# Small Data, Big Decisions: Model Selection in the Small-Data Regime

**Jörg Bornschein** [1] **Francesco Visin** [1] **Simon Osindero** [1]

## Abstract

Highly overparametrized neural networks can display curiously strong generalization performance – a phenomenon that has recently garnered a wealth of theoretical and empirical research in order to better understand it. In contrast to most previous work, which typically considers the performance as a function of the model size, in this paper we empirically study the generalization performance as the size of the training set varies over multiple orders of magnitude. These systematic experiments lead to some interesting and potentially very useful observations; perhaps most notably that training on smaller subsets of the data can lead to more reliable model selection decisions whilst simultaneously enjoying smaller computational overheads. Our experiments furthermore allow us to estimate Minimum Description Lengths for common datasets given modern neural network architectures, thereby paving the way for principled model selection taking into account Occams-razor.

## 1. Introduction

According to classical statistical learning theory, achieving an optimal generalisation loss requires selecting a model capacity that strikes the best balance between underfitting and overfitting, i.e., between not having enough capacity to model the training data accurately and having too much, and thus prone to adapt too closely to the training data at the expense of generalisation. Under this theory, the final generalisation loss plotted against model capacity should behave as a U-shaped curve – initially decreasing as the capacity increases (underfitting) to reach a minimum (optimal model size) and then finally increase again (overfitting).

Contrary to these results, it has long been observed that neural networks show a curiously good generalization performance (in terms of error) when applied to classification problems, even though the generalization cross-entropy exhibits all the characteristics of overfitting. There has recently been renewed interest in studying this phenomenon, both theoretically and empirically (Advani & Saxe, 2017; Spigler et al., 2018). Belkin et al. (2019) for example argue that beside the classical underfitting and overfitting regimes, a third one for massively overparameterised models exists. The transition into this regime is called the *interpolation threshold* and is characterized by a peak in generalization error, followed by a phase of further decrease. The peculiar shape of the generalization-error over model-size curve lends the term *double-descent* to this phenomenon. Work by Nakkiran et al. (2019) has sharpened this picture and shown that this behaviour can be observed when training modern neural network architectures on established, challenging datasets. Most of the work studying generalization of neural networks focuses on the relationship between generalization performance and model size. Instead we present an empirical study that investigates the generalization performance as a function of the training set size.

In the rest of the Introduction we outline the main contributions of our work.

### 1.1. Performance analysis

One of our key contributions is to gather performance curves as a function of the training set size for a wide range of architectures on ImageNet, CIFAR10, MNIST and EMNIST. We also perform an extensive sweep over a wide range of architectures, model sizes, optimizers and, crucially, we vary the size of the training sets over multiple orders of magnitude, starting from the full dataset down to only few examples per class. Our study covers even extreme cases – for example the training of oversized ResNet architectures with 10 or less examples per class. This is, to the best of our knowledge, the most extensive empirical analysis conducted on generalization for massively overparameterized models, and strengthens the emerging understanding of training regimes for modern deep learning.

### 1.2. A ranking-hypothesis

One salient observation we have made has not been described in the literature: overparameterized model archi-

---

[1] DeepMind, London, United Kingdom. Correspondence to: Jörg Bornschein <bornschein@google.com>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020.

Figure 1. Learning curves for ResNet-101 on subsets of the ImageNet dataset using a RMSProp with a cosine learning-rate schedule. Left: the uncalibrated generalization cross-entropy shows a strong overfitting signature (dashed line), the calibrated cross-entropy does not.

Figure 2. Post-convergence generalization error vs. generalization cross-entropy without early-stopping for a range of model architectures and training set sizes. The colours represent the different model architectures. We observe that the calibrated cross entropy is strongly correlated with the generalization error rate.

tectures seem to maintain their relative ranking in terms of generalization performance, when trained on arbitrarily small subsets of the training set.

This observation prompts us to hypothesize that: when (i) two sufficiently large neural network architectures A and B are trained with a well tuned optimizer on datasets of size N; and (ii) we observe that, in expectation, A performs better than B; then (iii) A will also perform better than B in expectation for all differently sized datasets drawn from the same underlying distribution, as long as we remain well beyond the interpolation threshold.

Unfortunately this is only an hypothesis, but if this conjecture is true it would have profound practical implications. Namely, it would mean that, for sufficiently large models, it would be possible to perform model selection or architecture-search using small subsets of the data, and expect that the decision regarding which model to prefer remains valid when applied to much larger datasets. Indeed, our experiments show that training on small or medium sized subsets of the training data leads in many cases not only to faster convergence, but also to a more robust signal for model selection than training on big datasets, and is therefore often preferable.

