

FOX-NAS: Fast, On-device and Explainable Neural Architecture Search

Chia-Hsiang Liu¹, Yu-Shin Han¹, Yuan-Yao Sung¹, Yi Lee¹, Hung-Yueh Chiang², Kai-Chiang Wu¹,
¹National Yang Ming Chiao Tung University, ²The University of Texas at Austin

{jacoblau.cs08g, yushinhan.eic09g}@nctu.edu.tw, hungyueh.chiang@utexas.edu,
 kcw@cs.nctu.edu.tw

Abstract

Neural architecture search can discover neural networks with good performance, and One-Shot approaches are prevalent. One-Shot approaches typically require a supernet with weight sharing and predictors that predict the performance of architecture. However, the previous methods take much time to generate performance predictors thus are inefficient. To this end, we propose FOX-NAS that consists of fast and explainable predictors based on simulated annealing and multivariate regression. Our method is quantization-friendly and can be efficiently deployed to the edge. The experiments on different hardware show that FOX-NAS models outperform some other popular neural network architectures. For example, FOX-NAS matches MobileNetV2 and EfficientNet-Lite0 accuracy with 240% and 40% less latency on the edge CPU. Search code and pre-trained models are released at <https://github.com/great8nctu/FOX-NAS>.¹

1. Introduction

Deep learning has been applied in various fields in the past decade, including image classification [21, 23], object detection [6, 18], semantic segmentation [15, 19], and natural language processing [22, 26]. Many exemplary architectures have been proposed in image classification. For example, AlexNet [12] and VGGNet [21] showed that the depth of convolutional neural networks is vital for achieving higher performance; ResNet [7] showed that identity-based skip connections are suitable for training deep neural networks; MobileNet [8, 9, 20] proposed the depthwise separable convolutions to build a lightweight model for edge devices.

With the success of deep neural nets, the demand for deploying deep learning algorithms to the edge rises rapidly. Compared with cloud platforms, edge devices have the ad-

¹FOX-NAS is the 3rd place winner of the 2020 Low-Power Computer Vision Challenge (LPCVC), DSP classification track. See all evaluation results at <https://lpcv.ai/competitions/2020>.

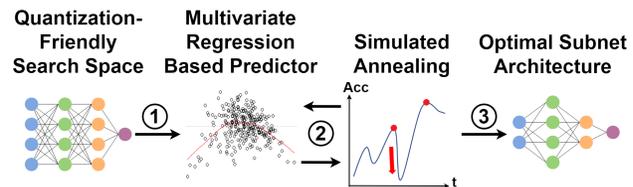


Figure 1. We define a quantization-friendly search space (cf. Section 3.2), predict the subnet’s performance by proposed multivariate regression (cf. Section 3.3), and use simulated annealing guided by our explainable predictors to avoid local optimal subnets and quickly find the global optimal subnet (cf. Section 3.4).

vantages of low cost, energy-saving, but with limited computation resources. Quantization [11] is an essential technique to make the neural network run more efficiently on edge devices. Previous works [10, 20, 28] address the issues and handcraft edged device-friendly models. However, since there are infinite candidate neural architectures, it is inefficient to find the optimal model relying on the manual trial and error method. As a result, neural architecture search (NAS) is proposed to find the optimal model architecture more efficiently using machine learning.

Neural architecture search is a technique for finding the optimal network architecture based on the search goal in the search space. The search goal can be accuracy, inference time, or any user-defined constraint. The most naïve method is to exhaustively instance models with different architectures from the search space and then train each model to estimate its performance. Since there are countless permutations and combinations of architectures and it is unlikely to train every candidate model, methods based on reinforcement learning [24, 29, 30] were proposed to do NAS. More NAS methods that effectively reduce the requiring time for NAS were then proposed. For example, Progressive NAS [13] uses the sequential model-based optimization (SMBO) method as the search strategy; AmoebaNet [17] proposed using the evolutionary algorithm to find the optimal network architecture. However, these NAS methods

take a lot of GPU time to complete the training.

