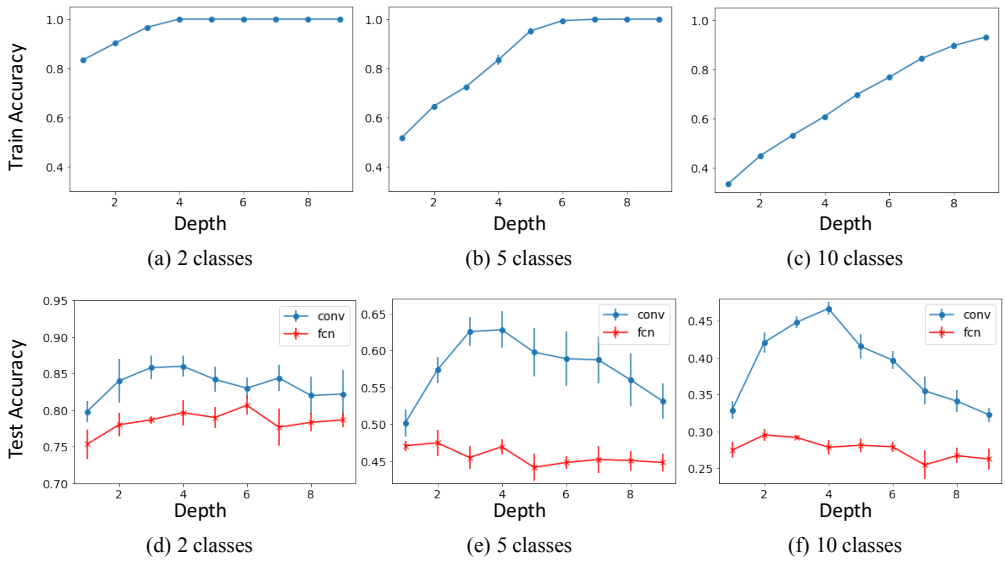
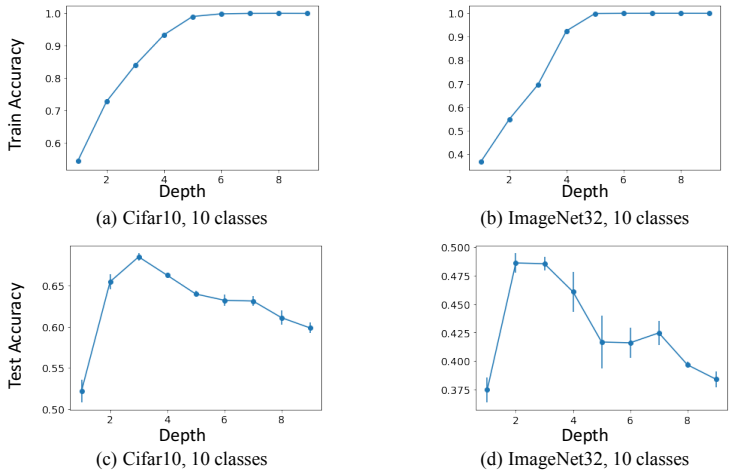


Figure 3: Train and test accuracy of the Fully-Conv Net as a function of depth, for ImageNet32 input images. Increasing depth beyond a critical threshold again leads to a decrease in test accuracy. Increasing depth beyond a critical value leads to a decrease in test accuracy. As depth increases, the performance of the Fully-Conv Net approaches that of a wide fully connected network (shown in red). All experiments are performed across 5 random seeds.

Remarks. When training to classify between 2 and 5 classes, the test accuracy continues to decrease even when increasing depth past the interpolation threshold, i.e. even after achieving 100% training accuracy. This in contrast to double descent where increasing model complexity beyond the interpolation threshold leads to an increase in test accuracy. Interestingly, as depth increases, the test accuracy approaches that of



(a) Cifar10, 10 classes (b) ImageNet32, 10 classes

(c) Cifar10, 10 classes (d) ImageNet32, 10 classes

Figure 4: Increasing depth well past the interpolation threshold leads to a decrease in test accuracy. All experiments are across 3 random seeds. (a) We trained Fully-Conv nets of increasing depth but fixed width 32 on 10 classes of CIFAR10. (b) We trained Fully-Conv nets of increasing depth but fixed width 32 on 10 classes of ImageNet32.

a fully connected network. While the Fully-Conv Nets were before or at the interpolation threshold for the 10 class setting in Figures 2 and 3, Figure 4 demonstrates that a similar decrease in test accuracy occurs also after the interpolation threshold for wider models which can interpolate the data.