### 1.3. Calibration & minimum description length

Independently of the model-selection hypothesis, we also show that it is possible to avoid some negative effects of overfitting by simply choosing an optimal softmax-temperature on a small held-out dataset; i.e., by calibrating the neural network (Guo et al., 2017). After calibration, the generalization cross-entropy becomes a stable and well-behaved quantity even when model sizes and training set sizes vary considerably. Being able to compute well-behaved generalization cross-entropy on small training sets allow us to compute reliable Minimum Description Length estimates, a quantity that is of interest for principled model selection taking into account Occam's-razor. We will discuss this more in depth in Section 5.

## 2. Related Work

The literature on generalization performance for learned predictors is vast and spans seminal work on the classical bias-variance-tradeoff (Geman et al., 1992; Domingos, 2000) all the way to theoretical and empirical work investigating the still poorly understood, but often strong generalization performance of neural networks. The latter has been approached from a number of different directions: For example pointing out that neural networks seem to perform implicit capacity control (Zhang et al., 2016), investigating the learning dynamics and properties of the loss landscape (Spigler et al., 2018) and interpreting stochastic gradient descent as an approximation of probabilistic inference (Mandt et al., 2017). Finally, two lines of work recently contributed to the understanding of generalization, the first studying infinitely wide neural networks (Jacot et al., 2018; Allen-Zhu et al., 2019) and the second focusing on the double-decent phenomenon (Belkin et al., 2019; Nakkiran et al., 2019).

The vast majority of these previous studies focus on the model-size dependency of the generalization performance. Notable exceptions are the work by Nakkiran et al. (2019), who investigates the neighbourhood of the interpolation threshold, and the work by Hestness et al. (2017), who first evaluated the generalization error-rate for language models and ResNets as a function of the training set size. They point out that the generalization error-rate can be well predicted by assuming a power law and interpolating from smaller training sets to bigger ones. In a similar fashion, our work shows that the relative model ranking trained on small datasets is maintained when trained on bigger ones, a result of practical importance for model selection and architecture search.

## 3. Methods

### 3.1. Temperature calibration

Overparameterized models are typically trained by minimizing either a regression loss or a categorical cross-entropy loss, but in contrast are then evaluated on their generalization performance measured on the error-rate. This is because the generalization cross-entropy can overfit severely for models that are sufficiently big.

We show here that it is possible to avoid the negative effects of overfitting when using the categorical cross-entropy by simply choosing an optimal softmax-temperature on a small heldout dataset; i.e., by calibrating the neural network in the way proposed by Guo et al. (2017). In practice we implement temperature calibration by performing gradient descent on the calibrated cross-entropy loss w.r.t. a single scalar temperature parameter. We interleave the regular model training steps with steps of gradient descent on the calibration loss on some held-out data, that we refer to as the *calibration dataset*. This allows us to track the generalization cross-entropy online during learning.

This calibration procedure can prevent the generalization cross-entropy from overfitting, even when the model size is order of magnitudes larger than the training set size, and without relying on early-stopping (see Figure 1 and Figure 2). Note that, just as with the post-convergence calibration proposed by Guo et al. (2017), the calibration procedure does not modify the other parameters, directly or indirectly.

Being able to compute well-behaved generalization cross-entropies is desirable because it is the loss we optimize for and because it allows us to compute MDL estimates, as explained in Section 5. That being said, we could express all other results in this paper in terms of error-rates instead of cross-entropies, and the observations and conclusions would still hold (see Supplementary material).

### 3.2. Datasets and models

Throughout this study we present a large number of experiments on subsets of different sizes of popular benchmarking datasets. We call the total set of datapoints available for a particular training run the *available dataset*, which is split into a *training set* and a *calibration set*. The former is used to optimize the connection weights, the batch-norm parameters and all other parameters that are considered part of the neural networks, while the latter to optimise the calibration temperature, as explained in Section 3.1, as well as to determine the optimal learning-rate and, potentially, to perform early-stopping. If not mentioned otherwise, we will use a 90%/10% training/calibration split of the *available dataset*. To assess a model's ability to generalise beyond the available-dataset, we then successively evaluate them on a separate held-out or *evaluation dataset*. We experimented with balanced subset sampling, i.e. ensuring that all subsets always contain an equal number of examples per class. But we did not observe any reliable improvements from doing so and therefore reverted to a simple i.i.d sampling strategy. We always pay particular attention to not to inadvertently leak data, i.e. use datapoints that have not been properly accounted for to select the hyperparameters.

We conduct experiments on the following datasets:

**MNIST** consists of 60k training and 10k test examples from 10 classes (LeCun, 1998). We train MLPs of various depth and width, with and without dropout, as well as standard ConvNets on this dataset. Unless otherwise noted, we use ReLU as the nonlinearity.

