
NPENAS: Neural Predictor Guided Evolution for Neural Architecture Search

Chen Wei

Xidian University
weichen_3@stu.xidian.edu.cn

Chuang Niu

Xidian University
niuchuang@stu.xidian.edu.cn

Yiping Tang

Xidian University
tangyiping@stu.xidian.edu.cn

Jimin Liang

Xidian University
jimleung@mail.xidian.edu.cn

Abstract

Neural architecture search (NAS) is a promising method for automatically finding excellent architectures. Commonly used search strategies such as evolutionary algorithm, Bayesian optimization, and Predictor method employs a predictor to rank sampled architectures. In this paper, we propose two predictor based algorithms NPUBO and NPENAS for neural architecture search. Firstly we propose NPUBO which takes a neural predictor with uncertainty estimation as surrogate model for Bayesian optimization. Secondly we propose a simple and effective predictor guided evolution algorithm(NPENAS), which uses neural predictor to guide evolutionary algorithm to perform selection and mutation. Finally we analyse the architecture sampling pipeline and find that mostly used random sampling pipeline tends to generate architectures in a subspace of the real underlying search space. Our proposed methods can find architecture achieves high test accuracy which is comparable with recently proposed methods on NAS-Bench-101 and NAS-Bench-201 dataset using less training and evaluated samples. Code will be publicly available after finish all the experiments.

1 Introduction

Deep learning with properly designed neural architectures has achieved significant advancement in different learning tasks, e.g. image classification [1, 2], object detection [3, 4], segmentation [5, 6] and language models [7]. Neural architecture directly affects the performance of deep learning and numerous researchers have devoted to design more efficient neural architectures. Neural architecture design requires expert knowledge and is laborious and time consuming. Therefore, automatic design of neural architecture has emerged and sprung up [8, 9, 10, 11, 12].

The goal of neural architecture search (NAS) is to find a architecture with minimum validation error in the search space with minimum time cost. Reinforcement learning [8, 13, 14], evolutionary algorithms [15, 16, 17, 18], Bayesian optimization [19, 20, 21, 22, 23, 24], gradient algorithm [9, 11, 25, 26] and predictor based algorithms [27] are the commonly used methods for architecture search. From the Bayesian optimization perspective, the validation error is a function f which takes an architecture as input and outputs the architecture's validation error. Since training an architecture is time consuming, in real application it's impossible to build the function f by training all the architectures in search space. Bayesian optimization [19, 20] uses a surrogate model to describe the distribution of f . Existing Bayesian optimization algorithms for NAS [21, 22, 24] take a Gaussian process or ensemble neural networks as surrogate model, which makes the surrogate model computational intensive and can not be trained end-to-end. In this paper, we propose an neural predictor with uncertainty

estimation which takes an architecture as input and output the mean and variance directly as surrogate model. Our proposed neural predictor avoids the calculating of inverse matrix and can be trained end-to-end. It illustrates top performance compared with the existing Bayesian optimization algorithms for NAS.

The procedure of Bayesian optimization for neural architecture is quite complex. In order to find the best architecture in minimum search steps, an acquisition function is needed to find the potential best locations for evaluation. Evolutionary algorithm is most commonly used to solve the acquisition function. To simplify the search procedure, a predictor guided evolutionary algorithm(NPENAS) for NAS is proposed. We design a spatial graph neural network based neural predictor to evaluate the performance of neural architecture. We evaluated the neural predictor guided evolutionary algorithm 600 trials and the average mean accuracy of searched architectures is 94.14% using only 150 samples in NAS-Bench-101 dataset [28], which is close to the best performance of 94.23% in the *ORACLE* setting [27].

The NAS-Bench-101 dataset [28] provides an implementation for architecture sampling by randomly generating adjacency matrix and node operations. We call this method as default sampling pipeline and demonstrate that it tends to generate architectures only in a subspace of NAS-Bench-101. Instead of using the default sample pipeline, we propose to sample the architectures directly from NAS-Bench-101 and prove that it is beneficial for performance improvement.

Our contribution can be summarized as follows:

- We propose an neural predictor with uncertainty estimation as the surrogate model for Bayesian optimization(NPUBO) for NAS. The predictor can be trained end-to-end in the Bayesian optimization framework and avoids the intensive inverse matrix calculation. The proposed method achieves best or comparable performance compared with the state-of-the-art Bayesian optimization NAS methods on NAS-Bench-101 and NAS-Bench-201 dataset.
- We propose a predictor guided evolutionary algorithm(NPENAS) for NAS which is simple compared with Bayesian optimization and achieves the best or comparable performance on NAS-Bench-101 and NAS-Bench-201 dataset.
- We investigate the drawback of the default architecture sampling pipeline and demonstrate that sampling architectures directly from the search space is beneficial for performance improvement.
- We design a new macro architecture search task *NAS-Bench-101-Macro* which utilize the cell level search space of NAS-Bench-101 and allow the predefined macro architecture in NAS-Bench-101 to use different cells. We evaluated our proposed methods on this task and an architecture with the top performance comparable with the architecture in NAS-Bench-101 was found. In addition, we collected the training details and constructed a new dataset NAS-Bench-101-Macro which is publicly available.

2 Related Works

2.1 Neural Architecture Search

NAS [29] employs some kind of search strategy to automate the architecture engineering. Mostly used search strategies include reinforcement learning [8, 13], gradient optimization [9, 11, 25], evolutionary algorithms [15, 16, 17], Bayesian optimization [19, 20, 21, 22, 23] and predictor-based method [27]. Reinforcement learning and gradient optimization optimize an agent which can learn how to sample best architectures in the search space. Evolutionary algorithms use the validation accuracy to rank the mutated architectures and push the top performance architecture in the population. Bayesian optimization and predictor-based method adopt a predictor to estimate the architecture accuracy.

Bayesian optimization uses a probabilistic surrogate model to predict the accuracy and uncertainty of an architecture. There are two steps in Bayesian optimization. First, a probabilistic surrogate model is built as a prior confidence of the object function. Next, an acquisition function is adopt to find the potential optimal position. In neural architecture search, Bayesian optimization mostly uses Gaussian process as the probabilistic surrogate model. NASBot [21] adopted Gaussian process as the surrogate model and proposed a distance metrics to calculate the kernel function. NASGBO [22]

and BONAS [24] used a graph neural network with complex local and global properties of graph nodes and edges to represent the neural architecture and built a Bayesian linear regression layer to output the uncertainty. BONAS [24] first sampled architecture to train an architecture embedding and used this embedding to build a design matrix which was needed by Bayesian linear regression. All of the methods of NASdBot [21], NASGBO [22] and BONAS [24] require computationally intensive matrix inversion operation. To avoid the kernel function calculation, BANANAS [23] employed a collection of identical neural networks to predict the accuracy of sampled architectures. However, the ensemble of several neural networks prohibits end-to-end training and leads to a new hyperparameter which is hard to determine.

