Large Neural Networks Learning from Scratch with Very Few Data and without Regularization *

Christoph Linse[0000−0002−7039−5189] and Thomas Martinetz[0000−0002−4539−4475]

Institute for Neuro- and Bioinformatics - University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany,
Email: {linse,martinetz}@inb.uni-luebeck.de
https://www.inb.uni-luebeck.de/en/home.html

Abstract. Recent findings have shown that Neural Networks generalize also in over-parametrized regimes with zero training error. This is surprising, since it is completely against traditional machine learning wisdom. In our empirical study we fortify these findings in the domain of fine-grained image classification. We show that very large Convolutional Neural Networks with millions of weights do learn with only a handful of training samples and without image augmentation, explicit regularization or pretraining. We train the architectures ResNet018, ResNet101 and VGG19 on subsets of the difficult benchmark datasets Caltech101, CUB_200_2011, FGVC.Aircraft, Flowers102 and StanfordCars with 100 classes and more, perform a comprehensive comparative study and draw implications for the practical application of CNNs. Finally, we show that VGG19 with 140 million weights learns to distinguish airplanes and motorbikes up to 95% accuracy with only 20 samples per class.

Keywords: Generalization · Small data sets · over-parameterized · Fine-grained image classification · Convolutional Neural Networks · Image recognition

1 Introduction

Deep Convolutional Neural Networks (CNNs) show excellent generalization capabilities in a variety of tasks, particularly in image recognition. Often, their performance can be boosted by increasing the depth or the width of the architectures. The good generalization capabilities are often attributed to a suitable bias within modern CNN architectures and effective regularization strategies like weight decay or dropout.

However, generalization is also present in highly over-parameterized regimes and without explicit regularization, where even random label assignments or even random image data can be memorized [14]. Belkin called it the "modern"

* The work of Christoph Linse was supported by the Bundesministerium für Wirtschaft und Klimaschutz through the Mittelstand-Digital Zentrum Schleswig-Holstein Project.
interpolation regime’ [2]. The generalization capabilities of large neural networks do not seem to be harmed by the millions of parameters in today’s popular architectures. Due to their large, inherent capacity the training error is zero after training by Empirical Risk Minimization (ERM), i.e. the training data is memorized, but, nevertheless, the test error can be unexpectedly low, as shown in [14]. This is in contrast to standard machine learning wisdom, where one would not expect any learning in these highly over-parameterized regimes. The data is highly over-fitted. Statistical learning theory based on uniform convergence requires a high probability that there is no bad solution with a large deviation between its empirical and its true risk, i.e. the training and the generalization error. This can only be expected if the number of training samples is (much) larger than the underlying VC-dimension of the classifier. However, this is not the case in over-parameterized regimes. In over-parameterized regimes always also bad solutions with zero training but large generalization error do exist. Interestingly, it seems that they hardly occur in general.

In this paper, we empirically study the generalization of large, highly over-parameterized networks with a focus on fine-grained image classification. Usually, fine-grained classification is tackled using various techniques to boost the generalization abilities, such as image augmentation, regularization, novel loss functions, semi-supervised learning or pretrained networks. In this paper, we want to study the extreme case of very large networks with 10-140 million weights and just a few training samples of 10-67 per class, with no augmentation, no pretraining and no regularization. We train models with random weight initialization and significantly more classes than samples per class, and choose complex, fine-grained classification problems. The generalization is measured for the popular, large models ResNet018, ResNet101 and VGG19 with increasing numbers of weights. We provide an extensive range of experiments and train hundreds of models on subsets of the Caltech101, CUB_200_2011, FGVC Aircraft, Flowers102 and StanfordCars dataset.

The paper is structured as follows. Section 2 provides all details about our experimental setup including the fine-grained classification datasets and the architectures. The results are presented and discussed in Section 3. The paper completes with our conclusions in Section 4.

