

---

# SURROGATE MODEL BASED HYPERPARAMETER TUNING FOR DEEP LEARNING WITH SPOT

---

A PREPRINT

Thomas Bartz-Beielstein, Frederik Rehbach, Amrita Sen, and Martin Zaeferrer \*

Institute for Data Science, Engineering, and Analytics

Technische Hochschule Köln

5164 Gummersbach, Germany

thomas.bartz-beielstein@th-koeln.de

July 20, 2021

## Abstract

A surrogate model based hyperparameter tuning approach for deep learning is presented. This article demonstrates how the architecture-level parameters (hyperparameters) of deep learning models that were implemented in Keras/tensorflow can be optimized. The implementation of the tuning procedure is 100% accessible from R, the software environment for statistical computing. With a few lines of code, existing R packages (tfruns and SPOT) can be combined to perform hyperparameter tuning. An elementary hyperparameter tuning task (neural network and the MNIST data) is used to exemplify this approach.

**Keywords** hyperparameter tuning · deep learning · hyperparameter optimization · surrogate model based optimization · sequential parameter optimization

## 1 Introduction

Deep Learning (DL) models require the specification of a set of architecture-level parameters, which are called *hyperparameters*. Hyperparameters are to be distinguished from the *parameters* of a model that are optimized in the initial loop, e.g., during the training phase via backpropagation. Hyperparameter values are determined before the model is executed—they remain constant during model development and execution whereas parameters are modified. We will consider Hyperparameter Tuning (HPT), which is much more complicated and challenging than parameter optimization (training the weights of a Neural Network (NN) model).

Typical questions regarding hyperparameters in DL models are as follows:

1. How many layers should be stacked?
2. Which dropout rate should be used?
3. How many filters (units) should be used in each layer?
4. Which activation function should be used?

Empirical studies and benchmarking suites are available, but to date, there is no comprehensive theory that adequately explains how to answer these questions. Recently, Roberts et al. [2021] presented a first attempt to develop a DLtheory.

In real-world projects, DL experts have gained profound knowledge over time as to what reasonable hyperparameters are, i.e., HPT skills are developed. These skills are based on human expert and domain knowledge

---

\*<https://www.spotseven.de>

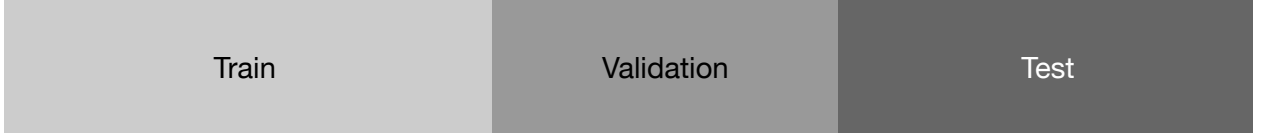


Figure 1: Dataset splitted into three parts: (i) a training set  $X^{(\text{train})}$  used to fit the models, (ii) a validation set  $X^{(\text{val})}$  to estimate prediction error for model selection, and (iii) a test set  $X^{(\text{test})}$  used for assessment of the generalization error

and not on valid formal rules. Figure 1 in Kedziora et al. [2020] nicely illustrates how data scientists select models, specify metrics, preprocess data, etc. Chollet and Allaire [2018] describe the situation as follows:

“If you want to get to the very limit of what can be achieved on a given task, you can’t be content with arbitrary [hyperparameter] choices made by a fallible human. Your initial decisions are almost always suboptimal, even if you have good intuition. You can refine your choices by tweaking them by hand and retraining the model repeatedly—that’s what machine-learning engineers and researchers spend most of their time doing. But it shouldn’t be your job as a human to fiddle with hyperparameters all day—that is better left to a machine.”

HPT develops tools to explore the space of possible hyperparameter configurations *systematically*, in a structured way. For a given space of hyperparameters  $\Lambda$ , a Deep Neural Network (DNN) model  $A$  with hyperparameters  $\lambda$ , training, validation, and testing data  $X^{(\text{train})}$ ,  $X^{(\text{val})}$  and  $X^{(\text{test})}$ , respectively, a loss function  $L$ , and a hyperparameter response surface function  $\psi$ , e.g., mean loss, the basic HPT process looks like this<sup>2</sup>:

- (HPT-1) Set  $t = 1$ . Parameter selection (at iteration  $t$ ). Choose a set of hyperparameters from the space of hyperparameters,  $\lambda(t) \in \Lambda$ .
- (HPT-2) DNN model building. Build the corresponding DNN model  $A_{\lambda(t)}$ .
- (HPT-3) DNN model training and evaluation. Fit the model  $A_{\lambda(t)}$  to the training data  $X^{(\text{train})}$  (see Figure 1) and measure the final performance, e.g., expected loss, on the validation data  $X^{(\text{val})}$ , i.e.,

$$\psi^{(\text{val})} = \frac{1}{|X^{(\text{val})}|} \sum_{x \in X^{(\text{val})}} L\left(x; A_{\lambda(t)}(X^{(\text{train})})\right), \quad (1)$$

where  $L$  denotes a loss function. Under  $k$ -fold Cross Validation (CV) the performance measure from Equation 1 can be written as

$$\psi_{\text{CV}}^{(\text{val})} = \frac{1}{k} \sum_{i=1}^k \frac{1}{|X^{(\text{val})}|} \sum_{x \in X_i^{(\text{val})}} L\left(x; A_{\lambda(t)}(X_i^{(\text{train})})\right), \quad (2)$$

because the training and validation set partitions are build  $k$  times.

- (HPT-4) Parameter update. The next set of hyperparameters to try,  $\lambda(t+1)$ , is chosen accordingly to minimize the performance, e.g.,  $\psi^{(\text{val})}$ .
- (HPT-5) Looping. Repeat until budget is exhausted.
- (HPT-6) Final evaluation of the best hyperparameter set  $\lambda^{(*)}$  on test (or development) data  $X^{(\text{test})}$ , i.e., measuring performance on the test (hold out) data

$$\psi^{(\text{test})} = \frac{1}{|X^{(\text{test})}|} \sum_{x \in X^{(\text{test})}} L\left(x; A_{\lambda^{(*)}}(X^{(\text{train} \cup \text{val})})\right). \quad (3)$$

Essential for this process is the HPT algorithm in (HPT-4) that uses the validation performance to determine the next set of hyperparameters to evaluate. Updating hyperparameters is extremely challenging, because it

<sup>2</sup>Symbols used in this study are summarized in Table 1.

requires creating and training a new model on a dataset. And, the hyperparameter space  $\Lambda$  is not continuous or differentiable, because it also includes discrete decisions. Standard gradient methods are not applicable in  $\Lambda$ . Instead, gradient-free optimization techniques, e.g., pattern search or Evolution Strategies (ESs), which sometimes are far less efficient than gradient methods, are applied.

The following HPT approaches are popular:

- manual search,
- simple random search, i.e., choosing hyperparameters to evaluate at random, repeatedly,
- grid and pattern search [Meignan et al. 2015] [Lewis et al. 2000] [Tatsis and Parsopoulos 2016],
- model free algorithms, i.e., algorithms that do not explicitly make use of a model, e.g., ESs [Hansen 2006] [Bartz-Beielstein et al. 2014],
- hyperband, i.e., a multi-armed bandit strategy that dynamically allocates resources to a set of random configurations and uses successive halving to stop poorly performing configurations [Li et al. 2016],
- Surrogate Model Based Optimization (SMBO) such as Sequential Parameter Optimization Toolbox (SPOT), [Bartz-Beielstein et al. 2005], and [Bartz-Beielstein et al. 2021].<sup>3</sup>

Manual search and grid search are probably the most popular algorithms for HPT. Interestingly, Bergstra and Bengio [2012] demonstrate empirically and show theoretically that randomly chosen trials are more efficient for HPT than trials on a grid. Because their results are of practical relevance, they are briefly summarized here: In grid search the set of trials is formed by using every possible combination of values, grid search suffers from the curse of dimensionality because the number of joint values grows exponentially with the number of hyperparameters.

“For most data sets only a few of the hyperparameters really matter, but that different hyperparameters are important on different data sets. This phenomenon makes grid search a poor choice for configuring algorithms for new data sets” [Bergstra and Bengio 2012].

The observation that only a few of the parameters matter can also be observed in the engineering domain, where parameters such as pressure or temperature play a dominant role. In contrast to DL, this set of important parameters does not change fundamentally in different situations. We assume that the high variance in the set of important DL hyperparameters is caused by confounding.

Let  $\Psi$  denote the space of hyperparameter response functions (as defined in the Appendix, see Definition 5). Bergstra and Bengio [2012] claim that random search is more efficient than grid search because a hyperparameter response function  $\psi \in \Psi$  usually has a low effective dimensionality; essentially,  $\psi$  is more sensitive to changes in some dimensions than others [Caffisch et al. 1997].

Due to its simplicity, it turns out in many situations, especially in high-dimensional spaces, that random search is the best solution. Hyperband should also be mentioned in this context, although it can result in a worse final performance than model-based approaches, because it only samples configurations randomly and does not learn from previously sampled configurations [Li et al. 2016]. Bergstra and Bengio [2012] note that random search can probably be improved by automating what manual search does, i.e., using SMBO approaches such as SPOT.

HPT is a powerful technique that is an absolute requirement to get to state-of-the-art models on any real-world learning task, e.g., classification and regression. However, there are important issues to keep in mind when doing HPT: for example, validation-set overfitting can occur, because hyperparameters are optimized based on information derived from the validation data.

Falkner et al. [2018] claim, that practical Hyperparameter Optimization (HPO) solutions should fulfill the following requirements:

- strong anytime and final performance,
- effective use of parallel resources,
- scalability, as well as robustness and flexibility.

---

<sup>3</sup>The acronym SMBO originated in the engineering domain [Booker et al. 1999], [Mack et al. 2007]. It is also popular in the Machine Learning (ML) community, where it stands for *sequential model-based optimization*. We will use the terms *sequential model-based optimization* and *surrogate model-based optimization* synonymously.

In the context of benchmarking, a treatment for these issues was proposed by Bartz-Beielstein et al. [2020]. Although their recommendations (denoted as (R-1) to (R-8)) were developed for benchmark studies in optimization, they are also relevant for HPT, because HPT can be seen as a special benchmarking variant.

- (R-1) Goals: what are the reasons for performing HPT? Improving an existing solution, finding a solution for a new, unknown problem, or benchmarking two models are only three examples with different goals.
- (R-2) Problems: how to select suitable problems? Can surrogates accelerate the tuning?
- (R-3) Algorithms: how to select a portfolio of DL algorithms to be included in the HPT study?
- (R-4) Performance: how to measure performance?
- (R-5) Analysis: how to evaluate results? Hypothesis testing, rank-based comparisons.
- (R-6) Design: how to set up a study, e.g., how many runs shall be performed? Tools from Design of Experiments (DOE) are highly recommended.
- (R-7) Presentation: how to describe results? Presentation for the management of publication in a journal?
- (R-8) Reproducibility: how to guarantee scientifically sound results and how to guarantee a lasting impact, e.g., in terms of comparability?

In addition to these recommendations, there are some specific issues that are caused by the DL setup. These will be discussed in Sec. 5.

Note, some authors used the terms HPT and HPO synonymously. In the context of our analysis, these terms have different meanings:

**HPO** develops and applies methods to determine the best hyperparameters in an effective and efficient manner.

**HPT** develops and applies methods that try to analyze the effects and interactions of hyperparameters to enable *learning and understanding*.

This article proposes a HPT approach based on SPOT that focuses on the following topics:

**Limited Resources.** We focus on situations, where limited computational resources are available. This may be simply due the availability and cost of hardware, or because confidential data has to be processed strictly locally.

**Understanding.** In contrast to standard HPO approaches, SPOT provides statistical tools for *understanding* hyperparameter importance and interactions between several hyperparameters.

**Transparency and Explainability.** Understanding is a key tool for enabling transparency, e.g., quantifying the contribution of DL components (layers, activation functions, etc.).