**EMNIST** provides 112,800 training and 18,800 test datapoints from 47 classes in its balanced subset (Cohen et al., 2017). We train the same family of model architectures we also train on MNIST.

**CIFAR10** consists of 50k training and 10k test examples from 10 classes (Krizhevsky et al., 2009). We train a wide range of models on this dataset, including simple architectures like MLPs and ConvNets; architectures that have been carefully optimized for image classification: ResNet-20 (He et al., 2016) and Wide ResNets (Zagoruyko & Komodakis, 2016) as well as a selection of architectures from the NASBench-101 (Ying et al., 2019) search space. The latter were chosen by disregarding the worst 10% and then picking 5 equally-spaced from the remaining ones. The Supplementary material contains the description of these architectures. The rationale is that we want to evaluate a range on non-optimal architectures and confirm their relative ranking is preserved when using smaller datasets too.

**ImageNet** contains 1.28M training and 50k validation examples from 1000 classes (Russakovsky et al., 2015). We train a selection of widely known standard models like VGG-16 (Simonyan & Zisserman, 2014), ResNets (He et al., 2016) and Inception (Szegedy et al., 2016). Additionally we consider S3TA (Zoran et al., 2019), a sequential and attention based model that takes multiple glimpses at various spacial locations in the image before emitting a prediction. It is composed of a ResNet-style feature extractor with a reduced

Figure 3. Cross-entropy performance profiles for the ResNet-101 architecture on ImageNet when trained with RMSProp. We use 90% of the available data for training, 10% for calibration and report the generalization performance on the unseen validation set. Note that even when training with as little as $2.3$ images per class, there is no harm in using a ResNet model with 4 more channels (16 more parameters) than the standard ResNet.

Figure 4. Performance profiles for a fully connected MLP with 3 hidden layers on CIFAR10 as a function of the hidden layer size. Red points mark the smallest models that approach a close to zero training error-rate.

number of strides, to keep the spatial resolution higher, and an LSTM equipped with a spatial attention mechanism to ingest features. The rationale for including it in our exploration is that it supposedly has a different inductive bias than pure conventional models, which could increase the chance of falsifying our consistent ranking hypothesis.

### 3.3. Training

To ensure our observations are not specific to a particular optimization method, we run experiments with different variants of gradient-descent. For each experiment we sweep over a fixed set of possible learning rates and pick the best one according to the calibration loss, independently for each model architecture and training set size.

Throughout this study we use the following optimizers:

Adam (Kingma & Ba, 2014) with fixed learning rates $\{10^{-4}; 3 \times 10^{-4}; 10^{-3}\}$ and 50 epochs.

Momentum SGD with initial learning rates $\{10^{-4}; 3 \times 10^{-4}; \ldots; 10^{-1}\}$ cosine-decaying over 50 epochs down to 0 (0.9 momentum and $= 10^{-4}$).

RMSProp + cosine schedule (Tieleman & Hinton, 2012) with initial learning rates of $\{0.03; 0.1; 0.3\}$ and cosine-decaying to 0 over 50 epochs. We choose the same hyperparameters used by (Ying et al., 2019) for their NASBench-101 experiments, with the exception of the number of epochs, which we reduced from 108 to 50 (momentum=0.9, $= 1$).

For all experiments we use a batch size of 256 examples. The term epoch always refers to the number of gradient steps required to go through the full-sized dataset once; i.e., on ImageNet an epoch is always $1.28M/256 = 5000$ gradient steps, regardless of the size of the actual training set used.

We evaluated many more combinations than those presented in this paper, a selection of which is contained in the Supplementary material. Throughout all these control runs our qualitative results were confirmed, which suggests that these results do not just emerge from the interaction of specific models and optimizers; or from a particularly sensible hyperparameter choice.

## 4. Experiments

### 4.1. Properties of calibration

Figure 1 visualizes typical learning curves when training a ResNet-101 model with cosine-decayed RMSProp on subsets of ImageNet. Depending on the training set size the model can memorise the training set and reach a zero training error-rate within a few thousand gradient steps. The uncalibrated cross-entropy on held-out data shows severe symptoms of overfitting in these cases but, as reported before, the error computed on the same set tends to instead flatline and not to degrade significantly.

This suggests that the models are not unfavourably adjusting their decision boundaries as training progresses into the overfitting regime, but are first and foremostly becoming unduly confident in their predictions. When using calibration we rectify this over-confidence and obtain stable cross-entropies that show the same kind of robust behaviour as the evaluation error-rate.

Early-stopping can still have a positive effect though: We regularly observe a small degradation in generalization performance just around the point where the rapid decrease in