In neural architecture search, Bayesian optimization method employs an acquisition function such as expected improvement (EI) [30], entropy search (ES) [31], upper confidence bound (UCB) [32] or Thompson sampling [33] to sample the potential optimal architecture in the search space. The acquisition function can better balance exploration and exploitation. NASBot [21], NASGBO [22] and BONAS [24] used the EI acquisition function, BANANAS [23] adopted independent Thompson sampling [23] acquisition function. As the architecture search space is too large and architectures in the search space exhibits locality [28], evolutionary algorithms are commonly used to optimize the acquisition function.

Neural predictor based methods employ a neural predictor to estimate the validation accuracy of neural architectures, and then use it to predict all the architectures in the search space. The appropriate prediction accuracy of the neural predictor is quite important. Too low or too high accuracy of the predictor may cause performance drop. Therefore, some neural predictor methods proposed two stage cascade model to predict the validation accuracy [27].

Compared with existing methods, we propose an neural predictor with uncertainty estimation as surrogate model which avoids the inverse matrix calculation and can be trained end-to-end. The proposed predictor outputs mean and variance directly. We also propose a simple neural predictor guided evolutionary algorithm (NPENAS) for architecture search, which contains only one neural predictor to rank the sampled architectures. Compared with the existing neural predictor algorithm [27] on NAS-Bench-101 dataset, which used a cascade graph neural network model, 172 training architectures and all the architectures in the search space, our proposed neural predictor only use 150 training architectures and 1500 mutated architectures to find the best architectures.

2.2 Neural Architecture Encoding

Neural architecture search always involves an architecture encoder which embeds the direct acyclic graph (DAG) architecture into a D -dimensional vector as the neural architecture representation. The D -dimensional vector is used by the performance predictor to estimate the validation accuracy of architectures. Architecture encoders fall into roughly three categories, recurrent network (RNN) encoder [34], vector encoder [23] and graph encoder [21, 22, 27]. RNN encoder takes a string sequence described architecture as input, and uses the hidden state of recurrent network as architecture embedding. Vector encoder combines the unrolled neural architecture adjacency matrix and operations or employs path-based encoding method [23] to represent the neural architectures and then uses a fully connected network (FCN) for architecture embedding. However, path-based encoding is not suitable for marco level search space because its encoding vector will increase exponentially [23]. Graph encoder represents the neural architecture via an adjacency matrix, node features and edge features, and then uses a graph convolution network to output vector formed embedding. Previous graph encoders [21, 22, 27, 24] adopt a spectral graph neural network and use complex edge and node features. In this paper, we encode the graph architecture via a spatial graph neural network as it can process the direct acyclic graph from a message passing [35] perspective.

2.3 Graph Neural Network

Graph neural network (GNN) achieves significant progress in non-euclideanly defined data structures, such as graphs and networks [36, 37]. GNN can be classified into two categories, the spectral-based methods [38, 39, 40] and spatial-based approaches [41, 42]. The spectral-based methods use spectral graph theory to build a real symmetric positive semidefinite matrix, i.e. graph Laplacian matrix. The orthonormal eigenvectors is known as the graph Fourier modes, and their associated real nonnegative eigenvalues as the frequencies [39]. The graph signals are filtered in the Fourier domain. ChebConv

[39] used the polynomial parametric filter and Chebyshev expansion [43] for fast filtering. GCN [40] limited the polynomial filter to linear model and employed renormalization trick to build a localized first-order approximation of spectral graph convolution. The spatial-based approaches leverages the spatial location of node features to conduct convolution. GraphSAGE [41] defined several aggregator functions to aggregate node features in the neighborhood of current node and used a pooling aggregator to generate the final graph embedding. GINConv [42] presented a theoretical framework for analyzing the expressive power of GNN and proposed a simple but powerful GNN architecture *Graph Isomorphism Network (GIN)* [42]. NASGBO [22], BONAS [24] and some neural predictors [27] methods use GCN [40] to embed the neural architecture. As the neural architecture is a directed graph, but GCN [40] assumes undirected graphs, the neural predictor proposed in [27] built two different graph convolutions using the normal adjacency matrix and the transpose of normal adjacency matrix respectively. In this paper, we use GNN to extract the neural architecture’s structural information. The spatial graph neural network GIN [42] is used to embed the neural architecture from the message passing perspective [35]. We only aggregate node features in the forward direction which is suitable for neural architectures.

3 Methodology

3.1 Problem Formulation

In neural architecture search, the search space defines the allowed neural network operations and connections between different operations. Most commonly used search space is the cell based search space [13, 44, 45, 14, 25, 46]. A cell is defined as a DAG that is comprised of an input node, an output node and some operation nodes. The final neural network is built by repeatedly stacking the same cells on top of each other sequentially [23]. The goal of neural architecture search is to find the best architecture s^* in search space S by optimizing a specific performance predictor f ,

$$s^* = \arg \min_{s \in S} f(s), \quad (1)$$

where predictor f takes a neural architecture s as input and outputs the performance measure (e.g. validation error for image classification). Each architecture in search space S is defined as a DAG $s = (V, E)$, where $V = \{v_i\}_{i=1:N}$ is the set of nodes representing operations in the neural network and $E = \{r_k, s_k\}_{k=1:M}$ represents the set of edges connecting different operations. Operation in each node is represented by a H -dimensional one-hot encoded node feature $h \in \mathbb{R}^H$, where H equals to the quantity of allowed operations in the search space.