**Reproducibility.** The software code used in this study is available in the open source R software environment for statistical computing and graphics (R) package SPOT via the Comprehensive R Archive Network (CRAN). SPOT is a well-established open-source software, that is maintained for more than 15 years [Bartz-Beielstein et al. 2005].

For sure, we are not seeking the overall best hyperparameter configuration that results in a NN which outperforms any other NN in every problem domain [Wolpert and Macready 1997]. Results are specific for one problem instance—their generalizability to other problem instances or even other problem domains is not self-evident and has to be proven [Haftka 2016].

This paper is structured as follows: Section 2 describes materials and methods that were used for the experimental setup. Experiments are described in Sec. 3. Section 4 presents results from a simple experiment. A discussion is presented in Sec. 5. The appendix contains information on how to set up the Python software environment for performing HPT with SPOT and Training Run Tools for TensorFlow (tfruns). Source code for performing the experiments will included in the R package SPOT. Further information are published on <https://www.spotseven.de> and, with some delay, on CRAN (<https://cran.r-project.org/package=SPOT>).

## 2 Materials and Methods

### 2.1 Hyperparameters

Typical hyperparameters that are used to define DNNs’ are as follows:

Table 1: Symbols used in this paper

Sym- bol	Name	Comment, Example
$A$	algorithm	
$G_x$	natural (ground truth) distribution	
$x$	data point	
$X$	data	usually partitioned into training, validation, and test data
$X^{(\text{train})}$	training data	
$X^{(\text{valid})}$	validation data	
$X^{(\text{test})}$	test data	
$t$	iteration counter	counter for the SPOT models, i.g., the $t$ -th SPOT metamodel will be denoted as $M(t)$
$\lambda$	hyperparameter configuration	
$\lambda_i$	$i$ -th hyperparameter configuration	used in SMBO
$\lambda^{(*)}$	best hyperparameter configuration	best configuration in theory
$\hat{\lambda}$	best hyperparameter configuration obtained by evaluating a finite set of samples	best configuration “in practice”
$\Lambda$	hyperparameter space	
$\Psi$	hyperparameter response space	
$\psi_i$	hyperparameter response surface function evaluated for the $i$ -th hyperparameter configuration $\lambda_i$	
$\psi^{(\text{train})}$	hyperparameter response surface function (on train data)	
$\psi^{(\text{test})}$	hyperparameter response surface function (on test data)	as defined in Equation 3
$\psi^{(\text{val})}$	hyperparameter response surface function (on validation data)	as defined in Equation 1

- optimization algorithms, e.g., Root Mean Square Propagation (RMSProp) (implemented in Keras as `optimizer_rmsprop()`) or ADaptive Moment estimation algorithm (ADAM) (`optimizer_adam()`). These will be discussed in Sec.2.2.1.
- loss functions, e.g., Mean Squared Error (MSE) (`loss_mean_squared_error()`), Mean Absolute Error (MAE) (`loss_mean_absolute_error()`), or Categorical Cross Entropy (CCE) (`loss_categorical_crossentropy()`). The actual optimized objective is the mean of the output array across all datapoints.
- learning rate
- activation functions
- number of hidden layers and hidden units
- size of the training batches
- weight initialization schemes
- regularization penalties
- dropout rates. Dropout is a commonly used regularization technique for DNNs. Applied to a layer, dropout consists of randomly setting to zero (dropping out) a percentage of output features of the layer during training [Chollet and Allaire 2018].
- batch normalization

**Example 1** (Conditionally dependent hyperparameters; Mendoza et al. [2019]). *This example illustrates that some hyperparameters are conditionally dependent on the number of layers. Mendoza et al. [2019] consider*

**Network hyperparameters**, e.g., batch size, number of updates, number of layers, learning rate,  $L_2$  regularization, dropout output layer, solver type (SGD, Momentum, ADAM, Adadelta, Adagrad, smorm, Nesterov), learning-rate policy (fixed, inv, exp, step)

**Parameters conditioned on solver type**, e.g.,  $\beta_1$  and  $\beta_2$ ,  $\rho$ , MOMENTUM,

**Parameters conditioned on learning-rate policy**, e.g.,  $\gamma$ ,  $k$ , and  $s$ ,

**Per-layer hyperparameters**, e.g., activation-type (sigmoid, tanH, ScaledTanH, ELU, ReLU, Leaky, Linear), number of units, dropout in layer, weight initialization (Constant, Normal, Uniform, Glorot-Uniform, Glorot-Normal, He-Normal), std. normal init., leakiness, tanh scale in/out.

For practical reasons, Mendoza et al. [2019] constrained the number of layers to be between one and six: firstly, they aimed to keep the training time of a single configuration low, and secondly each layer adds eight per-layer hyperparameters to the configuration space, such that allowing additional layers would further complicate the configuration process.

## 2.2 Hyperparameter: Features

This section considers some properties, which are specific to DNN hyperparameters.

### 2.2.1 Optimizers

Choi et al. [2019] considered RMSProp with momentum [Tieleman and Hinton 2012], ADAM [Kingma and Ba 2015] and ADAM [Dozat 2016] and claimed that the following relations holds:

$$\begin{aligned} \text{SGD} &\subseteq \text{MOMENTUM} \subseteq \text{RMSPROP} \\ \text{SGD} &\subseteq \text{MOMENTUM} \subseteq \text{ADAM} \\ \text{SGD} &\subseteq \text{NESTEROV} \subseteq \text{NADAM} \end{aligned}$$

**Example 2** (ADAM can approximately simulate MOMENTUM). *MOMENTUM can be approximated with ADAM, if a learning rate schedule that accounts for ADAM’s bias correction is implemented.*

Choi et al. [2019] demonstrated that these inclusion relationships are meaningful in practice. In the context of HPT and HPO, inclusion relations can significantly reduce the complexity of the experimental design. These inclusion relations justify the selection of a basic set, e.g., RMSProp, ADAM, and Nesterov-accelerated Adaptive Moment Estimation (NADAM).

### 2.2.2 Batch Size

Shallue et al. [2019] and Zhang et al. [2019] have shown empirically that increasing the batch size can increase the gaps between training times for different optimizers.

## 2.3 Performance Measures for Hyperparameter Tuning

### 2.3.1 Measures

Kedziora et al. [2020] state that “unsurprisingly”, accuracy<sup>4</sup> is considered as the most important performance measure. Accuracy might be adequate, if data is balanced. For unbalanced data, other measures are better. In general, there are many other ways to measure model quality, e.g., metrics based on time complexity and robustness or the model complexity (interpretability) [Bartz-Beielstein et al. 2020].

In contrast to classical optimization, where the same optimization function can be used for tuning and final evaluation, training of DNNs faces a different situation:

- training is based on the loss function,
- whereas the final evaluation is based on a different measure, e.g., accuracy.

The loss function acts as a surrogate for the performance measure the user is finally interested in. Several performance measures are used at different stages of the HPO procedures:

<sup>4</sup>Accuracy in binary classification is the proportion of correct predictions among the total number of observations [Metz 1978].

1. training loss, i.e.,  $\psi^{(\text{train})}$ ,
2. training accuracy, i.e.,  $f_{\text{acc}}^{(\text{train})}$ ,
3. validation loss, i.e.,  $\psi^{(\text{val})}$ ,
4. validation accuracy, i.e.,  $f_{\text{acc}}^{(\text{val})}$ ,
5. test loss, i.e.,  $\psi^{(\text{test})}$ , and
6. test accuracy, i.e.,  $f_{\text{acc}}^{(\text{test})}$ .

This complexity gives reason for the following question:

**Question:** Which performance measure should be used during the HPT (HPO) procedure?

Most authors recommend using test accuracy or test loss as the measure for hyperparameter tuning [Schneider et al. 2019]. In order to understand the correct usage of these performance measures, it is important to look at the goals, i.e., selection or assessment, of a tuning study.

### 2.3.2 Model Selection and Assessment

Hastie et al. [2017] stated that selection and assessment are two separate goals:

**Model selection**, i.e., estimating the performance of different models in order to choose the best one. Model selection is important *during* the tuning procedure, whereas model assessment is used for the *final* report (evaluation of the results).

**Model assessment**, i.e., having chosen a final model, estimating its prediction error (generalization error) on new data. Model assessment is performed to ascertain whether predicted values from the model are likely to accurately predict responses on future observations or samples not used to develop the model. Overfitting is a major problem in this context.

In principle, there are two ways of model assessment and selection: internal versus external. In the following,  $N$  denotes the total number of samples.

**External assessment** uses different sets of data. The first  $m$  data samples are for model training and  $N - m$  for validation. Problem: holding back data from model fitting results in lower precision and power.

**Internal Assessment** uses data splitting and resampling methods. The true error might be *underestimated*, because the same data samples that were used for fitting the model are used for prediction. The so-called in-sample (also apparent, or resubstitution) error is smaller than the true error.

In a data-rich situation, the best approach for both problems is to randomly divide the dataset into three parts:

1. a training set to fit the models,
2. a validation set to estimate prediction error for model selection, and
3. a test set for assessment of the generalization error of the final chosen model.

The test set should be brought out only at the end of the data analysis. It should not be used during the training and validation phase. If the test set is used repeatedly, e.g., for choosing the model with smallest test-set error, “the test set error of the final chosen model will underestimate the true test error, sometimes substantially.” [Hastie et al. 2017]

The following example 3 shows that there is no general agreement on how to use training, validation, and test sets as well as the associated performance measures.

**Example 3** (Basic Comparisons in Manual Search). *Wilson et al. [2017] describe a manual search. They allocated a pre-specified budget on the number of epochs used for training each model.*

- *When a test set was available, it was used to chose the settings that achieved the best peak performance on the test set by the end of the fixed epoch budget.*

- *If no explicit test set was available, e.g., for Canadian Institute for Advanced Research, 10 classes (CIFAR-10), they chose the settings that achieved the lowest training loss at the end of the fixed epoch budget.*

Theoretically, in-sample error is not usually of interest because future values of the hyperparameters are not likely to coincide with their training set values. Bergstra and Bengio [2012] stated that because of finite data sets, test error is not monotone in validation error, and depending on the set of particular hyperparameter values  $\lambda$  evaluated, the test error of the best-validation error configuration may vary, e.g.,

$$\psi_i^{(\text{train})} < \psi_j^{(\text{train})} \not\Rightarrow \psi_i^{(\text{test})} < \psi_j^{(\text{test})}, \tag{4}$$

where  $\psi_i^{(\cdot)}$  denotes the value of the hyperparameter response surface for the  $i$ -th hyperparameter configuration  $\lambda_i$ .

Furthermore, the estimator, e.g., for loss, obtained by using a single hold-out test set usually has high variance. Therefore, CV methods were proposed. Hastie et al. [2017] concluded

“that estimation of test error for a particular training set is not easy in general, given just the data from that same training set. Instead, cross-validation and related methods may provide reasonable estimates of the expected error.”

The standard practice for evaluating a model found by CV is to report the hyperparameter configuration that minimizes the loss on the validation data, i.e.,  $\hat{\lambda}$  as defined in Equation 10. Repeated CV is considered standard practice, because it reduces the variance of the estimator.  $k$ -fold CV results in a more accurate estimate as well as in some information about its distribution. There is, as always, a trade-off: the more CV folds the better the estimate, but more computational time is needed.

When different trials have nearly optimal validation means, then it is not clear which test score to report: small changes in the hyperparameter values could generate a different test error.