Recently, One-Shot NAS has been proposed to make NAS more efficient and effective. For example, ENAS [16] proposed to use the method of sharing weight in the training process of searching an architecture, so that the search time can be reduced to 16 GPU hours; DARTS [14] proposed an algorithm of gradient based optimization for differentiable NAS; ProxylessNAS [2] proposed an effective solution that can directly search the architectures for large-scale datasets and target hardware platforms. Additionally, Once-for-All [1] proposed a method that is different from the previous NAS. They decouple the training of the supernet from the architecture search and directly get a specialized subnet by selecting from the well-trained supernet without retraining. After the supernet training is completed, subnets are randomly sampled from the supernet to measure the performance, and then these data are used to do the architecture search. However, training the predictor required by the architecture search is time-consuming.

In this work, we propose a novel method for NAS named FOX-NAS, which has the advantages of being fast, on-device, and explainable. We continue the previous method [1] and reduce the time required for architecture search to complete the architecture search process directly on edge devices, as shown in Figure 1.

The contribution of this work has four aspects:

- 1) We adopt multivariate regression analysis as our predictors that reduce the time and data than the deep learning approach. In addition, the results of the parameters are explainable and controllable, which allows us to optimize the model for a variety of objectives, *e.g.*, power consumption, accuracy, latency.
- 2) We use simulated annealing as our search algorithm, which can complete the search within 1 minute and avoid suboptimal network structures.
- 3) We design quantization-friendly search spaces that adapt for CPU and TPU so that the resulting models are easily deployed to edge devices.
- 4) Our extensive experiments are shown in Table 1, demonstrating that the architecture we found is 4.2% higher than MobileNetV3-small [8] under the same latency, and in the edge TPU environment, the accuracy of our model is also higher than MobileNetEdgeTPU.

2. Preliminaries

NAS can be divided into three parts: search space, search strategy, and performance estimation.

Search Space. Since there are infinite combinations of neural network architectures, we first need to define the

architecture’s scope, called search space. In image classification, the backbone of the convolution model architecture, such as ResNet [7], MobileNet [8, 20], is used most frequently in NAS. In this work, we also adopt MobileNetV3 [8] as our backbone with quantization-friendly modules (*e.g.*, change the activation function to ReLU6).

Search Strategy. The search strategy is a key to an optimal network search since the search space is enormous (approximately 4×10^{22} in this work). In previous work, they used reinforcement learning [24, 29, 30], SMBO [13], and the evolution algorithm [17]. In this work, to avoid the local optimal solutions, we use the simulated annealing algorithm as our search strategy. Compared with other search strategies, simulated annealing has a higher probability of finding the global optimal solution. Moreover, coupled with the guidance of multivariate regression analysis, we can find the optimal solution more quickly.

Performance Estimation. We evaluate the performance of the architectures sampled by the search strategy to find the optimal network architecture. However, this is infeasible, especially when the search space and target dataset are large. Therefore, proxy tasks (*e.g.*, CIFAR-10) are often used to evaluate neural architecture performance. However, the optimal neural architectures found on the proxy tasks are not guaranteed to be the optimal architecture found in the target task [2]. Once-for-all [1] proposed using a neural network to generate accuracy and latency predictors, and it took 40 GPU hours to collect data. In this work, we propose to use multivariate regression to generate predictors, which can collect enough data for training in just 3.5 GPU hours.

3. Method

3.1. Problem Statement

Given a targeted latency on the specific hardware, we aim to find an optimal neural network, based on the neural architecture search (NAS) techniques, with the highest accuracy while meeting the constraints. Figure 2 is our flow chart.

3.2. Search Space

Regarding the famous CNN model architectures [8, 20], we also divide the CNN model into a sequence of units, and we have three types of CNN units, as shown in Figure 3. To make its operations faster and more efficient for different hardware, we design two different search spaces and replace the activation function with ReLU6 to make our neural architecture have the advantage of being quantization-friendly. We use the results of multivariate regression analysis, which allowed us to explain the impact of each control parameter on the performance of a neural network.