3.2 Neural Predictor with Uncertainty Estimation

Bayesian optimization takes the architecture performance predictor f as a black box function and uses a surrogate model to represent the prior confidence. Gaussian process is the mostly used surrogate model, which is defined as

$$f(s) \sim GP(\mu(s), k(s, s')), \quad s, s' \in S. \quad (2)$$

Gaussian process needs a distance measure to calculate the kernel function k and the calculation of kernel function includes the computational intensive operation of matrix inversion. In order to eliminate the matrix inversion and train the surrogate model end-to-end, we assume the neural architectures in the search space is independent identically distributed. Thereafter, the Gaussian process can be deduced to a collection of independent Gaussian random variables as

$$f(s) \sim N(\mu, \sigma), \quad s \in S. \quad (3)$$

We propose a neural predictor with uncertainty estimation as the surrogate model to describe the priori confidence of the performance predictor f , which takes a neural architecture s as input and outputs the estimation of the neural architecture’s validation error μ together with its uncertainty σ .

The neural architecture is encoded into graph representation. As exemplified in Fig.1, the connections between nodes are represented by an adjacency matrix and the operations of nodes are represented as one-hot vectors.

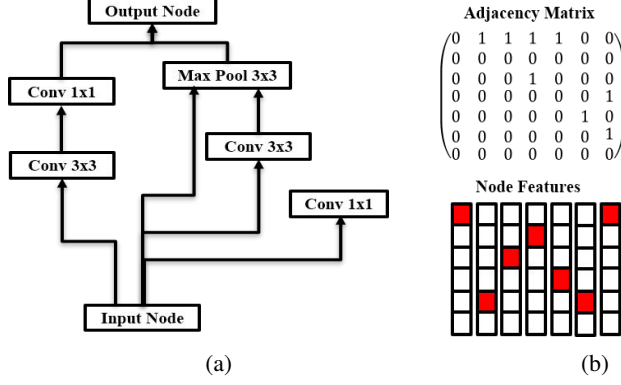


Figure 1: Graph representation of neural architecture. (a) An example of cell architecture. (b) Corresponding adjacency matrix and node features.

Our proposed neural predictor surrogate model contains two parts, the architecture encoder and the uncertainty performance predictor, as illustrated in Fig. 2.

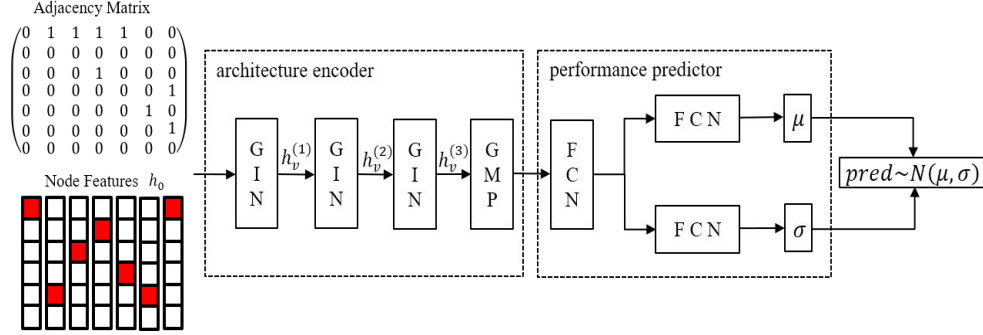


Figure 2: Neural predictor with uncertainty estimation

Three spatial graph neural network GINs [42] are sequentially connected to embed the neural architecture. Each GIN uses a *multi-layer perceptrons* (MLPs) to iteratively update a node by aggregating features of its neighbors,

$$h_v^{(k)} = MLP^{(k)}((1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)}), k = 1, 2, 3, \quad (4)$$

where $h_v^{(k)}$ is the k^{th} level feature of node v , $\epsilon^{(k)}$ is a fixed scalar or learnable parameter, $N(v)$ is the set of input nodes connected to v . The output of the final GIN is averaged over nodes by a *global mean pool* (GMP) layer to get the final embedding feature. The embed feature passes through several fully connect layers to generate the estimation of mean μ and variance σ of the validation error.

We use K randomly sampled neural architectures $\{s_1, s_2, \dots, s_K\}$ in the search space and their corresponding validation accuracy $\{y_1, y_2, \dots, y_K\}$ as the training dataset $D = \{(s_i, y_i), i = 1, 2, \dots, K\}$ to train the neural predictor, which is denoted as G hereafter. Instead of sampling from a continuous performance predictor f , we sample discrete values of f_{dis} from the multinomial independent Gaussian distribution as

$$f_{dis} \sim \prod_{i=1}^N N(G(s_i)), \quad \mu_i, \sigma_i = G(s_i), s_i \in S. \quad (5)$$

We use *maximum likelihood estimate* (MLE) loss to optimize the neural predictor. Denoting the parameters of G as w , the loss is defined as

$$w^* = \arg \max_w \prod_i^N P(y_i | \mu_i, \sigma_i), \quad \mu_i, \sigma_i = G(s_i), (s_i, y_i) \in D. \quad (6)$$

Bayesian optimization uses an acquisition function to find the potential best architectures. We adopt the *Thompson Sampling* (TS) [33] as our acquisition function. As we assume the neural architecture in search space is identical independent distributed, TS only samples the validation error prediction at given neural architectures from the surrogate model. We employ evolutionary algorithm to optimize the acquisition function, which was also adopted in NABOT [21], BANANAS [23] and BONAS [24].

As neural predictor with uncertainty estimation is taken as a surrogate model for Bayesian optimization, we denote this method as NPUBO. The complete procedure of NPUBO is shown in Algorithm.1.

Algorithm 1 Neural predictor with uncertainty estimation for Bayesian optimization (NPUBO)

Input: Search space S , number of initial training samples $init_num$, randomly sampled neural architectures and corresponding validation errors $D = \{(s_i, y_i), i = 1, 2, \dots, init_num\}$, neural predictor G , number of total training samples $total_num$, number of mutated architectures $mutate_num$, step size k .

Output: $s^* = \arg \min_{i=1,2,\dots,total_num} f(s_i)$.

For n from $init_num$ **to** $total_num$ **step size** k ,

1. Train neural predictor G using the dataset $D = \{(s_i, y_i), i = 1, 2, \dots, n\}$.
2. Perform selection and mutation on D to generate $mutate_num$ child architectures $M = \{(s_m), m = 1, 2, \dots, mutate_num\}$.
3. Using G to predict the mean μ and variance σ of the mutated child architectures M .
4. Sample discrete performance predictor values f_{dis} from the multinomial independent gaussian distribution $f_{dis} \sim \prod_{i=1}^{mutate_num} N(y_i | \mu_i, \sigma_i)$.
5. Select the top k architectures from M to generate the sampled dataset $M_k = \{(s_m), m = 1, 2, \dots, k\}$.
6. Train the sampled architectures in M_k to get their validation errors.
7. Add dataset M_k to the training dataset D .
8. Set $n = n + k$.