**Example 4** (Reporting the model assessment (final evaluation) [Bergstra and Bengio 2012]). *When reporting performance of learning algorithms, it can be useful to take into account the uncertainty due to the choice of hyperparameters values. Bergstra and Bengio [2012] present a procedure for estimating test set accuracy, which takes into account any uncertainty in the choice of which trial is actually the best-performing one. To explain this procedure, they distinguish between estimates of performance  $\psi^{(\text{val})}$  and  $\psi^{(\text{test})}$  based on the validation and test sets, respectively.*

*To resolve the difficulty of choosing a winner, Bergstra and Bengio [2012] reported a weighted average of all the test set scores, in which each one is weighted by the probability that its particular  $\lambda_s$  is in fact the best. In this view, the uncertainty arising from  $X^{(\text{valid})}$  being a finite sample of the natural (ground truth) distribution  $G_x$  makes the test-set score of the best model among  $\{\lambda_i\}_{i=1,2,\dots,S}$  a random variable,  $z$ .*

## 2.4 Practical Considerations

Unfortunately, training, validation, and test data are used inconsistently in HPO studies: for example, Wilson et al. [2017] selected *training loss*,  $\psi^{(\text{train})}$ , (and not validation loss) during optimization and reported results on the test set  $\psi^{(\text{test})}$ .

Choi et al. [2019] considered this combination as a “somewhat non-standard choice” and performed tuning (optimization) on the validation set, i.e., they used  $\psi^{(\text{val})}$  for tuning, and reported results  $\psi^{(\text{test})}$  on the test set. Their study allows some valuable insight into the relationship of validation and test error:

“For a *relative comparison* between models during the tuning procedure, in-sample error is convenient and often leads to effective model selection. The reason is that the relative (rather than absolute performance) error is required for the comparisons.” [Choi et al. 2019]

Choi et al. [2019] compared the final predictive performance of NN optimizers after tuning the hyperparameters to minimize validation error. They concluded that their “final results hold regardless of whether they compare final validation error, i.e.,  $\psi^{(\text{val})}$ , or test error, i.e.,  $\psi^{(\text{test})}$ ”. Figure 1 in Choi et al. [2019] illustrates that the relative performance of optimizers stays the same, regardless of whether the validation or the test error is used. Choi et al. [2019] considered two statistics: (i) the quality of the best solution and (ii) the speed of training, i.e., the number of steps required to reach a fixed validation target.



### 2.4.1 Some Considerations about Cross Validation

There are some drawbacks of  $k$ -fold CV: at first, the choice of the number of observations to be hold out from each fit is unclear: if  $m$  denotes the size of the training set, with  $k = m$ , the CV estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because the  $m$  “training sets” are similar to one another. The computational costs are relatively high, because  $m$  evaluations of the model are necessary. Secondly, the number of repetitions needed to achieve accurate estimates of accuracy can be large. Thirdly, CV does not fully represent variability of variable selection: if  $m$  subjects are omitted each time from set of  $N$ , the sets of variables selected from each sample of size  $N - m$  are likely to be different from sets obtained from independent samples of  $N$  subjects. Therefore, CV does not validate the full  $N$  subject model. Note, Monte-Carlo CV is an improvement over standard CV [Picard and Cook 1984].

## 2.5 Related Work

Before presenting the elements (benchmarks and software tools) for the experiments in Sec. 2.7, we consider existing, related approaches that might be worth looking at. This list is not complete and will be updated in forthcoming versions of this paper.

### 2.5.1 Hyperparameter Optimization Software and Benchmark Studies

SMBO based on Kriging (aka Gaussian processes or Bayesian Optimization (BO)) has been successfully applied to HPT in several works, e.g., Bartz-Beielstein and Markon [2004] propose a combination of classical statistical tools, BO (Design and Analysis of Computer Experiments (DACE)), and Classification and Regression Trees (CART) as a surrogate model. The integration of CART made SMBO applicable to more general HPT problems, e.g., problems with categorical parameters. Hutter et al. [2011] presented a similar approach by proposing Sequential Model-Based Optimization for General Algorithm Configuration (SMAC) as a tuner that is capable of handling categorical parameters by using surrogate models based on random forests. Similar to the Optimal Computing Budget Allocation (OCBA) approach in SPOT, Hutter et al. [2011] implemented an *intensification mechanism* for handling multiple instances. Early SPOT versions used a very simple intensification mechanism: (i) the best solution is evaluated in each iteration and (ii) new candidate solutions, that were proposed by the surrogate model, are evaluated as often as the current best solution. This simple intensification strategy was replaced by the more sophisticated OCBA strategy in SPOT [Bartz-Beielstein et al. 2011].

HPO developed very quickly, new branches and extensions were proposed, e.g., Combined Algorithm Selection and Hyperparameter optimization (CASH), Neural Architecture Search (NAS), Automated Hyperparameter and Architecture Search (AutoHAS), and further “Auto-\*” approaches [Thornton et al. 2013], [Dong et al. 2020]. Kedziora et al. [2020] analyzed what constitutes these systems and survey developments in HPO, multi-component models, NN architecture search, automated feature engineering, meta-learning, multi-level ensembling, dynamic adaptation, multi-objective evaluation, resource constraints, flexible user involvement, and the principles of generalization. The authors developed a conceptual framework to illustrate one possible way of fusing high-level mechanisms into an autonomous ML system. *Autonomy* is considered as the capability of ML systems to independently adjust their results even in dynamically changing environments. They discuss how Automated Machine Learning (AutoML) can be transformed into Autonomous Machine Learning (AutonoML), i.e, the systems are able to independently “design, construct, deploy, and maintain” ML models similar to the Cognitive Architecture for Artificial Intelligence (CAAI) approach presented by Strohschein et al. [2021]. Because Kedziora et al. [2020] already presented a comprehensive overview of this development, we will list the most relevant “highlights” in the following.

Snoek et al. [2012] used the CIFAR-10 dataset, which consists of 60,000  $32 \times 32$  colour images in ten classes, for optimizing the hyperparameters of a Convolutional Neural Networks (CNNs).

Bergstra et al. [2013] proposed a meta-modeling approach to support automated HPO, with the goal of providing practical tools that replace hand-tuning. They optimized a three layer CNN.

Eggenesperger et al. [2013] collected a library of HPO benchmarks and evaluated three BO methods. They considered the HPO problem under  $k$ -fold CV as a minimization problem of  $\psi^{(\text{val})}$  as defined in Equation 2.

Zoph et al. [2017] studied a new paradigm of designing CNN architectures and describe a scalable method to optimize these architectures on a dataset of interest, for instance the ImageNet classification dataset.

Balaprakash et al. [2018] presented DeepHyper, a Python package that provides a common interface for the implementation and study of scalable hyperparameter search methods.

Karmanov et al. [2018] created a “*Rosetta Stone*” of DL frameworks to allow data-scientists to easily leverage their expertise from one framework to another. They provided a common setup for comparisons across GPUs (potentially CUDA versions and precision) and for comparisons across languages (Python, Julia, R). Users should be able to verify expected performance of own installation.

Mazzawi et al. [2019] introduced a NAS framework to improve keyword spotting and spoken language identification models.

Mendoza et al. [2019] introduced Auto-Net, a system that automatically configures NN with SMAC by following the same AutoML approach as Auto-WEKA and Auto-sklearn. They achieved the best performance on two datasets in the human expert track of an AutoMLChallenge.

O’Malley et al. [2019] presented *Keras tuner*, a hyperparameter tuner for Keras with TensorFlow 2.0. They defined a model-building function, which takes an argument from which hyperparameters such as the units (hidden nodes) of the neural network. Available tuners are RandomSearch and Hyperband.

Because optimizers can affect the DNN performance significantly, several tuning studies devoted to optimizers were published during the last years: Schneider et al. [2019] introduced a benchmarking framework called Deep Learning Optimizer Benchmark Suite (DeepOBS), which includes a wide range of realistic DL problems together with standardized procedures for evaluating optimizers. Schmidt et al. [2020] performed an extensive, standardized benchmark of fifteen particularly popular DL optimizers.

A highly recommended study was performed by Choi et al. [2019], who presented a taxonomy of first-order optimization methods. Furthermore, Choi et al. [2019] demonstrated the sensitivity of optimizer comparisons to the hyperparameter tuning protocol. Optimizer rankings can be changed easily by modifying the hyperparameter tuning protocol. Their findings raised serious questions about the practical relevance of conclusions drawn from certain ways of empirical comparisons. They also claimed that tuning protocols often differ between works studying NN optimizers and works concerned with training NNs to solve specific problems.

Zimmer et al. [2020] developed Auto-PyTorch, a framework for Automated Deep Learning (AutoDL) that uses Bayesian Optimization HyperBand (BOHB) as a backend to optimize the full DL pipeline, including data preprocessing, network training techniques and regularization methods.

Mazzawi and Gonzalvo [2021] presented Google’s Model Search, which is an open source platform for finding optimal ML models based on TensorFlow (TF). It does not focus on a specific domain.

Wistuba et al. [2019] described how complex DL architectures can be seen as combinations of a few elements, so-called *cells*, that are repeated to build the complete network. Zoph and Le [2016] were the first who proposed a cell-based approach, i.e., choices made about a NN architecture is the set of meta-operations and their arrangement within the cell. Another interesting example are function-preserving morphisms implemented by the Auto-Keras package to effectively traverse potential networks Jin et al. [2019].

Tunability is an interesting concept that should be mentioned in the context of HPT [Probst et al. 2019]. The term *tunability*/ describes a measure for modeling algorithms as well as for individual hyperparameters. It is the difference between the model quality for default values (or reference values) and the model quality for optimized values (after HPT is completed). Or in the words of Probst et al. [2019]: “measures for quantifying the tunability of the whole algorithm and specific hyperparameters based on the differences between the performance of default hyperparameters and the performance of the hyperparameters when this hyperparameter is set to an optimal value”. Tunability of individual hyperparameters can also be used as a measure of their *relevance*, *importance*, or *sensitivity*. Hyperparameters with high tunability are accordingly of greater importance for the model. The model reacts strongly to (i.e., is sensitive to) changes in these hyperparameters. The hope is that identifying *tunable* hyperparameters, i.e., ones that model performance is particularly sensitive to, will allow other settings to be ignored, constraining search space. Unfortunately, tunability strongly depends on the choice of the dataset, which makes generalization of results very difficult.

To conclude this overview, we would like to mention relevant criticism of HPO: some publications even claimed that extensive HPO is not necessary.

1. Erickson et al. [2020] introduced a framework (AutoGluon-Tabular) that “requires only a single line of Python to train highly accurate machine learning models on an unprocessed tabular dataset such as a CSV file”. AutoGluon-Tabular ensembles several models and stacks them in multiple layers.

Table 2: The hyperparameters and architecture choices for the fully connected networks as defined in Falkner et al. [2018]

Hyperparameter	Lower Bound	Upper Bound	Log-transform
batch size	$2^3$	$2^8$	yes
dropout rate	0	0.5	no
initial learning rate	$1e - 6$	$1e - 2$	yes
exponential decay factor	-0.185	0	no
# hidden layers	1	5	no
# units per layer	$2^4$	$2^8$	yes

The authors claim that AutoGluon-Tabular outperforms AutoML platforms such as TPOT, H2O, AutoWEKA, auto-sklearn, AutoGluon, and Google AutoML Tables.

2. Yu et al. [2020] claimed that the evaluated state-of-the-art NAS algorithms do not surpass random search by a significant margin, and even perform worse in the Recurrent Neural Network (RNN) search space. Balaji and Allen [2018] reported a multitude of issues when attempting to execute automatic ML frameworks. For example, regarding the random process, the authors state that “one common failure is in large multi-class classification tasks in which one of the classes lies entirely on one side of the train test split”.
3. Liu [2018] remarks that “for most existent AutoML works, regardless of the number of layers of the outer-loop algorithms, the configuration of the outermost layer is definitely done by human experts”. Human experts are shifted to a higher level, and are still in the loop.
4. Li and Talwalkar [2019] stated that (i) better baselines that accurately quantify the performance gains of NAS methods, (ii) ablation studies (to learn about the NN by removing parts of it) that isolate the impact of individual NAS components, and (iii) reproducible results that engender confidence and foster scientific progress are necessary.

### 2.5.2 Artificial Toy Functions

Because BO does not work well on high-dimensional mixed continuous and categorical configuration spaces, Falkner et al. [2018] used a simple counting ones problem to analyze this problem. Zaefferer and Bartz-Beielstein [2016] discussed these problems in greater detail. How to implement BO for discrete (and continuous) optimization problems was analyzed in the seminal paper by Bartz-Beielstein and Zaefferer [2017].