3.2 Image Classification with Modern Deep Learning Models

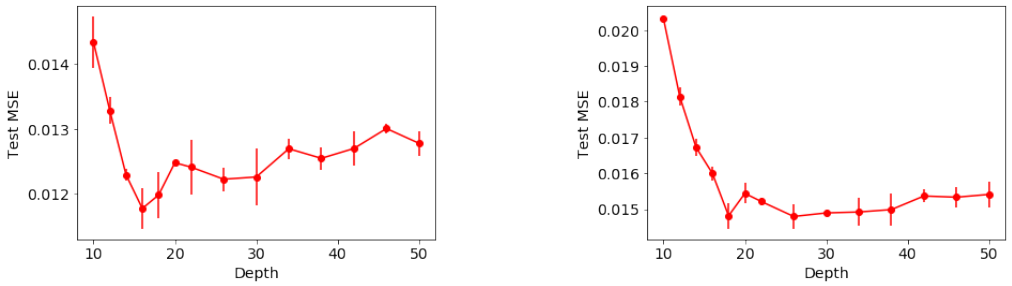
To understand the effect of increasing depth in modern machine learning models, we analyzed variants of ResNet trained on CIFAR10. ResNet-18 and ResNet-34 consist of 4 stages, each of which are connected by a downsampling operation. Each stage is comprised of a number of *basic blocks*, which are two layers with a residual connection. There is a convolutional layer before the first stage, and a fully connected layer after the last stage. ResNet-18 uses 2 basic blocks in each stage, while ResNet-34 uses $(3, 4, 6, 3)$ blocks in each stage respectively. By varying the number of blocks in each stage, we constructed a variety of ResNet models of different depths; in particular, by choosing (n_1, n_2, n_3, n_4) blocks in each stage, we can construct a ResNet model of depth $2 + 2 \cdot (n_1 + n_2 + n_3 + n_4)$. The width w of a model is defined to be the number of filters in the first stage; there are then $(w, 2w, 4w, 8w)$ filters in each stage respectively. See Figure 9 in the Appendix for a diagram. We trained models up to depth 50, see Appendix C for a more detailed description of the models used.

For Figure 1 and Figure 5, we trained each of our ResNet models using MSE loss on a random subset of 25,000 training images from CIFAR10, and plotted the test loss as a function of depth. Our experimental methodology is based on that of [25]. We trained for 500 epochs using SGD with learning rate 0.1 and momentum 0.9, and we decreased the learning rate by a factor of 10 every 200 epochs. Note that in the width 64 and 32 models, test loss increases beyond a critical depth, and in the width 16 model, test loss does not improve beyond a certain depth. Plots of the train and test losses and accuracies of all models are given in the Appendix in Figure 11.

Remarks. In Appendix D, we provide the training and test accuracies for the ResNets presented in this work. We also demonstrate that increasing depth in later blocks of ResNet leads to a more drastic increase in test error than increasing depth in earlier blocks.

4 The Role of Depth in Linear Neural Networks

In the previous section, we observed that past a critical depth threshold the performance of CNNs degrades with depth. For Fully-Conv networks, as depth increased, test performance appeared to approach that of a



(a) Width = 32

(b) Width = 16

Figure 5: Test loss for ResNet models of width 32 (a) and width 16 (b) increases with depth.

fully connected network. In this section, we aim to better understand this phenomenon, and in particular determine whether the solution learned by CNNs of increasing depth does indeed approach the solution learned by a fully-connected network. We turn to the setting of linear neural networks to simplify this analysis. Linear neural networks are useful for analyzing deep learning phenomena since they represent linear operators but have non-convex optimization landscapes. Furthermore, the solution learned by a single layer fully connected network is well understood as it is simply the minimum Frobenius norm solution [6][1].

In this section, we conduct experiments on linear CNNs of increasing depth in both the classification and autoencoding settings. We show that linear CNNs of increasing depth consistently produce solutions of decreasing Frobenius norm, thus approaching the solution learned by a single layer fully connected network. To connect this to generalization, we provide a specific example of a classification setting where the solution learned by a fully-connected network, and thus linear CNNs of large depth, generalizes poorly, yet shallow CNNs generalize well. We propose that a similar mechanism could also explain the decrease in test accuracy beyond a critical threshold in the non-linear setting.