End For

3.3 Neural Predictor Guided Evolutionary Algorithm

To simplify the complex procedure of Bayesian optimization based NAS. We propose a neural predictor guided evolutionary algorithm for architecture search. The neural predictor guided evolutionary algorithm contains a neural predictor which takes neural architecture as input and output the performance prediction. Evolutionary algorithm utilizes performance prediction to perform selection and mutation. We use a spatial graph neural network to embed architectures which is the same as the neural predictor in Section 3.2. The predictor outputs performance prediction and the architecture of neural predictor is shown in Fig.3.

We take training dataset $D = \{(s_i, y_i), i = 1, 2, \dots, K\}$ and use *Mean Square Error*(MSE) as loss function to train the neural predictor G and the loss function is shown in Eqn.7. The parameters of neural predictor is denoted as w .

$$w^* = \arg \min_w \sum_i^K (G(s_i) - y_i)^2, \quad (s_i, y_i) \in D \quad (7)$$

The neural predictor guided evolutionary algorithm is shown in Algorithm.2.

3.4 Architectures Sampling Pipeline

Mostly used random architecture sampling pipeline on NAS-Bench-101 [28] dataset is shown in Fig.4 which is provided by NAS-Bench-101 [28]. This pipeline first randomly generate a 7×7 adjacency

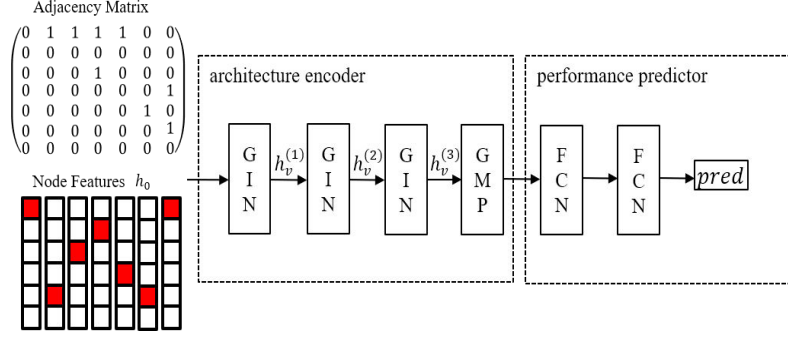


Figure 3: Neural predictor used by NPENAS

Algorithm 2 Neural Predictor Guided Evolutionary (NPENAS)

Input: Search space S , number of initial training samples $init_num$, randomly sampled neural architectures and corresponding validation errors $D = \{(s_i, y_i), i = 1, 2, \dots, init_num\}$, neural predictor G , number of total training samples $total_num$, number of mutated architectures $mutate_num$, step size k .

Output: $s^* = \arg \min_{i=0,1,\dots,total_num} f(s_i)$.

For n from $init_num$ **to** $total_num$ **step size** k ,

1. Train neural predictor G using dataset $D = \{(s_i, y_i), i = 1, 2, \dots, n\}$.
2. Perform selection and mutation on D to generate $mutate_num$ child architectures $M = \{(s_m), m = 1, 2, \dots, mutate_num\}$.
3. Using G to predict performance of mutated child architectures M .
4. Select the top k architectures from M to generate the sampled dataset $M_k = \{(s_m), m = 1, 2, \dots, k\}$.
5. Train the sampled architectures in M_k to get their validation errors.
6. Add dataset M_k to the training dataset D .
7. Set $n = n + k$.

End For

matrix s_{adj} , as this adjacency matrix may not a valid architecture in NAS-Bench-101, adjacency matrix s_{adj} is pruned to a new adjacency matrix s_{adj_n} which is in NAS-Bench-101 dataset and take s_{adj_n} and its corresponding operations as key to query validation accuracy y_i and test accuracy. After finding validation accuracy and test accuracy, neural architecture search algorithms use (s_{adj}, y_i) to search best architecture. We call this method as default sampling pipeline and sampled architectures as default sampled architectures. This pipeline has two negative effect for neural architecture search:

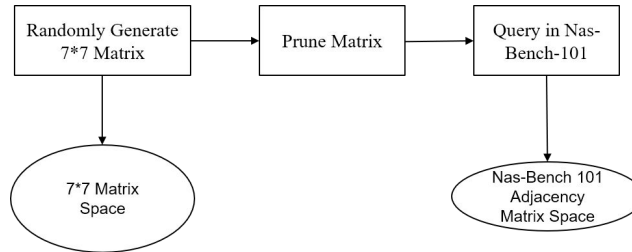


Figure 4: Default sampling pipeline

1. After pruning, many randomly generated different 7×7 adjacency matrixes are mapped to the same architecture in search space. This will make some vector encoder based neural architecture search algorithms hard to learn.

2. This pipeline tends to generate architectures in a subspace of the real underlying search space.

The first effect is explicitly, we mainly analysis the second effect. Path-based encoding [23] is a vector based architecture representation method. Path-based encoding encodes the input to output paths in a cell. Each input to output path has a unique value, so this method can eliminate isolate nodes. The cell level search space in NAS-Bench-101 dataset contains 364 unique paths. So each cell architecture can be represented by a 364- D vector. For more detailed information can be found in BANANAS [23]. If a path appears in cell then the corresponding position is set to 1 otherwise set to 0. As path-based encoding take a cell as composed by several input to output paths, so the distribution of paths can be used to represent the sampled architecture’s distribution. We use the default sampling pipeline to randomly sample $5k$ architectures and plot the distribution of paths to represent the default sampled architecture’s distribution, as shown in Fig.5b. We also use path-based encoding to encode all cell architectures in NAS-Bench-101 search space and plot the distribution of paths, we call this distribution as ground truth distribution, as show in Fig.5a.