### 2.5.3 Experiments on Surrogate Benchmarks

Falkner et al. [2018] optimized six hyperparameters that control the training procedure of a fully connected DNN (initial learning rate, batch size, dropout, exponential decay factor for learning rate) and the architecture (number of layers, units per layer) for six different datasets gathered from OpenML [Vanschoren et al. 2014], see Table 2.

Falkner et al. [2018] used a surrogate DNN as a substitute for training the networks directly, . To build a surrogate, they sampled 10,000 random configurations for each data set, trained them for 50 epochs, and recorded classification error after each epoch, and total training time. Two independent random forests models were fitted to predict these two quantities as a function of the hyperparameter configuration used. Falkner et al. [2018] noted that Hyperband (HB) initially performed much better than the vanilla BO methods and achieved a roughly three-fold speedup over random search.

Instead of using a surrogate network, we will use the original DNNs. Our approach is described in Sec.3.

## 2.6 Stochasticity

Results from DNN tuning runs are noisy, e.g., caused by random sampling of batches and initial parameters. Repeats to estimate means and variances that are necessary for a sound statistical analysis require substantial computational costs.

## 2.7 Software: Keras, Tensorflow, tfruns, and SPOT

### 2.7.1 Keras and Tensorflow

Keras is the high-level Application Programming Interface (API) of TF, which is developed with a focus on enabling fast experimentation. TF is an open source software library for numerical computation using data flow graphs [Abadi et al. 2016]. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them [O’Malley et al. 2019]. The `tensorflow` R package provides access to the complete TF API from within R.

### 2.7.2 The R Package tfruns

The R package `tfruns`<sup>5</sup> provides a suite of tools for tracking, visualizing, and managing TF training runs and experiments from R. `tfruns` enables tracking the hyperparameters, metrics, output, and source code of every training run and comparing hyperparameters and metrics across runs to find the best performing model. It automatically generates reports to visualize individual training runs or comparisons between runs. `tfruns` can be used without any changes to source code, because run data is automatically captured for all Keras and TF models.

### 2.7.3 SPOT

The SPOT package for R is a toolbox for tuning and understanding simulation and optimization algorithms [Bartz-Beielstein et al. 2021]. SMBO investigations are common approaches in simulation and optimization. Sequential parameter optimization has been developed, because there is a strong need for sound statistical analysis of simulation and optimization algorithms. SPOT includes methods for tuning based on classical regression and analysis of variance techniques; tree-based models such as CART and random forest; BO (Gaussian process models, aka Kriging), and combinations of different meta-modeling approaches.

SPOT implements key techniques such as exploratory fitness landscape analysis and sensitivity analysis. SPOT can be used for understanding the performance of algorithms and gaining insight into algorithm’s behavior. Furthermore, SPOT can be used as an optimizer and for automatic and interactive tuning. SPOT finds improved solutions in the following way:

1. Initially, a population of (random) solutions is created.
2. A set of surrogate models is specified.
3. Then, the solutions are evaluated on the objective function.
4. Next, surrogate models are built.
5. A global search is performed on the surrogate model(s) to generate new candidate solutions.
6. The new solutions are evaluated on the objective function, e.g., the loss is determined.

These steps are repeated, until a satisfying solution has been found as described in Bartz-Beielstein et al. [2021].

**SPOT Surrogate Models** SPOT performs model selection during the tuning run: training data  $X^{(\text{train})}$  is used for fitting (training) the models, e.g., the weights of the DNNs. Each trained model  $A_{\lambda_i}(X^{(\text{train})})$  will be evaluated on the validation data  $X^{(\text{val})}$ , i.e., the loss is calculated as

$$\psi_i^{(\text{val})} = \frac{1}{|X^{(\text{val})}|} \sum_{x \in X^{(\text{val})}} L(x; A_{\lambda_i}(X^{(\text{train})})). \quad (5)$$

Based on  $(\lambda_i, \psi_i^{(\text{val})})$ , a surrogate model  $M(t)$  is fitted, e.g., a BO (Kriging) model using SPOT’s `buildKriging` function. Figure 2 shows one example.

For each hyperparameter configuration  $\lambda_i$ , SPOT reports information about the related DNN models  $A_{\lambda_i}$

1. training loss,  $\psi^{(\text{train})}$ ,

---

<sup>5</sup><https://cran.r-project.org/package=tfruns>, <https://tensorflow.rstudio.com/tools/tfruns>

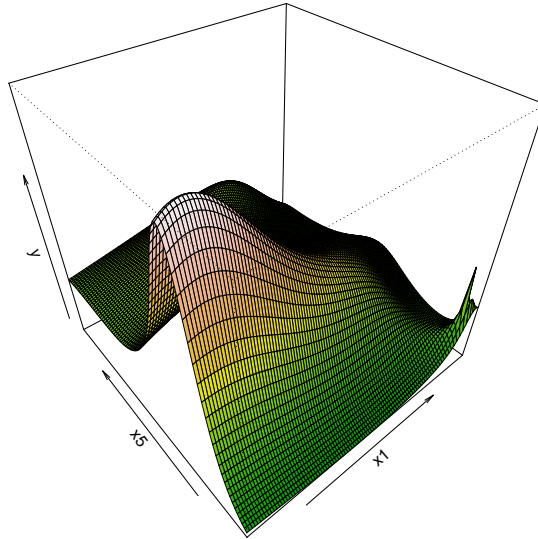


Figure 2: Perspective plot of the surrogate model used by SPOT in this study

2. training accuracy,  $f_{\text{acc}}^{(\text{train})}$ ,
3. validation (testing) loss,  $\psi^{(\text{val})}$ , and
4. validation (testing) accuracy,  $f_{\text{acc}}^{(\text{val})}$ .

Output from a typical run is show in Figure 3.

### 3 Experiments: Tuning Hyperparameters with SPOT

How the software packages (Keras, TF, tfruns, and SPOT) can be combined in a very efficient and effective manner will be exemplified in this section. The general DNN workflow is as follows: first the training data, `train_images` and `train_labels` are fed to the DNN. The DNN will then learn to associate images and labels. Based on the Keras parameter `validation_split`, the training data will be partitioned into a (smaller) training data set and a validation data set. The corresponding code is shown in Sec. 3.1. The trained DNN produces predictions for validations.

#### 3.1 The Data Set: MNIST

The DNN in this example uses the Keras R package to learn to classify hand-written digits from the Modified National Institute of Standards and Technology (MNIST) data set. This is a supervised multi-class classification problem, i.e., grayscale images of handwritten digits ( $28 \times 28$  pixels) should be assigned to ten categories (0 to 9). MNIST is a set of 60,000 training and 10,000 test images. The MNIST data set is included in Keras as `train` and `test` lists, each of which includes a set of images (`x`) and associated labels (`y`): `train_images` and `train_labels` form the training set, the data that the DNN will learn from. The DNN can be tested on the  $X^{(\text{test})}$  set (`test_images` and `test_labels`). The images are encoded as as 3D arrays, and the labels are a 1D array of digits, ranging from 0 to 9.

Before training the DNN, the data are preprocessed by reshaping it into the shape the DNN can process. The natural (original) training images were stored in an array of shape  $(60000, 28, 28)$  of type `integer` with values in the  $[0, 255]$  interval. They are transformed into a double array of shape  $(60000, 28 \times 28)$  with Red, Green, and Blue color space (RGB) values between 0 and 1, i.e., all values will be scaled that they are in the  $[0, 1]$  interval. Furthermore, the labels are categorically encoded.

```
mnist <- dataset_mnist()
x_train <- mnist$train$x
```

```

y_train <- mnist$train$y
x_test <- mnist$test$x
y_test <- mnist$test$y

dim(x_train) <- c(nrow(x_train), 784)
dim(x_test) <- c(nrow(x_test), 784)

x_train <- x_train / 255
x_test <- x_test / 255

y_train <- to_categorical(y_train, 10)
y_test <- to_categorical(y_test, 10)

```

### 3.2 The Neural Network

The DNN consists of a sequence of two dense (fully connected) layers. The second layer is a ten-way softmax layer that returns an array of ten probability scores. Each score represents the probability that the input image belongs to one of the ten MNIST digit classes.

```

model <- keras_model_sequential()
model %>%
  layer_dense(units = 256, activation = 'relu', input_shape = c(784)) %>%
  layer_dropout(rate = 0.4) %>%
  layer_dense(units = 128, activation = 'relu') %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 10, activation = 'softmax')

```

Finally, (i) the *loss function*, which determines how the DNN good a prediction is based on the training data, (ii) the *optimizer*, i.e., the update mechanism of the network, which adjusts the weights using backpropagation, and (iii) the *metrics*, which monitor the progress during training and testing, are specified using the `compile` function from Keras.

```

model %>% compile(
  loss = 'categorical_crossentropy',
  optimizer = optimizer_rmsprop(lr = 0.01),
  metrics = c('accuracy')
)

```

The DNN training can be started as follows (using Kera's `fit` function).

```

history <- model %>% keras::fit(
  x_train, y_train,
  batch_size = 128,
  epochs = 20,
  verbose = 0,
  validation_split = 0.2
)

```

Figure 3 shows the quantities that are being displayed during training:

- (i) the *loss* of the network over the training and validation data,  $\psi^{(\text{train})}$  and  $\psi^{(\text{val})}$ , respectively, and
- (ii) the *accuracy* of the network over the training and validation data,  $f_{\text{acc}}^{(\text{train})}$  and  $f_{\text{acc}}^{(\text{val})}$ , respectively.

This figure illustrates that an accuracy greater than 95 percent on the training data can be reached quickly. Finally, using Kera's `evaluate` function, the DNN model performance can be checked on  $X^{(\text{test})}$ .

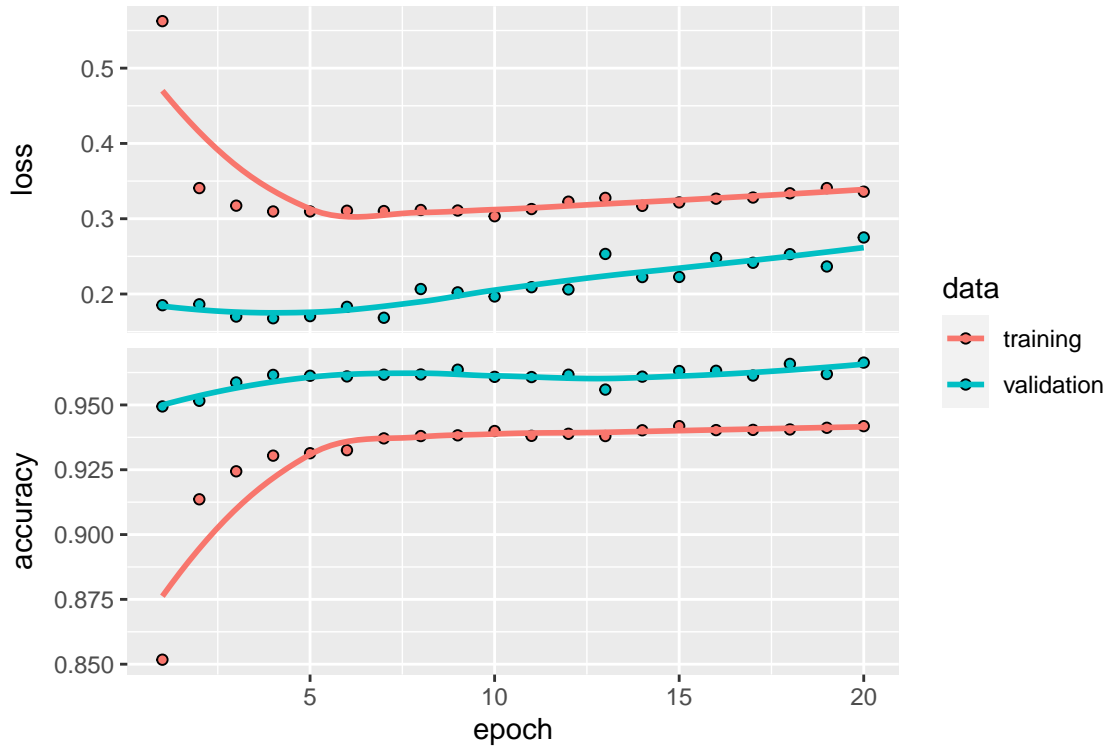


Figure 3: Training and validation data. Loss and accuracy plotted against epochs.

```
score <- model %>% evaluate(
  x_test, y_test,
  verbose = 0
)
cat('Test loss:', score[[1]], '\n')
cat('Test accuracy:', score[[2]], '\n')
```

```
Test loss: 0.2828377
Test accuracy: 0.9638
```

The relationship between  $\psi^{(\text{train})}$ ,  $\psi^{(\text{val})}$ , and  $\psi^{(\text{test})}$  as well as between  $f_{\text{acc}}^{(\text{train})}$ ,  $f_{\text{acc}}^{(\text{val})}$ , and  $f_{\text{acc}}^{(\text{test})}$  can be analyzed with SPOT.