2 Materials and Methods

2.1 Datasets

The fine-grained classification datasets are summarized in Table 1 where the number of classes, the size of the datasets and statistics about the number of images per class are presented. Figure 1 provides example images from every dataset showing 2 images for 4 classes, respectively.

Caltech101 was published in 2004. It contains 101 categories and a background class, which we ignore in our experiments. The categories offer a wide degree of diversity including objects like laptops, smartphones and motorbikes,
Table 1. Detailed structure of the benchmark datasets.

| Dataset       | Mode | Classes | Number of images in/of... | dataset | smallest class | largest class | avg. per class |
|---------------|------|---------|---------------------------|---------|----------------|---------------|---------------|
| Caltech101    | -    | 101     |                           | 8677    | 31             | 800           | 85.91         |
| Caltech2      | -    | 2       |                           | 1596    | 798            | 798           | 798.00        |
| CUB_200_2011  | train| 200     |                           | 5994    | 29             | 30            | 29.97         |
|               | test |         |                           | 5794    | 11             | 30            | 28.97         |
| FGVC Aircraft | train| 100     |                           | 6667    | 66             | 67            | 66.67         |
|               | test |         |                           | 3333    | 33             | 34            | 33.33         |
| Flowers102    | train| 102     |                           | 2040    | 20             | 20            | 20.00         |
|               | test |         |                           | 6149    | 20             | 238           | 60.28         |
| Stanford Cars | train| 196     |                           | 8144    | 24             | 68            | 41.55         |
|               | test |         |                           | 8041    | 24             | 68            | 41.03         |

but also animal species and plants. 8677 images were collected manually from the internet and scaled to have a width of about 300 pixels. Sometimes, the objects are shown in a cluttered environment and sometimes they are seen in front of a white background.

_Caltech2_ is a two-class subset of Caltech101 with 798 images showing motorbikes and 798 images showing airplanes.

_CUB_200_2011_ is the Caltech-UCSD Birds-200-2011 dataset [11] from 2011, which extends the Caltech-UCSD Birds 200 dataset [12]. It contains a total of 5994 train and 5794 test images with a mean width of about 470 pixels. There are 200 different bird categories and each category is represented with about 30 training images. Variations in color of the plumage, lighting, perspectives and poses introduce a large intra-class variance while the differences between bird categories can be subtle, which makes this data set particularly challenging.

_FGVC Aircraft_ or the Fine-Grained Visual Classification of Aircraft dataset [7] was published in 2013 and contains 6667 training and 3333 test images showing 100 different airplane models. The images typically have wide aspect ratios and a width of about 1100 pixels. The data is split into an official training and test set with about 67 and 33 images per class, respectively. Airplanes are rigid objects and therefore do not pose the challenge of deformation. However, there is a notable degree of intra-class variance in the visual appearance due to advertisement, airlines and perspective.

_Flowers102_ was published in 2008 as the 102 Category Flower dataset [8]. In average the images are 630 pixels wide and show one or several blossom instances. In order to be consistent with the other datasets, which have training and test sets only, we train with the combination of the training and validation set of Flowers102, leading to 20 training images for each flower type and 2040 images in total. For testing there are 6149 images. While flowers can have small inter-class variations, visual properties like scale, pose, light and also flower subtypes (e.g.
Fig. 1. Fine-grained image classification datasets. Two examples of four different classes are shown, respectively. For Caltech2 with its two classes four examples per class are depicted.

Color variants) can lead to strong intra-class variation. Therefore, Flowers102 is considered to be a challenging dataset.

StanfordCars [5] from Stanford University was presented in 2013 and is composed of 196 different car models. The 8144 training and 8041 test images have an average width of 700 pixels. Typically, only one car is shown at once. The imaging conditions and the locations of the cars are not restricted to a single environment. Sometimes, a car is on a road, other images have been taken indoors and with a stylized background. Also, the perspective is not constrained, such that the image recognition system has to learn features from the front, side and rear of a car model.