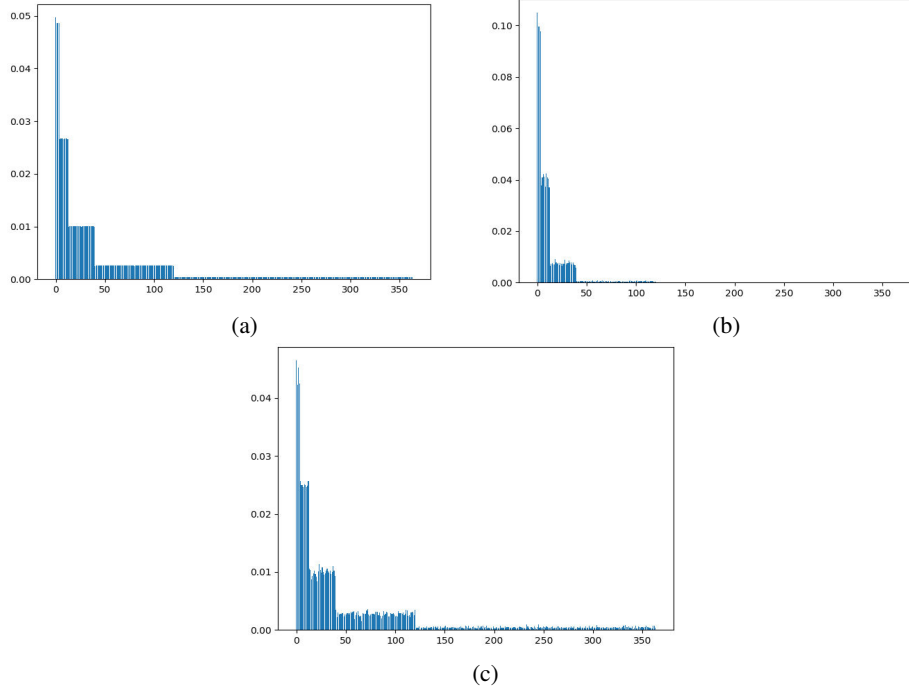


Figure 5: The architecture’s paths distributions. 5a is the ground truth paths distribution of NAS-Bench-101. 5b is the paths distribution of default sampling pipeline and 5c is the paths distribution of new sampling pipeline.

Compare the paths distribution of default sampled architectures and the ground truth paths distribution we find paths in default sampled architectures tend to have low path value. We also calculated the *KL-divergence* of this two distributions. The *KL-divergence* is defined in Eqn.8 and *KL-divergence* of the default sampled architecture’s paths distribution and the ground truth paths distribution is 0.3115. As *KL-divergence* does not satisfies commutative law, we take the default sampled architecture’s paths distribution as first term and the ground truth paths architecture’s distribution as second term.

$$D_{KL}(A \parallel B) = \sum_i P_A(x_i) \log\left(\frac{P_A(x_i)}{P_B(x_i)}\right) \quad (8)$$

In order to eliminate the above two negative effect, we sample architectures directly from search space. We call this pipeline new sampling pipeline and sampled architectures as new sampled architectures. Using new sampling pipeline, we randomly sample $5k$ architectures and plot the architecture’s paths distribution in Fig.5c and the *KL-divergence* of new sampled architecture’s paths distribution and the ground truth architecture’s paths distribution is 0.0127 which is approximately 24.5x more similar than default sampling pipeline.

3.5 NAS-Bench-101-Macro

For the sake of reduce search space, cell level architecture search uses a predefined macro architecture as final training architecture which is always composed by sequentially stack the same searched cell. This paradigm can find good enough architectures, but there may be some better architectures out of this paradigm *e.g.* architectures with sequentially stacked different cells. In order to verify the search ability of our proposed methods and analysis the performance of architectures composed by sequentially stacking different cells, we define an open domain search task *NAS-Bench-101-Macro* which is based on the larges benchmark dataset NAS-Bench-101. *NAS-Bench-101-Macro* uses predefined macro architecture as NAS-Bench-101 but can sequentially stack different cells and the cell search space is the same as NAS-Bench-101. As the predefined macro architecture in NAS-Bench-101 contains 9 cells and the search space of NAS-Bench-101 is $432k$, the search space of *NAS-Bench-101-Macro* is $(432k)^9$.

We represent the macro architecture using a adjacency matrix and node features. As the macro architecture is composed by stacking cells, one cell’s output is the subsequent cell’s input. We build an adjacency matrix to represent this relationship and use node feature to represent each node’s operation which is illustrated in Fig.6. This adjacency matrix and node features is taken as input to our proposed neural predictors.

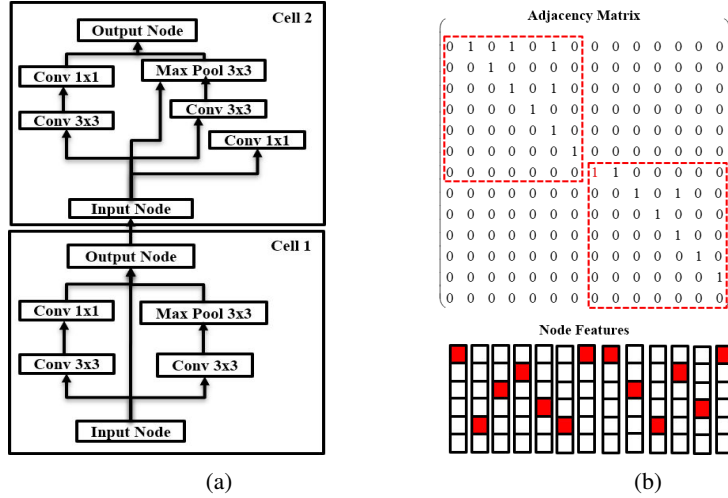


Figure 6: Two cell neural architectures represented by adjacency matrix and node features. The red 1 in 6b means the connection between cell 1 and cell 2 (The output of cell 1 as input to cell 2).The top-left dotted rectangle is cell architecture 1 and bottom-right dotted rectangle is cell architecture 2.

We gather commonly used information during architecture search, and build a dataset called *NAS-Bench-101-Macro* which is publicly available¹. This dataset contains the macro architecture’s structure information, training details, validation accuracy and testing accuracy. The items is shown in the following list:

- Lists of cells in this macro architecture.
- Training loss.
- Training accuracy.
- Validation accuracy.
- Testing accuracy.

4 Experiments and Analysis

In this section, we report the empirical performances of our proposed NPUBO and NPENAS. We first demonstrate superiority of our proposed neural predictors over existing method on correctly predict

architecture’s validation error. Secondly, we evaluate the effectiveness of our proposed methods on closed and open domain search tasks. Finally, we conduct some transfer learning analysis and ablation studies.