Running the DNN model as a standalone process *before* starting the tuning process is strongly recommended. As shown in this section, the default DNN model seems to work fine.

### 3.3 Interfacing tfruns from SPOT

After testing the model as a standalone implementation, the model can be combined with the SPOT framework. A wrapper function is used to connect tfruns to the SPOT tuner. The setup requires a few lines of R code only. Instead of two hyperparameters, `var1` and `var2`, that are passed to TF as shown in the following code example, an arbitrary amount of hyperparameters can be passed.

```
funTfrunsSingle <- function(x) {
  runs <- tuning_run("kerasModel.R",
    flags = list(var1 = x[1], var2 = x[2]),
    confirm = FALSE)
  runs$metric_val_loss[[1]]
}
funTfruns <- wrapFunction(funTfrunsSingle)
```

Table 3: The hyperparameters and architecture choices for the first DNN example: fully connected networks

Variable Name	Hyperparameter	Type	Default	Lower Bound	Upper Bound
$x_1$	first layer dropout rate	numeric	0.4	$1e-6$	1
$x_2$	second layer dropout rate	numeric	0.3	$1e-6$	1
$x_3$	units per first layer	integer	256	16	512
$x_4$	units per second layer	integer	128	4	256
$x_5$	learning rate	numeric	0.001	0.0001	0.1
$x_6$	training epochs	integer	20	5	25
$x_7$	batch size	integer	64	8	256
$x_8$	rho	numeric	0.9	0.5	0.999

The first line defines the R function `funTfrunsSingle()` for a single hyperparameter configuration. It calls the code from the `kerasModel.R` file, which implements the DNN described in Sec. 3.1.

In order to evaluate several hyperparameter configurations during one single function call, SPOT’s `wrapFunction()` is applied to the `funTfrunsSingle()` function. Note, that SPOT operates on `matrix` objects.

### 3.4 Hyperparameter Tuning with SPOT

The following hyperparameters will be tuned:

- $x_1, x_2$  the *dropout rates*. The dropout rates of the first and second layer will be tuned individually.
- $x_3, x_4$  the *number of units*, i.e., the number of single outputs from a single layer. The number of units of the first and second layer will be tuned individually.
- $x_5$  the *learning rate*, which controls how much to change the DNN model in response to the estimated error each time the model weights are updated.
- $x_6$  the *number of training epochs*, where a training epoch is one forward and backward pass of a complete data set.
- $x_7$  the *batch size*, and
- $x_8$  `optimizer_rmsprop()`’s *decay factor*.

These hyperparameters and their ranges are listed in Table 3. Using these parameter specifications, we are ready to perform the first SPOT HPT run:

```
res <- spot(
  x = NULL,
  fun = funTfruns,
  lower <- c(1e-6, 1e-6, 16, 16, 1e-9, 10, 16, 0.5),
  upper <- c(0.5, 0.5, 512, 256, 1e-2, 50, 512, 1-1e-3),
  control = list(
    funEvals = 480,
    types = c(
      "numeric",
      "numeric",
      "integer",
      "integer",
      "numeric",
      "integer",
      "integer",
      "numeric"
    )
  )
)
```

The budget is set to  $n = 480$  evaluations (20 times the number of hyperparameters (8), multiplied by the number of repeats (3)), i.e., the total number of DNN training and testing iterations.



Figure 4: Internal ‘list’ structure of the result object ‘res’ from the SPOT run.

```
List of 9
$ xbest : num [1, 1:8] 4.67e-01 1.48e-01 4.31e+02 9.20e+01 1.97e-04 ...
$ ybest : num [1, 1] 0.0685
$ x      : num [1:480, 1:8] 0.433 0.134 0.339 0.222 0.215 ...
$ y      : num [1:480, 1] 0.8908 0.0837 0.1272 0.1418 0.1203 ...
$ logInfo : logi NA
$ count   : int 480
$ msg      : chr "budget exhausted"
$ modelFit:List of 33
..$ thetaLower : num 1e-04
..$ thetaUpper : num 100
..$ types      : chr [1:8] "numeric" "numeric" "integer" "integer" ...
...
..$ min        : num 0.0717
..- attr(*, "class")= chr "kriging"
$ ybestVec: num [1:280] 0.0824 0.0824 0.0824 0.0824 0.0824 0.0824 0.0824 0.0824 0.0824 0.0824 0.0824 ...
```

SPOT provides several options for adjusting the HPT parameters, e.g., type of the SMBO model and optimizer as well as the size of the initial design. These parameters can be passed via the `spotControl` function to SPOT. For example, instead of the default model, which is BO, a random forest can be chosen. A detailed description of the SPOT tuning algorithm can be found in Bartz-Beielstein et al. [2021].

## 4 Results

While discussing the hyperparameter tuning results, HPT does not look the the final, best solution only. For sure, the hyperparameter practitioner is interested in the best solution. But even from this *greedy* point of view, considering the *route to the solution* can is also of great importance, because analysing this route enables *learning* and can be much more efficient in the long run compared to a greedy strategy.

**Example 5.** Consider a classification task that has to be performed several times in a different context with similar data. Instead of blindly (automatically) running the HPO procedure individually for each classification task (which might also require a significant amount of time and resources, even when it is performed automatically) a few HPT procedures are performed. Insights gained from HPT might help to avoid pitfalls such as ill specified parameter ranges, too short run times, etc.

In addition to an effective and efficient way to determine the optimal hyperparameters, SPOT provides tools for *learning and understanding*.<sup>6</sup>

The HPT experiment from Sec. 3 used  $n = 480$  DNN evaluations, i.e., SPOT generated a *result list* (`res`) with the information shown in Fig.4.

*Plots.* First of all, the `res` list information will be used to visualize the *route to the solution*: in Fig. 5, loss function values are plotted against the number of iterations. This figure reveals that some hyperparameter configurations should be investigated further, because these configurations result in relatively high loss function values. Using the default hyperparameter configuration results in a loss value of 0.28. The related hyperparameters values are shown in Table 4.

Table 4: Worse configurations

	dropout1	dropout2	units0	units1	lr	epochs	batchSize	rho
red	0.03	0.42	133.00	198.00	0.01	48.00	43.00	0.58
cyan	0.05	0.05	295.00	163.00	0.00	10.00	449.00	0.51

*Box plots.* Secondly, looking at the relationship between interesting hyperparameter configurations from this experiment might be insightful: therefore, Fig. 6 visualizes (using box plots) the ranges of the

<sup>6</sup>Or, as Bartz-Beielstein [2006] wrote: “[SPOT] provides means for understanding algorithms’ performance (we will use datascopes similar to microscopes in biology and telescopes in astronomy).”

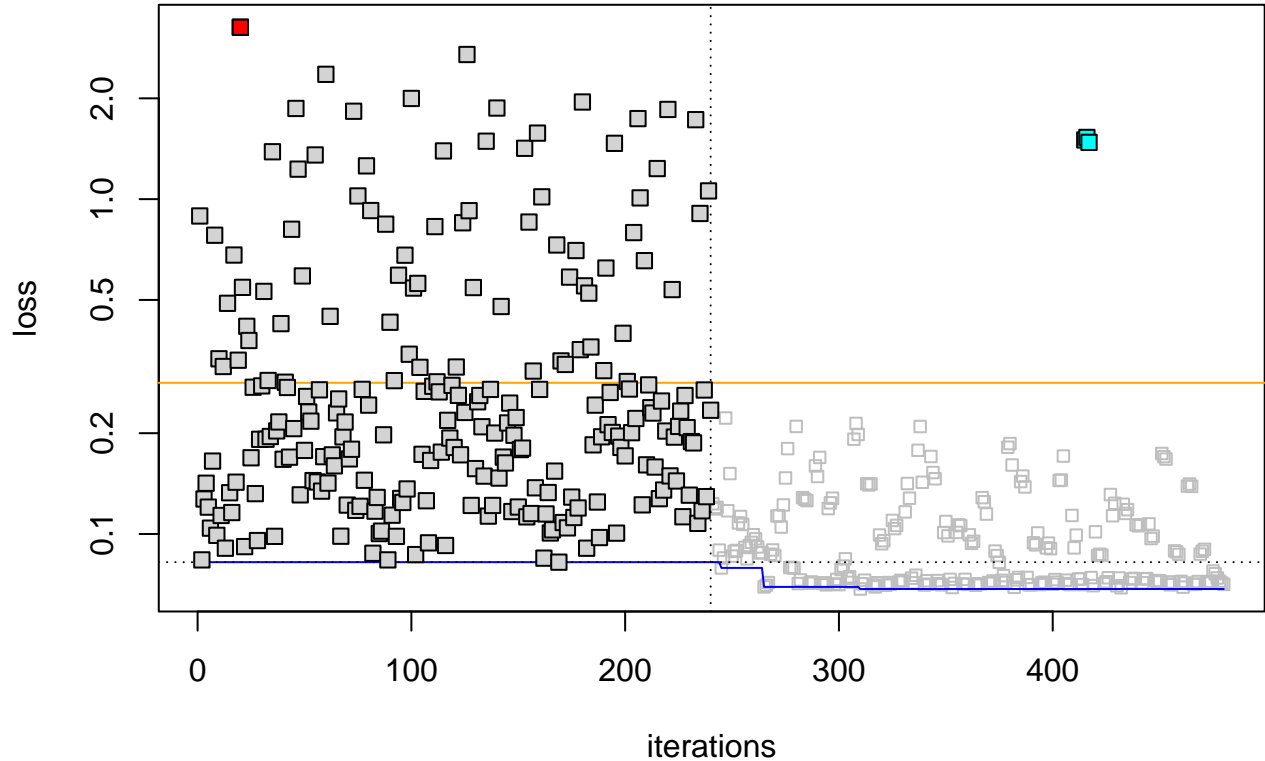


Figure 5: Loss function values plotted against the number of iterations. The **orange** line represents the loss obtained with the default DNN hyperparameters. The dotted **black** line represents the best loss value from the initial design. Initial design points have black boxes. The **blue** line shows the best function value found during the tuning procedure. **Grey** squares represent the values generated during the hyperparameter tuning procedure. The **red** square shows one large value, and **cyan** colored dots indicate worse configurations that occurred during the tuning procedure. These values should be investigated further. Note: loss values plotted on a log scale

eight hyperparameters from the complete HPT experiment. Figure 6 shows information about the best hyperparameter configuration (colored in blue), the worst configuration (red), and the worst configuration from the tuning phase (cyan).

*Regression trees.* Thirdly, to analyze effects and interactions between hyperparameters, a simple regression tree can as shown in Fig. 7 can be used. The regression tree supports the observations, that hyperparameter values for  $x_1$ , i.e., the dropout rate (first layer),  $x_5$ , i.e., the learning rate, and  $x_7$ , i.e., the batch size are relevant. To conclude this first analysis, interactions will be visualized. SPOT provides several tools for the analysis of interactions. Highly recommended is the use of contour plots as shown in Fig. 8.

