

A Survey on Neural Architecture Search

Martin Wistuba

MARTIN.WISTUBA@IBM.COM

IBM Research AI

IBM Technology Campus, Damastown Ind. Park

Dublin, D15HN66, Ireland

Ambrish Rawat

AMBRISH.RAWAT@IE.IBM.COM

IBM Research AI

IBM Technology Campus, Damastown Ind. Park

Dublin, D15HN66, Ireland

Tejaswini Pedapati

TEJASWINIP@US.IBM.COM

IBM Research AI

IBM T.J. Watson Research Center

Yorktown Heights, NY 10598, USA

Abstract

The growing interest in both the automation of machine learning and deep learning has inevitably led to the development of a wide variety of automated methods for neural architecture search. The choice of the network architecture has proven to be critical, and many advances in deep learning spring from its immediate improvements. However, deep learning techniques are computationally intensive and their application requires a high level of domain knowledge. Therefore, even partial automation of this process helps to make deep learning more accessible to both researchers and practitioners. With this survey, we provide a formalism which unifies and categorizes the landscape of existing methods along with a detailed analysis that compares and contrasts the different approaches. We achieve this via a comprehensive discussion of the commonly adopted architecture search spaces and architecture optimization algorithms based on principles of reinforcement learning and evolutionary algorithms along with approaches that incorporate surrogate and one-shot models. Additionally, we address the new research directions which include constrained and multi-objective architecture search as well as automated data augmentation, optimizer and activation function search.

Keywords: Neural Architecture Search, Automation of Machine Learning, Deep Learning, Reinforcement Learning, Evolutionary Algorithms, Constrained Optimization, Multi-Objective Optimization

1. Introduction

Deep learning methods are very successful in solving tasks in machine translation, image and speech recognition. This success is often attributed to their ability to automatically extract features from unstructured data such as audio, image and text. We are currently witnessing this paradigm shift from the laborious job of manual feature engineering for unstructured data to engineering network components and architectures for deep learning methods. While architecture modifications do result in significant gains in the performance of deep learning methods, the search for suitable architectures is in itself a time-consuming, arduous and error-prone task. Within the last two years there has been an insurgence

in research efforts by the machine learning community that seeks to automate this search process. On a high level, this automation is cast as a search problem over a set of decisions that define the different components of a neural network. The set of feasible solutions to these decisions implicitly defines the search space and the search algorithm is defined by the optimizer. Arguably, the works by Zoph and Le (2017) and Baker et al. (2017) mark the beginning of these efforts where their works demonstrated that good architectures can be discovered with the use of reinforcement learning algorithms. Shortly thereafter, Real et al. (2017) showed that similar results could also be achieved by the hitherto well studied approaches in neuroevolution (Floreano et al., 2008). However, both these search approaches consumed hundreds of GPU hours in their respective computations. Consequently, many of the subsequent works focused on methods that reduce this computational burden. The successful algorithms along this line of research leverage from the principle of reusing the learned model parameters, with the works of Cai et al. (2018a) and Pham et al. (2018) being the notable mentions.

The design of the search space forms a key component of neural architecture search. In addition to speeding up the search process, this influences the duration of the search and the quality of the solution. In the earlier works on neural architecture search, the spaces were designed to primarily search for chain-structured architectures. However, with branched handcrafted architectures surpassing the classical networks in terms of performance, appropriate search spaces were proposed shortly after the initial publications (Zoph et al., 2018) and these have since become a norm in this field.

In parallel to these developments, researchers have broadened the horizons of neural architecture search to incorporate objectives that go beyond reducing the search time and generalization error of the found architectures. Methods that simultaneously handle multiple objective functions have become relevant. Notable works include methods that attempt to limit the number of model parameters or the like, for efficient deployment on mobile devices (Tan et al., 2018; Kim et al., 2017). Furthermore, the developed techniques for architecture search have been extended for advanced automation of other related components of deep learning. For instance, the search for activation functions (Ramachandran et al., 2018) or suitable data augmentation (Cubuk et al., 2018a).

Currently, the automation of deep learning in the form of neural architecture search is one of the fastest developing areas of machine learning. With new papers emerging on [arXiv.org](https://arxiv.org) each week and major conferences publishing a handful of interesting work, it is easy to lose track. With this survey, we provide a formalism which unifies the landscape of existing methods. This formalism allows us to critically examine the different approaches and understand the benefits of different components that contribute to the design and success of neural architecture search. Along the way, we also highlight some popular misconceptions pitfalls in the current trends of architecture search. We supplement our criticism with suitable experiments.

Our review is divided into several sections. In Section 2, we discuss various architecture search spaces that have been proposed over time. We use Section 3 to formally define the problem of architecture search. Then we identify four typical types of optimization methods: reinforcement learning, evolutionary algorithms, surrogate model-based optimization, and one-shot architecture search. We define these optimization procedures and associate them to existing work and discuss it. Section 6 highlights the architecture search, consider-

ing constraints, multiple objective functions and model compression techniques. Alternate approaches that are motivated from the use of transfer learning are discussed in Section 5. Similarly, the class of methods that use early termination to fasten the search process are detailed in Section 4. Finally, the influence of search procedures on related areas is discussed in Section 7. The discussion is supported with extensive illustrations to elucidate the different methods under a common formalism and relevant experimentation that examines the different aspects of neural architecture search methods.

2. Neural Architecture Search Space

From a computational standpoint, neural networks represent a function that transforms input variables \mathbf{x} to output variables $\hat{\mathbf{y}}$ through a series of operations. This can be formalized in the language of computational graphs (Bauer, 1974) where neural networks are represented as directed acyclic graphs with a set of nodes Z . Each node $\mathbf{z}^{(k)}$ represents a tensor and is associated with an operation $o^{(k)} \in O$ on its set of parent nodes $I^{(k)}$ (Goodfellow et al., 2016). The only exception is the input node \mathbf{x} which has neither a set of parent nodes nor an operation associated to it and is only considered as an input to other nodes. The computation at a node k amounts to

$$\mathbf{z}^{(k)} = o^{(k)}(I^{(k)}). \quad (1)$$

The set of operations includes unary operations such as convolutions, pooling, activation functions or multivariate operations such as concatenation or addition. For notational convenience, we adopt a convention wherein we fix the unary operations to use concatenation as an implicit merge operation when acting on a set of multiple inputs and often omit it in the representation. Note that for a neural network, any representation that specifies the set of parents and the operation for each node completely defines the architecture of the network. We refer to such a representation as α and use this to unify and outline the different neural architecture search methods under the same framework.

A *neural architecture search space* is a subspace of this general definition of neural architectures. Its space of operations can be limited and certain constraints may be imposed on the architectures. In the rest of this survey, we use search space to refer to the set of feasible solutions of a neural architecture search method. Most works on neural architecture search address the automated search of architectures for the task of image recognition and are consequently concerned with convolutional neural networks (CNNs) (LeCun et al., 1998). The corresponding search spaces in these works can be broadly classified into two categories. The first category of search spaces is defined for the graphs that represent an entire neural architecture which we refer to as the global search space and discuss in Section 2.1. We discuss the second category in Section 2.2 which assumes that an architecture is a combination of few cells which are repeated to build the complete network. Although in principle the global search space encompasses the cell-based search space, we make this distinction to highlight some of the nuanced aspects of both these spaces. An objective comparison of these search spaces is provided in Section 2.3. Another common task tackled with neural architecture search is language modeling where search methods seek to find the architecture for recurrent neural networks (RNNs). We discuss the commonly used search space in this context in Section 2.4.

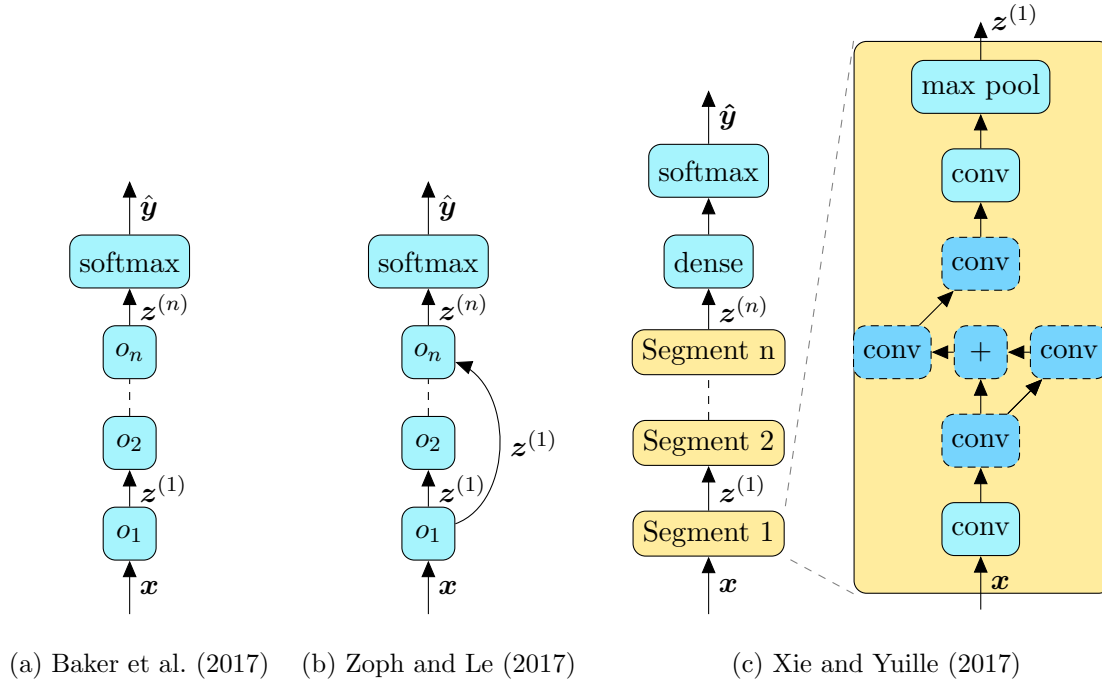


Figure 1: Global search spaces: (a) chain-structured, (b) with skips, (c) architecture template, only the connections between the dark blue (dashed) operations are not fixed.

2.1 Global Search Space

Instances belonging to the global search space admit large degrees of freedom regarding the arrangement of operations. An *architecture template* may be assumed which limits the freedom of admissible structural choices within an architecture definition. This template often fixes certain aspects of the network graph. For instance, it may divide the architecture graph into several segments or enforce specific constraints on operations and connections both within and across these segments, thereby limiting the type of architectures that belong to a search space. Figure 1 illustrates examples of architectures from such template-constrained search spaces. Here, we exclude constraints which enforce a predetermined repeating pattern of subgraphs and dedicate a separate section for the discussion of such search spaces in Section 2.2.

The simplest example of a global search space is the *chain-structured search space* as shown in Figure 1a. In its most primitive form, this search space consists of architectures that can be represented by an arbitrary sequence of ordered nodes such that for any node $z^{(k)}$, $z^{(k-1)}$ is its only parent. The data flow for neural networks with such architectures simplifies to

$$z^{(k)} = o^{(k)} \left(\left\{ z^{(k-1)} \right\} \right). \quad (2)$$

Baker et al. (2017) explore this search space. They consider a set of operations that includes convolutions, pooling, linear transformations (dense layers) with activation, and global average pooling with different hyperparameter settings such as number of filters, kernel size, stride and pooling size. Furthermore, they consider additional constraints so as to ex-

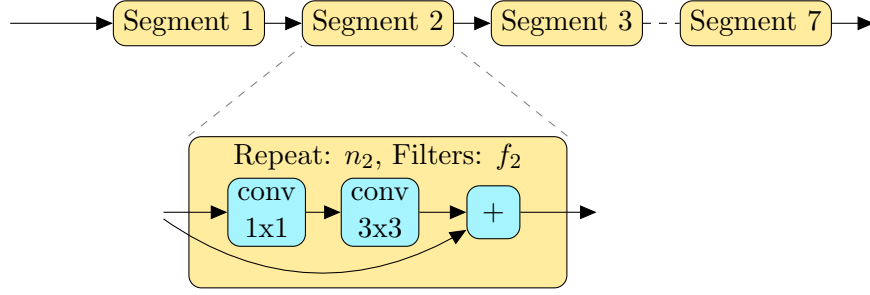


Figure 2: Tan et al. (2018) propose to decompose the architecture into different segments. Each segment i has its own pattern (blue operations) which is repeated n_i times and has f_i filters.

clude certain architectures that correspond to patently poor or computationally expensive neural network models. For instance, architectures with pooling as the first operation do not belong to their search space. Furthermore, architectures with dense (fully connected) layers applied as high-resolution feature maps, or as feature transformations before other operations like convolutions are excluded from their search space.

In a parallel work around the same time, Zoph and Le (2017) define a relaxed version of the chain-structured search space. By permitting arbitrary skip connections to exist between the ordered nodes of a chain-structured architecture, members belonging to this search space exhibit a wider variety of designs. However, in their definition the operation set is restricted to use only convolutions with different hyperparameter settings. While they adhere to the basic recursive formulation of the chain-structured search space, they extend it with the inclusion of skip connections. For an architecture in this search space, the set of parents of a node $\mathbf{z}^{(k)}$ necessarily contains $\mathbf{z}^{(k-1)}$ with possible inclusion of other ancestor nodes,

$$\mathbf{z}^{(k)} = o^{(k)} \left(\left\{ \mathbf{z}^{(k-1)} \right\} \cup \left\{ \mathbf{z}^{(i)} \mid \alpha_{i,k} = 1, i < k - 1 \right\} \right). \quad (3)$$

For such nodes, the merge operation is fixed as concatenation. Figure 1b provides an example of such a skip connection. As mentioned earlier, we omit the explicit representation of the concatenation.

Xie and Yuille (2017) consider a similar search space which uses summation instead of concatenation as the merging operation. Furthermore, the architectures in this search space are no longer sequential in the sense that the set of parents for a node $\mathbf{z}^{(k)}$ does not necessarily contain the node $\mathbf{z}^{(k-1)}$. Moreover, this search space incorporates a template which separates architectures into sequentially connected segments (e.g. three segments are used for image recognition on the CIFAR-10 dataset (Krizhevsky, 2009)). Each segment is parameterized by a set of nodes with convolutions as their operation. As part of the template, the segments begin with a convolution and conclude with a max pooling operation with a stride of two to reduce feature dimensions (see Figure 1c).

Additionally, the maximum number of convolution operations along with their number of filters for each segment is also fixed as part of the template (for the case of CIFAR-10 this is fixed to three convolutions of filter size 64 for the first, four convolutions of filter size 128 for the second, and five convolutions of filter size 256 for the last segment). With the

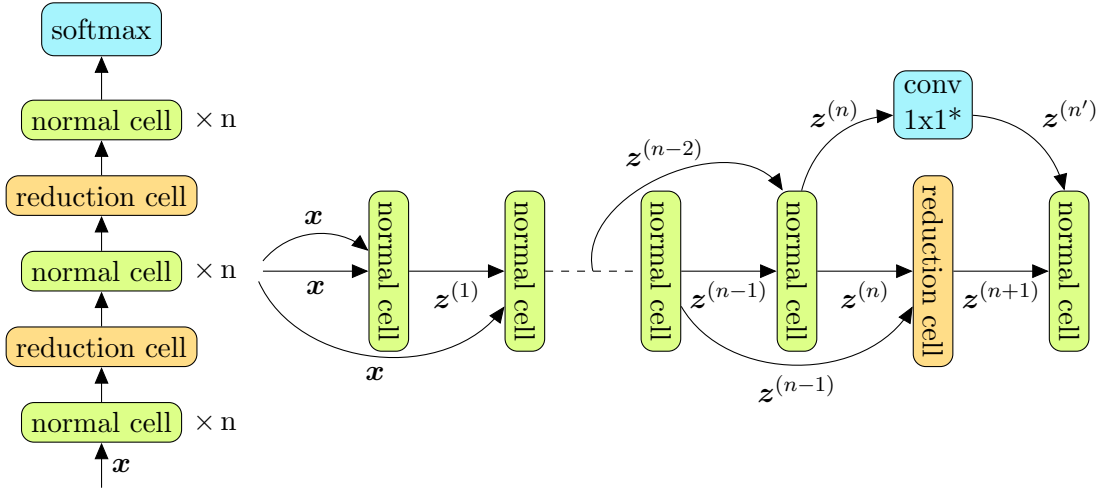


Figure 3: Structure of the NASNet search space instances. n normal cells followed by a reduction cell. This sequence is repeated several times, the reduction cell might be repeated. This decision is a hyperparameter and depends on the image resolution. The $1 \times 1^*$ convolution is a special operation which converts $z^{(n)}$ to match the shape of $z^{(n+1)}$.

operations fixed, an adjacency matrix defines the connections in the directed acyclic graphs corresponding to each segment.