4.1 Prediction Analysis

Setup. We compare the mean average percent error of our proposed neural predictor with meta neural network in BANANAS [23] on testing set under different training set size and architecture generation pipeline. We compare two different architecture generation pipelines, the default sampling pipeline and new sampling pipeline. Meta neural network in BANANAS [23] adopts path-based encoding or unrolled adjacency matrix with concatenated one-hot encoded operation vector as architecture representation. The training set size is 20, 100 and 500. We use 5 ensemble meta neural network and each neural network has 10 layers and each layer has width 20. Each meta neural network is trained for 200 epochs the full details can found in BANANAS [23]. NAS-Bench-101 [28] is utilize to perform comparison.

Dataset. NAS-Bench-101 [28] is the largest benchmark dataset for NAS which contains 423k unique architectures for image classification. All the architectures is trained and evaluated multiple times on CIFAR-10. NAS-Bench-101 employs a cell level search space which is defined as a direct acyclic graph. The maximum allowed nodes is 7 and edges is 9. Three operations are allowed to use. The predefined macro architecture contains sequentially connected normal cells and reduction cells and the same cell architecture is used in normal cells. All the architectures trained on CIFAR-10 with a annealed cosine decay learning rate. Training is performed via RMSProp [47] on the cross-entropy loss with $L2$ weight decay. The best architecture found in this dataset achieved a mean test error of 94.32%. The architecture with the highest validation accuracy attain mean test accuracy of 94.23% with corresponding mean validation accuracy 95.06%.

Details of Neural Predictor. As shown in Fig.2 and 3 our proposed neural predictors have three sequentially connected GIN [42] layers with hidden layer size 32. The output of the last GIN[42] layer sequentially passes through a global mean pooling layer and 2 fully connected layers with hidden layer size 16. CELU[48] layer and batch normalization layer[49] is inserted after each GIN and fully connect layer. A drop out layer is used after the first fully connect layer with dropout rate 0.1. Neural predictor used by NPENAS replace CELU layer to ReLU. We employ Adam optimizer [50] with initial learning rate $5e-3$ and weight decay $1e-4$. Neural predictor with uncertainty estimation is trained 1000 epochs with batch size 16. Neural predictor used by NPENAS is trained 300 epochs and utilize $1e-3$ as initial learning rate. We use a consine schedule [51] gradually decay the learning rate to zero.

Analysis of Resutls. From Table1 we find that our proposed neural predictors can achieve the best mean accuracy when training set size is 100 and 500 which means our proposed methods can better embed the input neural architectures. Our proposed neural predictor with uncertainty estimation use mean output as architecture’s error prediction.

From Table1 we also find the test error of path-based encoding under new sampling pipeline is nearly 2 times large than default sampling pipeline. As default sampling pipeline tends to sample architecture in a sub space, the *KL-divergence* of training architecture’s paths distribution and testing architecture’s paths distribution is 0.126 which is approximately 5 times smaller than new sampling pipeline which is 0.652. The divergence can be found in Fig.7 which illustrates the architecture’s paths distributions with sampled 100 training dataset and 500 testing dataset.

4.2 Closed Domain Search

We compare algorithms on NAS-Bench-101 [28] and NAS-Bench-201 [52]. NAS-Bench-201 is a newly proposed NAS dataset which contains $15.6k$ architectures and all architectures are trained on CIFAR-10, CIRAR-100 and ImageNet dataset. We adopts the architecture’s CIFAR-10 information to compare algorithms. On NAS-Bench-201 CIFAR-10 setting, the best architecture achieves 91.52% mean test accuracy and the architecture with highest validation accuracy has a mean test accuracy of 91.08%. We use the same experiment settings as BANANAS [23]. Each algorithm is given a budget of 150 queries. Every 10 iterations, each algorithm returns the architecture with the lowest validation

Table 1: Results of mean average percent error. We compare performance under two different architecture sampling pipelines, *default* means default sampling pipeline, *new* means new sampling pipeline. We compare four different methods, adjacency matrix means unrolled adjacency matrix concatenated with one-hot encoded operation vector, path-based encoding is defined in BANANAS [23], and our proposed neural predictors.

Training set size	20		100		500	
	Train	Test	Train	Test	Train	Test
Adjacency matrix(default)	0.313	2.722	0.683	2.484	0.532	2.369
Path-based encoding(default)	0.223	2.002	0.408	1.272	0.28	0.965
Neural Predictor with Uncertainty(default)	0.77	2.811	0.835	1.426	0.714	0.943
Neural Predictor without Uncertainty(default)	0.624	2.337	0.181	1.493	0.368	1.082
Adjacency matrix(new)	0.25	2.363	0.303	1.67	0.497	1.77
Path-based encoding(new)	0.347	2.59	0.419	1.943	0.327	2.161
Neural Predictor with Uncertainty(new)	0.362	2.77	1.158	1.627	1.62	1.423
Neural Predictor without Uncertainty(new)	0.607	2.239	0.645	1.575	0.916	1.412

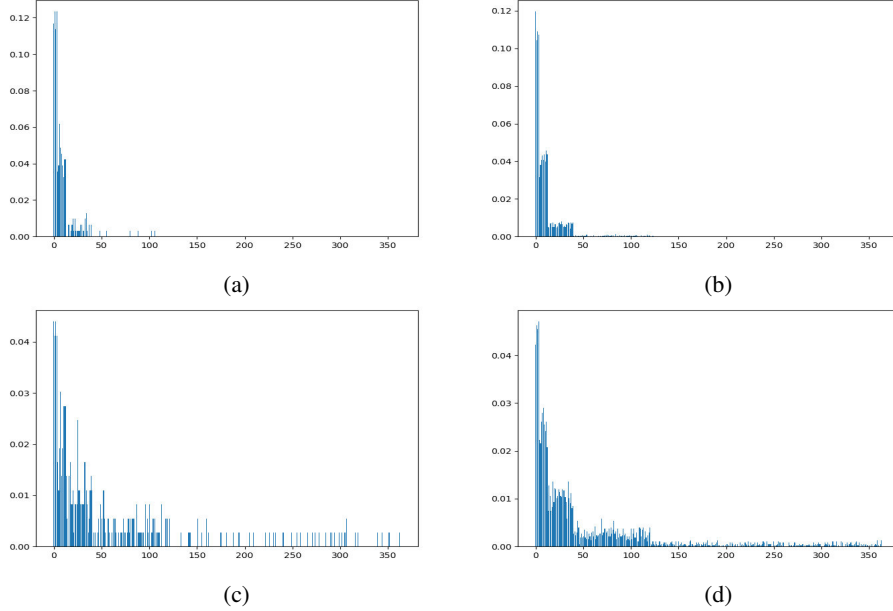


Figure 7: Architecture’s paths distribution. Upper row demonstrates default sampling pipeline, 7a is the paths distribution of training set and 7b is the paths distribution of testing set. Bottom row demonstrates the new sampling pipeline, 7c is the paths distribution of training set and 7d is the paths distribution of testing set.

error so far and its corresponding test accuracy is reported, so there are 15 best architectures in total. We run 600 trials for each algorithms and report the averaged results. If using above experiment setting NAO [34] and Neural Predictor [27] will have poor performance, so we compare the searched best architecture’s mean testing accuracy of our proposed algorithms with NAO and Neural Predictor. To demonstrate the effective of our proposed methods, we compare our approaches with the following algorithms.