Figure 8 supports the observations, that hyperparameters  $x_1$  and  $x_5$  have significant effects on the loss function.

```
summary(result$y)
```

```
##          V1
## Min.     :0.06850
## 1st Qu.:0.08563
## Median  :0.12405
## Mean    :0.25914
## 3rd Qu.:0.21495
## Max.    :3.26130
```

*Linear models.* Finally, a simple linear regression model can be fitted to the data. Based on the data from SPOT's `res` list, this can be done as follows:

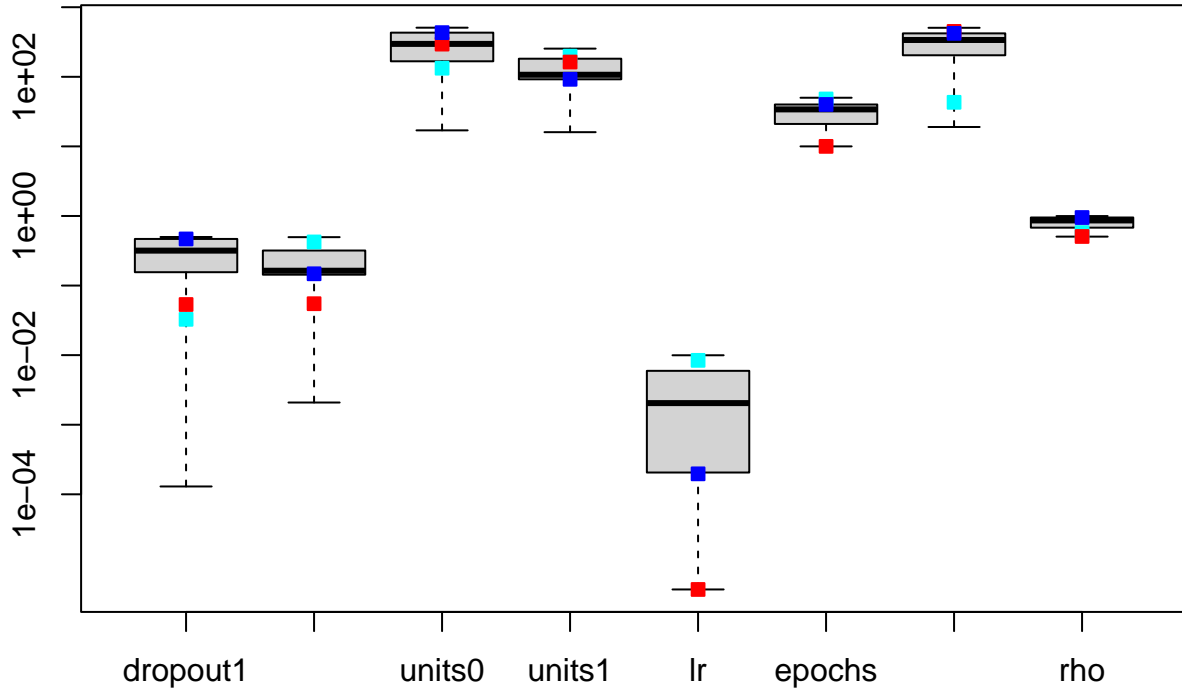


Figure 6: Eight box plots, i.e., each plot represents the values of one parameter (plotted on a log scale). The red square represent the worst value, the blue one show the settings of the best value, and the cyan one show the worst value from the tuning phase.

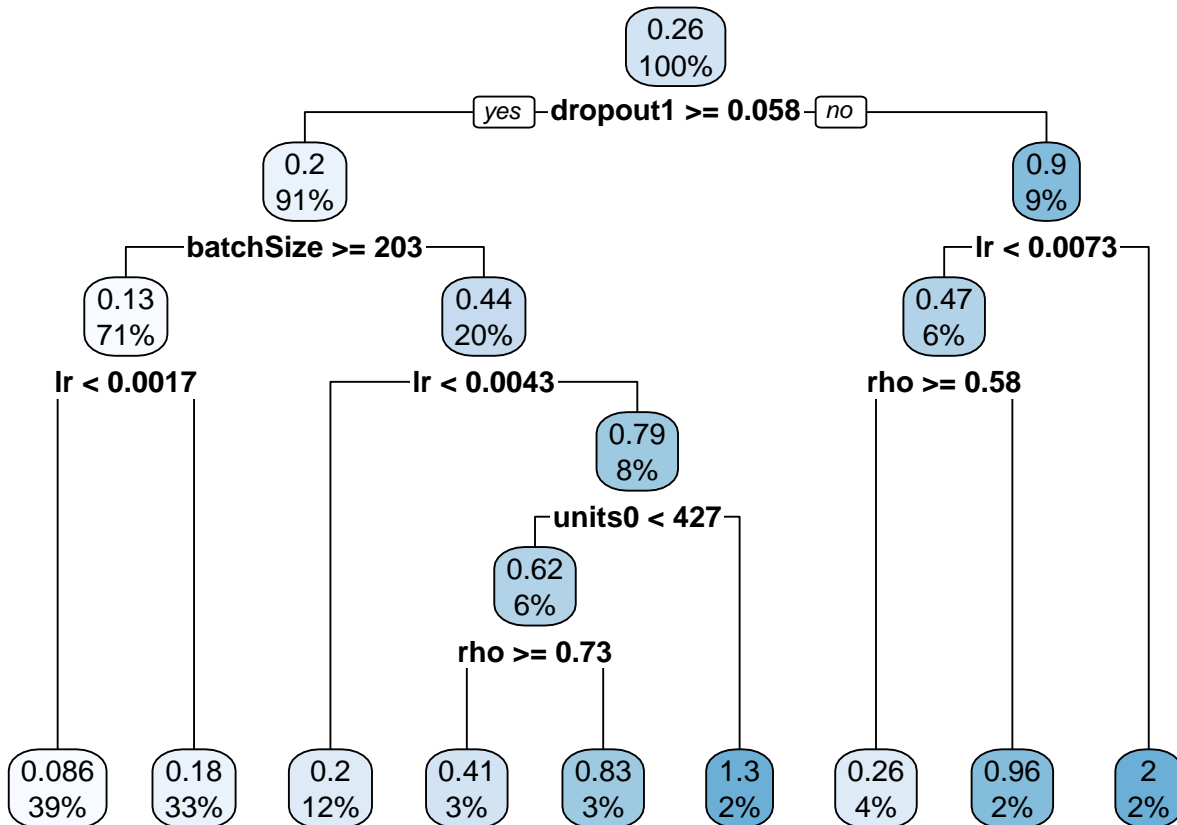
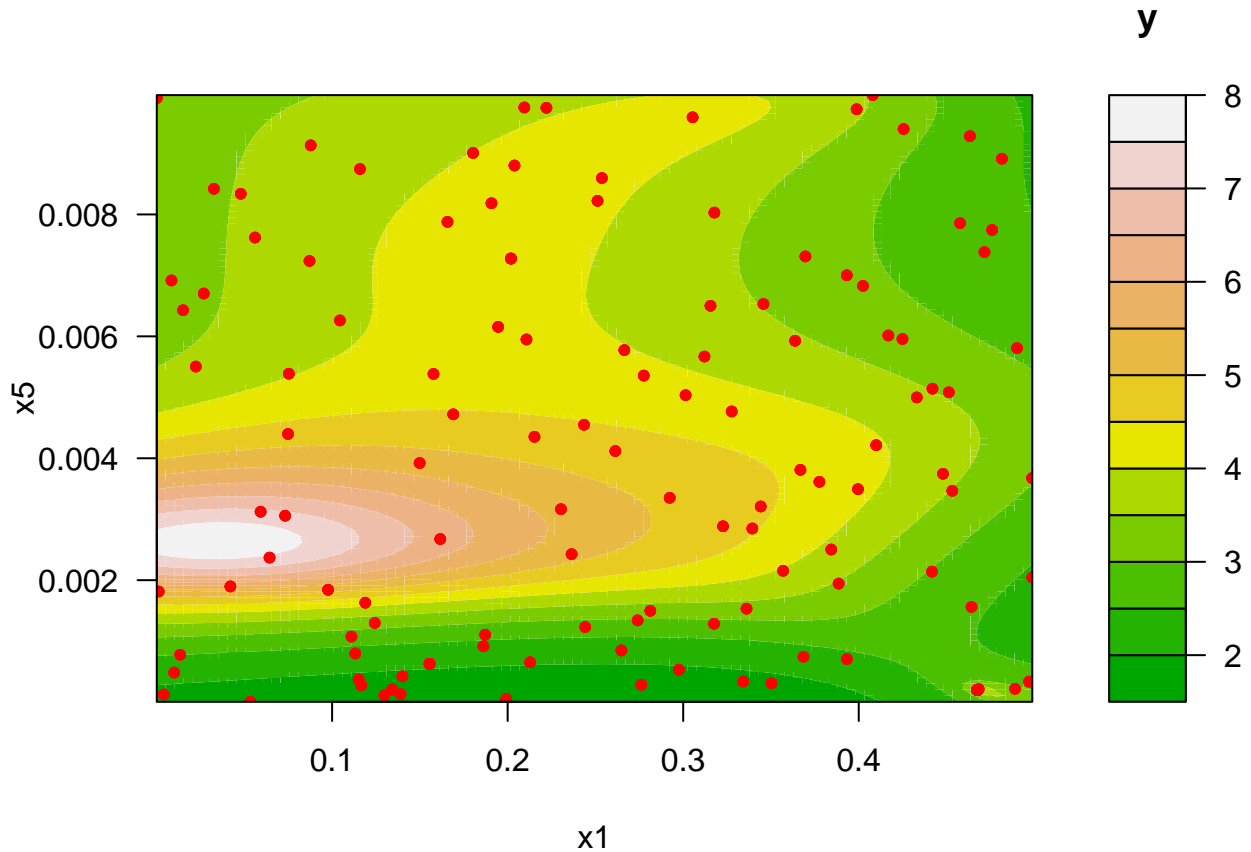


Figure 7: Regression tree based on the first run with 600 evaluations. Apparently, hyperparameter values from  $x_1$ ,  $x_5$  and  $x_7$  are important. This result supports the previous analysis.

Figure 8: Surface plot: learning rate  $x_5$  plotted against dropout1  $x_1$ .

```
lm.res <- lm(res$y ~ res$x)
summary(lm.res)
```

```
##
## Call:
## lm(formula = res$y ~ res$x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56771 -0.14752 -0.01971  0.11414  2.13325
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.687e-01  1.005e-01   6.653 7.97e-11 ***
## res$x1       -4.783e-01  9.330e-02  -5.126 4.32e-07 ***
## res$x2        2.856e-01  1.007e-01   2.837  0.00475 **
## res$x3        4.506e-04  1.034e-04   4.357 1.62e-05 ***
## res$x4       -1.798e-04  2.181e-04  -0.824  0.41010
## res$x5       4.716e+01  4.353e+00  10.832 < 2e-16 ***
## res$x6        7.589e-03  1.224e-03   6.201 1.23e-09 ***
## res$x7       -8.828e-04  9.885e-05  -8.931 < 2e-16 ***
## res$x8       -6.831e-01  9.154e-02  -7.462 4.14e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2763 on 471 degrees of freedom
```

```
## Multiple R-squared:  0.5149, Adjusted R-squared:  0.5066
## F-statistic: 62.49 on 8 and 471 DF,  p-value: < 2.2e-16
```

Although this linear model requires a detailed investigation (a misspecification analysis is necessary) it also is in accordance with previous observations that hyperparameters  $x_1$ ,  $x_5$  and  $x_7$  (and in addition to the previous observations, also  $x_8$ ) have significant effects on the loss function.

## 5 Discussion and Conclusions

This study briefly explains, how HPT can be used as a datascope for the optimization of DNN hyperparameters. The results and observations presented in Sec.4 can be stated as hypotheses, e.g.,

**(H-1):** hyperparameter  $x_1$ , i.e., the dropout rate, has a significant effect on the loss function. Its values should larger than zero.