4.1 Preliminaries

Linear networks are models of the form $f(x) = W_d \cdots W_1 x$, where $W_i \in \mathbb{R}^{k_i \times k_{i-1}}$ is a weight matrix. CNNs are a special type of a constrained neural network, where each weight matrix W_i belongs to a subspace $\mathcal{S}_i \subset \mathbb{R}^{k_i \times k_{i-1}}$ of dimension r_i. Since the product of linear operators is linear, these neural networks represent linear functions. In particular, let $W = W_d \cdots W_1$ represent the operator of a linear network. We will analyze the Frobenius norm (square root of the sum of squares of the entries) of W and the stable rank (a surrogate for the rank) of W in our analysis.

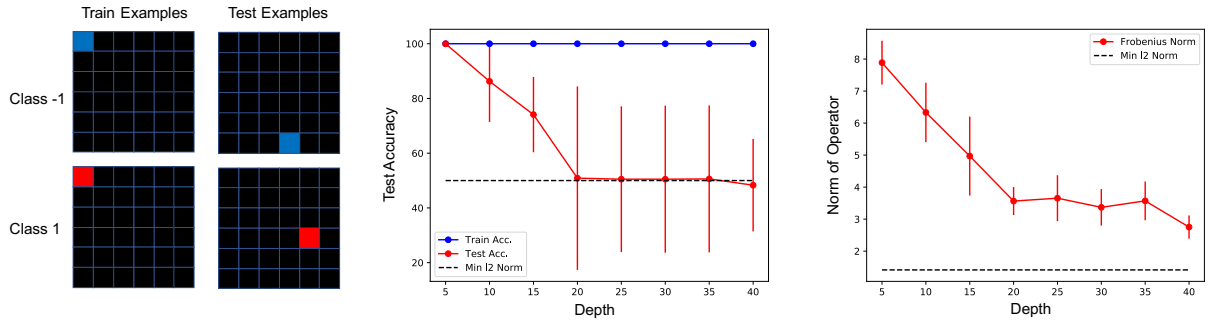
Definition 1. *[23, Chapter 7] Given a matrix $W \in \mathbb{R}^{d_1 \times d_2}$, the **stable rank** of W is given by:*

$$\frac{\|W\|_F^2}{\|W\|^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2};$$

where $\{\sigma_i\}$ denote the singular values of W.

Computing the Linear Operator Efficiently. Given a linear neural network, it is possible to compute the linear operator by first constructing the matrix representation for each layer and then multiplying [20]. Instead, we use the following simpler approach for producing the operator. Given a network implementing a map from an $s \times s$ image to $\mathbb{R}^d$, we reshape the $s^2 \times s^2$ identity matrix as an $s^2 \times s \times s$ tensor (so the batch size is s^2) and feed this through the network. We then reshape the output matrix to be of size $d \times s^2$ to obtain the resulting operator.

[1] This assumes zero initialization for the single layer network. When the network has 1 output, the minimum Frobenius norm solution is the minimum ℓ_2 norm solution.



(a) (b) (c)

Figure 6: A toy example demonstrating that increasing depth in linear convolutional networks leads to operators of decreasing ℓ_2 norm, which manifests as a decrease in test accuracy. (a) A visualization of samples from our toy dataset. (b) The training and test performance of linear convolutional networks of varying depth across 5 random seeds. The test accuracy of the minimum ℓ_2 norm solution for this problem is shown as a dashed black line. (c) The ℓ_2 norm of the operator with varying depth. The norm of the minimum ℓ_2 norm solution for this problem is shown as a dashed black line.

4.2 Linear Convolutional Classifiers

The following experiment provides an example where fully connected networks do not generalize, and in which linear convolutional classifiers of increasing depth converge to this non-generalizing solution. Consider a toy dataset of 6×6 color images as shown in Figure 6a. We use 2 training examples to represent two classes: Class label 1 has a blue pixel in the upper left hand corner and class label -1 has a red pixel in the upper left hand corner. We then construct 200 test examples with 100 having a red pixel and 100 having a blue pixel in a randomly selected location in the lower right 3×3 quadrant of the square.