- **Random Search:** This search method is a competitive baseline [53]. Random search randomly sample several architectures in search space, and take the best architecture as search result.
- **Evolution Algorithm:** Evolution algorithm takes a pool of population architectures and use evolution strategies to search a collection of parent architectures. Using evolution strategy to mutate the parent architectures to generate child architecture. The old or bad architectures

are removed from population and newly and good child architectures are added to population. The best architecture in population is taken as search result.

- **BANANAS with and without path-based encoding [23]:** This algorithm adopts a ensemble meta neural network as surrogate model and propose a path-based encoding method to represent neural architecture. They also proposed a *Independent Thompson Sampling*(ITS) as acquisition function and this function also optimized by a evolution strategy. The output architecture from (ITS) with highest accuracy is taken as search result.
- **AlphaX [54]:** AlphaX explores the search space with a Monte Carlo Tree Search and a Meta-Deep Neural Network. Meta-DNN predicts the network performance to speed up exploration. The architecture with lowest error in MCTS search path is taken as search result.
- **NAO [34]:** NAO composed by a neural encoder, neural performance predictor and a neural decoder. NAO encodes the discrete architecture into a continuous representation vector. Then this continuous vector is optimized via gradient ascent and the optimized vector is transformed into a new architecture via neural decoder. The architecture with best performance is taken as search result.
- **Neural Predictor for Neural Architecture Search[27]:** Neural predictor for NAS trains a neural predictor and using this predictor to rank all the architectures in search space. The top k architectures are selected and trained. The trained architectures with the lowest validation error is taken as search result. In order to enhance the neural predictor prediction accuracy a cascade neural predictor is proposed. For simplify compare we denoted the neural predictor for neural architecture search as NPNAS and the cascade neural predictor for neural architecture search as CNPNAS.

Results of comparison on NAS-Bench-101 is shown in Fig.8. As illustrated in Fig.8 our proposed NPUBO and NPENAS is better than other algorithms and NPENAS achieves the best performance. Fig.8b is the comparison of algorithms using new sampling pipeline. From Fig.8b we find our proposed NPENAS have the best performance and the vanilla unrolled adjacency matrix concatenate one-hot encoded vector is better than path-based encoding which against the results in BANANAS [23]. As the new sampling pipeline can adequately explore the search space, algorithms using this pipeline have low variance and perform slightly better than using default sampling pipeline. Our proposed NPUBO performs slightly better than BANANAS [23] with path-based encoding which also uses bayesian optimization, and our proposed NPENAS using 150 training samples can get mean test accuracy of 94.14% averaged over 600 trials. In *ORACLE* setting [27] the best mean accuracy is 94.23% on NAS-Bench-101 dataset.

Results of comparison on NAS-Bench-201 is shown in Fig.8c. As there are only $15.6k$ architectures in NAS-Bench-201 which is relatively small, algorithms can find a good architecture using less training samples compared with NAS-Bench-101. Our proposed methods NPUBO and NPENAS is better than other algorithms and NPUBO achieves the best performance.

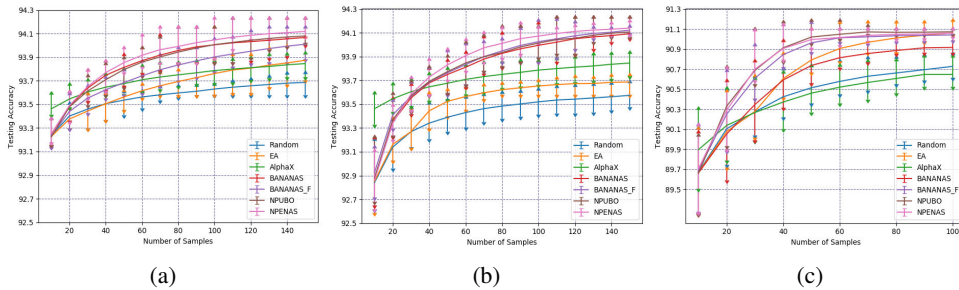


Figure 8: Performance of our proposed NPUBO and NPENAS compare to other algorithms on NAS-Bench-101 and NAS-Bench-201. The error bars represent the 30% to 70% percentile range. Fig.8a is the comparison of algorithm on NAS-Bench-101 via default sampling pipeline. Fig.8b is the comparison of algorithms on NAS-Bench-101 via new sampling pipeline. Fig.8c is the comparison of algorithms on NAS-Bench-201 via new sampling pipeline.

Table 2: Comparison of our proposed methods with NAO [34] and Neural Predictor [27]

Dataset	Model	Training Samples	Evaluated Samples	Mean Test Accuracy(%)
NAS-Bench-101	NAO [34]	1,000	–	93.73
	NPNAS [27]	172	432,000	94.12
	CNPNAS [27]	172	432,000	94.17
	NPUBO	150	1,500	94.11
	NPENAS	150	1,500	94.14
NAS-Bench-201	NAO [34]	800	–	90.86
	NPNAS [27]	172	432,000	91.09
	CNPNAS [27]	172	432,000	91.09
	NPUBO	100	1,000	91.07
	NPENAS	100	1,000	91.06

The comparison of our proposed methods with NAO [34] and Neural Predictor [27] can be found in Table 2. We run NAO [34] nigh times on NAS-Bench-101 with initial randomly sampled 600 architectures and set seed architectures to 50, and the results is averaged over 10 trials. On NAS-Bench-201 dataset we run NAO [34] five times with initial randomly sampled 200 architectures and set seed architectures to 50 and the results also averaged over 10 trials. We use same experiment setting as reported in Neural Predictor [27] on NAS-Bench-101 and NAS-Bench-201.