This hypothesis requires further investigations. The results scratch on the surface of the HPT set of tools, e.g., the role and impact of noise was not considered. SPOT provides very powerful tools such as OCBA to handle noisy function evaluations efficiently [Chen et al. 1997] [Bartz-Beielstein et al. 2011].

Furthermore, there seems to be an upper limit for the values of the loss function: no loss function values are larger than 3.

Considering the research goals stated in Sec. 1, the HPT approach presented in this study provides many tools and solutions. Whereas in ML and optimization, standard workflows are available, e.g., Cross-Industry Standard Process for Data Mining (CRISP-DM) and DOE, the situation in DL is different. It might take some time until a Cross-Industry Standard Process for Deep Learning (CRISP-DL) will be established, because several, fundamental questions are not fully answered today.

In addition to the research goals (R-1) to (R-8) from Sec. 1, important goals that are specific for HPT in DNN should be mentioned. We will discuss problem and algorithm designs separately:

**Problem Design.** The *problem design* comprehends the set of parameters that related to the problem. In HPT and HPO, regression or classification tasks are often considered. In our study, the MNIST data set was chosen.

- Selection of an adequate performance measure: Kedziora et al. [2020] claimed that “research strands into ML performance evaluation remain arguably disorganised, [...]. Typical ML benchmarks focus on minimising both loss functions and processing times, which do not necessarily encapsulate the entirety of human requirement.”
- A sound test problem specification is necessary, i.e., train, validation, and test sets should be clearly specified.
- Initialization (this is similar to the specification of starting points in optimization) procedures should be made transparent.
- Usage of surrogate benchmarks should be considered (this is similar to the use of CFD simulations in optimization)
- Repeats (power of the test, severity), i.e., how many runs are feasible or necessary?
- What are meaningful differences (w.r.t. specification of the loss function or accuracy)?
- Remember: scientific relevance is not identical to statistical significance.
- Floor and ceiling effects should be avoided.
- Comparison to baseline (random search, random sampling, mean value ...) is a must.

**Algorithm Design.** The *algorithm design* in HPT and HPO refers to the model, i.e., DNNs. In our study, the neural network from Sec. 3.2 was chosen.

- A sound algorithm (neural network) specification us required.
- Initialization, pre-training (starting points in optimization). Pre-tuning should be explained.
- Hyperparameter (ranges, types) should be clearly specified.
- Are there any additional (untunable) parameters?
- How is noise (randomness, stochasticity) treated?
- How is reproducibility ensured (and by whom)?

- Last but not least: open source code and open data should be provided.

To conclude: differences between HPT and HPO were discussed. A HPT approach based on SMBO was introduced and exemplified. It combines two packages from the statistical programming environment R: `tfruns` and `SPOT`, hence providing a HPT environment that is fully accessible from R.

Although HPT can be performed with R functions, the underlying TF (Keras) and Python environment has to be installed. This installation is explained in the Appendix.

## 6 Appendix

### 6.1 Software Installations

#### 6.1.1 Installing Python

**Create and Activate a Python Environment in the Project** It is recommended that one Python virtual environment is used per experiment. Navigate into the project directory by using the following command:

```
cd <project-dir>
```

Create a new virtual environment in a folder called `python` within the project directory using the following command:

```
virtualenv python
```

The `virtualenv` can be activated using the following command in a terminal:

```
source python/bin/activate
```

To verify that the correct version of Python was activated the following command can be executed in a terminal:

```
which python
```

**Install Python packages in the Environment** Python packages such as `numpy`, `pandas`, `matplotlib`, and other packages can be installed in the Python `virtualenv` by using `pip install`:

```
pip install numpy pandas matplotlib tensorflow
```

**Install and Configure `reticulate` to use the Correct Python Version** Install the `reticulate` package using the following command in the R console (e.g., from within RStudio):

```
install.packages("reticulate")
```

To configure `reticulate` to point to the Python executable in the `virtualenv python` from above, create a file in the project directory called `.Rprofile` with the following contents:

```
Sys.setenv(RETICULATE_PYTHON = "python/bin/python")
```

R (or RStudio) must be restarted for the setting to take effect. To check that `reticulate` is configured for the correct version of Python the following command can be used in the R (or RStudio) console:

```
reticulate::py_config()
```

#### 6.1.2 Installing Keras

To get started with Keras, the Keras R package, the core Keras library, and a backend tensor engine (such as TF) must be installed. This can be done as follows from within R (or RStudio):

```
install.packages("tensorflow")
install.packages("keras")
library(keras)
install_keras()
```

This will provide you with a default CPU-based installation of Keras and TF. To install a GPU-based version of the TF backend engine, the corresponding command reads as follows:

```
install_keras(tensorflow = "gpu")
```

### 6.1.3 Installing SPOT

The following commands can be used to install the most recent version of SPOT and the additional package Sequential Parameter Optimization Toolbox – Miscellaneous Functions (SPOTMisc) from CRAN:

```
install.packages("SPOT")
install.packages("SPOTMisc")
```

Further information about the most recent SPOT versions will be published on <https://www.spotseven.de/spot/>.

## 6.2 The Hyperparameter Optimization Problem

For convenience, we include the definitions of “learning algorithms” and “hyperparameter optimization” from Bergstra and Bengio [2012]. The objective of a learning algorithm  $A$  is to find a function  $f$  that minimizes some expected loss  $L(x; f)$  over i.i.d. samples  $x$  from a natural (ground truth) distribution  $G_x$ .

**Definition 1** (Learning algorithm; [Bergstra and Bengio 2012]). A *learning algorithm*  $A$  is a functional that maps a data set  $X^{(\text{train})}$  (a finite set of samples from  $G_x$ ) to a function  $f$ .

A learning algorithm can estimate  $f$  through the optimization of a training criterion with respect to a set of parameters  $\theta$ . The learning algorithm itself often has hyperparameters  $\lambda$ , and the actual learning algorithm is the one obtained after choosing  $\lambda$ , which can be denoted  $A_\lambda$ , and  $f = A_\lambda(X^{(\text{train})})$  for a training set  $X^{(\text{train})}$ .

Practitioners are interested in a way to choose  $\lambda$  so as to minimize generalization error, which can be defined as follows.

**Definition 2** (Generalization error; [Bergstra and Bengio 2012]).

$$E_{x \sim G_x} \left[ L(x; A_\lambda(X^{(\text{train})})) \right]. \quad (6)$$

The computation performed by  $A$  itself often involves an inner optimization problem (optimizing the weights of a NN), which is usually iterative and approximate.

The problem of identifying a good value for hyperparameters  $\lambda$  is called the problem of hyperparameter optimization. The outer-loop optimization problem, which is of great practical importance in empirical machine learning work, can be stated as follows:

$$\lambda^{(*)} = \arg \min_{\lambda \in \Lambda} E_{x \sim G_x} \left[ L(x; A_\lambda(X^{(\text{train})})) \right]. \quad (7)$$

**Definition 3** (Hyperparameter optimization problem; [Bergstra and Bengio 2012]). The *hyperparameter optimization problem* can be stated in terms of a hyperparameter response function,  $\psi \in \Psi$  as follows:

$$\lambda^{(*)} \approx \arg \min_{\lambda \in \Lambda} \frac{1}{|X^{(\text{val})}|} \sum_{x \in X^{(\text{val})}} L(x; A_\lambda(X^{(\text{train})})) \quad (8)$$

$$= \arg \min_{\lambda \in \Lambda} \psi(\lambda) \quad (9)$$

$$\approx \arg \min_{\{\lambda^{(i)}\}_{i=1,2,\dots,S}} \psi(\lambda) = \hat{\lambda}. \quad (10)$$

We can define HPO as a minimization problem:

**Definition 4** (Hyperparameter optimization; [Bergstra and Bengio 2012]). Hyperparameter optimization is the minimization of

$$\psi(\lambda) \text{ over } \lambda \in \Lambda. \quad (11)$$

**Definition 5** (Hyperparameter surface). Similar to the definition in DOE, the function  $\psi \in \Psi$  is referred to as the hyperparameter *response surface*.

Different data sets, tasks, and learning algorithm families give rise to different sets  $\Lambda$  and functions  $\Psi$ . Knowing in general very little about  $\Psi$  or the search space  $\Lambda$ , the dominant strategy for finding a good  $\lambda$  is to choose some number ( $S$ ) of trial points  $\{\lambda^{(i)}\}_{i=1,2,\dots,S}$ , to evaluate  $\psi(\lambda)$  for each one, and return the  $\lambda^{(i)}$  that worked the best as  $\tilde{\lambda}$ . This strategy is made explicit by Equation 10.

Whereas  $\lambda$  denotes an arbitrarily chosen hyperparameter configuration, important hyperparameter configurations will be labeled as follows:  $\lambda_i$  is the  $i$ -th hyperparameter configuration,  $\lambda^{(*)}(t)$  is the best hyperparameter configuration at iteration  $t$ , and  $\lambda^{(*)}$  is the final best hyperparameter configuration.

**Definition 6** (Low effective dimension; [Bergstra and Bengio 2012]). If a function  $f$  of two variables could be approximated by another function of one variable ( $f(x_1, x_2) \approx g(x_1)$ ), we could say that  $f$  has a *low effective dimension*.

## Glossary

**ADAM** ADaptive Moment estimation algorithm.

**API** Application Programming Interface.

**AutoDL** Automated Deep Learning.

**AutoHAS** Automated Hyperparameter and Architecture Search.

**AutoML** Automated Machine Learning.

**AutonoML** Autonomous Machine Learning.

**BO** Bayesian Optimization.

**BOHB** Bayesian Optimization HyperBand.

**CAAI** Cognitive Architecture for Artificial Intelligence.

**CART** Classification and Regression Trees.

**CASH** Combined Algorithm Selection and Hyperparameter optimization.

**CCE** Categorical Cross Entropy.

**CIFAR-10** Canadian Institute for Advanced Research, 10 classes.

**CNN** Convolutional Neural Network.

**CRAN** Comprehensive R Archive Network.

**CRISP-DL** Cross-Industry Standard Process for Deep Learning.

**CRISP-DM** Cross-Industry Standard Process for Data Mining.

**CV** Cross Validation.

**DACE** Design and Analysis of Computer Experiments.

**DeepOBS** Deep Learning Optimizer Benchmark Suite.

**DL** Deep Learning.

**DNN** Deep Neural Network.

**DOE** Design of Experiments.

**ES** Evolution Strategy.

**HB** Hyperband.

**HPO** Hyperparameter Optimization.



**HPT** Hyperparameter Tuning.

**MAE** Mean Absolute Error.

**ML** Machine Learning.

**MNIST** Modified National Institute of Standards and Technology.

**MSE** Mean Squared Error.

**NADAM** Nesterov-accelerated Adaptive Moment Estimation.

**NAS** Neural Architecture Search.

**NN** Neural Network.

**OCBA** Optimal Computing Budget Allocation.

**R** R software environment for statistical computing and graphics.

**RGB** Red, Green, and Blue color space.

**RMSProp** Root Mean Square Propagation.

**RNN** Recurrent Neural Network.

**SMAC** Sequential Model-Based Optimization for General Algorithm Configuration.

**SMBO** Surrogate Model Based Optimization.

**SPOT** Sequential Parameter Optimization Toolbox.

**SPOTMisc** Sequential Parameter Optimization Toolbox – Miscellaneous Functions.

**TF** TensorFlow.

**tfruns** Training Run Tools for TensorFlow.

## References

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv e-prints*, art. arXiv:1603.04467, Mar. 2016.
- A. Balaji and A. Allen. Benchmarking Automatic Machine Learning Frameworks, 2018. URL <http://arxiv.org/abs/1808.06492>.
- P. Balaprakash, M. Salim, T. D. Uram, V. Vishwanath, and S. M. Wild. DeepHyper: Asynchronous hyperparameter search for deep neural networks. In *25th IEEE International Conference on High Performance Computing (HiPC18)*. IEEE, 2018. doi: 10.1109/hipc.2018.00014.
- T. Bartz-Beielstein. *Experimental Research in Evolutionary Computation—The New Experimentalism*. Natural Computing Series. Springer, Berlin, Heidelberg, New York, 2006. ISBN 3-540-32026-1. doi: 10.1007/3-540-32027-X. URL <http://dx.doi.org/10.1007/3-540-32027-X>.
- T. Bartz-Beielstein and S. Markon. Tuning Search Algorithms for Real-World Applications: A Regression Tree Based Approach. In G. W. Greenwood, editor, *Proceedings 2004 Congress on Evolutionary Computation (CEC’04)*, Portland OR, pages 1111–1118, Piscataway NJ, 2004. IEEE. ISBN 0-7803-8515-2. doi: 10.1109/CEC.2004.1330986. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1330986>.
- T. Bartz-Beielstein and M. Zaefferer. Model-based methods for continuous and discrete global optimization. *Applied Soft Computing*, 55:154 – 167, 2017. ISSN 1568-4946. doi: 10.1016/j.asoc.2017.01.039. URL <http://www.sciencedirect.com/science/article/pii/S1568494617300546>.
- T. Bartz-Beielstein, C. Lasarczyk, and M. Preuss. Sequential Parameter Optimization. In B. McKay et al., editors, *Proceedings 2005 Congress on Evolutionary Computation (CEC’05)*, Edinburgh, Scotland, pages 773–780, Piscataway NJ, 2005. IEEE Press. ISBN 0-7803-9363-5. doi: 10.1109/CEC.2005.1554761.