2.2 Network Architectures

ResNet018 [4] is employed as a baseline model. It is moderately sized with 11.2 million weights, but still in the over-parameterized regime with regard to the small-scale datasets in the study. Then, we use ResNet101 as a much deeper model with 101 layers and almost four times as many weights as ResNet018. Third, we use the VGG19 [10] architecture with a total of 140.0 million weights. A trained VGG19 model needs more than half a gigabyte when saved on the disk.
The VGG19 consists of a set of 16 convolutional layers and 3 fully connected ones. The majority of the weights are located at the fully connected part of the network. More detailed information about the architectures is provided in Table 2, where the number of parameters, the number of numerical operations, the model size and the memory consumption on the GPU when performing a forward pass are summarized.

| Model   | Parameters($10^6$) | Mult-Adds($10^9$) | Size(MB) | GPU Memory Forward Pass(MB) |
|---------|--------------------|-------------------|----------|-----------------------------|
| ResNet018 | 11.2               | 1.8               | 43.9     | 1825                        |
| ResNet101 | 42.7               | 7.8               | 167.5    | 5119                        |
| VGG19   | 140.0              | 19.6              | 546.9    | 6009                        |

2.3 Training Set-up

The network weights are initialized using Kaiming initialization [3]. Training minimizes the cross-entropy loss for 300 epochs. At the end of the training procedure the classification error on the training set is always zero. The Lamb optimizer [13] is used to get a stable training process with smooth learning curves. The automatic differentiation framework of Pytorch is used [9]. The initial learning rate of $0.0015$ is decreased until $1.5 \cdot 10^{-6}$ in a step-wise manner. The batch size is 32. No image augmentation is performed. Also, there is no explicit regularization like weight decay or dropout. In our data pre-processing pipeline the input images are reshaped to have $224 \times 224$ pixels. In a normalization step the mean RGB value of the entire dataset is subtracted from each image and we divide by the standard deviation. 5 models are trained in each setting to enable a statistical validation of their performance metrics. Depending on the dataset size, it takes between 30 minutes and a few hours to train a single model on a single NVIDIA GTX 2080 Ti graphics card.

3 Results and Discussion

The models are trained on different subsets with increasing number of training samples per class. Table 3 and Table 4 show the obtained accuracies on the test sets. For the Caltech101 dataset we have a separate table, since there is no official split in training and test set and we experimented with different splits.

In Figure 2 the results are illustrated more clearly. All networks with their 10-140 million parameters do learn with these handful of training data per class. Before training, their accuracy is 0.01-0.005 (random guessing), and after training the accuracy has increased by a factor of 10-70. It is astonishing, that a
Table 3. Test accuracy of different classifiers on the benchmark datasets with various number of training samples per class. For each configuration 5 models were trained with different seeds and initializations. The results show the mean test performance and its standard deviation.