An alternative work by Tan et al. (2018) is motivated to look for neural network models for mobile devices that perform efficiently on multiple fronts which include accuracy, inference time and number of parameters. They design a suitable search space for this purpose that consists of architectures with a hierarchical representation (Figure 2). Architectures in this search space are also formed by sequentially connecting segments. Segments are further structured to have repeating patterns of operations and each segment is parameterized by the choice of operations and the number of repetitions of the patterns. The structure of the patterns itself is simplified to be sequentially connected operations with an optional skip connection and an optional choice of stride. Although they simplify the per-segment search space, they allow segments to be different which differentiates this from the cell-based search space discussed in the following section (Section 2.2). They argue that this flexibility provides them the necessary apparatus to tackle multi-objective designs.

2.2 Cell-Based Search Space

A cell-based search space builds upon the observation that many effective handcrafted architectures are designed with repetitions of fixed structures. Such architectures often consist of smaller-sized graphs that are stacked to form a larger architecture. Across the literature, these repeating structures have been interchangeably referred to as cells, blocks or units. In the context of this survey, we refer to them as *cells*. Not only is this design known to yield high performing models, it also enables easy generalization across datasets and tasks as these units can be flexibly stacked to build larger or smaller networks. Zoph et al. (2018) were one of the first to explore such a search space leading to the popular

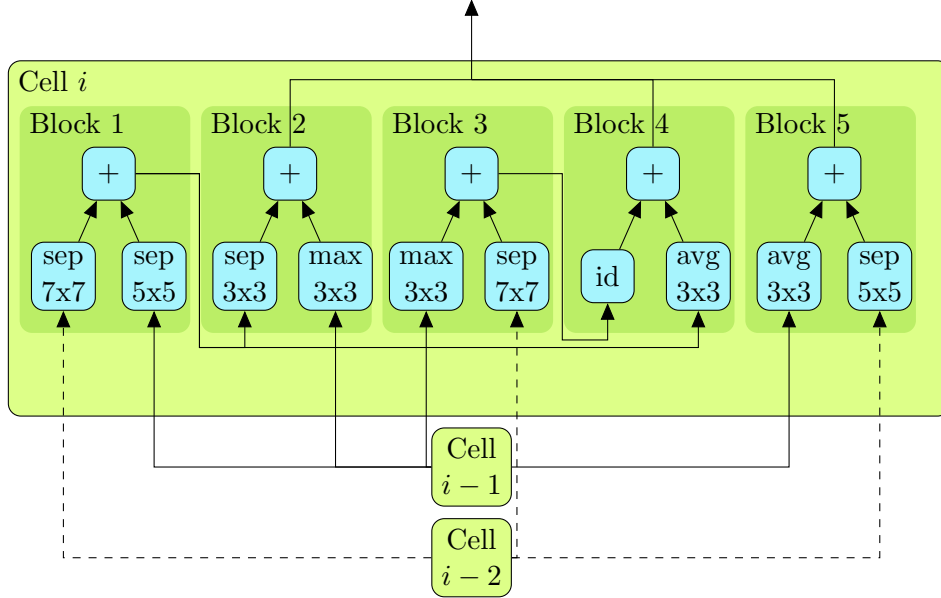


Figure 4: Reduction cell of the NASNet-A architecture (Zoph et al., 2018) as one example how a cell in the NASNet search space can look like. Blocks can be used as input for other blocks (e.g. block 1 and 3), unused blocks are concatenated and are the output of the cell.

architecture called NASNet. Post the development of this work, other cell-based search spaces have been proposed. Most of these broadly stick to the structure proposed by Zoph et al. (2018) with small modifications to the choice of operations and cell-combination strategies. In the remaining section we define and discuss these different approaches to cell-based search spaces.

In the cell-based search space a network is constructed by repeating a structure called a cell in a prespecified arrangement as determined by a template (see Figure 3). A cell is often a small directed acyclic graph representing a feature transformation. While the topology of the cell is maintained across the network, its hyperparameters are often varied. For instance the number of filters for the convolution operations can differ across different cells and a template usually specifies the fixed policy for varying these across the network structures. One aspect of cell-based search spaces that demands attention is their handling of tensor dimensions when building the entire network. While some approaches have dedicated cells to take care of dimension reduction like the reduction cells in the work by Zoph et al. (2018), others achieve this by introducing pooling operations between different units in the network (Zhong et al., 2018). The prespecified template for connecting the various cells localizes the search to structures within a cell. Often the template also includes a set of initial convolution layers which capture low-level features for the given dataset. It is also worth remarking that most approaches that seek to learn a topology limit the hyperparameter choice to smaller values during the search process thereby making the search process efficient. The final proposed network is trained with a larger number of filters (f) and more repetitions of the cells (n) with an aim to achieve better performance (Zoph et al., 2018).

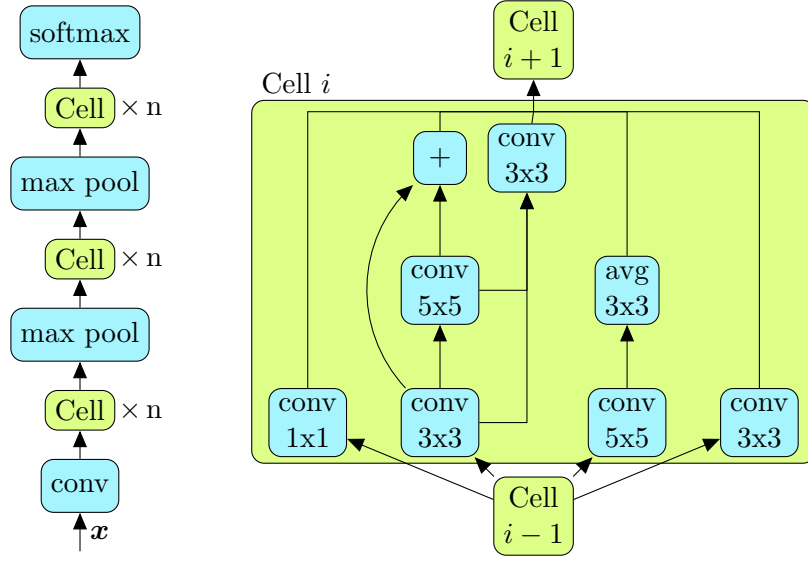


Figure 5: Architecture template (left) and the Block-QNN-B cell discovered by Zhong et al. (2018) (right).

Zoph et al. (2018) are one of the first to propose a cell-based approach, popularly referred to as the *NASNet search space*. The architectures in this search space consider a template of cells as visualized in Figure 3. This search space distinguishes between two cell types, namely normal cells and reduction cells, which handle the feature dimensions across the architecture. Operations in a normal cell have a stride of one and do not change the dimensions of the feature maps, whereas the first set of operations in a reduction cell have a stride of two and it consequently halve the spatial dimensions of the feature maps. Each cell consists of b blocks, where b is a search space hyperparameter (with $b = 5$ being the most popular choice). Each block is completely defined by its two inputs and the corresponding operations. The computation flow for a block is defined by

$$z^{(\text{block}_i)} = o^{(i_1)} \left(\left\{ z^{(i_1)} \right\} \right) + o^{(i_2)} \left(\left\{ z^{(i_2)} \right\} \right). \quad (4)$$

An example for the cell structure is visualized in Figure 4. In their original formulation, Zoph et al. (2018) also consider concatenation as a possible merge operation of $o^{(i_1)}$ and $o^{(i_2)}$. However in their experiments they note that architectures with summation operation dominate those with concatenation in terms of accuracy. Therefore, later works such as Liu et al. (2018a) fix this decision to summation. For a block, the set of possible inputs include the output of one of the previous two cells and the output of a previously defined block within a cell. By including the output from previous two cells as candidates for the input to the block, this search space includes instances with skip connections across cells. The output of a cell is determined as concatenation of the outputs of all blocks within the cell which do not serve as inputs to any other block. Cells from this search space differ in their choice of the inputs and the operations for the different blocks.

Concurrently with Zoph et al. (2018), Zhong et al. (2018) propose a similar cell-based search space as shown in Figure 5. The key differences lie in their definition of cells and the

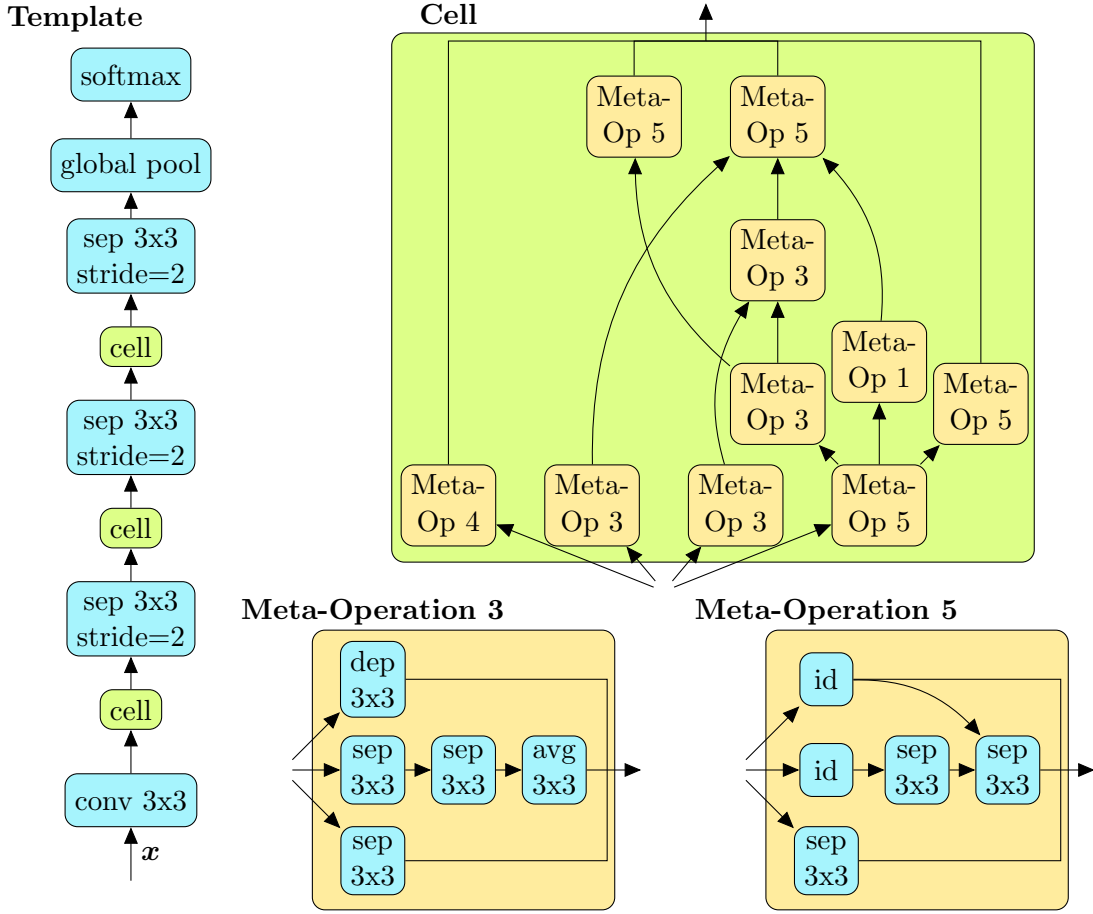


Figure 6: The search space proposed by Liu et al. (2018b) is based on an architecture template (left) which defines the sequence of cells and reduction operations. The choices made about an architecture is the set of meta-operations and their arrangement within the cell (right).

use of fixed max-pooling layers as opposed to reduction cells for handling feature dimensions in the architecture template. Since they do not decompose the cell structure any further, graphs comprising of arbitrary connections between different nodes are admissible as cells in this search space. This alternate definition of cell admits higher degrees of freedom than the normal cell of the NASNet search space, for instance the number of operations within each cell is not fixed. In addition to the number of operations, the instances of cells also differ in the type of operations and their input. This search space formulation does not allow skip connections across different cells. An example architecture discovered is presented in Figure 5.

Liu et al. (2018b) propose a search space similar to the one proposed by Zhong et al. (2018). An architecture template that describes their high-level definition of an architecture as shown in Figure 6. The main difference is that the search space is decomposed into a hierarchical search space. The first level of hierarchy defines a fixed numbers of meta-

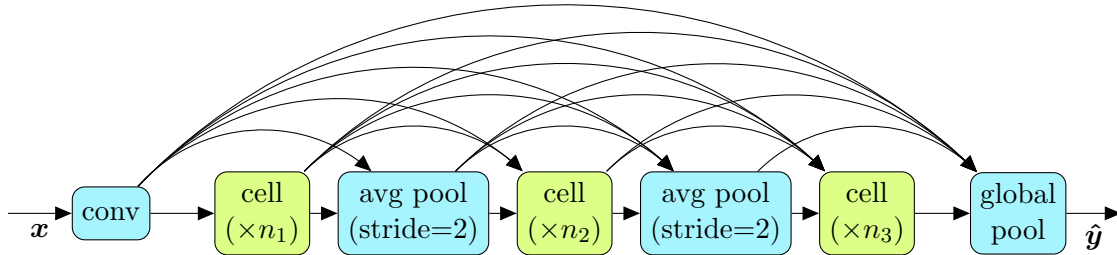


Figure 7: Mobile search spaces as used by Dong et al. (2018). The entire network including the cells are densely connected.

operations. A meta-operation is the connection of few operations to a larger segment (Figure 6). The second level represents the cell by connections between the meta-operations.

The cell-based design paradigm has also been used for defining search spaces that are suitable for mobile devices. Dong et al. (2018) propose a search space specifically aimed at meeting such requirements which include objectives like fewer parameters and fast inference time. Architectures in this search space comprise of cells without branches and a fixed internal structure that alternates two normalization (e.g. batch normalization) and convolution layers. All operations are densely connected (Huang et al., 2017) as shown in Figure 7 which also applies for those in the cell. The number of choices is significantly smaller than the previously discussed cell-based search spaces, making this a less challenging optimization task.

2.3 Global vs. Cell-Based Search Space

So far, there has been no detailed comparative study of the different search spaces proposed in the vast literature of neural architecture search. However, cell-based search spaces, in particular the NASNet search space, are the most popular choices for developing new methods. Most works that examine both search spaces, the global and the cell-based space, support this choice by providing empirical evidence that well-performing can be obtained in the cell-based search space (Pham et al., 2018). Regardless, cell-based search spaces benefit from the advantage that discovered architectures can be easily transferred across datasets. Moreover, the complexity of the architectures can be varied almost arbitrarily by changing the number of filters and cells. In general, the architectures belonging to the global search space do not exhibit all these properties, but specific cases may benefit from some of them. For instance, architectures can be naturally modified by varying the number of filters, but transferring a discovered architecture to a new datasets with different input shapes or deepening the architecture is not trivial. Interestingly, Tan et al. (2018) endorse the use of the global search space when searching for mobile architectures. They base this argument on a hypothesis that diversity of layers is critical for achieving both high accuracy and low latency for deployment in mobile devices and that these cannot be provided by conventional cell-based search spaces. Interestingly, Hu et al. (2018) remark

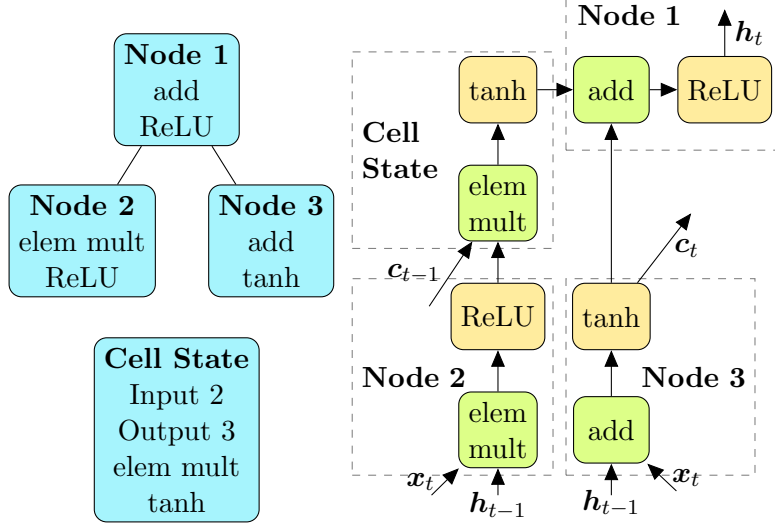


Figure 8: On the left is an example for a full binary tree describing a recurrent cell. Each node has a merging and an activation function associated. The cell state information is stored independently. It points to two nodes from the binary tree which are associated to the input and output as well as a merging and activation function. The right shows how this representation is translated to a computational graph.

on the importance of selecting initial architectures for searching in the global search space and show that architectures with comparable performance to the ones discovered in cell-based search spaces can be easily arrived at with appropriate initial conditions. This is an interesting insight which might revive the use of global search space in the development of new methods. However, the choice of useful initial architecture might be task-dependent and the guidelines for this selection remain unclear.