While simple, this classification setting is useful for comparing the performance of convolutional networks and fully connected networks. It is set up to be trivial to solve using the convolution operation, but is such that the minimum Frobenius norm solution would not be able to generalize. Indeed, as demonstrated in Figure 6b, a linear fully convolutional network with 5 layers and 32 filters per layer is able to consistently get 100% test accuracy across 5 random seeds. On the other hand, performing linear regression to learn the minimum Frobenius norm solution yields a 50% test accuracy.

In Figure 6b, we see that increasing the depth of the fully convolutional network leads to a degradation in test accuracy. In particular, the average test accuracy across 5 random seeds is approximately 50% for networks of depth 20 or larger, which all obtain 100% training accuracy. In Figure 6c, we compare the Frobenius norm of the corresponding linear operator across depths and see that the Frobenius norm decreases with increasing depth. In fact, the norm approaches that of the minimum Frobenius norm solution (shown in black) as depth increases.

This simple example demonstrates that the performance of CNNs of increasing depth approaches that of a fully connected network, which does not generalize for many image classification tasks.

4.3 Linear Autoencoders

Similar to the case of linear convolutional classifiers, we now demonstrate that increasing depth in linear convolutional autoencoders leads to a decrease in Frobenius norm and stable rank. As done in [20, 28], we begin with the simple example of a linear convolutional autoencoder trained on a single image. When there is only a single layer, the authors in [20, 28] prove that the solution must be full rank due to the sparsity and weight sharing in a convolutional layer. This is demonstrated in Figure 7d,e where the depth 1 solution has large stable rank.

In Figure 7a,b,d,e, we demonstrate that increasing depth in linear convolutional autoencoders leads to solutions of decreasing Frobenius norm and stable rank. In particular, we observe that the norm of the trained networks decreases to approximately that of the minimum Frobenius norm solution, which is a projection onto the training examples [20, 28].

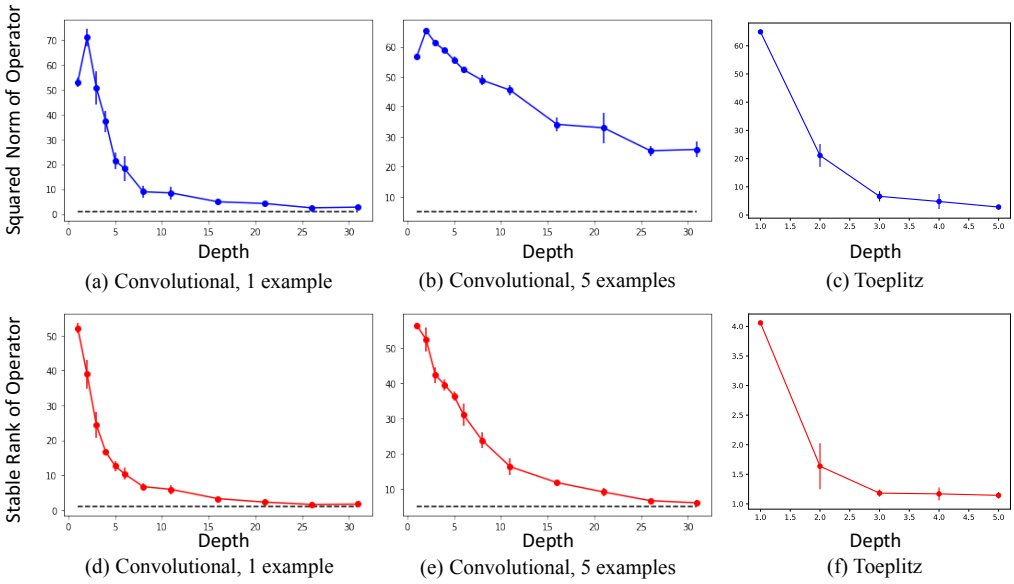


Figure 7: Training layer-constrained linear autoencoders of increasing depth leads to a decrease in the Frobenius norm and stable rank of the resulting operator. (a,d) We train convolutional networks of varying depth on 1 example. (b,e) We train convolutional networks of varying depth on 5 examples. (c,f) We train a network with Toeplitz layers to autoencode a 5 dimensional vector.