| Dataset          | Classes | Model          | Samples p. class | Accuracy | Std |
|------------------|---------|----------------|------------------|----------|-----|
| CUB_200_2011     | 200     | ResNet018      | 20               | 15.04    | 0.60|
|                  |         | ResNet101      | 20               | 22.45    | 0.58|
|                  |         | VGG19          | 20               | 22.66    | 0.75|
|                  |         | ResNet018 (all)| 30               | 29.96    | 1.03|
|                  |         | ResNet101 (all)| 30               | 24.99    | 0.97|
|                  |         | SVM (HOG + RGB)[1] | 30 (all) | 5.00     | -   |
| FGVC Aircraft    | 100     | ResNet018      | 20               | 16.03    | 1.09|
|                  |         | ResNet101      | 20               | 25.01    | 0.46|
|                  |         | VGG19          | 20               | 36.63    | 0.85|
|                  |         | ResNet018 (all)| 67               | 49.88    | 0.59|
|                  |         | ResNet101 (all)| 67               | 58.91    | 0.24|
|                  |         | VGG19          | 67               | 68.78    | 0.26|
|                  |         | SVM, Spatial Pyramid[7] | 67 (all) | 48.69    | -   |
| Flowers102       | 102     | ResNet018      | 10               | 37.21    | 0.82|
|                  |         | ResNet101      | 10               | 40.45    | 2.07|
|                  |         | VGG19          | 10               | 36.99    | 1.06|
|                  |         | ResNet018 (all)| 20               | 54.66    | 1.52|
|                  |         | ResNet101 (all)| 20               | 54.87    | 3.40|
|                  |         | VGG19          | 20               | 53.05    | 0.76|
|                  |         | SVM (HSV)[8]   | 10               | 43.0     | -   |
| StanfordCars     | 196     | ResNet018      | 20               | 5.50     | 0.28|
|                  |         | ResNet101      | 20               | 6.79     | 0.53|
|                  |         | VGG19          | 20               | 4.72     | 1.09|
|                  |         | ResNet018 (24-68 p.cl.) | 42     | 20.33    | 0.66|
|                  |         | ResNet101 (24-68 p.cl.) | 42     | **24.34** | **1.27**|
|                  |         | VGG19 (24-68 p.cl.) | 42     | 16.14    | 4.75|
|                  |         | SVM, Spatial Pyramid[5] | 42 (24-68 p.cl.) | 69.5     | -   |

Classifier with 140 million parameters (VGG19) learns to discriminate 100 different aircraft models with only 67 example images per class up to 70% accuracy, which is an increase by a factor of 70. It is correct in more than 2 of 3 cases in a task which is definitely difficult also for humans. This is in great contrast to what traditional machine learning wisdom would have expected. Similarly for flowers. With only 20 example images per class the classifier is able to learn to discriminate 102 flowers up to an accuracy of more than 50%. Instead of being correct in every 100th case (before training), after training it is correct in every second case. The increase in accuracy is by a factor of 50. For the 200 birds ResNet101 achieves the best results with an increase in accuracy by a factor of 60 by training with 30 images per class. The variance is always below 5%. For these large networks there exist bad zero training error solutions, even solutions with zero test accuracy. However, these bad solutions did not occur. The accuracies for different seeds hardly vary.

On the datasets we studied, the most successful model is ResNet101, which is also the deepest one in this study. It is the best generalizing model in 6 of 8 cases. VGG19, the model with the highest number of weights, is the best
Table 4. Performance of ResNet018, ResNet101 and VGG19 on the Caltech101 dataset for different splits into training and test set. The results show the mean performance and standard deviation of 5 trained models.

| Model     | Train samples          | Test samples | Accuracy | Std  |
|-----------|------------------------|--------------|----------|------|
| ResNet018 | 10 per class           | 20 per class | 27.24    | 1.12 |
| ResNet101 | 32.79                  |              |          |      |
| VGG19     |                        |              | 39.41    | 0.57 |
| ResNet018 | 20 per class           | 10 per class | 42.30    | 0.98 |
| ResNet101 | 43.96                  |              |          |      |
| VGG19     |                        |              | 54.38    | 1.01 |
| ResNet018 | 80.00% (25-640 p.cl.) | 20.00% (6-160 p.cl.) | 65.90    | 0.52 |
| ResNet101 | 65.27                  |              |          |      |
| VGG19     |                        |              | 71.11    | 0.83 |
| ResNet018 | rest (21-790 p.cl.)   | 10 per class | 66.08    | 1.24 |
| ResNet101 | 66.32                  |              |          |      |
| VGG19     |                        |              | 71.52    | 0.59 |
| DAGSVM    | 15 per class           | rest (16-785 p.cl.) | 59.05$^1$ | 0.56 |
| DAGSVM    | 30 per class           | rest (1-770 p.cl.) | 66.23$^1$ | 0.48 |

$^1$ Mean recognition rate per class.

generalizing model in 2 of 8 cases. The smallest model, ResNet018, has the worst results on the fine-grained image classification tasks, even though it has the least number of parameters and should generalize better when trained on small datasets according to traditional machine learning wisdom.