Search Space for CPU. The CPU-like hardware uses the first and second types of CNN units, as shown in Fig-

Model	Search Strategy	Search Space	Performance Estimation Strategy	Training Cost (GPU hours)	Search Cost (GPU hours)
NASNet [30]	Reinforcement learning	arch	Train and evaluate	48000N	
MnasNet [24]	Reinforcement learning	arch	Train and evaluate	40000N	
AmoebaNet [17]	Evolution algorithm	arch	Train and evaluate	75600N	
DARTS [14]	Gradient optimization	arch	Train and evaluate	250N	96N
ProxylessNAS [2]	Gradient optimization	arch	Train and evaluate	300N	200N
Once-for-All [1]	Supernet/ Evolution algorithm	arch	Performance predictor	1200 + 40	≈ 0
FBNetV3 [4]	Supernet/ NARS	arch/recipe	Performance predictor	10700	≈ 0
FOX-NAS	Supernet/ Simulated annealing	Quantization friendly arch	Explainable performance predictor	1200 + 3.5	≈ 0

Table 1. The search method comparison between FOX-NAS and the state-of-the-art NAS on ImageNet. We propose new methods in all three parts of NAS. FOX-NAS only needs to train the supernet once, which takes 1200 GPU hours, and then it only takes 3.5 hours to train performance predictors.

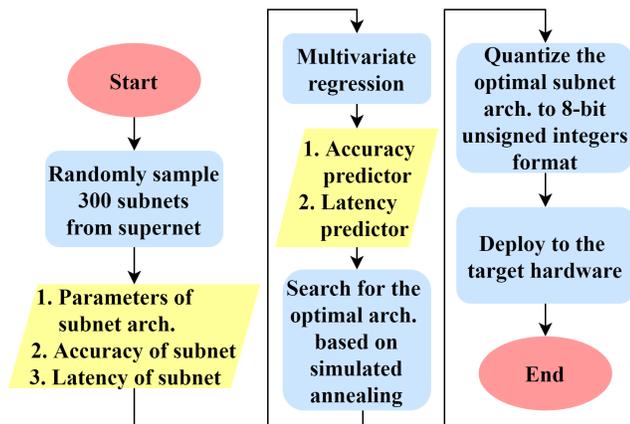


Figure 2. Flowchart of FOX-NAS. We collect data of subnet information from our quantization-friendly supernet first, then use multivariate regression to generate performance predictors, and then use the simulated annealing algorithm to find the optimal neural network architecture. After quantized the searched subnet into 8-bit integer format, it can directly deploy the subnet to the specific edge device.

ure 3a and Figure 3b. The memory access and computation of CPU-like hardware are expensive, so the separable convolution is needed to reduce the computation of CNN. Therefore, we refer to the backbone of MobileNetV3 [8] and the supernet of Once-for-All [1]. Table 2 lists the can-

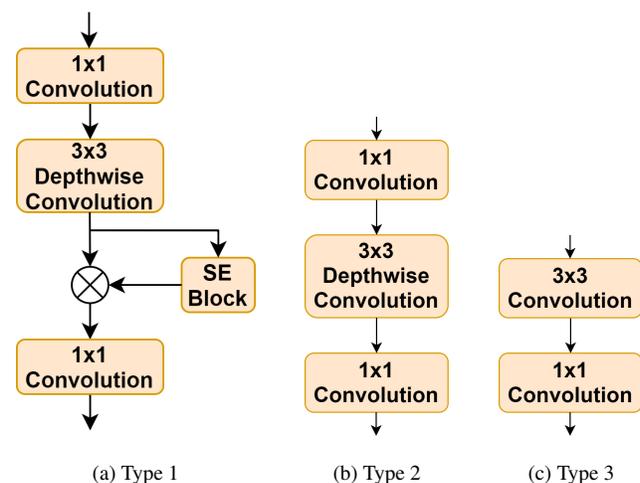


Figure 3. CNN units. FOX-NAS has three CNN units for different hardware.

didate of all the architecture parameters in our search space. We found that the number of channels of the previous units significantly impacted the latency on the CPU-like hardware through the multivariate regression analysis. As a result, when designing the search space, we set the minimum expansion ratio to 2 while making more choices for our subnet. In addition, to find the subnet with a broader accuracy range, our input image size is changed from 128×128 to 320×320 . Finally, we adopt the quantization-