From Table 2 we find on NAS-Bench-101 and NAS-Bench-201 CNPNAS [27] achieves the best testing accuracy, but this method adopts a two stage GCN and have to evaluate all the architectures in search space. Our proposed methods NPUBO and NPENAS have comparable performance with CNPNAS [27] using less training and evaluated samples.

4.3 Open Domain Search

We perform algorithm comparison on *NAS-Bench-101-Macro* task. As macro architecture search is time consuming, algorithms are given a budget of 600 queries and run one trial. Sampled architectures from NAS-Bench-101-Macro are trained on CIFAR-10 using the same hyperparameter setting as NAS-Bench-101.

Comparison of algorithms on NAS-Bench-101-Macro are illustrated in Table. 3. Using our proposed NPUBO, we find an architecture achieve 94.22% test accuracy which is comparable with the best architecture in NAS-Bench-101 dataset.

Table 3: Comparison of algorithms on NAS-Bench-101-Macro task. Baseline represents the best test accuracy in NAS-Bench-101[28].

Model	Training Samples	Test Accuracy(%)
Baseline [28]	–	94.23
NPUBO	600	94.22

5 Acknowledgments

The research was supported by the National Natural Science Foundation of China (61976167, U19B2030, 61571353) and the Science and Technology Projects of Xi’an, China (201809170CX11JC12).

References

- [1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

- [3] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.
- [4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, 2016.
- [5] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:640–651, 2014.
- [6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.
- [8] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017.
- [9] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Conference on Computer Vision and Pattern Recognition*, 2019.
- [10] Vladimir Nekrasov, Hao Chen, Chunhua Shen, and Ian Reid. Fast neural architecture search of compact semantic segmentation models via auxiliary cells. In *Conference on Computer Vision and Pattern Recognition*, 2019.
- [11] Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four gpu hours. In *Conference on Computer Vision and Pattern Recognition*, 2019.
- [12] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Conference on Computer Vision and Pattern Recognition*, 2019.
- [13] Barret Zoph, V. Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, 2017.
- [14] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *ArXiv*, abs/1802.03268, 2018.
- [15] 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 *International Conference on Machine Learning*, 2017.
- [16] Masanori Suganuma, Shinichi Shirakawa, and Tomoharu Nagao. A genetic programming approach to designing convolutional neural network architectures. In *GECCO '17*, 2017.
- [17] Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical representations for efficient architecture search. *ArXiv*, abs/1711.00436, 2017.
- [18] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Efficient multi-objective neural architecture search via lamarckian evolution. In *International Conference on Learning Representations*, 2018.
- [19] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104:148–175, 2016.
- [20] Peter I. Frazier. A tutorial on bayesian optimization. *ArXiv*, abs/1807.02811, 2018.
- [21] Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabás Póczos, and Eric P. Xing. Neural architecture search with bayesian optimisation and optimal transport. *ArXiv*, abs/1802.07191, 2018.

- [22] Lizheng Ma, Jiaxu Cui, and Bo Yang. Deep neural architecture search with deep graph bayesian optimization. *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 500–507, 2019.
- [23] Colin. White, Willie Neiswanger, and Yash Savani. Bananas: Bayesian optimization with neural architectures for neural architecture search. *ArXiv*, abs/1910.11858, 2019.
- [24] Han Shi, Renjie Pi, Hang Xu, Zhenguo Li, James T. Kwok, and Tong Zhang. Efficient sample-based neural architecture search with learnable predictor. 2019.
- [25] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *ArXiv*, abs/1806.09055, 2018.
- [26] Albert Shaw, Daniel Hunter, Forrest Landola, and Sammy Sidhu. Squeezenas: Fast neural architecture search for faster semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
- [27] Wei Wen. Neural predictor for neural architecture search. volume abs/1912.00848, 2019.
- [28] Chris Ying, Aaron Klein, Esteban Real, Eric L. Christiansen, Kevin Murphy, and Frank Hutter. Nas-bench-101: Towards reproducible neural architecture search. *ArXiv*, abs/1902.09635, 2019.
- [29] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *J. Mach. Learn. Res.*, 20:55:1–55:21, 2018.
- [30] Jonas Mockus. On bayesian methods for seeking the extremum. In *Optimization Techniques*, 1974.
- [31] Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, 13:1809–1837, 2011.
- [32] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2009.
- [33] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11:1–96, 2017.
- [34] Renqian Luo, Fei Tian, Tao Qin, and Tie-Yan Liu. Neural architecture optimization. In *NeurIPS*, 2018.
- [35] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *ArXiv*, abs/1704.01212, 2017.
- [36] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çağlar Gülçehre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey R. Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Manfred Otto Heess, Daan Wierstra, Pushmeet Kohli, Matthew M Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *ArXiv*, abs/1806.01261, 2018.
- [37] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *ArXiv*, abs/1901.00596, 2019.
- [38] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30:83–98, 2013.
- [39] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Neural Information Processing Systems*, 2016.
- [40] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2016.
- [41] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, 2017.
- [42] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *ArXiv*, abs/1810.00826, 2018.
- [43] David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *ArXiv*, abs/0912.3848, 2009.

- [44] Liang-Chieh Chen, Maxwell D. Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jonathon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in Neural Information Processing Systems*, 2018.
- [45] Chenxi Liu, Barret Zoph, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan L. Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. *ArXiv*, abs/1712.00559, 2017.
- [46] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *ArXiv*, abs/1812.00332, 2018.
- [47] Tieleman Tijmen and Geoffrey E Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 2012.
- [48] Jonathan T. Barron. Continuously differentiable exponential linear units. *ArXiv*, abs/1704.07483, 2017.
- [49] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.
- [50] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [51] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.
- [52] Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. *ArXiv*, abs/2001.00326, 2020.
- [53] Christian Sciuto, Kaicheng Yu, Martin Jaggi, Claudiu Musat, and Mathieu Salzmann. Evaluating the search phase of neural architecture search. *ArXiv*, abs/1902.08142, 2019.
- [54] Linnan Wang, Yiyang Zhao, and Yuu Jinnai. Alphax: exploring neural architectures with deep neural networks and monte carlo tree search. *ArXiv*, abs/1805.07440, 2019.