Of course, these results are not state-of-the-art but can be further improved by standard tricks like data augmentation, regularization or pretrained weights. But this was not our goal in these experiments. The goal was to show that learning takes place easily far outside the orbits of traditional machine learning wisdom, often just as good or even better than it was state-of-the-art a few years ago within the orbits of traditional thinking. For this purpose, we have added some state-of-the-art results from then in the Tables 3 and 4, usually achieved with Support-Vector Machines on elaborated handcrafted features.

Another interesting observation in Figure 2 is that the accuracy curves of different architectures plotted against the number of training samples per class have no intersections. This means that the performance ranking of the architectures is consistent over the number of training samples. For CUB_200_2011, Flowers102 and StanfordCars the best generalizing model is ResNet101, disregarding the size of the training set. For FGVCAircraft the VGG19 net consistently provides the best results. Thus, the selection of the best model depends on the dataset and not on the number of training images. In other words, the data distribution chooses the architecture, not the size of the training dataset. Of course, it is not clear to what extent this relationship can be generalized.
Fig. 2. Performance of ResNet018, ResNet101 and VGG19 on different splits of fine-grained image classification datasets. ResNet018 is denoted as circles, ResNet101 as triangles and VGG19 as x. The colors correspond to the different datasets. The results show the mean performance of 5 trained models.

3.1 Further Experiments on the Caltech101 Dataset

In order to substantiate the findings further experiments are performed on the Caltech101 dataset. Table 4 presents the accuracy of ResNet018, ResNet101 and VGG19 using different splits of the Caltech101 dataset, which lacks an official split. Interestingly, the network with the highest number of parameters, VGG19, has always the highest generalization ability with a maximum of $71.52(59)\%$ test accuracy. Again, the relative order of the models in terms of generalization does not depend on the choice of the training split.

This finding is studied more closely by ablating the Caltech101 dataset. We built a subset of Caltech101 by selecting 798 motorbike and 798 aircraft images as they are the classes with the largest number of samples. We trained the three architectures on this binary classification problem using different training and test splits and kept the previous hyperparameters. Figure 3 shows the performances of 5 models trained on each configuration. The test accuracy is plotted against the number of training images. Again, the models always have zero training error.

Just as in the case of the complete Caltech101 dataset, VGG is the best generalizing model. It achieves a test accuracy of about 95% with only 20 training samples per class, which is highly unexpected in such a highly over-parameterized regime. The curves of ResNet018 and ResNet101 are very close and cannot be distinguished from each other. Only for 20 training images per class ResNet101
Fig. 3. Test accuracy on different splits of the Caltech2 dataset consisting of images showing airplanes and motorbikes. The results show the mean performance and standard deviation of 5 models.

has an about 5% higher test accuracy than ResNet018. Again, the architecture with more weights performs better in this regime, and again the accuracy curves of single architectures with respect to the number training samples do not intersect, which is in line with our previous observations. Already for 200 training samples all three models converge to an almost perfect classification accuracy of 99%. The variance is below 1%. Again, bad solutions, which exist, do not occur.

4 Conclusion and Outlook

We studied the generalization abilities of deep, high capacity Convolutional Neural Networks in the context of fine-grained image classification with very small training data sets. We trained the frequently used CNNs ResNet018, ResNet101 and VGG19 without augmentation, explicit regularization and pre-trained weights in order to measure their genuine generalization ability in highly over-parameterized regimes on the challenging benchmark datasets Caltech101, CUB_200_2011, FGVC Aircraft, Flowers102 and StanfordCars. We were able to show that the generalization error is remarkably low already with very few training data, and that the larger CNNs performed consistently better than the smallest one. For a binary classification problem we could show that a network with 140 million weights (VGG19) reaches 95% accuracy with only 20 training samples per class. This is against the narrative that Deep Neural Networks require large training data sets. We observed, that certain architectures work
better on certain datasets, no matter the size of the training split. The relative generalization abilities between architectures do not seem to depend on the number of training images. Problems prefer certain models, but models do not prefer a specific training set size more than other models.