Search Space	CPU	TPU
	Backbone	Backbone
Image Size Candidate	128~320	128~320
Kernel Size Candidate	3, 5, 7	3
Expansion Ratio Candidate	2, 3, 4, 6	4, 6, 8
Depth Candidate	2, 3, 4	3, 4, 5
Type of CNN Unit	1, 2	2, 3

Table 2. FOX-NAS has two search spaces for different hardware. The adjustable architecture parameters of the search space include image size, kernel size, expansion ratio, depth, and type of CNN unit.

friendly activation function to achieve a better quantization effect, which enables our subnet to control the accuracy loss within 1% when quantizing. The number of different neural network architectures in our CPU search space is approximately 4×10^{22} .

Search Space for TPU. The TPU-like hardware uses the second and third types of CNN units, as shown in Figure 3b and Figure 3c. The TPU-like hardware has a high degree of parallelism in matrix operations, so the higher the processor utilization, the higher the computing efficiency. Therefore, we refer to the model architecture of Mobilenet-EdgeTPU and replace the separable convolution of the previous unit with the traditional convolution. As shown in Table 2, we fix the kernel size of search space at 3×3 because the 5×5 and 7×7 kernel sizes are not friendly for the edge TPU. TPU has excellent parallelism, so we change the candidate of expanding ratio to 4, 6, 8. The number of different neural network architectures in our TPU search space is approximately 3×10^{13} . The effect of different backbones on different hardware is different, and the related experimental results and analysis are in Section 4.1.

3.3. Performance Prediction Based on Multivariate Regression

Multivariate regression analysis is a machine learning algorithm based on supervised learning, which can analyze multiple variables (*e.g.*, the impact of each layer’s kernel size and expansion ratio on performance). We adopt multivariate regression to generate performance predictors. This method is not data-hungry, and it is fast and explainable. Compared with the method based on deep learning, multivariate regression analysis can better understand the effect of each variable on the results and create promising predictors with much fewer data. Furthermore, the results of multivariate regression analysis make the parameters of a network architecture explainable, which provides a hint of twisting the architecture for target constraints. We trained a total of 7 predictors for different image sizes, and each predictor only needs 300 pieces of data to train. Using a

consumer-level GPU only takes 3.5 GPU hours (the previous method needs to collect 16K data, a total of 40 GPU hours [1]). To collect the training data for multivariate regression, we first randomly sample different sub-networks from the super-network and record their network architectures, characterized by 25 variables, including the number of layers, widths, and kernel sizes. We then use 50K validation data to measure the accuracy of each sub-network. Meanwhile, we run each sub-network on the target hardware to collect the latency data.

Assuming that the estimated multivariate linear regression model is:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \dots + \widehat{\beta}_k X_k \quad (1)$$

β_i are slope parameters or called correlation coefficients, representing the impact of the variable, Y is the dependent variable, X_i are independent variables, the symbol \wedge indicates an estimate for the variables. Moreover, taking our CPU search space as an example, the performance prediction model equation can be expressed as follows:

$$\begin{aligned} \widehat{Perf.} = & \widehat{\beta}_0 + \widehat{\beta}_{D_1} D_1 + \widehat{\beta}_{E_1^{avg}} E_1^{avg} + \widehat{\beta}_{K_1^{avg}} K_1^{avg} + \\ & \widehat{\beta}_{E_{1,2}} E_{1,2} + \widehat{\beta}_{K_{1,2}} K_{1,2} + \widehat{\beta}_{E_1^{total}} E_1^{total} + \\ & \widehat{\beta}_{E_1^{total} \cdot D_1} E_1^{total} \cdot D_1 + \widehat{\beta}_{K_1^{avg} \cdot D_1} K_1^{avg} \cdot D_1 + \dots + \\ & \widehat{\beta}_{K_m^{avg} \cdot D_m} K_m^{avg} \cdot D_m \end{aligned} \quad (2)$$