2.4 Search Space for Recurrent Cells

Recurrent neural networks (RNNs) are the popular class of deep learning models used for sequence-to-sequence modelling tasks in natural language processing and other domains. Zoph and Le (2017) were the first and only to define a search space for recurrent cells. A recurrent cell (g) can be described by

$$\mathbf{h}_t = g_{\theta, \alpha}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}), \quad (5)$$

where θ is the vector of trainable parameters, α the structure of g , \mathbf{h}_t the hidden state, \mathbf{c}_t the cell state, and \mathbf{x}_t the input at step t . The architecture α is defined by a full binary tree where every node is associated with a binary merge operation (e.g. addition, element-wise multiplication) and an activation function (e.g. tanh, sigmoid) as shown in Figure 8. Every node i represents a parameterized function

$$g_{\theta}^{(i)}(\mathbf{z}^{(i_1)}, \mathbf{z}^{(i_2)}) = o^{(\text{act})}\left(o^{(\text{merge})}\left(\mathbf{z}^{(i_1)}, \mathbf{z}^{(i_2)}\right)\right), \quad (6)$$

where $\mathbf{z}^{(i_1)}$ and $\mathbf{z}^{(i_2)}$ are the outputs computed by the child nodes of node i . Each leaf node takes as input \mathbf{x}_t and \mathbf{h}_{t-1} , the cell state is considered independently. A node within the tree has to be selected to provide \mathbf{c}_t . \mathbf{c}_{t-1} is used in combination with an arbitrary node using the selected merge operation and activation function.

3. Optimization Methods

With a comprehensive coverage of the search spaces, we are now in a position to formally define the problem of neural architecture search. We denote the space of all datasets as D , the space of all deep learning models as M , and the architecture search space as A . Given this setup, a general deep learning algorithm Λ is defined as the mapping

$$\Lambda : D \times A \rightarrow M . \quad (7)$$

In the context of this definition, an architecture of the search space $\alpha \in A$ defines more than just the topology. It further encodes all properties required to train a network on a dataset which includes the choice of model parameter optimization algorithm, regularization strategies and all other hyperparameters.

Given a dataset d , which is split into a training partition d_{train} and a validation partition d_{valid} , the general deep learning algorithm Λ estimates the model $m_{\alpha, \theta} \in M_{\alpha}$. This model is estimated by minimizing a loss function \mathcal{L} which is penalized with a regularization term \mathcal{R} with respect to the training data. That is,

$$\Lambda(\alpha, d) = \arg \min_{m_{\alpha, \theta} \in M_{\alpha}} \mathcal{L}(m_{\alpha, \theta}, d_{\text{train}}) + \mathcal{R}(\theta) . \quad (8)$$

Neural architecture search is the task of finding the architecture α^* which maximizes an objective function \mathcal{O} on the validation partition d_{valid} . Formally,

$$\alpha^* = \arg \max_{\alpha \in A} \mathcal{O}(\Lambda(\alpha, d_{\text{train}}), d_{\text{valid}}) = \arg \max_{\alpha \in A} f(\alpha) . \quad (9)$$

The objective function \mathcal{O} can be the same as the negative loss function \mathcal{L} . For the classification problem it is often the case that the loss is the negative cross-entropy and the objective function the classification accuracy. This problem definition falls under the broader scope of hyperparameter optimization (Bergstra and Bengio, 2012).

Optimizing the *response function* f is a global black-box optimization problem. In the following, we discuss several optimization strategies based on reinforcement learning, evolutionary algorithms and others.

3.1 Reinforcement Learning

Reinforcement learning (RL) approaches are useful for modeling a process of sequential decision making where an agent interacts with an environment with the sole objective of maximizing its future return. The agent learns to improve its behavior through multiple episodes of interaction with the environment. At every step t , the agent executes an action $a^{(t)}$ and receives an observation of the state $s^{(t)}$ along with a reward $r^{(t)}$. The environment on the other hand receives the agent's action, accordingly transits to a new state, and emits

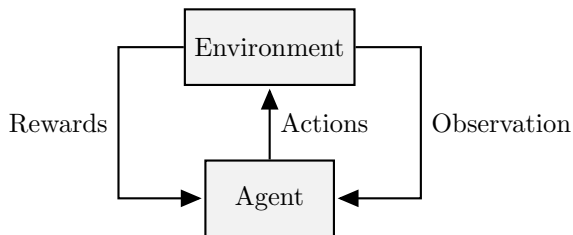


Figure 9: A general framework for reinforcement learning algorithms.

the corresponding observation and scalar reward (Figure 9). The objective of the agent lies in maximizing its return which is expressed as a discounted sum of rewards over T decision steps as defined by $\sum_{t=0}^T \gamma r^{(t)}$, where γ is the discounting factor. Naturally, such approaches are well suited for neural architecture search where the agent, namely the search algorithm, takes decisions to modify the system’s state, i.e. the architecture, so as to maximize the long term objective of high performance which is often measured in terms of accuracy.

Reinforcement learning systems are often modeled as a *Markov decision process* encompassing cases where the environment is fully observable. Most approaches consider simplified scenarios where the set of actions and states is finite and the setup consists in a finite horizon problem where episodes terminate after a finite number of steps. Even in this setting the number of possible options is combinatorially high. An RL agent seeks to learn a policy π which serves as the mapping from a state to a probability distribution over the set of actions. While one class of RL methods indirectly learns this policy with the use of value functions and state-action value functions, other approaches directly learn a parameterized policy. The state-value function $v_{\pi}(s)$ defines the expected return for the agent given that the starting state is s and the policy π is followed thereafter. The value of taking an action a in state s under a policy π is similarly defined by q_{π} . Both value functions satisfy the consistency equation defined by the Bellman equation (Sutton and Barto, 1998). Furthermore, the value function induces a partial order over policies and the optimal policy is the one that ascribes largest values to all states. The goal of an agent is to find this optimal policy.

Temporal Difference Learning Approaches like SARSA (Rummery and Niranjan, 1994), TD- λ (Sutton and Barto, 1998), and Q-learning (Watkins, 1989) attempt to find this policy implicitly via approximating the optimal value functions. The optimal policy is subsequently determined as the greedy-policy with respect to the optimal value function. The optimal value functions $v^*(s)$ and $q^*(a, s)$ satisfy the Bellman optimal criterion. Q-learning (Watkins, 1989) learns an action-value function $Q(s, a)$ that directly approximates q^* . The agent learns Q independently from the policy it follows. The Q-learning algorithm uses the following update rule for the action-value function Q ,

$$Q(s^{(t)}, a^{(t)}) \leftarrow (1 - \eta) Q(s^{(t)}, a^{(t)}) + \eta \left[r^{(t)} + \gamma \max_{a'} Q(s^{(t+1)}, a') \right]. \quad (10)$$

Here, η is the Q-learning rate. The learned policy and correspondingly the proposed architecture is subsequently derived by greedily selecting the actions for every state which maximize $Q(s, a)$.

Policy Gradient Methods An alternative approach in RL consists in policy gradient methods which do not appeal to value functions and instead directly learn policies as defined by a collection of parameters, $\pi_{\theta}(a|s)$. These methods select actions without explicitly consulting a value function. The parameters of the policy θ are adjusted so as to move in the direction of the agent’s performance measure via classical gradient ascent updates. The required gradient is often not directly available as the performance depends both on the actions selected and the distribution of states under which the actions is taken. Hence an empirical estimate of the gradient is used to perform the necessary update. REINFORCE (Williams, 1992) is a classical algorithm that estimates this gradient as,

$$\mathbb{E}_{\pi} \left[\sum_{t=0}^T \nabla_{\theta} \ln \pi_{\theta}(a^{(t)}|s^{(t)}) G^{(t)} \right]. \quad (11)$$

where $G^{(t)}$ is the return from step t , $\left(\sum_{t=0}^T \gamma r^{(t)}\right)$. However these empirical gradient estimates are often noisy and additional tricks like inclusion of a baseline are incorporated for useful learning. Moreover, alternative formulations of this gradient as derived through objective functions of importance sampling have led to other approximations. These approaches include Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) and Proximal Policy Optimization (PPO) (Schulman et al., 2017). It is worth noting that these methods are on-policy methods as the agents follow the policy they seek to learn.

Optimization with Q-Learning Baker et al. (2017) were one of the first to propose the use of RL-based algorithms for neural architecture search. They use a combination of Q-learning, ϵ -greedy, and experience replay in the design of their algorithm. The actions in their approach are the choice of different layers to add to an architecture as well as the option to terminate building the architecture and declare it as finished (we discussed this search space in Section 2.1). Subsequently, the states are the premature architectures. The trajectories sampled from this state space correspond to models which are subsequently trained to compute the validation accuracy. The Q function is appropriately updated with experience replay. In order to trade off between exploration and exploitation, they incorporate an ϵ -greedy strategy where random trajectories are sampled with a probability of ϵ . This is an off-policy method where the agent does not use the optimal policy during the episodes. Additionally, the action-value function utilized in this case is deterministic. The algorithm works as follows: after initializing the action-value function, the agent samples a trajectory which comprises of multiple decision steps, eventually leading to a terminal state. The algorithm then trains the model corresponding to the trajectory and updates the action-value function as defined in the Q-learning algorithm. Zhong et al. (2018) also use a Q-learning-based algorithm to search for network architectures. However, they perform the search in a cell-based search space which we discussed in Section 2.2.

Optimization with Policy Gradient Methods Alternate approaches based on policy gradient method have also been used for neural architecture search. The work by Zoph and Le (2017) was the first to consider this modeling approach. In their approach they directly model a controller whose predictions can be considered as actions that are used to construct a neural architecture. The controller parameterized by θ defines the stochastic policy $\pi_{\theta}(a|s)$. They incorporate an autoregressive controller which predicts the action

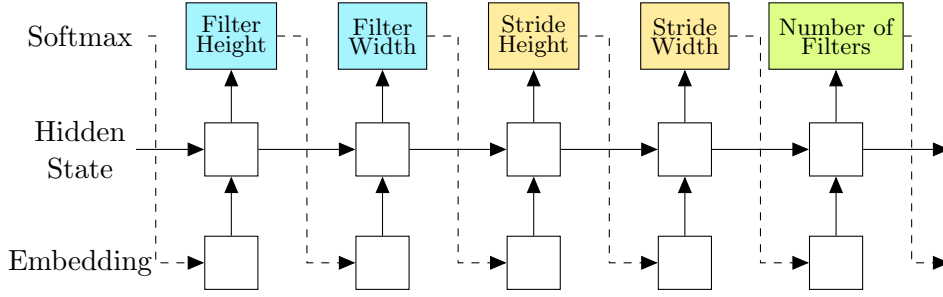


Figure 10: The controller used by Zoph and Le (2017) (predictions for skip connections are not shown) to predict configuration of one layer.

based on previous actions, and model it with a Recurrent Neural Network (RNN). During an episode, every action is sampled from the probability distribution implied by a softmax operation and then fed into the next time step as input (Figure 10). The RNN-controller in their approach samples layers which are sequentially appended to construct the final network. The final network is trained to obtain the validation accuracy and the parameters of the controller are updated as per the REINFORCE update rule (Equation (11)) which involves scaling the gradient with the obtained validation accuracy. As discussed in Section 2.1, they consider a space of chain-structured networks parameterized by only convolutional layers with different filter width, filter height, stride parameters, and number of filters. Additionally, they include anchor points in their encoding to allow for skip connections.

Policy gradient approaches have also been explored to learn controllers for cell-based search spaces. Zoph et al. (2018) define an RNN-controller that outputs actions which sequentially determines the inputs and operations for a prespecified number of blocks within a NASNet search space cell as defined in Section 2.2. The sampled architectures are trained and the parameters of the controller are updated according to the PPO update rule.

Cai et al. (2018a) also model a controller to parameterized the policy and train it with REINFORCE. However the action space of the controller differs and it requires an initial architecture to start with. There are two types of actions, widening a layer or inserting a new layer to deepen the network. Both actions are performed by means of *function-preserving transformations* (Chen et al., 2016). This means that although the structure of the architecture changes, the function that it models remains unchanged. Thus, the accuracy of this network does not change at first. This is ensured by initializing the additionally added parameters in a special way. In the example of adding a layer, it is initialized to correspond to the identity function. At every step and every layer, the controller predicts whether it wants to change it by means of widening or deepening it. As this involves predicting transformation decisions for a variable length sequence, they incorporate a bi-directional recurrent network (Schuster and Paliwal, 1997) with an embedding layer to learn a low-dimensional representation of the network which is subsequently fed into actors that compute the transformation decisions. For the predicted actions the function-preserving transformations are computed which enable efficient reuse of weights for training the newly sampled architecture which reduces the overall search duration.

The work by Cai et al. (2018b) builds upon these transformations and proposes new transformations to include branches in the network architectures. A convolutional layer or an identity map is first converted to its equivalent multi-branch motif such that the functionality is preserved. This is achieved by splitting or replicating the input to a convolutional or identity map into multiple branches, applying the corresponding transformation (convolutional or identity) to each branch, and respectively concatenating or adding the outputs of each branch. The function-preserving transformations are subsequently applied on branches to explore more complex architectures. Similar to the work by Cai et al. (2018a), the meta-controller in their approach also uses an encoder network to learn a low-dimensional representation of the architecture and provide a distribution over transformation actions. They propose to learn the meta-controller with REINFORCE.

Monte Carlo Tree Search An alternative way of solving the reinforcement learning problem lies in the use of Monte Carlo methods which are based on averaging sample returns (Sutton and Barto, 1998). A popular algorithm in this domain is the UCT algorithm (Kocsis and Szepesvári, 2006) which has been adopted in different ways for neural architecture search. This algorithm is designed for a tree-structured state-action space which is incrementally explored and expanded. A policy based on upper confidence bound (Auer, 2002) is utilized to explore this tree. The work of Negrinho and Gordon (2017) adopts the vanilla UCT algorithm with bisection for neural architecture search. Wistuba (2018b) builds on this work and employs function-preserving operations to speed up the training process and uses surrogate models to predict the reward for unexplored states. Similarly, Wang et al. (2019) make use of a surrogate model to predict the state reward and search on the NASNet search space and use the weight sharing paradigm (Section 3.4).

3.2 Evolutionary Algorithms

Evolutionary algorithms (EA) are population-based global optimizer for black-box functions which consist of following essential components: initialization, parent selection, recombination and mutation, survivor selection (Jong, 2006). The initialization defines how the first generation of the population is generated. After the initialization the optimizer repeats the following steps until termination (see Figure 11):

1. Select parents from the population for reproduction.
2. Apply recombination and mutation operations to create new individuals.
3. Evaluate the fitness of the new individuals.
4. Select the survivors of the population.

In this broad class of algorithms, the choices for mutation and recombination operators along with the choice of fitness and parent selection functions guide the overall search process. The choice of operators used for recombination and mutation is motivated to trade off diversity and similarity in the population, akin to the exploration and exploitation trade-off in reinforcement learning-based search algorithms. Similarly, the choice of fitness functions reflects the optimization objective and the choice of survivor selection enables competition between the individuals of the population.

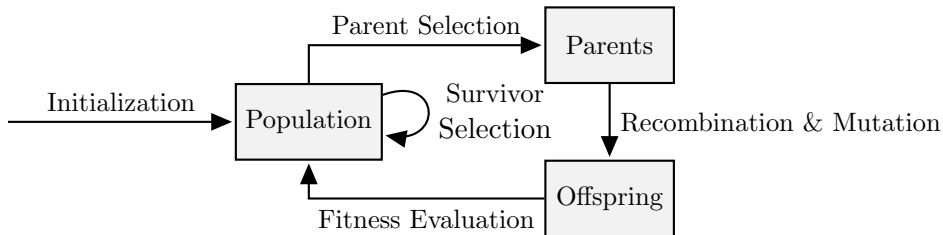


Figure 11: A general framework for evolutionary algorithms.

In the context of neural architecture search, the population consists of a pool of network architectures. A parent architecture or a pair of architectures is selected in step 1 for mutation or recombination, respectively. The steps of mutation and recombination refer to operations that lead to novel architectures in the search space which are evaluated for fitness in step 3 and the process is repeated till termination. Often only mutation operators are used in the domain of neural architecture search. There is no indication that a recombination operation applied to two individuals with high fitness would result into an offspring with similar or better fitness. On the contrary, oftentimes the fitness of the resultant offspring is much lower. For this reason and other reasons of simplicity, this part is often omitted.