We conclude, that very large Neural Networks can learn with very few data, also without any data augmentation, regularization or pretraining. Performance even can benefit from an increased number of network weights in these highly over-parameterized regimes. Training up to zero training error, i.e., overfitting up to memorizing the training data, does not harm generalization. All this is against traditional machine learning wisdom and the narrative, that Deep Neural Networks require many training data. At least in case of binary image classification problems, only a few hundred training samples can be sufficient also for very large CNNs to provide best possible solutions. In principle, bad solutions do exist in these highly over-parameterized regimes, but they do not occur.

Our empirical findings are restricted to CNNs and image classification tasks. We expect, that these findings extrapolate to other task domains and network architectures, but this remains to be shown. The observation, that the accuracy curves of different architectures do not intersect could be validated for large scale datasets in future work. Moreover, the study could be enriched with more recent architectures of different sizes.

Still, a lot has to be done in order to understand the unexpectedly good generalization abilities of large Neural Networks in connection with small training data sets. We need a new understanding of learning, which extends classical statistical learning theory into over-parameterized regimes. This will require abandoning or at least expanding the concept of uniform convergence and traditional capacity measures of classifiers. The usual narratives of traditional machine learning wisdom of "overfitting" and "the larger the network the more data is needed" do not apply in this context.

References

1. Alter, A.L., Wang, K.M.: An exploration of computer vision techniques for bird species classification (2017)
2. Belkin, M., Hsu, D., Ma, S., 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 (Aug 2019). https://doi.org/10.1073/pnas.1903070116, https://pnas.org/doi/full/10.1073/pnas.1903070116
3. He, K., Zhang, X., Ren, S., Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1026–1034. IEEE, Santiago, Chile (Dec 2015). https://doi.org/10.1109/ICCV.2015.123, http://ieeexplore.ieee.org/document/7410480/
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA
5. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D Object Representations for Fine-Grained Categorization. In: 2013 IEEE International Conference on Computer Vision Workshops. pp. 554–561. IEEE, Sydney, Australia (Dec 2013). https://doi.org/10.1109/ICCVW.2013.77, http://ieeexplore.ieee.org/document/6755945/ 

6. Li Fei-Fei, Fergus, R., Perona, P.: Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop. pp. 178–178. IEEE, Washington, DC, USA (2004). https://doi.org/10.1109/CVPR.2004.383, https://ieeexplore.ieee.org/document/1384978 

7. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-Grained Visual Classification of Aircraft. arXiv:1306.5151 [cs] (Jun 2013), http://arxiv.org/abs/1306.5151, arXiv: 1306.5151 

8. Nilsback, M.E., Zisserman, A.: Automated Flower Classification over a Large Number of Classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE, Bhubaneshwar, India (Dec 2008). https://doi.org/10.1109/ICVGIP.2008.47, http://ieeexplore.ieee.org/document/4756141/ 

9. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing System. pp. 1–4 (2017) 

10. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] (Apr 2015), http://arxiv.org/abs/1409.1556, arXiv: 1409.1556 

11. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 

12. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010) 

13. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Hsieh, C.J.: Large batch optimization for deep learning: Training bert in 76 minutes. In: International Conference on Learning Representations (2020) 

14. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv:1611.03530 [cs] (Feb 2017), http://arxiv.org/abs/1611.03530, arXiv: 1611.03530 

15. Zhang, H., Berg, A., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). vol. 2, pp. 2126–2136 (2006). https://doi.org/10.1109/CVPR.2006.301