FOX-NAS-CPU has five types of CNN units with different input channels, so m equals 5 in this case. j represents the type of CNN units, D_j are the number of depth of each CNN unit, E_j^{avg} are the average expansion ratio of each CNN unit, K_j^{avg} are the average kernel size of each CNN unit, $E_{j,j+1}$ and $K_{j,j+1}$ are the expand ratio and kernel size of each CNN unit with different input and output channels, E_j^{total} are the total number of expansion ratio of each CNN unit, and $E_j^{total} \cdot D_j$ and $K_j^{avg} \cdot D_j$ represent the interaction between total expansion ratio and number of depth of each CNN unit and the interaction between the average kernel size and number of depth of each CNN unit, respectively.

In addition, we can analyze and explain our neural network architecture through some statistics. $S_{\widehat{\beta}_i}$ are the standard deviation of estimate coefficients $\widehat{\beta}_i$, called the standard error. The t-value is a commonly used statistic, and its formula is as follows:

$$t = \frac{\widehat{\beta}_i}{S_{\widehat{\beta}_i}} \quad (3)$$

The p-value is the probability density value \geq t-value under the T-distribution, representing whether the impact of the variable on the output variable is highly correlated. For example, the p-value less than 0.05 indicates that the

variable is highly correlated with the output variable. In addition, we can look up the t-value in the T-distribution table with the given degrees of freedom to get the p-value. The R^2 value measures the percentage of the variation in Y being explained by the fitted regression model. Thus, the larger the R^2 value, the better the fit of the regression model, and its formula is as follows:

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (4)$$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (5)$$

$$R^2 = 1 - \frac{SSE}{TSS} \quad (6)$$

TSS is the total sum of squares, SSE is sum of squares errors, \bar{Y} is the mean of Y . When adding more independent variables, R^2 will be larger, showing an overestimation phenomenon. Adjusted R^2 has been adjusted in degrees of freedom to avoid the expansion, and its formula is as follows:

$$R_{adj}^2 = 1 - \frac{SSE \times (n - 1)}{TSS \times (n - k)} \quad (7)$$

n is the number of collected data, k is the number of correlation coefficients. The adjusted R^2 value of the regression model we generate can reach above 92, indicating that our predictor is accurate. The variables in the predictor are highly correlated and explainable, so we can more effectively control performance by adjusting the architecture parameters.

When an optimal subnet is proposed from our predictors, we can further twist the architecture to make the performance of the subnet meet our target constraint since the impact of each architecture parameter on the subnet performance is explainable. We choose the architecture parameter with a sufficiently small p-value and then compare the correlation coefficients of the parameter to make a precise and efficient adjustment based on the target constraint.

For example, suppose that the latency constraint we set is 60 ms, but the subnet we found through the predictor is 60.3 ms, which exceeds the constraint of 0.3 ms. We need to adjust the parameters of this subnet architecture manually. Suppose we find the p-value of E_1^{avg} is close to 0 in both the accuracy and latency predictors, which means that the impact of E_1^{avg} on the performance is highly correlated. Moreover, suppose the correlation coefficients $\hat{\beta}$ of E_1^{avg} is very small in the accuracy predictor, which means that E_1^{avg} only has little impact on accuracy. Suppose the $\hat{\beta}$ of E_1^{avg} is 0.4 in the latency predictor, which means that the latency

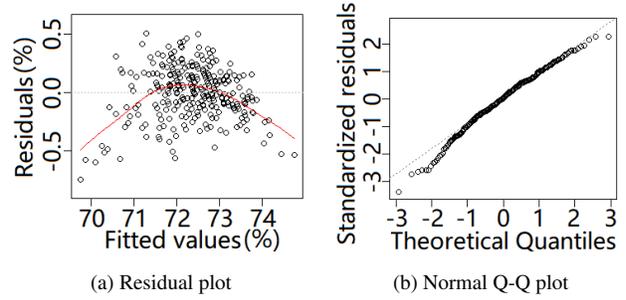


Figure 4. (a) The red line in the residual plot is our model, and the points in the figure are the collected data. The fitted value is between 70 and 74, and the maximum residual is only 0.5, which is a small residual, less than 1% of the fitted value. (b) There is a 45° straight dashed line in the normal Q-Q plot. If the data is normally distributed, the point will fall on the 45° reference line. The bottom end of the Q-Q plot deviates from the straight line, but the upper end does not deviate, then we can say that it is a left-skewed distribution.

can be reduced by 0.4 ms when E_1^{avg} is reduced by 1. As a result, we can manually twist E_1^{avg} to achieve the target latency.