The most common parent selection method in neural architecture search is *tournament selection* (Goldberg and Deb, 1990). This method selects the individuals for further evolution in two steps. First, it selects k individuals from the population at random. Then it iterates over them in the descending order of their fitness while selecting individuals for further steps with some (prespecified) high probability p . An alternative to tournament selection is *fitness proportionate selection*. In this approach an individual is selected proportional to its fitness. Thus, in a population $\{\alpha_1, \dots, \alpha_N\}$, the i -th individual is selected with probability $\frac{f(\alpha_i)}{\sum_{j=1}^N f(\alpha_j)}$, where $f(\alpha_i)$ is the fitness of the individual α_i .

After the recombination and mutation step, the population has grown. The intent of the survivor selection component is to reduce the population size and enable competition within the individuals. Several different policies are used to achieve this, ranging from selecting only the best (*elitist selection*) to selecting all individuals. One class of evolutionary algorithm that has been widely adopted for neural architecture search is genetic algorithms. These approaches bear their name from the representation of individuals in their methodology which is done by means of a fixed-length encoding, called the *genotype*. Mutation and recombination operations are performed directly on this representation. Further, a genotype is used to create the physical appearance of the individual, which in the case of neural architecture search is the architecture. This materialization is called the *phenotype*. It is important to note that parts of the genotype can be conditionally active or inactive. The information of the genotype is referred to as *active* if its modification results in a change in the phenotype (given all other information remains the same). For an example see Figure 13.

In the remaining section we discuss the popular choices for search space, mutation operators and selection functions that have been utilized for neural architecture search. We note that EA-based neural architecture search methods include a set of highly diverse approaches which have benefited from the varied choices of encoding the search space along

Table 1: High-level details of various evolutionary algorithms for neural architecture search.

Method	Search Space	Init	Parent Sel.	Survivor Sel.
Real et al. (2017)	global	simple	tournament	tournament
Xie and Yuille (2017)	global	random	all	fitness prop.
Suganuma et al. (2017)	global	random	n/a	elitist
Liu et al. (2018b)	cell-based	random	tournament	all
Real et al. (2019)	cell-based	random	tournament	youngest
Elsken et al. (2018)	global	simple	n/a	elitist
Wistuba (2018a)	cell-based	simple	tournament	all

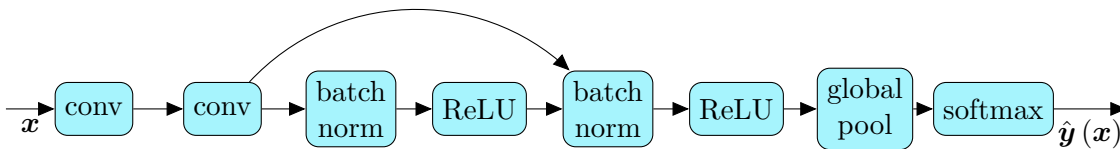


Figure 12: A possible architecture discovered by the algorithm described by Real et al. (2017). Noticeable are the redundant operations such as two convolutions in a row without an activation.

with the choices of mutation operators and selection functions. In the context of this work we describe six notable works in EA-based neural architecture search. A broad overview of these approaches is provided in Table 1.

Real et al. (2017) Real et al. (2017) were one of the first to propose an evolutionary algorithm to find competitive convolutional neural network architectures for image classification. Their approach begins with one thousand copies of the simplest possible architecture, i.e. a global pooling layer followed by the output layer. In the parent selection step of their approach they propose to sample a pair of architectures for further processing. While the better of the two architectures is copied (along with the weights), mutated, trained for 25,600 steps with a batch size 50, and added to the population, the other one is removed from the population. This is equivalent to using tournament selection with $k = 2$ and $p = 1$. The set of mutations consists of simple operations such as adding convolutions (possibly with batch normalization or ReLU activation) at arbitrary locations before the global pooling layer, along with mutations that alter kernel size, number of channels, stride and learning rate, add or remove skip connections, remove convolutions, reset weights, and a mutation that corresponds to no change. Since no further domain constraints are enforced, architectures with redundant components can be potentially sampled in this approach (see Figure 12).

Xie and Yuille (2017) Xie and Yuille (2017) made their work available only one day later. In contrast to Real et al. (2017), their work develops a genetic algorithm over a more structured search space which comprises of a sequence of segments with possibly parallel convolution operations. In their approach they consider networks to be composed of three

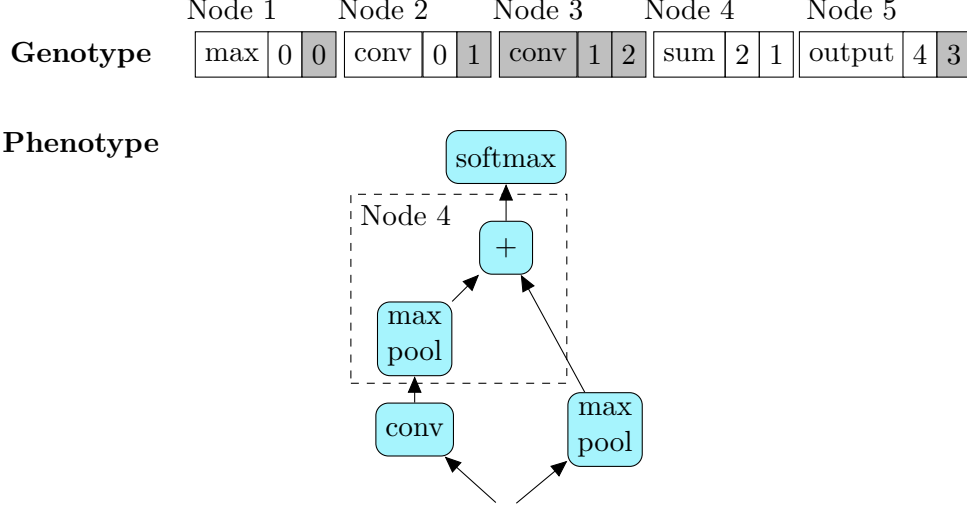


Figure 13: The genetic encoding used by Suganuma et al. (2017). Gray shaded parts of the genotype are inactive. Inactive parts such as Node 3 are not present in the phenotype. The additional max pooling layer introduced by Node 4 is an example how not explicitly encoded operations in the genotype can appear in the phenotype.

segments, each consisting of multiple convolution layers (see Section 2.1 for more details). Each segment is described by an adjacency matrix which defines a directed acyclic graph over the layers within the segment. The genotype of the entire network is obtained as a fixed length binary string of the adjacency matrices of the three segments. The algorithm begins with a population of 20 random samples from the search space. In the parent selection step, all pairs (α_i, α_{i+1}) , $i \bmod 2 = 1$ are considered for a cross-over operation. With a probability p , the cross-over operation swaps segments between the two selected networks. As the next step, all individuals which have undergone no modifications in the previous step are considered for mutation. In their algorithm, mutations are defined as random flip operations on the adjacency matrices that define the segments. Finally, the obtained offspring is trained from scratch and its fitness is evaluated. The fitness of an individual is defined as the difference between its validation accuracy and the minimum accuracy among all individuals of the population. As fitness proportionate selection is used in their approach, this modification ensures that the weakest individual survives with a probability of zero. It is worth remarking that this is one of the first works to demonstrate a successful transfer of an architecture automatically discovered on a smaller dataset to a larger one, namely from CIFAR-10 to ImageNet (Russakovsky et al., 2015).

Suganuma et al. (2017) Suganuma et al. (2017) present another optimizer based on genetic algorithms but consider a wider set of operations in their search space, which includes convolutions and pooling along with concatenation and summation of vectors. In their approach, they encode the entire network as a sequence of blocks represented by a triple that defines the operation and inputs of the block. The string obtained by concatenating the block encoding constitutes the genotype. Inactive parts are admissible in such a definition

of a genotype, for instance unused inputs for certain operations or disconnected blocks. The inactive parts do not materialize in the phenotype (for an example see Figure 13). They propose to use the $(1+\lambda)$ -evolutionary strategy ($\lambda = 2$) (Beyer and Schwefel, 2002) approach to guide the evolution where mutation and selection operators are applied in a loop until termination. Beginning with a genotype that corresponds to the parent individual (random for the first iteration), λ offspring are generated by forcefully modifying the active parts of the genotype. A mutation is also applied to the parent on one of the inactive parts of its genotype. However this does not result in any change of the phenotype and hence does not add to the overall training cost. This action ensures progress of the search even in cases where the parent of this generation will become the parent of the next generations again. The offspring networks are trained to obtain validation accuracies and the parent for the next iteration is selected using elitist selection strategy.

Liu et al. (2018b) Liu et al. (2018b) propose another evolutionary algorithm which uses the hierarchically organized search space as described in Section 2.2. At every step a mutation selects the hierarchy level to modify along with the modification to the representations at that level. The population is initialized with 200 trivial genotypes which are diversified by applying 1000 random mutations. Furthermore, parents are selected using tournament selection (5% of the population) and no individuals are removed during the evolutionary process. In their setting, mutations can add, alter and remove operations from the genotype.

Real et al. (2019) The follow-up work by Real et al. (2019) is one of the most significant works in the direction of using evolutionary algorithms for architecture search. It is primarily known for finding the architectures AmoebaNet-B and AmoebaNet-C which set new records for the task of image classification on CIFAR-10 and ImageNet dataset. However, their search process used a total of 3,150 GPU hours. Their evolutionary search algorithm operates on the NASNet search space (details in Section 2.2) with tournament selection used to select parents at each iteration. The selected individuals are mutated with a random change in an operation or a connection in either a normal or a reduction cell and are trained for 25 epochs. As opposed to previous works, their algorithm does not solely rely on validation accuracy and incorporates age in the selection of survivors. This is motivated to restrain repeated selection of well-performing models for mutation and introduce diversity to the population. This basically adds a regularization term to the objective function which makes sure that we are searching for architectures which are not only capable of reaching high validation accuracy once but every time.

The previously discussed evolutionary algorithms are mostly concerned about finding a good performing architecture ignoring GPU budget limitations. The shortest search time used among those methods is still seventeen GPU days. The two works discussed in the following are more concerned about efficiency and report comparable results within a day. The method to accomplish this is a combination of mutations which are function-preserving transformations (see Section 3.1) and a more aggressive, greedy evolution.

Elsken et al. (2018) Elsken et al. (2018) propose a simple yet efficient evolutionary algorithm inspired by the function-preserving transformations. Similar to Suganuma et al. (2017) they follow the $(1+\lambda)$ evolutionary strategy. A parent is selected with elitist selection

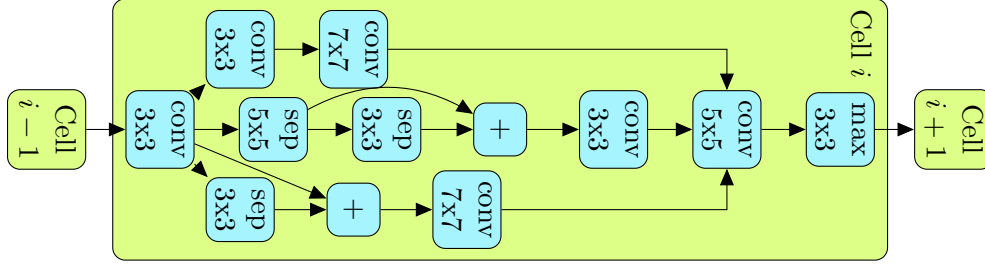


Figure 14: Exemplary outcome of the evolutionary algorithm proposed by Wistuba (2018a).

strategy, eight different offspring are generated with function-preserving mutations, and each is trained for a total of 17 epochs. The initial architecture comprises of three convolution and two max pooling layers. As discussed earlier, the function-preserving operations significantly reduce the training time per architecture and therefore the entire search duration.

Wistuba (2018a) In a similar line of work, Wistuba (2018a) also utilizes function-preserving mutations but uses a cell-based search space. The proposed search method begins with a simple network and performs function-preserving mutations to grow the population. The initial population is diversified by generating fifteen additional children from the base network. During the evolutionary algorithm, parents are selected using tournament selection and the offspring is trained for fifteen epochs prior to their fitness evaluation. Similar to Liu et al. (2018b), all individuals survive the evolution process to ensure diversity. Discovered cells have arbitrary structures, an example is provided in Figure 14.

3.3 Surrogate Model-Based Optimization

As the name implies, surrogate model-based optimizers use a surrogate model \hat{f} to approximate the response function f (Equation (8)) (Jones, 2001). For the case of neural architecture search this is motivated to get an approximate response for an architecture by circumventing the time-consuming training step and thereby improve the overall efficiency of the search process. The surrogate itself is modeled as a machine learning model and is trained on a meta-dataset which contains architecture descriptions along with their response values, which are gathered during the architecture search:

$$H = \{(\alpha_1, f(\alpha_1)), (\alpha_2, f(\alpha_2)), \dots\} \quad (12)$$

Surrogate models are generally trained to minimize the squared error:

$$\sum_{(\alpha, f(\alpha)) \in H} \left(\hat{f}(\alpha) - f(\alpha) \right)^2. \quad (13)$$

However, often in practice only a ranking for the architectures is desired and in such cases a low loss value is not necessitated as long as the surrogates provide a useful ranking.

The predictions from surrogate model are often used to identify promising candidate architectures. These candidates are evaluated, their corresponding new meta-instances are added to H and the surrogate model is accordingly updated (Figure 15). These steps are executed until a convergence criterion is reached. We describe three different approaches

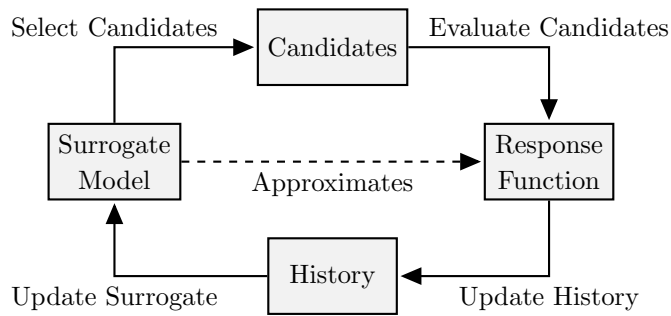


Figure 15: A general framework for surrogate model-based optimization.

to surrogate model-based optimization that have been used with an intent to improve the efficiency of neural architecture search.

Kandasamy et al. (2018) Kandasamy et al. (2018) cast this as a black-box optimization problem and tackle it with Bayesian optimization (Snoek et al., 2012). Bayesian optimization methods use a combination of a probabilistic surrogate model and an acquisition function to obtain suitable candidates. The acquisition function measures the utility by accounting for both, the predicted response and the uncertainty in the prediction. In their approach they model the surrogate with a Gaussian process (Rasmussen and Williams, 2006) and use expected improvement as the acquisition function (Mockus et al., 1978) for the optimization. The architecture with the highest expected utility is selected for evaluation and the consequent meta-instances are added to H . The surrogate model is updated and the previous steps are repeated. The authors render two changes to the standard Bayesian optimization algorithm. First, a novel kernel function to compute similarity between two network architectures is proposed and second, an evolutionary algorithm is used to maximize the acquisition function. The kernel value computation in this work is modeled as an optimal transport program (Villani, 2008). In essence, the graph topology, the location and frequency of operations as well as the number of feature maps are used to compute the similarity between two architectures.

Liu et al. (2018a) Liu et al. (2018a) also incorporate a surrogate model and search for architectures in the NASNet search space (Section 2.2). They explore this search space by progressively increasing the number of blocks in a cell, which, as previously discussed, is a hyperparameter and fixed in the template. While searching for models with an additional block, they simultaneously train a surrogate model which can predict the validation accuracy of the architecture given the encoding of the cell. The overall search process is designed as a beam search where at every step, a new set of candidates is generated by expanding the current pool of cells with an additional block and the top k of those are trained as per the predictions of the surrogate model. Further, in each step the surrogate model is updated by considering the newly obtained k architectures in the meta-dataset. Given the design of their setup, the surrogate model is required to handle variable sized inputs. In addition to an RNN and an Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) which are natural choices for this modeling, they also investigate a multi-layer perceptron (MLP) for this model.