While the previous experiments have considered networks with convolutional layers, in Figure 7c, we present an example of a linear autoencoder with alternate layer constraints (Toeplitz layers) for which increasing depth again decreases the operator's Frobenius norm and stable rank. The key similarity between this layer constraint and that of a convolutional layer is that both increase representational power through depth. Indeed, the authors in [26] proved that every matrix can be decomposed as a product of Toeplitz matrices, and thus, neural networks with Toeplitz layers are expressive through depth. We note that showing that every matrix can be written as a product of convolutional layers remains open, but a simple parameter counting argument as given in [20] implies that representational power increases with depth. This experiment thus provides empirical evidence that the phenomenon of decreasing norm with increasing depth occurs generally for networks with layer constraints that are more expressive through depth.

5 Discussion

In this work, we presented an empirical analysis of the impact of depth on generalization in CNNs. We first demonstrated that in modern non-linear CNNs, increasing depth past a critical threshold led to a decrease in test accuracy. This result is in stark contrast to the role of width in CNNs, as explained by double descent. Furthermore, for Fully-Conv Nets, we observed that increasing depth led to performance comparable to that of fully connected networks.

To better understand this phenomenon, we analyzed the operators learned by linear CNNs and demonstrated that increasing depth in these networks led to a decrease in Frobenius norm and stable rank of the learned operator. Moreover, as depth increased, we observed that the norm of the operator approached that of the minimum Frobenius norm solution, which is the solution learned by fully connected network. As demonstrated by our example in Section 4, in settings where the minimum Frobenius norm solution does not generalize, we observe that deep convolutional networks do not generalize. Understanding whether the norm of the functions learned by deep non-linear CNNs approaches that of functions learned by non-linear fully connected networks is an important direction of future work that could explain the poor generalization of deep CNNs observed in this work.

Throughout this work, we consistently observed that increasing depth in CNNs beyond the interpolation

threshold led to a decrease in test accuracy. Hence, our findings imply that practitioners should decrease depth in these settings to obtain better test performance. An interesting direction for future work is to understand where the critical depth threshold occurs as a function of width and number of classes. Importantly, if, as initial evidence in this paper suggests, the critical depth is correlated with the number of classes, then practitioners should use shallow, wide convolutional networks for problems with few classes.

Acknowledgements

The authors were supported by the National Science Foundation (DMS-1651995), Office of Naval Research (N00014-17-1-2147 and N00014-18-1-2765), IBM, and a Simons Investigator Award to C. Uhler. The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

[1] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[2] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[3] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv:1903.07571*, 2019.

[4] K Bibas, Y. Fogel, and M. Feder. A new look at an old problem: A universal learning approach to linear regression. *arXiv preprint arXiv:1905.04708*, 2019.

[5] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[6] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

[7] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1. Springer, 2001.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference in Machine Learning (ICML)*, 2015.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 2012.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] Partha P. Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*, 2019.

[16] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[17] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference in Learning Representations (ICLR)*, 2020.

[18] Benham Neyshabur. Towards learning convolutions from scratch. *arXiv preprint arXiv:2007.13657*, 2020.

[19] Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep cnns. In *International Conference in Machine Learning (ICML)*, 2018.

[20] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Memorization in overparameterized autoencoders. In *ICML Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[22] Gregor Urban, Krzysztof J. Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shenjie Wang, Abdelrahman Mohamed, Matthai Philipose, Matt Richardson, and Rich Caruana. Do deep convolutional nets really need to be deep and convolutional? In *International Conference on Learning Representations (ICLR)*, 2017.

[23] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 1. Cambridge University Press, 2018.

[24] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical Evaluation of Rectified Activations in Convolution Network, 2015. arXiv:1505.00853.

[25] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking Bias-Variance Trade-off for Generalization of Neural Networks. In *International Conference in Machine Learning (ICML)*, 2020.

[26] Ke Ye and Lek-Heng Lim. Every matrix is a product of toeplitz matrices. *Foundations of Computational Mathematics*, 16(3):577–598, 2016.

[27] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[28] Chiyuan Zhang, Samy Bengio, Moritz Hardt, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. In *International Conference on Learning Representations (ICLR)*, 2020.