Figure 4a shows the residual plot of our accuracy predictor. A fitted value \hat{Y} is the model's prediction of the response value when we input the values of the predictor. The residuals e are equal to the difference between the ground truth and the fitted value:

$$e = Y - \hat{Y} \quad (8)$$

The normal quantile-quantile plot is used to evaluate whether our residuals are normally distributed. Figure 4b shows our normal Q-Q plot. These analysis graphs represent the high reliability of our model.

3.4. Search Strategy Based on Simulated Annealing

Simulated annealing is an algorithm based on probability to find the optimal solution under the objective function. It is usually used when the search space is discrete. Because it can accept unsatisfactory results during the search process, it can find the global optimal solution rather than the local optimal solution compared to other algorithms, as shown in Figure 5.

We use simulated annealing as our search strategy. As shown in Figure 2, we can find the optimal neural architecture in one minute after using multivariate regression to generate the performance predictors, combined with the simulated annealing algorithm. The simulated annealing method we used is summarized in Algorithm 1. Under the constraint of computational resources, we use the guidance of multivariate regression analysis to the sample neural architecture, aiming to find the neural architecture with the best

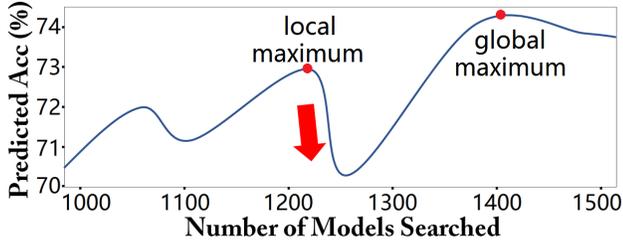


Figure 5. When we find the local maximum, if the local maximum is far away from the global maximum, we need to go through a period of inferior results before we can find the global maximum solution. The simulated annealing algorithm accepts unsatisfactory results, so it has a higher probability of finding the global maximum than the local maximum compared to other algorithms.

performance. In the process of searching based on simulated annealing, we accept the neural architectures with inferior performance, but as the number of architectures searched increases, we reduce the acceptance of unsatisfactory architectures until the algorithm converges. This is also a feature of simulated annealing, which has the advantage of a higher probability of finding the global optimal solution.

With the guidance of multivariate regression analysis, we can find the optimal model more quickly. We give each architecture parameter a weight, representing the probability of selecting this parameter when sampling a subnet architecture. In the original method, the weight of each parameter is 1, which means that the probability of selecting each parameter is equal. However, through multivariate regression analysis, we know the impact of each architecture parameter on the performance of the subnet. Accordingly, in the search process, we have two sets of weights. In the early stage of the search, we use the first set of weights, which are very large for a few architecture parameters. The purpose is to find a preliminary solution first. Then we change to the second set of weights, which are almost the same for all architecture parameters, to find the global optimal solution. The related experimental results and analysis are in Section 4.2.

4. Experiments

In this section, we compare FOX-NAS on different hardware with some popular neural networks. In addition, we compare the performance of FOX-NAS with different backbones on different hardware. Finally, there is a performance comparison between different search methods. Our experiment is performed on image classification using ImageNet [5].