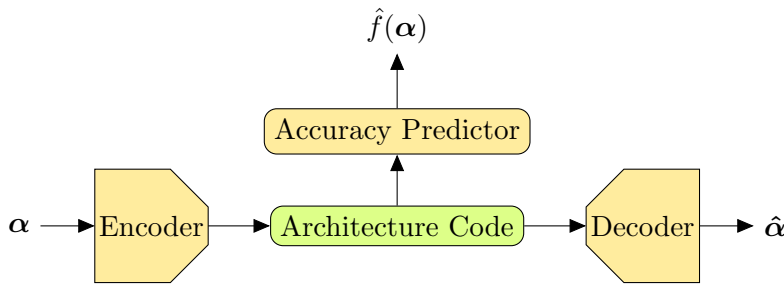


Figure 16: Luo et al. (2018) propose to combine an autoencoder with the surrogate model. This model is jointly learned to achieve $\alpha \approx \hat{\alpha}$ and $f(\alpha) \approx \hat{f}(\alpha)$.

Luo et al. (2018) In an interesting approach, Luo et al. (2018) jointly learn an autoencoder for the architecture representation with the surrogate model which uses the continuous encoding provided by the autoencoder, the architecture code, as its input (Figure 16). A key difference lies in their search algorithm which uses the surrogate model to sample new architectures by taking gradient steps with respect to the architecture code. The architecture code is changed by moving in the direction which yields better accuracy according to the surrogate model. The new architecture is obtained by mapping the potentially better architecture code back using the decoder learned as part of the autoencoder. At every step, these samples are trained, the meta-dataset is expanded, and the surrogate model and autoencoder are updated accordingly. The encoder, decoder and the surrogate model are jointly trained to minimize a weighted sum of the reconstruction loss and the performance prediction loss.

3.4 One-Shot Architecture Search

We define an architecture search method as one-shot if it trains a single neural network during the search process.¹ This neural network is then used to derive architectures throughout the search space as candidate solutions to the optimization problem. Most of these one-shot architectures are based on an over-parameterized network (Saxena and Verbeek, 2016; Pham et al., 2018; Liu et al., 2019b; Xie et al., 2019b; Casale et al., 2019). The advantage of this family of methods is the relatively low search effort which is only slightly greater than the training costs of one architecture in the search space. As we describe later, this methodology can be conveniently combined with many of the previously discussed optimization methods.

Weight Sharing Pham et al. (2018) search in a subspace of the NASNet search space (see Section 2.2) and operate on an over-parameterized network that covers the entire search space. They use reinforcement learning to learn a controller to sample architectures (which are only subgraphs of the complete architecture) from the search space (see Figure 17). The weights of the controller and a part of the over-parameterized network are alternately updated by gradient descent. During the network parameter update step, a batch is selected and an architecture is sampled from the search space. Only the parameters of this

1. Not to be confused with the one-shot learning paradigm which refers to algorithms that seek to learn from one instance.

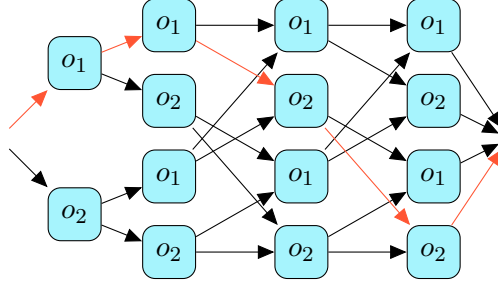


Figure 17: Example for weight sharing with only two operations in the chain-structured search space. Every box has its own weights, every path (e.g. the red one: $o_1 \rightarrow o_1 \rightarrow o_2 \rightarrow o_2$) is one architecture in the search space. Thus, weights are shared across different architectures.

architecture are updated, the remaining parameters and the parameters of the controller remain unchanged. During the update of the controller, architectures are sampled and evaluated on a validation batch. The obtained reward is used to update the controller similar to Zoph and Le (2017) (details provided in Section 3.1). This idea is referred to as weight sharing, as it corresponds to training multiple networks with shared weights. Bender et al. (2018) suggest to replace the controller with uniform sampling. Casale et al. (2019) learn the parameters of an independent categorical distribution to sample architectures. In all these variations, after the training step, the final architecture is selected according to the one-shot model’s validation performance and trained from scratch.

Differentiable Architecture Search By introducing a set of binary variables $\{\alpha_{i,j,k}\}$, an equivalent formulation of Equation (1) can be obtained as

$$\mathbf{z}^{(k)} = \sum_{j \in |O|} o^{(j)} \left(\left\{ \mathbf{z}^{(i)} \mid \alpha_{i,j,k} = 1, i < k \right\} \right). \quad (14)$$

For notational convenience, we will assume that each operation which determines $\mathbf{z}^{(k)}$ can only use a single input from $\mathcal{I}^{(k)}$. Then, this definition simplifies to

$$\mathbf{z}^{(k)} = \sum_{i \in \mathcal{I}^{(k)}} \sum_{j \in |O|} \alpha_{i,j,k} \cdot o^{(j)} \left(\left\{ \mathbf{z}^{(i)} \right\} \right), \quad (15)$$

with $\alpha_{i,j,k} \in \{0, 1\}$. In the works discussed thus far, the assumption has been that every operation is either part of the network or not. Liu et al. (2019b) relax this assumption and instead assumes a linearly weighted combination where the $\alpha_{i,j,k}$ can assume any real value in the range of 0 to 1, which allows for softer decisions on paths. They parameterize α by structural parameters β ,

$$\alpha_{i,j,k} = \frac{\exp(\beta_{i,j,k})}{\sum_{i \in \mathcal{I}^{(k)}} \sum_{j \in |O|} \exp(\beta_{i',j',k})}. \quad (16)$$

This softmax operation ensures that for every k

$$\sum_{i \in \mathcal{I}^{(k)}} \sum_{j \in |O|} \alpha_{i,j,k} = 1. \quad (17)$$

From Equation (9) a new, differentiable loss function for both structural parameters and model parameters θ can be derived:

$$\min_{\alpha(\beta) \in A} \mathcal{L} \left(\arg \min_{m_{\alpha(\beta), \theta} \in M_{\alpha(\beta)}} \mathcal{L}(m_{\alpha(\beta), \theta}, d_{\text{train}}) + \mathcal{R}(\theta), d_{\text{valid}} \right). \quad (18)$$

Liu et al. (2019b) propose an alternating optimization method which learns the model parameters by minimizing the loss on the training set and the structural parameters by minimizing the loss on the validation set using gradient-based optimization methods. After training this network, the final architecture is chosen based on the values of $\alpha_{i,j,k}$, the larger the better. This method is a very elegant solution that makes all parameters differentiable. However, it has a significant drawback: all parameters must be kept in memory all the time. Since the network covers the entire search space, this is a significant disadvantage. The following two works overcome this by introducing update rules which require to keep only a part of the network in memory at any one time.

The work by Xie et al. (2019b) builds upon the work of Liu et al. (2019b), but has three major differences. First, they represent the architecture by $p(\beta)$, where $p(\beta)$ is fully factorizable and is modeled with a concrete distribution (Maddison et al., 2017). Choosing this distribution results in only one path of the over-parameterized network being selected for training. This path corresponds to a feasible solution in the search space and accordingly requires less memory and can be easily kept in memory. Second, they minimize the loss of d_{train} with respect to the two types of parameters β and θ . The validation partition d_{valid} is used only to select the final architecture. And finally, gradients are estimated from Monte Carlo samples rather than analytic expectation. Cai et al. (2019) achieve the same effect with the use of binary gates (Courbariaux et al., 2015). Each binary gate chooses a part of the path based on a learned probability and in this sense is very similar to the idea of Xie et al. (2019b). Both these works significantly reduce the memory overload by not loading the entire over-parameterized model in the memory. Discrete architectures are derived by selecting the most likely operation and the top two predecessors in the over-parameterized network. This architecture is then trained again from scratch.

Hypernetworks Brock et al. (2018) propose the use of dynamic hypernetworks (Ha et al., 2017), a neural network which generates the weights for another network conditioned on a variable, i.e. in this case the architecture description. The hypernetwork is trained to generate network weights for a variety of architectures. It can be used to rank different architectures and derive a final architecture which is then trained from scratch. This method also shares weights, but most of them are shared in the hypernetwork. Hypernetworks save training time of candidate architectures during the search as the weights are obtained as predictions from these hypernetworks.

Zhang et al. (2019) extend this idea by combining it with a graph neural network (Scarselli et al., 2009) which directly operates on a graph defined by nodes and edges. Every node is a recurrent neural network, and messages between these networks can be propagated using the edges. Each node stores its own state, which is updated by means of message propagation. For each architecture, a graph neural network is generated which is homomorphic. This means we have one node for each operation in the architecture, and for each connection between operations we have an edge in the graph neural network. The hypernetwork

is then conditioned on the states of the graph neural network to infer the weights of the architecture.

Discussion All of the one-shot approaches share the weights across different sampled architectures, and rely on the hypothesis that the trained over-parameterized network can be used to rank architectures for quality. This belief is supported by Bender et al. (2018) who show a correlation between the validation error obtained for a sample of the over-parameterized network and the one obtained after training the corresponding architecture from scratch. However, they acknowledge that training the over-parameterized network is not trivial and that the practical approaches of batch normalization, dropout rate and regularization play an important role. Zhang et al. (2019) conduct a similar experiment using a separate implementation and confirm the correlation between the two variables. The correlation for the hypernetwork-based optimizers has been independently confirmed by Brock et al. (2018) and Zhang et al. (2019). Thus, we have sufficient empirical evidence that weight sharing is indeed a useful tool for the search of CNN architectures. As a result, the search only takes up insignificantly additional time than training a single candidate in the search space and is consequently incredibly efficient compared to other optimization methods. Sciuto et al. (2019) challenge this widespread belief for its applicability for searching recurrent cells. They show that the ranking of architectures implied by the various methods (Pham et al., 2018; Liu et al., 2019b; Luo et al., 2018) does not correlate with the true ranking. The reason that the search methods still deliver good results is solely due to the very limited search space. They provide empirical evidence that a random search outperforms many of the previously described methods in the search for an RNN cell.

3.5 Conclusions

In this section, we reviewed various optimization algorithms based on methods such as reinforcement learning, evolutionary algorithms, surrogate model-based optimization, and one-shot models. Naturally, the imminent question is - which method should be prescribed? In order to answer that we need to inspect if the results can be fairly compared. In our opinion, this is not the case. The different experiments differ drastically in terms of search space, search duration and data augmentation. However, we can safely conclude that the NASNet search space is the most commonly used search space, presumably because the imposed restrictions in its definition favor the discovery of well-performing architectures. Similarly, strategies that reuse parameters (through function-preserving transformations or weight sharing) are effective in reducing the search duration. However, we would like to re-emphasize the questionable benefits of weight sharing for RNNs as noted in our previous discussion.

At this point we need to criticize the lack of fair baselines. Although architecture search can be considered as a special way of optimizing hyperparameters, most of the related work is disregarded. In particular, random search which has proven to be an extremely strong baseline. In fact, Sciuto et al. (2019) show that random search finds better RNN cells than any other optimizer. Li and Talwalkar (2019) also confirm this result and additionally show that random search finds architectures that perform at least as well as the ones obtained from established optimizers for CNNs. It is worth noting that these random optimizers work on a search space that is known to sample well-performing architectures. Xie et al. (2019a)

Table 2: We give an overview of the results obtained and the search time required on CIFAR-10 by the various search algorithms discussed. In addition, we list the results of various random searches and human-designed architectures.

	Reference	Error (%)	Params (Millions)	GPU Days
RL	Baker et al. (2017)	6.92	11.18	100
	Zoph and Le (2017)	3.65	37.4	22,400
	Cai et al. (2018a)	4.23	23.4	10
	Zoph et al. (2018)	3.41	3.3	2,000
	Zoph et al. (2018) + Cutout	2.65	3.3	2,000
	Zhong et al. (2018)	3.54	39.8	96
	Cai et al. (2018b)	2.99	5.7	200
	Cai et al. (2018b) + Cutout	2.49	5.7	200
EA	Real et al. (2017)	5.40	5.4	2,600
	Xie and Yuille (2017)	5.39	N/A	17
	Suganuma et al. (2017)	5.98	1.7	14.9
	Liu et al. (2018b)	3.75	15.7	300
	Real et al. (2019)	3.34	3.2	3,150
	Elsken et al. (2018)	5.2	19.7	1
	Wistuba (2018a) + Cutout	3.57	5.8	0.5
SMBO	Kandasamy et al. (2018)	8.69	N/A	1.7
	Liu et al. (2018a)	3.41	3.2	225
	Luo et al. (2018)	3.18	10.6	200
One-Shot	Pham et al. (2018)	3.54	4.6	0.5
	Pham et al. (2018) + Cutout	2.89	4.6	0.5
	Bender et al. (2018)	4.00	5.0	N/A
	Casale et al. (2019) + Cutout	2.81	3.7	1
	Liu et al. (2019b) + Cutout	2.76	3.3	4
	Xie et al. (2019b) + Cutout	2.85	2.8	1.5
	Cai et al. (2019) + Cutout	2.08	5.7	8.33
	Brock et al. (2018)	4.03	16.0	3
	Zhang et al. (2019)	2.84	5.7	0.84
Random	Liu et al. (2018b)	3.91	N/A	300
	Luo et al. (2018)	3.92	3.9	0.3
	Liu et al. (2019b) + Cutout	3.29	3.2	4
	Li and Talwalkar (2019) + Cutout	2.85	4.3	2.7
Human	Zagoruyko and Komodakis (2016)	3.87	36.2	-
	Gastaldi (2017) (26 2x32d)	3.55	2.9	-
	Gastaldi (2017) (26 2x96d)	2.86	26.2	-
	Gastaldi (2017) (26 2x112d)	2.82	35.6	-
	Yamada et al. (2016) + ShakeDrop	2.67	26.2	-

Table 3: Results obtained on ImageNet by architectures discovered by neural architecture search methods as well as human-designed architectures.

	Reference	Top 1/Top 5 Accuracy (%)	Params (Millions)	Image Size (squared)	GPU Days
RL	Zoph et al. (2018)	82.7/96.2	88.9	331	2,000
	Zhong et al. (2018)	77.4/93.5	N/A	224	96
	Cai et al. (2018b)	74.6/91.9	594	224	200
EA	Xie and Yuille (2017)	72.1/90.4	156	224	17
	Liu et al. (2018b)	79.7/94.8	64.0	N/A	300
	Real et al. (2019)	82.8/96.1	86.7	331	3,150
	Real et al. (2019)	83.9/96.6	469	331	3,150
SMBO	Liu et al. (2018a)	74.2/91.9	5.1	224	225
	Liu et al. (2018a)	82.9/96.2	86.1	331	225
One-Shot	Bender et al. (2018)	75.2/N/A	11.9	224	N/A
	Casale et al. (2019)	74.0/91.6	5.6	N/A	1
	Liu et al. (2019b)	73.3/91.3	4.7	224	4
	Xie et al. (2019b)	72.7/90.8	4.3	224	1.5
	Cai et al. (2019)	75.1/92.5	N/A	224	8.33
	Brock et al. (2018)	61.4/83.7	16.2	32	3
	Zhang et al. (2019)	73.0/91.3	6.1	224	0.84
Human	Howard et al. (2017)	70.6/89.5	4.2	224	-
	Zhang et al. (2018b)	70.9/89.8	5.0	224	-
	Xie et al. (2017)	80.9/95.6	83.6	320	-
	Zhang et al. (2017)	81.3/95.8	92.0	331	-
	Chen et al. (2017)	81.5 95.8	79.5	320	-

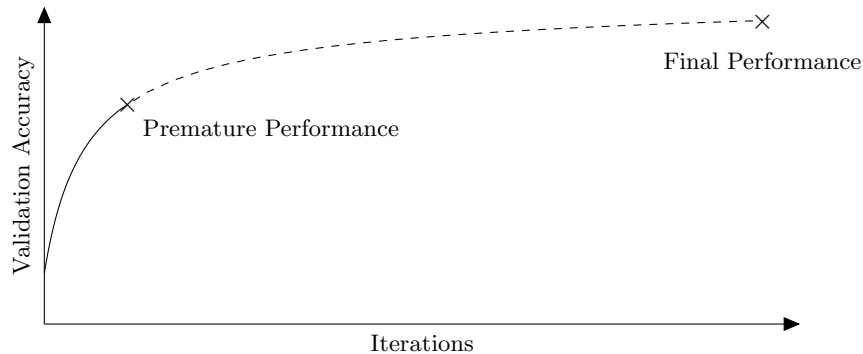


Figure 18: Learning curve prediction tries to predict from the premature learning curve (solid line) the final performance. This can significantly reduce the number of required training iterations.

go one step further and use a graph generator to generate random graph structures that no longer adhere to the rules of established search spaces. Although these graph generators do not have any deep learning specific prior, the generated architectures perform better than those found by complex architecture optimizers. Architectures designed by humans often serve as a motivation to justify the search for neural architectures. However, state-of-the-art architectures are often not considered for comparison or not trained under equivalent experimental conditions. Thus, the effect of learning rate decay strategies, augmenting techniques (cutout (DeVries and Taylor, 2017)), regularization tricks (ScheduledDropPath (Zoph et al., 2018)), and other nuanced strategies that effect training dynamics gets overshadowed. We include the results for the state-of-the-art architectures for CIFAR-10 in Table 2 and report additional results for models trained with the popular data-augmentation technique of cutout in Table 4. Results for ImageNet are reported in Table 3. We note that the performance gap between human-designed and discovered architectures is smaller than what is often claimed or reported, in particular for CIFAR-10.