Appen

A Full

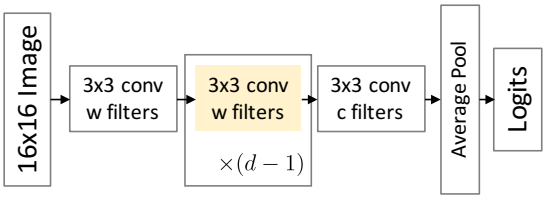


Figure 8: A diagram of the Fully-Conv Net used with width w, depth d, and c classes. Each convolutional layer (except for the last one) is followed by batch norm and LeakyReLU.

B Res

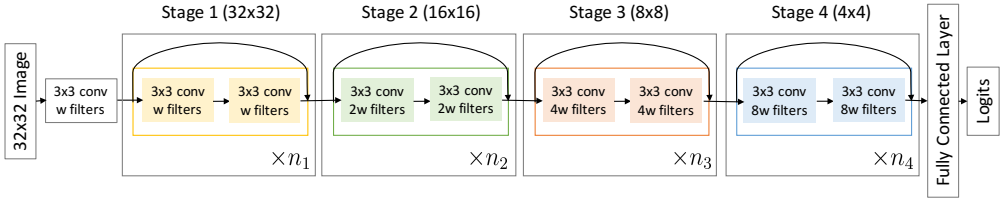


Figure 9: A diagram of the custom ResNet models used. The ith block is repeated n_i times in stage i. The first convolutional layer in stages 2, 3, and 4 downsamples the input using a stride of 2, while the remaining layers have a stride of 1.

C Experimental Details

All models were trained on an NVIDIA TITAN RTX GPU using the PyTorch library. A repository with code used for this paper can be found here: `https://github.com/eshnich/deep_cnn`.

Dataset	Network	Width	Activation	Optimizer	Init.	Train Size	# Epochs	Loss	Trials	Figure
CIFAR10	Fully-Conv	16	LeakyReLU	Adam lr=1e-4	Kaiming Normal	5000 /class	2000, or 100% accuracy	Cross Entropy	5	2
						5000 /class			5	
CIFAR10	FCN	4096								
ImageNet32	Fully-Conv	16				1300 /class			5	3
ImageNet32	FCN	4096				1300 /class			5	
CIFAR10	Fully-Conv	32				5000 /class			3	4
						1300 /class				
ImageNet32	Fully-Conv	32							3	
CIFAR10	ResNet	64	ReLU	SGD w/ momentum LR schedule from Yang et al (2020)	Default Pytorch	2500/class	500	MSE	3	1, 11, 12ae
	ResNet	32							3	1, 6a, 11, 12bf
	ResNet	16							3	1, 6b, 11, 12cf
	ResNet	8							3	1, 12dh, 13
	ResNet Increase Stage 1	32							1	14
	ResNet Increase Stage 3	32							1	14
	ResNet10 ResNet18 Varied kernel size	32							1	16
	ResNet	32				500/class	2000		1	15
Toy classification	Linear Conv	32	None	Adam lr=1e-4	Xavier Uniform	2	Until Convergence	MSE	5	7
Cifar10	Linear Conv	16	None	Adam lr=1e-4	Xavier Normal	1, 5	Until Convergence	MSE	5	8

Figure 10: Experimental details for all experiments conducted.

Model Depth	**Blocks per Stage**
10	1, 1, 1, 1
12	1, 1, 2, 1
14	1, 2, 2, 1
16	2, 2, 2, 1
18	2, 2, 2, 2
20	2, 2, 3, 2
22	2, 3, 3, 2
26	3, 3, 3, 3
30	3, 4, 4, 3
34	3, 4, 6, 3
38	3, 4, 8, 3
42	3, 4, 10, 3
46	3, 4, 12, 3
50	3, 4, 14, 3

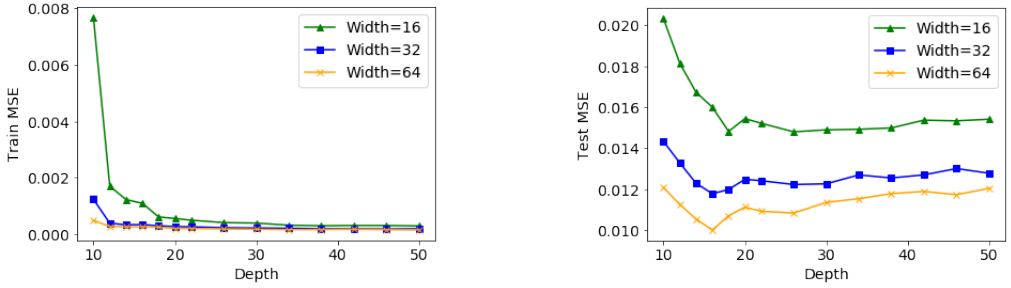
Table 1: Stage breakdown for all ResNet models used.