- T. Bartz-Beielstein, M. Friese, M. Zaefferer, B. Naujoks, O. Flasch, W. Konen, and P. Koch. Noisy optimization with sequential parameter optimization and optimal computational budget allocation. In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*, pages 119–120, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0690-4. doi: 10.1145/2001858.2001926.
- T. Bartz-Beielstein, J. Branke, J. Mehnen, and O. Mersmann. Evolutionary algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(3):178–195, 2014. ISSN 1942-4795. doi: 10.1002/widm.1124. URL <http://dx.doi.org/10.1002/widm.1124>.
- T. Bartz-Beielstein, C. Doerr, J. Bossek, S. Chandrasekaran, T. Eftimov, A. Fischbach, P. Kerschke, M. Lopez-Ibanez, K. M. Malan, J. H. Moore, B. Naujoks, P. Orzechowski, V. Volz, M. Wagner, and T. Weise. Benchmarking in optimization: Best practice and open issues. arXiv, 07 2020. URL <https://arxiv.org/abs/2007.03488>.
- T. Bartz-Beielstein, M. Zaefferer, and F. Rehbach. In a Nutshell – The Sequential Parameter Optimization Toolbox. *arXiv e-prints*, art. arXiv:1712.04076, Dec. 2021.
- J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13: 281–305, 2012. ISSN 1532-4435; 1533-7928/e.
- J. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML’13*, pages I–115–I–123. JMLR.org, 2013.
- A. J. Booker, J. E. Dennis Jr, P. D. Frank, D. B. Serafini, V. Torczon, and M. W. Trosset. A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization*, 17(1):1–13, 1999. doi: 10.1007/BF01197708. URL <http://link.springer.com/10.1007/BF01197708>.
- R. E. Caflisch, W. Morokoff, and A. Owen. Valuation of mortgage backed securities using brownian bridges to reduce effective dimension, 1997.
- H. C. Chen, C. H. Chen, L. Dai, and E. Yücesan. New Development of Optimal Computing Budget Allocation for Discrete Event Simulation. In S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, editors, *Proceedings of the 1997 Winter Simulation Conference*, pages 334–341, Piscataway NJ, 1997. IEEE Computer Society.
- D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl. On Empirical Comparisons of Optimizers for Deep Learning, 2019. URL <http://arxiv.org/abs/1910.05446>.
- F. Chollet and J. Allaire. *Deep Learning with R*. Manning, 2018.
- X. Dong, M. Tan, A. W. Yu, D. Peng, B. Gabrys, and Q. V. Le. AutoHAS: Efficient Hyperparameter and Architecture Search, 2020. URL <http://arxiv.org/abs/2006.03656>.
- T. Dozat. Incorporating nesterov momentum into adam. 2016.
- K. Eggenberger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. H. Hoos, and K. Leyton-brown. Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In *In NIPS Workshop on Bayesian Optimization in Theory and Practice*, 2013.
- N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data, 2020. URL <http://arxiv.org/abs/2003.06505>.
- S. Falkner, A. Klein, and F. Hutter. BOHB: Robust and Efficient Hyperparameter Optimization at Scale, 2018. URL <http://arxiv.org/abs/1807.01774>.
- R. T. Haftka. Requirements for papers focusing on new or improved global optimization algorithms. *Structural and Multidisciplinary Optimization*, 54(1):1–1, 2016. ISSN 1615-1488. doi: 10.1007/s00158-016-1491-5. URL <http://dx.doi.org/10.1007/s00158-016-1491-5>.
- N. Hansen. The CMA evolution strategy: a comparing review. In J. A. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2017. doi: 10.1007/978-0-387-84858-7. 12th printing.
- F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In C. A. C. Coello, editor, *Learning and Intelligent Optimization*, pages 507–523, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25566-3.

- H. Jin, Q. Song, and X. Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '19*, pages 1946–1956, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330648. URL <https://doi.org/10.1145/3292500.3330648>.
- I. Karmanov, M. Salvaris, M. Fierro, and D. Dean. Comparing deep learning frameworks: A rosetta stone approach. Microsoft docs, Microsoft, 2018. Last accessed 29-May-2021. <https://docs.microsoft.com/de-de/archive/blogs/machinelearning/comparing-deep-learning-frameworks-a-rosetta-stone-approach>.
- D. J. Kedziora, K. Musial, and B. Gabrys. AutoML: Towards an Integrated Framework for Autonomous Machine Learning, 2020. URL <http://arxiv.org/abs/2012.12600>.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- R. M. Lewis, V. Torczon, and M. W. Trosset. Direct search methods: Then and now. *Journal of Computational and Applied Mathematics*, 124(1–2):191–207, 2000.
- L. Li and A. Talwalkar. Random Search and Reproducibility for Neural Architecture Search. Conference on Uncertainty in Artificial Intelligence (UAI), 2019, 2019. URL <http://arxiv.org/abs/1902.07638>.
- L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *arXiv e-prints*, art. arXiv:1603.06560, Mar. 2016.
- B. Liu. A Very Brief and Critical Discussion on AutoML, 2018. URL <http://arxiv.org/abs/1811.03822>.
- Y. Mack, T. Goel, W. Shyy, and R. Haftka. *Surrogate Model-Based Optimization Framework: A Case Study in Aerospace Design*, pages 323–342. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-49774-5. doi: 10.1007/978-3-540-49774-5\_14. URL [https://doi.org/10.1007/978-3-540-49774-5\\_14](https://doi.org/10.1007/978-3-540-49774-5_14).
- H. Mazzawi and X. Gonzalvo. Introducing model search: An open source platform for finding optimal ML models. <https://ai.googleblog.com/2021/02/introducing-model-search-open-source.html?m=1>, 2 2021.
- H. Mazzawi, X. Gonzalvo, A. Kracun, P. Sridhar, N. A. Subrahmanya, I. Lopez-Moreno, H. Park, and P. Violette. Improving keyword spotting and language identification via neural architecture search at scale. In *INTERSPEECH*, 2019.
- D. Meignan, S. Knust, J.-M. Frayet, G. Pesant, and N. Gaud. A Review and Taxonomy of Interactive Optimization Methods in Operations Research. *ACM Transactions on Interactive Intelligent Systems*, Sept. 2015.
- H. Mendoza, A. Klein, M. Feurer, J. T. Springenberg, M. Urban, M. Burkart, M. Dippel, M. Lindauer, and F. Hutter. *Towards Automatically-Tuned Deep Neural Networks*, pages 135–149. Springer International Publishing, Cham, 2019. ISBN 978-3-030-05318-5. doi: 10.1007/978-3-030-05318-5\_7. URL [https://doi.org/10.1007/978-3-030-05318-5\\_7](https://doi.org/10.1007/978-3-030-05318-5_7).
- C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, VIII(4):283–298, October 1978.
- T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, et al. Keras Tuner, 2019. <https://github.com/keras-team/keras-tuner>.
- R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984. doi: 10.1080/01621459.1984.10478083. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10478083>.
- P. Probst, A.-L. Boulesteix, and B. Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019.
- D. A. Roberts, S. Yaida, and B. Hanin. The Principles of Deep Learning Theory, 2021. URL <http://arxiv.org/abs/2106.10165>.
- R. M. Schmidt, F. Schneider, and P. Hennig. Descending through a Crowded Valley – Benchmarking Deep Learning Optimizers, 2020. URL <http://arxiv.org/abs/2007.01547>.
- F. Schneider, L. Balles, and P. Hennig. DeepOBS: A Deep Learning Optimizer Benchmark Suite, 2019. URL <http://arxiv.org/abs/1903.05499>.
- C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the Effects of Data Parallelism on Neural Network Training. *Journal of Machine Learning Research* 20 (2019) 1–49, 2019. URL <http://arxiv.org/abs/1811.03600>.

- J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv e-prints*, art. arXiv:1206.2944, June 2012.
- J. Strohschein, A. Fischbach, A. Bunte, H. Faeskorn-Woyke, N. Moriz, and T. Bartz-Beielstein. Cognitive capabilities for the caai in cyber-physical production systems. *The International Journal of Advanced Manufacturing Technology*, 2021. doi: 10.1007/s00170-021-07248-3. URL <https://doi.org/10.1007/s00170-021-07248-3>.
- V. Tatsis and K. Parsopoulos. Grid search for operator and parameter control in differential evolution. In *9th Hellenic Conference on Artificial Intelligence (SETN 2016)*, 2016.
- C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13*, pages 847–855, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2487629. URL <https://doi.org/10.1145/2487575.2487629>.
- T. Tieleman and G. Hinton. Lecture 6.5-RMSProp: Divide the gradient by running average of its recent magnitude. 4(2):26–31, 2012.
- J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *arXiv e-prints*, art. arXiv:1407.7722, July 2014.
- A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/81b3833e2504647f9d794f7d7b9bf341-Paper.pdf>.
- M. Wistuba, A. Rawat, and T. Pedapati. A Survey on Neural Architecture Search, 2019. URL <http://arxiv.org/abs/1905.01392>.
- D. H. Wolpert and W. G. Macready. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, Apr. 1997.
- K. Yu, C. Sciuto, M. Jaggi, C. Musat, and M. Salzmann. Evaluating the search phase of neural architecture search. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=H1loF2NFwr>.
- M. Zaefferer and T. Bartz-Beielstein. Efficient global optimization with indefinite kernels. In J. Handl, E. Hart, P. R. Lewis, M. López-Ibáñez, G. Ochoa, and B. Paechter, editors, *Parallel Problem Solving from Nature – PPSN XIV: 14th International Conference, Edinburgh, UK, September 17-21, 2016, Proceedings*, pages 69–79. Springer International Publishing, Cham, 2016. ISBN 978-3-319-45823-6. doi: 10.1007/978-3-319-45823-6\_7. URL [http://dx.doi.org/10.1007/978-3-319-45823-6\\_7](http://dx.doi.org/10.1007/978-3-319-45823-6_7).
- G. Zhang, L. Li, Z. Nado, J. Martens, S. Sachdeva, G. E. Dahl, C. J. Shallue, and R. Grosse. Which Algorithmic Choices Matter at Which Batch Sizes? Insights From a Noisy Quadratic Model, 2019. URL <http://arxiv.org/abs/1907.04164>.
- L. Zimmer, M. Lindauer, and F. Hutter. Auto-PyTorch Tabular: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL, 2020. URL <http://arxiv.org/abs/2006.13799>.
- B. Zoph and Q. V. Le. Neural Architecture Search with Reinforcement Learning, 2016. URL <http://arxiv.org/abs/1611.01578>.
- B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning Transferable Architectures for Scalable Image Recognition, 2017. URL <http://arxiv.org/abs/1707.07012>.