We recognize that in recent years many interesting and creative architecture search methods have been developed, but we argue that the source of performance gains still remains unclear. This remains a highly relevant and unanswered question for this research area.

4. Early Termination of Training Processes

Although expensive, training is an inevitable step to obtain the validation accuracy on a dataset. For neural architecture search, however, it is not necessary to know the validation accuracy, it is sufficient to know whether an architecture candidate is better or worse than the current best solution. Human experts estimate the likelihood if the current candidate can surpass the current best solution by analyzing the behavior of its learning curve. We refer to the function that maps the number of iterations used for training a model to its validation accuracy at that time as its learning curve (Figure 18). If there is a high probability that a current training process does not lead to an improved solution, the process is terminated. The time thus saved can now be used for other training processes. Learning

curve prediction methods are automated methods that try to estimate this probability. The previously discussed surrogate models (Section 3.3) are related to these methods. The main difference is that learning curve prediction methods require additional training of a candidate and take the learning curve into account.

Many approaches in neural architecture search do not train models to completion. Instead, the validation accuracy of this premature model is used as an approximation for the final performance. One popular hyperparameter optimization which relies on this premature performance is Hyperband (Li et al., 2017). In this work, a random set of candidate models is trained in parallel and after every few iterations (the considered number of iterations is increased exponentially), training of a large fraction of candidates is terminated based on the premature performance. In this way the number of candidates reduces exponentially, while at the same time the training time per candidate is also increased at an exponential rate. Eventually only one candidate remains which serves as the output of the search algorithm.

In the following we discuss more sophisticated approaches which try to predict the final performance based on architecture and data information as well as the premature learning curve. All these methods use a model to predict the final performance of a training job. This model provides the probability that a premature learning curve will exceed a particular accuracy. This is used to determine the likelihood that the current training job will provide a network with higher accuracy than the current best one. If this likelihood does not exceed a predetermined threshold, training is terminated.

Domhan et al. (2015) Typically learning curves are increasing and saturating functions. Domhan et al. (2015) build on this observation and select a set of eleven different increasing and saturating function families and combine them by a weighted sum to probabilistically model a learning curve. However, since it does not learn across learning curves, it requires a relatively long premature learning curve in order to yield accurate predictions.

Klein et al. (2017) Klein et al. (2017) extend the work by Domhan et al. (2015) to learn the learning curve prediction model across different learning curves. For this purpose, the function families are combined with a Bayesian neural network which infers the parameters of the functions and their weights conditioned on the premature learning curve. Parameters of the Bayesian neural network are learned on a set of mature learning curves.

Chandrashekar and Lane (2017) The two previously discussed works make strong assumptions regarding the learning curves by using a set of function families which might not apply generally. Moreover, noisy optimization or learning rate schedules may violate the assumption made about the learning curve behavior. For instance, a step-wise decay of the learning rate violates the assumption that the learning curve is a quickly saturating function and a cyclic learning rate schedule violates the assumption that the learning curve is an increasing function. Chandrashekar and Lane (2017) overcome this problem by using mature learning curves observed during training rather than a predefined set of functions. An affine transformation is used to reduce the squared error between all mature learning curves and a premature learning curve. The prediction model is formed as an average of the best fits of mature learning curves.

Baker et al. (2018) All previously discussed learning curve predictors are based only on the premature learning curve as well as machine learning models specifically designed for learning curve prediction. Baker et al. (2018) approach this problem differently. First, they enrich the representation and include additional features like the first and second derivative of the premature learning curve, properties of the architecture (number of parameters and layers) and chosen hyperparameters (e.g. initial learning rate and weight decay). Second, they study a set of traditional regression models to solve the prediction task. Empirically, they find that the ν -Support Vector Regression (Schölkopf et al., 2000) performs best among other alternatives such as random forest or Bayesian linear regression. Furthermore, they report that the premature performance can be a strong competitor method which in most cases outperforms the work by Domhan et al. (2015) and Klein et al. (2017). While Domhan et al. (2015) did not compare against the premature performance, Klein et al. (2017) only report results with respect to the mean squared error. However, the squared error is no relevant measure for this task and it is by no means surprising that the premature performance has a high squared error. Instead, the ranking of learning curves is more important and thus a measure like R^2 (as used by Baker et al. (2018)) is more appropriate. Finally, Baker et al. (2018) are also the first and only to use a more sophisticated learning curve prediction method in combination with a neural architecture search method. They build upon their previous work (Baker et al., 2017), report a speed up by a factor of 3.8, and find equally good neural architectures.

5. Transfer Learning

Transfer learning refers to the idea of reusing knowledge gained when solving one task in order to solve a different but related task more efficiently.

In the context of neural architectures search, many simple techniques have proven to be effective. Some works (Baker et al., 2017) demonstrate that an architecture discovered for one dataset can be easily applied (without introducing significant changes) to a dataset with the same image dimensions. Zoph et al. (2018) demonstrate easy transferability of architectures discovered on datasets with few data points and a small resolution (CIFAR-10) to datasets with significantly more data and a higher resolution (ImageNet). Furthermore, architectures which perform well across a diverse set of tasks provide a competitive solution (Wistuba and Pedapati, 2019). Other similar paradigms include searching on a subset of the data (Sabharwal et al., 2016) and searching on a down-scaled version of the data (Hinz et al., 2018). It is also a common practice to search for networks with few cell repetitions and for cells that use only few filters. Then, the final model is obtained by increasing the number of filters and number of cell repetitions significantly to build larger models with better performance.

Transferring knowledge across tasks provides another alternative avenue that can be leveraged to improve the performance of neural architecture search methods. Most search algorithms train multiple sample architectures during the search process. In the same spirit as the use of surrogate models, training of these sampled architectures can be made more efficient with the use of additional information gained from training architectures on other datasets of tasks. In order to learn dataset-specific properties, features which describe

properties of the dataset, also called meta-features (Reif et al., 2014), are considered. We discuss some of these approaches in this section.

Wong et al. (2018) Wong et al. (2018) use transfer learning to automate the fine-tuning of architectures across tasks. They extend the controller of Zoph and Le (2017) to include task-specific embeddings as input to the controller to account for the different tasks in the policy. The controller generates a sequence of discrete actions that specify design choices for fine-tuning a network architecture. This includes choosing the embeddings, the structure of the classifier and various hyperparameters. The parameters of the controller are updated with REINFORCE. In their work they explore transfer learning for both image classification and natural language processing tasks.

Istrate et al. (2019) The approach by Istrate et al. (2019) is based on a surrogate model which predicts the performance of neural architectures for a new dataset based on performance observations of neural architectures on other datasets. They use as *land-marker* (Pfahringer et al., 2000), the accuracy of one computationally inexpensive a CNN model on new dataset, which is their only meta-feature. This meta-feature is used to determine the most similar datasets. It is also used together with the architecture encoding as input for the surrogate model. The observations on datasets with similar validation accuracy are used to train the surrogate model. An evolutionary algorithm with a setup similar to Real et al. (2017) is used to obtain the the architecture that maximizes the surrogate model’s prediction. This architecture is the outcome of the search process and is trained from scratch.

Kokiopoulou et al. (2019) Kokiopoulou et al. (2019) propose to automate fine-tuning for transferring models across different text classification tasks by using a surrogate model. This model is trained to predict the validation accuracy for a given set of meta-features and architecture design choice. Contrary to other works, they propose to learn the meta-features from the raw data rather than relying on manually engineered meta-features or embeddings. The surrogate model consists of two components. While the first component learns to compute predictive meta-features given a small amount of data, the second component predicts the validation accuracy for an architecture based on the architecture description and the meta-features. These components are trained across multiple datasets and then used for a new dataset. Instead of just predicting the accuracy of several architectures, the authors propose to search it by means of gradient ascent. For this purpose they use a continuous parametrization for the neural architectures.

Xue et al. (2019) Xue et al. (2019) assume a lifelong neural architecture search setting with an incoming stream of datasets. For each new dataset, a set of landmarks is trained to establish a dataset representation. Based on this representation, it is decided whether the dataset resembles a previously handled one, or whether it is a fundamentally different one. For the former case, they avoid a search and reuse the solution for the previously handled dataset. Otherwise, a new search with any arbitrary optimizer is initiated for the dataset.

6. Constraints and Multiple Objectives

While it is important to find networks that yield high accuracy, sometimes for a practitioner it is also imperative to consider other objectives, such as the number of model parameters, the number of floating point operations, and device-specific statistics like the latency of the model. There exist two different approaches which account for these additional objectives. In the first approach, the conditions are added as constraints to the optimization problem so as to enforce requirements like fewer parameters or faster inference time. The exact form of the constraints and the trade-off between different constraints can be adjusted as per practical requirements. In the second approach, the problem is tackled as a multi-objective function optimization problem which yields a set of proposals as a solution (Deb, 2014). This separates the decision making into two steps. In the first step, a set of candidates is obtained without considering any trade-offs between the different objectives and then the decision for a solution is made in the second step. This separation allows for reusing the set of candidate solution for different specifications.

6.1 Constrained Optimization

For some tasks modeled with deep learning, specific constraints g_i , such as, thresholds on inference time or memory requirement are explicitly stated. For such cases the single-objective optimization problem defined in Equation (9) then turns into a constrained optimization problem as defined by

$$\max_{\alpha \in A} f(\alpha) \quad (19a)$$

$$\text{subject to} \quad g_i(\alpha) \leq c_i \quad \forall i \in I. \quad (19b)$$

The optimization methods discussed in Section 3 cannot be directly adopted for this scenario as they were designed to solve an unconstrained optimization problem. However, the classical technique of penalty methods (Fiacco and McCormick, 1990) can be used to form the unconstrained optimization given by

$$\max_{\alpha \in A} f(\alpha) \cdot \prod_{i \in I} \lambda(g_i(\alpha), c_i). \quad (20)$$

The function λ is a penalty function, that punishes any violations of constraints. Once a penalty function has been selected, most optimization method discussed in Section 3 can be directly applied to solve the problem. In the following, we discuss the various penalty functions considered in the literature along with the specific optimizer used by different approaches.

Tan et al. (2018) use

$$\lambda(g_i(\alpha), c_i) = \left[\frac{g_i(\alpha)}{c_i} \right]^{w_i}, \quad (21)$$

as a penalty function, where w is treated as a hyperparameter and is set to suit the desired trade-off. They use the reinforcement learning optimizer proposed by Zoph and Le (2017) in the factorized hierarchical search space as discussed in Section 2.2. Zhou et al. (2018) propose to use

$$\lambda(g_i(\alpha), c_i) = \phi^{\max\{0, (g_i(\alpha) - c_i)/c_i\}}, \quad (22)$$

where ϕ is a penalization constant in the range of 0 to 1 and use a similar approach to Cai et al. (2018a) to find the final architecture. Specifically, a policy is learned that decides whether to add, remove or keep a layer as well as whether to alter its number of filters. Hsu et al. (2018) use a harder penalty function which returns 0 when the constraint is violated and 1 otherwise. They use a reinforcement learning optimizer similar to the one proposed by Zoph and Le (2017) which predicts the hyperparameters of different layers. Thus, the reward received by the controller is the accuracy for cases which do not violate any constraint and is 0 otherwise. The authors evaluate their optimizer to select the layer-wise hyperparameters (number of filters, kernel size) of an AlexNet (Krizhevsky et al., 2012) and the cell-wise hyperparameters (stage, growth rate) of a CondenseNet (Huang et al., 2018a). We refer to Section 3.1 for more details on the optimization methods.

6.2 Multi-Objective Optimization

Another approach to handle multiple fronts lies in the formalism of multi-objective optimization problem (Hwang and Masud, 2012) defined as

$$\max_{\alpha \in A} f_1(\alpha), f_2(\alpha), \dots, f_n(\alpha) . \quad (23)$$

However, often there is no single optimal solution that simultaneously minimizes all these functions. As some of the objectives can be conflicting, an optimal solution with respect to one objective may not be optimal with respect to another. Therefore, in such scenarios, the task boils down to finding a set of solutions which are *Pareto optimal*. A solution is Pareto optimal if none of the objectives can be improved without worsening at least one other objective. This means that for a Pareto optimal solution α , no other solution $\alpha' \in A$ exists such that $f_i(\alpha) \geq f_i(\alpha')$ with $f_j(\alpha) > f_j(\alpha')$ for some j . Otherwise, we say α' dominates α , $\alpha' \prec \alpha$. Finally, it is up to the user to select a model from the set of Pareto optimal solutions, the Pareto front, maybe depending on different deployment scenarios.

Decomposition Methods One way to solve this problem is the decomposition approach. A parameterized aggregation function h is used to transform the multi-objective optimization problem into a single-objective optimization problem

$$\max_{\alpha \in A} h((f_1(\alpha), f_2(\alpha), \dots, f_n(\alpha)), \mathbf{w}) . \quad (24)$$

Examples for h are the weighted sum, weighted exponential sum, weighted min-max or weighted product (Santiago Pineda et al., 2014). However, solving this problem for a fixed setting of the weights \mathbf{w} will in most cases not find all Pareto optimal solutions. Therefore, the problem is solved for different \mathbf{w} . In neural architecture search this requires multiple optimization runs for several different weight vectors which is prohibitive. Therefore, a common approach is to select an aggregation function and fix the weight vector according to domain knowledge and the desired trade-off between objectives. Consequently, the multi-objective becomes a single-objective optimization problem which can be solved by any method discussed in Section 3. Hsu et al. (2018) propose to use the weighted sum as the aggregation function and use a reinforcement learning approach (Zoph and Le, 2017) to solve this problem. Unfortunately, not every objective function is differentiable which is a requirement for some of the optimization methods discussed in Section 3.4. Cai et al. (2019)

demonstrate one way to overcome this obstacle. They propose to replace non-differentiable objective functions with differentiable surrogate models. Specifically, they use a surrogate model to predict the latency of an operation.

NSGA-II An alternative approach is to estimate architectures which are not dominated by the current solution set and evaluate their performance. NSGA-II (Deb et al., 2002) is an elitist evolutionary algorithm that is utilized predominantly. The initial population is selected at random. Candidate solutions are sorted into different lists, called fronts. All the non-dominated solutions belong to the first front. The solutions in the i^{th} front are only dominated by all the solutions in the $1, \dots, i - 1$ fronts. Solutions within a front are sorted according to their crowding distance which is a measure for the density of solutions within this solution’s region. It is computed by the sum of all the neighborhood distances across all the objectives and ensures that the algorithm explores diverse solutions in the search space. This yields a ranking of all solutions where the solution in front 1 with lowest density is the best one. Now, the top k solutions are selected as parents, mutated, recombined and evaluated.

Kim et al. (2017) and Lu et al. (2018) both employ NSGA-II in order to tackle the problem. The early work of Kim et al. (2017) encodes the genotype by means of layer type and number of outputs leading to only chain-structured networks. In their experiments Lu et al. (2018) report results on both NASNet search space and the genotype representation by Xie and Yuille (2017) (see Section 2 for details) and therefore are not limited to chain-structured networks. They use the vanilla NSGA-II for modeling the evolutionary process. However the sampling step in their approach comprises of two phases - exploration and exploitation. For the exploration phase, the offspring is generated by applying mutations and cross-over. In the exploitation phase new samples are obtained from a Bayesian network that models the distribution of previously trained architectures.

Elsken et al. (2019) Elsken et al. (2019) propose another evolutionary algorithm which shares many common aspects with their earlier work (Elsken et al., 2018) (see Section 3.2). No cross-over operations are used and mutations are function-preserving. They extend the pool of mutations and include new ones that shrink an architecture by means of layer removal or filter reduction. Similar to NSGA-II, their algorithm tries to achieve diverse solutions. They take advantage of the fact that the objective functions can be divided into *cheap-to-evaluate objective functions* (e.g. number of parameters, inference time etc.) and *expensive-to-evaluate objective functions* (e.g. validation accuracy). A kernel density estimator is used to model the density of candidates with respect to the cheap-to-evaluate objectives in order to select candidates from low-density regions. Finally, the expensive-to-evaluate objectives for the candidates are evaluated and the Pareto front is updated.