Algorithm 1: NAS based on Simulated Annealing

Input: latency constraint L
Output: optimal model M

- 1 // When sampling the model architecture, we select T architecture parameters to make changes.
- 2 // We have two sets of weights representing the probability of selecting each architecture parameter when sampling the model architecture.
- 3 // We have a probability formula to evaluate whether to accept the architecture with inferior estimation performance.

```

4
5 Function SampleModel ( $M, T, weight$ ) :
6    $tmp.lat \leftarrow \infty$ 
7   while  $tmp.lat \geq L$  do
8      $tmp.arch \leftarrow \text{SampleArch}(M, T, weight)$ 
9      $tmp.acc \leftarrow \text{AccPredictor}(tmp.arch)$ 
10     $tmp.lat \leftarrow \text{LatPredictor}(tmp.arch)$ 
11  return  $tmp$ 
12
13 Function Main:
14  Get an initial model  $M$ 
15  Get an initial temperature  $T > 0$ 
16  Get a counter  $C = 0$ 
17  Get the constants  $k, n$ 
18   $weight \leftarrow$  weights of the early stage (guided by
19    our explainable predictors)
20  while  $T > 0$  do
21     $C \leftarrow C + 1$ 
22     $r \leftarrow \text{Random}(0,1)$ 
23     $M_{new} \leftarrow \text{SampleModel}(M, T, weight)$ 
24     $\Delta \leftarrow M_{new}.acc - M.acc$ 
25    if  $\Delta > 0$  or  $r > \exp(-\Delta / kT)$  then
26       $M \leftarrow M_{new}$ 
27    else
28      if  $C \% n == 0$  then
29         $\lfloor$  Decrease the temperature  $T$ 
30      if  $C == 2n$  then
31         $weight \leftarrow$  new set of weights searched
32  return  $M$ 

```

4.1. Comparison of Model Performance with Different Hardware and Constraints

We compare the performance of the models on different hardware, including cloud GPU, edge TPU, and edge CPU. In the experiment on GPU, the latency is measured with batch size 64 and 32-bit floating-point format on Nvidia 2080Ti with Pytorch 1.8.1 + CUDA 11.0. As for edge CPU and edge TPU, we use the quantization tool of Tensorflow

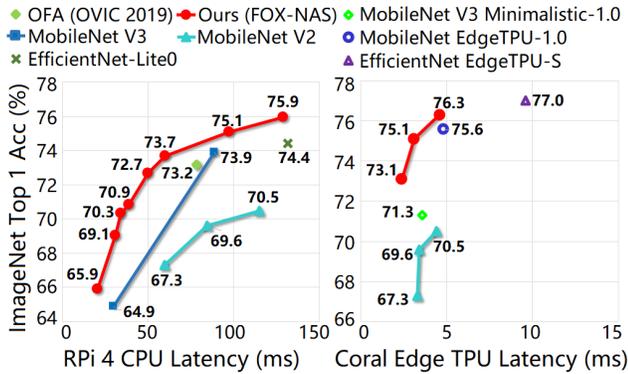


Figure 6. FOX-NAS achieves SOTA performance on ARM CPU and edge TPU.

Model	Top-1 (%)	GPU latency (ms)	Image size	Params (M)
MNASNet-0.5 [24]	68.9	13	224	2.1
MNASNet-0.75 [24]	73.3	20	224	2.9
MNASNet-1.0 [24]	75.2	24	224	3.9
MobileNetV2 [20]	72	40	224	3.5
MobileNetV3-S [8]	67.5	10	224	2.9
MobileNetV3-L [8]	75.2	25	224	5.4
EfficientNet-B0 [25]	76.1	47	224	5.3
FBNetV2-F3 [27]	73.2	18	224	6.9
FBNetV2-F4 [27]	76	25	224	7.0
FBNetV2-L1 [27]	77.2	31	224	8.5
FBNetV3-A [4]	79.1	33	224	8.6
FairNAS-A [3]	75.3	28	224	4.6
FairNAS-A-SE [3]	77.5	34	224	5.9
ProxlessNAS [2]	75.1	25	224	5.1
OFA-1080Ti-15 [1]	73.8	13	144	6.0
OFA-1080Ti-22 [1]	75.3	17	188	5.2
OFA-1080Ti-27 [1]	76.4	22	188	5.2
FOX-NAS-TPU-A	73.9	12	192	4.1
FOX-NAS-TPU-B	75.3	17	192	5.3
FOX-NAS-TPU-C	76.3	22	224	5.3

Table 3. Comparison with popular models on Nvidia 2080Ti GPU. On the Nvidia 2080Ti GPU, FOX-NAS can achieve the best accuracy under the three different latency constraints.