# Classes	**Classes used**
2	dog, cat
5	bird, cat, deer, dog, horse
10	all classes

Table 2: CIFAR10 classes considered in this work.

# Classes	**Classes used**
2	kit_fox, English_setter
5	2 classes + Siberian_husky, Australian_terrier, English_springer
10	5 classes + Egyptian_cat, Persian_cat, malamute, Great_Dane, Walker_hound

Table 3: ImageNet32 classes considered in this work.



(a) Train Losses of ResNet models with widths 16, 32, and 64.

(b) Test Losses of ResNet models with widths 16, 32, and 64.

Figure 11: Train and Test losses of the ResNet models for all widths.

D Additional Experiments

D.1 Additional ResNet Plots

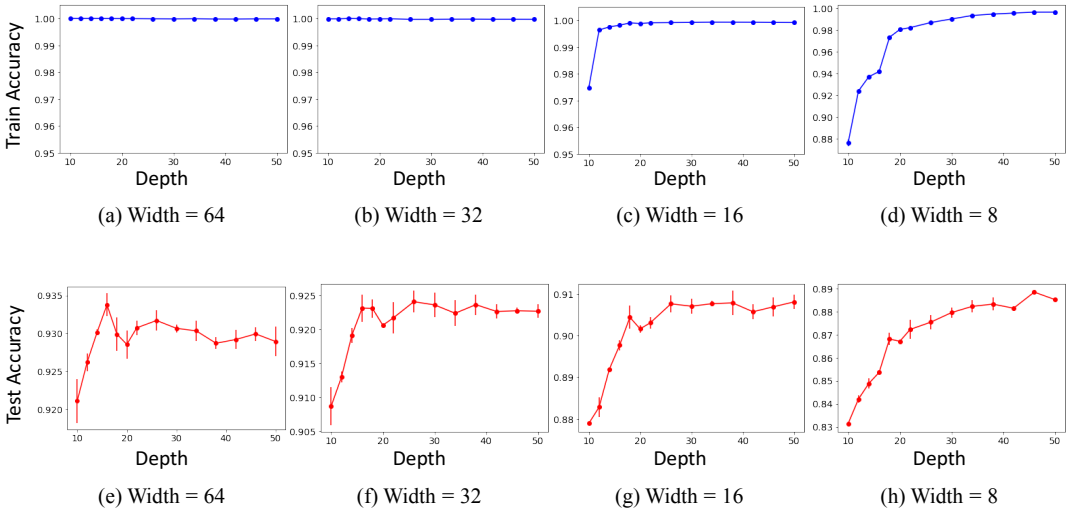
In Figure 11, we plot the train and test losses of all ResNet models used (for widths 16, 32, 64). Additionally, in Figure 12 we plot the accuracies of all ResNet models.

D.2 Effect of Depth in models with small widths

The only model in which test loss continues to increase is the width 8 model. We argue this is because the width 8 model is not sufficiently over-parameterized; in fact, in Figure 13, we see that the width 8 model is unable to reach zero training loss, while all the other models are after sufficient depth.

D.3 Effect of Downsampling

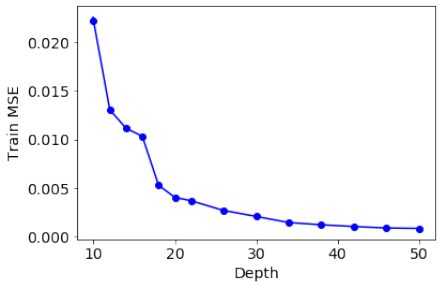
In Figure 14 we compare a ResNet model where we increase the number of blocks in the first stage versus a model where we increase the number of blocks in the third stage. We observe that the model where the third stage blocks are increased performs worse. This is likely because adding a block in a later stage, after downs



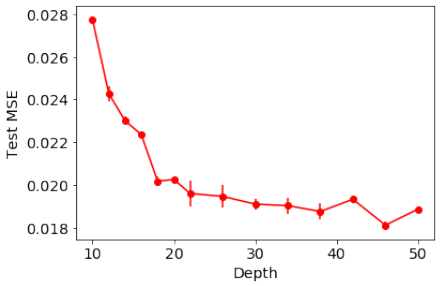
(a) Width = 64 (b) Width = 32 (c) Width = 16 (d) Width = 8

(e) Width = 64 (f) Width = 32 (g) Width = 16 (h) Width = 8

Figure 12: Train and Test accuracies for the ResNet models of width 8, 16, 32, and 64, for increasing depths.