Smithson et al. (2016) Smithson et al. (2016) propose a surrogate model-based approach to search for a chain-structured architecture. The search algorithm starts with a random parameter setting. New candidates are sampled from a Gaussian distribution around previously evaluated architectures. A surrogate model (artificial neural network) is used to predict the values for the expensive-to-evaluate objective functions, cheap-to-evaluate objectives are evaluated exactly. Based on both the predictions and evaluations, it is determined whether this solution is dominated by an existing solution which will in

turn determine the probability the candidate is accepted. Every accepted candidate will be evaluated and thereafter the surrogate model is updated.

Dong et al. (2018) Dong et al. (2018) is another surrogate model-based approach which extends the work by Liu et al. (2018a) (see Section 3.3) to solve multi-objective optimization problems. The only difference of the optimization method is how models are selected. Instead of only considering model accuracy, they also consider all other objectives. Feasible candidates are those which do not violate any constraints and are not dominated (according to the prediction of the surrogate model) by any known solution.

6.3 Model Compression

Motivated by the efficiency of network deployment on mobile devices, model compression has surfaced as a relevant goal for architecture design. Neural architecture search methods in this context seek to find policies to prune a model so as to satisfy budget constraints in terms of number of parameters, without sacrificing a great deal in accuracy. To this end, we discuss three important works that propose methods to automate model compression.

He et al. (2018) In their work, He et al. (2018) use a reinforcement controller to prune a trained network in a layer-wise fashion. The controller takes as input a representation for each layer and is trained to output the desired sparsity ratio of this layer after compression. They use deep deterministic policy gradient (Lillicrap et al., 2016) to update the parameters of the controller. During the exploration phase, they evaluate the reward accuracy of the pruned model without fine-tuning and argue that it serves as a good proxy for the fine-tuned accuracy.

Ashok et al. (2018) Ashok et al. (2018) use reinforcement learning to learn two policies which sequentially remove and shrink layers for any given initial architecture with the aim of compressing it. The architecture altered by the policies is trained from scratch with knowledge distillation (Hinton et al., 2015), where the initial architecture serves as the teacher. REINFORCE is used to train the reinforcement learning controllers using following reward function with resembles the penalty function introduced in Equation (20),

$$\frac{f(\boldsymbol{\alpha})}{f(\boldsymbol{\alpha}_{\text{init}})} g(\boldsymbol{\alpha}) (2 - g(\boldsymbol{\alpha})) , \quad (25)$$

where $g(\boldsymbol{\alpha})$ is the compression ratio, i.e. the number of parameters of $\boldsymbol{\alpha}$ divided by number of parameters of the initial architecture $\boldsymbol{\alpha}_{\text{init}}$.

Cao et al. (2019) Cao et al. (2019) maximize Equation (25) by using Bayesian optimization (Section 3.3) with a Gaussian process built using a deep kernel (Wilson et al., 2016). The acquisition function is maximized by random exploration of the search space. Candidates are derived from the initial architecture by randomly removing and shrinking layers as well as adding skip connections. The authors highlight that the consideration of adding skip connection is crucial for the discovery of compressed architectures which are comparable to the initial architecture with respect to classification accuracy.

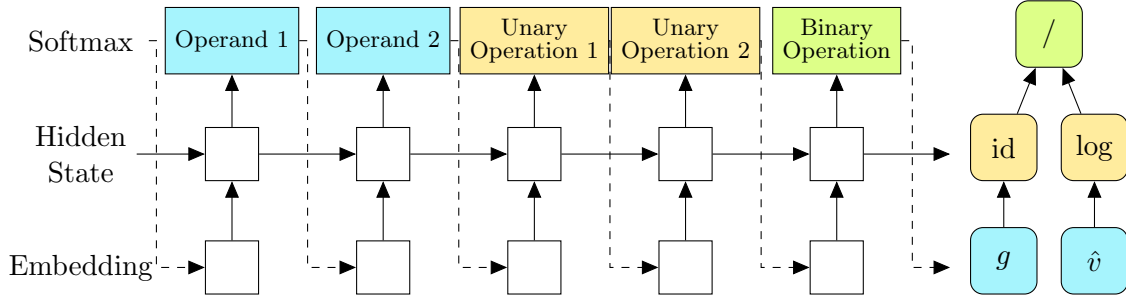


Figure 19: Controller output for one group. Each group consists of two operands and the unary operations applied to it as well as the binary operation which combines the outcome of the unary operations.

7. Other Related Methods

Methodologies originally developed for architecture search have been extended to automate other aspects of deep learning. For instance, Bello et al. (2017) encode a parameter optimization method used in training of a deep learning model by means of a graph structure that defines the update rules akin to classical gradient-based optimization. An update is defined by several blocks as shown in Figure 19. A block consists of two operands, one unary function for each of the operands and a binary function. The unary functions are applied on the operands and the binary function is applied to combine the respective outcomes. The output of this block becomes a possible operand itself and can be used by other blocks. This structure is stacked up to a depth of three. Possible operands include different orders of the gradient (g, g^2, g^3), its bias-corrected running estimates, differently scaled versions of the parameter, random noise, constants, and the sign of the gradient. Additionally the classical updates of Adam and RMSProp optimizer are also encapsulated as operands. Examples for unary functions are the identity function, clipping functions, the sign function, functions that modify the gradient to zero with a certain probability, as well as logarithmic and exponential function. Binary functions include addition, subtraction, multiplication, and division. A reinforcement controller, similar to the one proposed by Zoph and Le (2017) (see Section 3.1), is applied to search for an optimal structure. The main difference in the work of Bello et al. (2017) is that they use Trust Region Policy Optimization (Schulman et al., 2015) instead of REINFORCE for updating the controller parameters. A deep-learning optimizer sampled from the controller is used to train a simple two-layer convolutional neural network for only five epochs and the validation accuracy is used as the reward.

Ramachandran et al. (2018) uses the aforementioned setup and search space structure to search for activation functions. They learn the controller with Proximal Policy Optimization (Schulman et al., 2017). As they seek to search for activation functions, their setup comprises of only one operand which is the input of the activation function and a different set of unary and binary operations. Examples for unary operations are powers of different degree, various sigmoid functions and the sine function. The set of possible binary functions on the input (x_1, x_2) is extended with additional functions like $x_1 \cdot \sigma(x_2)$, maximum and minimum, and a weighted average. The reward is estimated by training a ResNet-20 with

Table 4: Classification error of various architectures with different augmentation strategies. Whenever two numbers are reported, the left one are the results obtained by us then the right one the results reported by Cubuk et al. (2018a). Cubuk et al. (2018a) do not report results for mixup.

Model	Baseline	Cutout	Mixup	Cutout +Mixup	Auto- Augment	Auto- Augment +Mixup
Wide-ResNet (28-10)	3.93/3.87	2.89/3.08	3.24	2.95	2.71/2.68	2.60
Shake-Shake (26 2x32d)	3.73/3.55	3.03/3.02	3.13	3.10	2.59/2.47	2.85
Shake-Shake (26 2x96d)	3.15/2.86	2.50/2.56	2.37	2.15	2.16/1.99	1.84
Shake-Shake (26 2x112d)	3.13/2.82	2.55/2.57	2.37	2.12	2.09/1.89	1.83
PyramidNet +ShakeDrop	2.77/2.67	2.09/2.31	1.90	1.51	1.62/1.48	1.33

the sampled activation function on CIFAR-10 for 10,000 steps. The authors report the discovery of the following activation function

$$\text{swish}(x) = x \cdot \sigma(x), \quad (26)$$

where σ is the logistic function. In their experiments, this particular activation function turned out to be more effective than other commonly used activation functions and has been used by many other authors since. However, it is interesting to note that the discovered function closely resembles the previously proposed swish activation in an earlier work by the authors.

Cubuk et al. (2018a) define a similar controller and train it with Proximal Policy Optimization to optimize data augmentation policies for image classification. The controller predicts five sub-policies, each consists of two augmentation operations and always uses horizontal flipping, random crops and a cutout operation. Examples for augmentation operations are rotation, sheering, control the brightness, contrast and sharpness of the image. For every image, a random sub-policy is selected and applied. The validation accuracy is used as the reward signal. Finally, the best sub-policies of five policies are concatenated to the final augmentation policy which consists of 25 sub-policies.

Experiment We conduct an experiment to better understand the impact of established augmentation schemes and contrast it with the found policies. As part of this we reconducted the experiments by Cubuk et al. (2018a) using their publicly available code. A Wide-ResNet (Zagoruyko and Komodakis, 2016), three versions of Shake-Shake (Gastaldi, 2017) and a PyramidNet (Yamada et al., 2016) with ShakeDrop (Yamada et al., 2018) are trained with different augmentation strategies as detailed in the following. *Baseline* is horizontal flipping and random crops and *cutout* additionally uses cutout (DeVries and

Taylor, 2017). *AutoAugment* extends the cutout policy by augmentation sub-policies found during the search. We extend this setup further by considering mixup (Zhang et al., 2018a) as another augmentation technique in order to see whether the AutoAugment policy can improve over standard state-of-the-art augmentation techniques. As we see in Table 4, the combination of cutout with mixup in many cases is better than using just one of them. Since AutoAugment is cutout plus other augmentation techniques, perhaps it is not surprising to notice a significant lift over cutout. We consider the combination of cutout and mixup as a useful baseline which it is not able to outperform. However, adding mixup to AutoAugment does not hurt the performance in most cases as expected. In cases where the combination of cutout and mixup performed better than AutoAugment, this addition provides a significant improvement resulting in the best scores. We are able to achieve an error of 1.33% on the CIFAR-10 object recognition task. To the best of our knowledge, the only work claiming a lower error on this task is the work by Huang et al. (2018b) which claims 1% error. However, they use a pretrained model on ImageNet which has 557 million parameters, while the PyramidNet used in our experiments has “only” 26.2 million parameters and can be trained on a single GPU. More importantly, we do not use any pretrained weights and train the model from scratch. Finally, we select eight augmentation policies at random and compare them to the discovered AutoAugment policies. The results on Wide-ResNet yield a mean error of 2.81 and ranges from 2.67 to 2.97. Thus, the mean error across eight random policies is worse than AutoAugment, however, some of them are better. We observe a similar behavior for Shake-Shake (26 2x32d) and (26 2x96d) where we obtain a mean error of 2.70 and 2.24, respectively. The errors range from 2.46 to 2.85 and 2.02 to 2.40. Two of the eight policies sampled at random yield better results than AutoAugment for every architecture, yet again, shows the importance to consider random as a baseline. The search space for augmentation policies seems to be defined in such a way that any augmentation policy will perform better than cutout. Only for Wide-ResNet the worst random policy considered by us is marginally worse.

8. Outlook and Future Applications of Neural Architecture Search Methods

In the broader context of automated deep learning, neural architecture search only tackles one component of the pipeline. Moreover, most of the current approaches in neural architecture search only focus on CNNs for solving the task of object recognition and RNN cells for language modeling. However, other interesting areas have already surfaced some of which have been discussed in sections 5 to 7. For example, many recent works address the idea of using knowledge across different datasets to speed up the optimization processes (Section 5). Similarly, the search for architectures under constraints, with multiple objectives, and the automation of model compression has garnered significant attention in the past year (Section 6). The alignment of architecture search algorithms with other graph structures such as parameter optimization algorithms and activation functions or the search for data augmentation are exciting developments (Section 7). Recently there have also been promising developments along other dimensions that address more complex tasks with image data such as object detection and segmentation (Zoph et al., 2018; Tan et al., 2018; Liu et al., 2019a; Weng et al., 2019), as well as works that address safety-critical issues

such as the discovery of architectures that are robust against adversarial attacks (Cubuk et al., 2018b; Sinn et al., 2019). Furthermore, the existing techniques are extended to apply the architecture search for other types of networks. For instance, there are first attempts to automatically optimize the architecture of autoencoders (Suganuma et al., 2018), transformers (So et al., 2019) and graph CNNs (Gao et al., 2019). In general, these techniques are increasingly being used to automate all components of the data science workflow. The big remaining challenge is the joint optimization of all configuration parameters of the deep learning workflow. So far, the individual components have been treated independently with some components still relying on manual configuration. However, for true automation, it is inevitable that all sensitive parameters must be learned or searched for. This includes data augmentation, the various search space hyperparameters, the choice of optimizer, the actual architecture choice and possibly even the final model compression. Moreover, further development of any of these search methods needs to incorporate sample efficiency in its design, an aspect that is currently not handled by neural architecture search.

Given the large amounts of resources dedicated to the research in neural architecture search, it is imperative to question its meaningfulness and usefulness to research. Although, one might hope to get insights into the understanding of the space of deep learning models or perhaps even the interpretability of deep learning in general, so far there have not been any dedicated efforts channelled into this direction. Another important question, which we believe is often overlooked, pertains to the credibility of neural architecture search as a practically useful field of research. It is crucial to emphasize the role of search space designs in the performance of obtained architectures, some of which are heavily inspired by the design of existing architectures. Many architectures originate from the NASNet search space, which were found by various search algorithms. Therefore, a legitimate question is to ask whether the search algorithms have really discovered new or innovative architectures, or whether they owe their success to the design and properties of the NASNet search space. Li and Talwalkar (2019) as well as the inventors of the NASNet search space, partially answer this question by showing that even a random search yields good, if not better, results. Furthermore, these architectures ultimately do not seem to be significantly better than those developed by humans. The human innovations that contribute to the invention of new architectures are difficult to discover given the restrictions of search spaces. This brings us to the question of what to expect from the future developments of architecture search algorithms. If we really seek to find new architectures with innovative elements that advance research in general, then perhaps we have to let go of the restrictive search spaces. While a general automation of deep learning might answer most of these concerns, perhaps for now we need to ask ourselves why an architecture discovered by a search algorithm is better than the choice of an existing architecture in combination with hyperparameter optimization. In terms of usefulness, the answer can not be that each dataset needs its own architecture. Many architecture search papers show that found architectures can be easily transferred to other datasets, even other tasks, which is a well known property of deep learning architectures. So why then should a practitioner commit to the use of a search algorithm for finding a new architecture, in particular since it does not free her from the optimization of the search space hyperparameters?

The general automation of deep learning is still in its infancy, and many of the aforementioned concerns remain unanswered for the time being. However, this remains an exciting domain and the future works will certainly highlight its practical relevance and usefulness.

References

- Anubhav Ashok, Nicholas Rhinehart, Fares Beainy, and Kris M. Kitani. N2N learning: Network to network compression via policy gradient reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL <https://openreview.net/forum?id=B1hcZZ-AW>.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002. URL <http://jmlr.org/papers/v3/auer02a.html>.
- Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=S1c2cvqee>.
- Bowen Baker, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Accelerating neural architecture search using performance prediction. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. URL <https://openreview.net/forum?id=HJqk3N1vG>.
- Friedrich L Bauer. Computational graphs and rounding error. *SIAM Journal on Numerical Analysis*, 11(1):87–96, 1974.
- Irwan Bello, Barret Zoph, Vijay Vasudevan, and Quoc V. Le. Neural optimizer search with reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 459–468, 2017. URL <http://proceedings.mlr.press/v70/bello17a.html>.
- Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and simplifying one-shot architecture search. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 550–559, Stockholm, Sweden, 10–15 Jul 2018. PMLR. URL <http://proceedings.mlr.press/v80/bender18a.html>.
- James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012. URL <http://dl.acm.org/citation.cfm?id=2188395>.
- Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies - A comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002. doi: 10.1023/A:1015059928466. URL <https://doi.org/10.1023/A:1015059928466>.

- Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. SMASH: one-shot model architecture search through hypernetworks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL <https://openreview.net/forum?id=rydeCEhs->.
- Han Cai, Tianyao Chen, Weinan Zhang, Yong Yu, and Jun Wang. Efficient architecture search by network transformation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2787–2794, 2018a. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16755>.
- Han Cai, Jiacheng Yang, Weinan Zhang, Song Han, and Yong Yu. Path-level network transformation for efficient architecture search. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 677–686, 2018b. URL <http://proceedings.mlr.press/v80/cai18a.html>.
- Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA, 2019*. URL <https://openreview.net/forum?id=Hy1VB3AqYm>.
- Shengcao Cao, Xiaofang Wang, and Kris M. Kitani. Learnable embedding space for efficient neural architecture compression. In *Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA, 2019*. URL <https://openreview.net/forum?id=S1xLN3C9YX>.
- Francesco Paolo Casale, Jonathan Gordon, and Nicolo Fusi. Probabilistic neural architecture search. *CoRR*, abs/1902.05116, 2019. URL <http://arxiv.org/abs/1902.05116>.
- Akshay Chandrashekar and Ian R. Lane. Speeding up hyper-parameter optimization by extrapolation of learning curves using previous builds. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, pages 477–492, 2017. doi: 10.1007/978-3-319-71249-9_29. URL https://doi.org/10.1007/978-3-319-71249-9_29.
- Tianqi Chen, Ian J. Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.05641>.
- Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4470–4478, 2017. URL <http://papers.nips.cc/paper/7033-dual-path-networks>.

- Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3123–3131, 2015. URL <http://papers.nips.cc/paper/5647-binaryconnect-training-deep-neural-networks-with-binary-weights-during-propagations>.
- Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018a. URL <http://arxiv.org/abs/1805.09501>.
- Ekin Dogus Cubuk, Barret Zoph, Samuel S. Schoenholz, and Quoc V. Le. Intriguing properties of adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018b. URL <https://openreview.net/forum?id=Skz1zaRLz>.
- Kalyanmoy Deb. Multi-objective optimization. In *Search methodologies*, pages 403–449. Springer, 2014.
- Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017. URL <https://doi.org/10.1109/4235.996017>.
- Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. URL <http://arxiv.org/abs/1708.04552>.
- Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3460–3468, 2015. URL <http://ijcai.org/Abstract/15/487>.
- Jin-Dong Dong, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. Dpp-net: Device-aware progressive search for pareto-optimal neural architectures. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 540–555. Springer, 2018. doi: 10.1007/978-3-030-01252-6_32. URL https://doi.org/10.1007/978-3-030-01252-6_32.
- Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Simple and efficient architecture search for convolutional neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. URL <https://openreview.net/forum?id=H1hymrkDf>.

- Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Efficient multi-objective neural architecture search via lamarckian evolution. In *Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA, 2019*. URL <https://openreview.net/forum?id=ByME42AqK7>.
- Anthony V Fiacco and Garth P McCormick. *Nonlinear programming: sequential unconstrained minimization techniques*, volume 4. Siam, 1990.
- Dario Floreano, Peter Dürri, and Claudio Mattiussi. Neuroevolution: from architectures to learning. *Evolutionary Intelligence*, 1(1):47–62, 2008. URL <https://doi.org/10.1007/s12065-007-0002-4>.
- Yang Gao, Hong Yang, Peng Zhang, Chuan Zhou, and Yue Hu. Graphnas: Graph neural architecture search with reinforcement learning. *CoRR*, abs/1904.09981, 2019. URL <http://arxiv.org/abs/1904.09981>.
- Xavier Gastaldi. Shake-shake regularization. *CoRR*, abs/1705.07485, 2017. URL <http://arxiv.org/abs/1705.07485>.
- David E. Goldberg and Kalyanmoy Deb. A comparative analysis of selection schemes used in genetic algorithms. In Gregory J. E. Rawlins, editor, *Proceedings of the First Workshop on Foundations of Genetic Algorithms. Bloomington Campus, Indiana, USA, July 15-18 1990.*, pages 69–93. Morgan Kaufmann, 1990.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=rkpACe1lx>.
- Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. AMC: automl for model compression and acceleration on mobile devices. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 815–832, 2018. doi: 10.1007/978-3-030-01234-2_48. URL https://doi.org/10.1007/978-3-030-01234-2_48.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL <http://arxiv.org/abs/1503.02531>.
- Tobias Hinz, Nicolás Navarro-Guerrero, Sven Magg, and Stefan Wermter. Speeding up the hyperparameter optimization of deep convolutional neural networks. *International Journal of Computational Intelligence and Applications*, 17(2):1850008, 2018. doi: 10.1142/S1469026818500086. URL <https://doi.org/10.1142/S1469026818500086>.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL <https://doi.org/10.1162/neco.1997.9.8.1735>.

- Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL <http://arxiv.org/abs/1704.04861>.
- Chi-Hung Hsu, Shu-Huan Chang, Da-Cheng Juan, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Shih-Chieh Chang. MONAS: multi-objective neural architecture search using reinforcement learning. *CoRR*, abs/1806.10332, 2018.
- Hanzhang Hu, John Langford, Rich Caruana, Eric Horvitz, and Debadeepta Dey. Macro neural architecture search revisited. In *Workshop on Meta-Learning at NeurIPS 2018, MetaLearn 2018, 3-8 December 2018, Montréal, Canada.*, 2018. URL http://metalearning.ml/2018/papers/metalearn2018_paper16.pdf.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.243. URL <https://doi.org/10.1109/CVPR.2017.243>.
- Gao Huang, Shichen Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Condensenet: An efficient densenet using learned group convolutions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2752–2761. IEEE Computer Society, 2018a. doi: 10.1109/CVPR.2018.00291. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Huang_CondenseNet_An_Efficient_CVPR_2018_paper.html.
- Yanping Huang, Yonglong Cheng, Dehao Chen, HyounkJoong Lee, Jiquan Ngiam, Quoc V. Le, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *CoRR*, abs/1811.06965, 2018b. URL <http://arxiv.org/abs/1811.06965>.
- C-L Hwang and Abu Syed Md Masud. *Multiple objective decision making methods and applications: a state-of-the-art survey*, volume 164. Springer Science & Business Media, 2012.
- Roxana Istrate, Florian Scheidegger, Giovanni Mariani, Dimitrios S. Nikolopoulos, Costas Bekas, and A. Cristiano I. Malossi. TAPAS: train-less accuracy predictor for architecture search. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, 2019*.
- Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. Global Optimization*, 21(4):345–383, 2001. doi: 10.1023/A:1012771025575. URL <https://doi.org/10.1023/A:1012771025575>.
- Kenneth A. De Jong. *Evolutionary computation - a unified approach*. MIT Press, 2006. ISBN 978-0-262-04194-2.
- Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabás Póczos, and Eric P. Xing. Neural architecture search with bayesian optimisation and optimal transport. In

- Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2020–2029, 2018. URL <http://papers.nips.cc/paper/7472-neural-architecture-search-with-bayesian-optimisation-and-optimal-transport>.
- Ye-Hoon Kim, Bhargava Reddy, Sojung Yun, and Chanwon Seo. Nemo : Neuro-evolution with multiobjective optimization of deep neural network for speed and accuracy. In *AutoML Workshop at ICML 2017*, 2017.
- Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with bayesian neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=S11KBYclx>.
- Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings*, volume 4212 of *Lecture Notes in Computer Science*, pages 282–293. Springer, 2006. URL https://doi.org/10.1007/11871842_29.
- Efi Kokiopoulou, Anja Hauth, Luciano Sbaiz, Andrea Gesmundo, Gabor Bartok, and Jesse Berent. Fast task-aware architecture inference. *CoRR*, abs/1902.05781, 2019. URL <http://arxiv.org/abs/1902.05781>.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012. URL <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 0018-9219. doi: 10.1109/5.726791.
- Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search. *CoRR*, abs/1902.07638, 2019. URL <http://arxiv.org/abs/1902.07638>.
- Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:185:1–185:52, 2017. URL <http://jmlr.org/papers/v18/li16-558.html>.
- Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement

- learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1509.02971>.
- Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan L. Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 19–35, 2018a. doi: 10.1007/978-3-030-01246-5_2. URL https://doi.org/10.1007/978-3-030-01246-5_2.
- Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *CoRR*, abs/1901.02985, 2019a. URL <http://arxiv.org/abs/1901.02985>.
- Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical representations for efficient architecture search. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018b. URL <https://openreview.net/forum?id=BJQRKzbA->.
- Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA*, 2019b. URL <https://openreview.net/forum?id=S1eYHoC5FX>.
- Zhichao Lu, Ian Whalen, Vishnu Boddeti, Yashesh D. Dhebar, Kalyanmoy Deb, Erik D. Goodman, and Wolfgang Banzhaf. NSGA-NET: A multi-objective genetic algorithm for neural architecture search. *CoRR*, abs/1810.03522, 2018. URL <http://arxiv.org/abs/1810.03522>.
- Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. Neural architecture optimization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 7827–7838, 2018. URL <http://papers.nips.cc/paper/8007-neural-architecture-optimization>.
- Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=S1jE5L5gl>.
- Jonas Mockus, Vytautas Tiešis, and Antanas Žilinskas. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117–129, 1978.
- Renato Negrinho and Geoffrey J. Gordon. Deeparchitect: Automatically designing and training deep architectures. *CoRR*, abs/1704.08792, 2017. URL <http://arxiv.org/abs/1704.08792>.

- Bernhard Pfahringer, Hilan Bensusan, and Christophe G. Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 743–750, 2000.
- Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4095–4104, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL <http://proceedings.mlr.press/v80/pham18a.html>.
- Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. URL <https://openreview.net/forum?id=Hkuq2EkPf>.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X. URL <http://www.worldcat.org/oclc/61285753>.
- Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V. Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2902–2911, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL <http://proceedings.mlr.press/v70/real17a.html>.
- Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Aging evolution for image classifier architecture search. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, 2019*.
- Matthias Reif, Faisal Shafait, Markus Goldstein, Thomas M. Breuel, and Andreas Dengel. Automatic classifier selection for non-experts. *Pattern Anal. Appl.*, 17(1):83–96, 2014. doi: 10.1007/s10044-012-0280-z. URL <https://doi.org/10.1007/s10044-012-0280-z>.
- G. A. Rummery and M. Niranjan. On-line q-learning using connectionist systems. Technical report, 1994.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL <https://doi.org/10.1007/s11263-015-0816-y>.
- Ashish Sabharwal, Horst Samulowitz, and Gerald Tesauro. Selecting near-optimal learners via incremental data allocation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2007–2015, 2016. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12524>.

- Alejandro Santiago Pineda, Hector Joaquin Fraire Huacuja, Bernabé Dorronsoro, Johnatan E. Pecero, Claudia Gómez Santillán, Juan Javier González Barbosa, and Jos’e Carlos Soto Monterrubio. A survey of decomposition methods for multi-objective optimization. In *Recent Advances on Hybrid Approaches for Designing Intelligent Systems*, pages 453–465. Springer, 2014. doi: 10.1007/978-3-319-05170-3_31. URL https://doi.org/10.1007/978-3-319-05170-3_31.
- Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4053–4061, 2016. URL <http://papers.nips.cc/paper/6304-convolutional-neural-fabrics>.
- Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605. URL <https://doi.org/10.1109/TNN.2008.2005605>.
- Bernhard Schölkopf, Alexander J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000. doi: 10.1162/089976600300015565. URL <https://doi.org/10.1162/089976600300015565>.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1889–1897, 2015. URL <http://jmlr.org/proceedings/papers/v37/schulman15.html>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL <http://arxiv.org/abs/1707.06347>.
- Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093. URL <https://doi.org/10.1109/78.650093>.
- Christian Sciuto, Kaicheng Yu, Martin Jaggi, Claudiu Musat, and Mathieu Salzmann. Evaluating the search phase of neural architecture search. *CoRR*, abs/1902.08142, 2019. URL <http://arxiv.org/abs/1902.08142>.
- Mathieu Sinn, Martin Wistuba, Beat Buesser, Maria-Irina Nicolae, and Minh Tran. Evolutionary search for adversarially robust neural networks. *Safe Machine Learning Workshop at ICLR 2019, New Orleans, Louisiana, USA*, 2019.
- Sean C. Smithson, Guang Yang, Warren J. Gross, and Brett H. Meyer. Neural networks designing neural networks: multi-objective hyper-parameter optimization. In *ICCAD*, page 104. ACM, 2016.
- Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages

- 2960–2968, 2012. URL <http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms>.
- David So, Quoc Le, and Chen Liang. The evolved transformer. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5877–5886, 2019. URL <http://proceedings.mlr.press/v97/so19a.html>.
- Masanori Suganuma, Shinichi Shirakawa, and Tomoharu Nagao. A genetic programming approach to designing convolutional neural network architectures. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2017, Berlin, Germany, July 15-19, 2017*, pages 497–504, 2017. doi: 10.1145/3071178.3071229. URL <https://doi.org/10.1145/3071178.3071229>.
- Masanori Suganuma, Mete Ozay, and Takayuki Okatani. Exploiting the potential of standard convolutional autoencoders for image restoration by evolutionary search. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4778–4787, 2018. URL <http://proceedings.mlr.press/v80/suganuma18a.html>.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 0262193981. URL <http://www.worldcat.org/oclc/37293240>.
- Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. *CoRR*, abs/1807.11626, 2018. URL <http://arxiv.org/abs/1807.11626>.
- Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.
- Linnan Wang, Yiyang Zhao, Yuu Jinnai, Yuandong Tian, and Rodrigo Fonseca. Alphax: exploring neural architectures with deep neural networks and monte carlo tree search. *CoRR*, abs/1903.11059, 2019. URL <http://arxiv.org/abs/1903.11059>.
- Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King’s College, Cambridge, UK, May 1989. URL http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf.
- Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019. doi: 10.1109/ACCESS.2019.2908991. URL <https://doi.org/10.1109/ACCESS.2019.2908991>.
- Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL <https://doi.org/10.1007/BF00992696>.
- Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence*

- and Statistics, *AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 370–378, 2016. URL <http://proceedings.mlr.press/v51/wilson16.html>.
- Martin Wistuba. Deep learning architecture search by neuro-cell-based evolution with function-preserving mutations. In *ECML/PKDD (2)*, volume 11052 of *Lecture Notes in Computer Science*, pages 243–258. Springer, 2018a. URL <http://www.ecmlpkdd2018.org/wp-content/uploads/2018/09/108.pdf>.
- Martin Wistuba. Practical deep learning architecture optimization. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 263–272, 2018b. doi: 10.1109/DSAA.2018.00037. URL <https://doi.org/10.1109/DSAA.2018.00037>.
- Martin Wistuba and Tejaswini Pedapati. Inductive transfer for neural architecture optimization. *CoRR*, abs/1903.03536, 2019. URL <http://arxiv.org/abs/1903.03536>.
- Catherine Wong, Neil Houlsby, Yifeng Lu, and Andrea Gesmundo. Transfer learning with neural automl. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 8366–8375, 2018. URL <http://papers.nips.cc/paper/8056-transfer-learning-with-neural-automl>.
- Lingxi Xie and Alan L. Yuille. Genetic CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1388–1397. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.154. URL <https://doi.org/10.1109/ICCV.2017.154>.
- Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995, 2017. doi: 10.1109/CVPR.2017.634. URL <https://doi.org/10.1109/CVPR.2017.634>.
- Saining Xie, Alexander Kirillov, Ross B. Girshick, and Kaiming He. Exploring randomly wired neural networks for image recognition. *CoRR*, abs/1904.01569, 2019a. URL <http://arxiv.org/abs/1904.01569>.
- Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. SNAS: stochastic neural architecture search. In *Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA, 2019b*. URL <https://openreview.net/forum?id=rylqooRqK7>.
- Chao Xue, Junchi Yan, Rong Yan, Stephen M. Chu, Yonggang Hu, and Yonghua Lin. Transferable automl by model sharing over grouped datasets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- Yoshihiro Yamada, Masakazu Iwamura, and Koichi Kise. Deep pyramidal residual networks with separated stochastic depth. *CoRR*, abs/1612.01230, 2016. URL <http://arxiv.org/abs/1612.01230>.

- Yoshihiro Yamada, Masakazu Iwamura, and Koichi Kise. Shakedrop regularization. *CoRR*, abs/1802.02375, 2018. URL <http://arxiv.org/abs/1802.02375>.
- Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016. URL <http://www.bmva.org/bmvc/2016/papers/paper087/index.html>.
- Chris Zhang, Mengye Ren, and Raquel Urtasun. Graph hypernetworks for neural architecture search. In *Proceedings of the International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA, 2019*. URL <https://openreview.net/forum?id=rkgW0oA9FX>.
- Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018a. URL <https://openreview.net/forum?id=r1Ddp1-Rb>.
- Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6848–6856, 2018b. doi: 10.1109/CVPR.2018.00716. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_ShuffleNet_An_Extremely_CVPR_2018_paper.html.
- Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3900–3908, 2017. doi: 10.1109/CVPR.2017.415. URL <https://doi.org/10.1109/CVPR.2017.415>.
- Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Practical block-wise neural network architecture generation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2423–2432, 2018. doi: 10.1109/CVPR.2018.00257. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhong_Practical_Block-Wise_Neural_CVPR_2018_paper.html.
- Yanqi Zhou, Siavash Ebrahimi, Sercan Ömer Arik, Haonan Yu, Hairong Liu, and Greg Diamos. Resource-efficient neural architect. *CoRR*, abs/1806.07912, 2018. URL <http://arxiv.org/abs/1806.07912>.
- Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=r1Ue8Hcxg>.
- Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City,*

UT, USA, June 18-22, 2018, pages 8697–8710, 2018. doi: 10.1109/CVPR.2018.00907. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zoph_Learning_Transferable_Architectures_CVPR_2018_paper.html.