(a) Train losses for the width 8 ResNet model, for increasing depths.



(b) Test losses for the width 8 ResNet model, for increasing depths.

Figure 13: Train and test losses for the width 8 ResNet model. We see that test loss decreases as model depth increases, but train loss has still not reached 0, even for large depths.

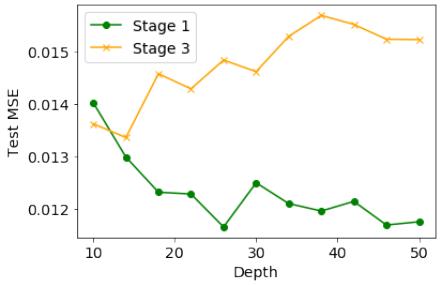


Figure 14: Train losses for the width 32 ResNet model where the 1st stage blocks are increased versus the model where the 3rd stage blocks are increased..

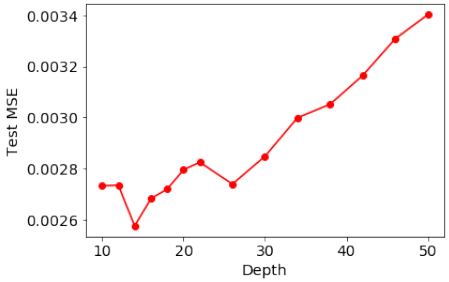


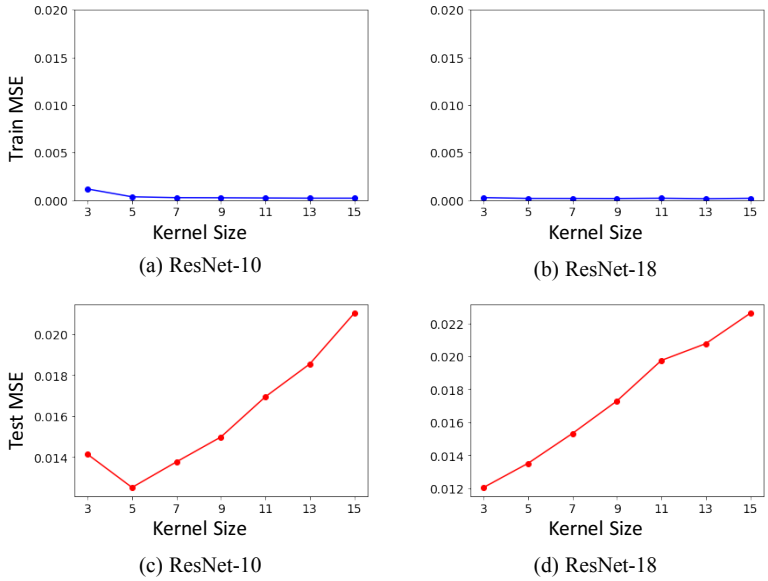
Figure 15: Test losses for the width 32 ResNet samples, trained on 500 samples per class.

D.4 Effect of Number of Samples

In Figure 15 is a plot of test losses when training the width 32 ResNet model on 500 samples per class ($\frac{1}{10}$ of CIFAR10). Number of training epochs is increased accordingly. We observe that test loss increases as depth increases, showing that this phenomenon is robust to change in sample size.

D.5 Effect of Kernel Size

Another form of overparameterization is increasing the kernel size for convolutional filters. In Figure 16, we train ResNet-10 and ResNet-18 of width 32 and varying kernel sizes, and observe that as kernel size increases, test loss increases. This is consistent with our proposed explanation based on expressivity, since increasing kernel size increases representational power independent of depth.



(a) ResNet-10 (b) ResNet-18

(c) ResNet-10 (d) ResNet-18

Figure 16: Train and test losses for width 32 models for increasing kernel size. We observe that test loss increases as kernel size increases.