2, and the latency is measured with batch size 1 and 8-bit unsigned integer format on the Raspberry pi 4 ARM Cortex-A72 CPU and Coral USB TPU accelerator.

Latency Comparison on Edge CPU. Figure 6 shows the performance comparison between FOX-NAS and other famous and open-source integer neural network models. The hardware used for the measurement is the ARM CPU

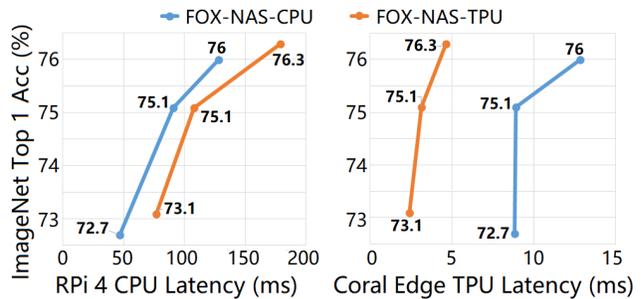


Figure 7. Latency comparison between CPU backbone and TPU backbone on different hardware. The backbone of the CPU runs more efficiently on the CPU than the backbone of the TPU. On the contrary, the backbone of the TPU runs more efficiently on edge TPU.

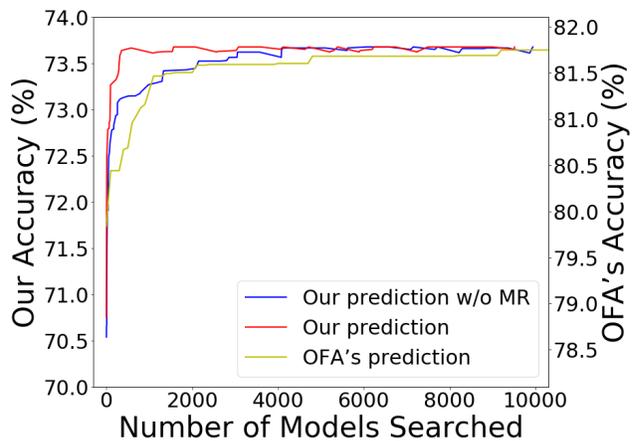


Figure 8. Performance comparison of search strategy between OFA and FOX-NAS. FOX-NAS uses simulated annealing and multivariate regression analysis to make our search process converge faster. The label on the left y-axis is the prediction of our accuracy predictor, and the label on the right is the prediction of OFA's predictor. OFA's predictors are provided by their GitHub. Because OFA and we have different hardware constraints, there will be a gap between the predicted results. To facilitate comparison, we shift the predictions.

on the Raspberry Pi 4. FOX-NAS can generate subnets with an extensive accuracy range, from 76% to 66%, and each subnet performs well. Under the same accuracy level, FOX-NAS can be 240% faster than MobileNetV2 [20], 50% faster than MobileNetV3 [8], and 40% faster than EfficientNet-Lite0 [25]. Under the same latency level, FOX-NAS can be 6.4% more accurate than MobileNetV2, 4.2% more accurate than MobileNetV3, and 1.5% more accurate than EfficientNet-Lite0.

Latency Comparison on Different Hardware. Unlike the hardware architecture of edge CPU, architecture

such as edge TPU and cloud GPU has powerful parallel computing capabilities. Therefore, we propose different search spaces for different hardware to make the neural network operation more efficient. Figure 6 shows the performance comparison between FOX-NAS and other popular models on edge TPU. The Edge TPU Compiler version used in our experiment is 2.0.291256449. Under the same latency level, FOX-NAS is 5.8% more accurate than MobileNetV2 [20] and performs better than MobileNetEdgeTPU. Because the cache on the Coral USB TPU accelerator is small, only about 6 MB, and models over 6 MB will cause much damage to the latency, so we did not search for a larger model to evaluate.

Table 3 show the performance comparison of FOX-NAS and other models on GPU. Because the edge TPU has some restrictions on neural network models, many models cannot be directly deployed on this device. To compare with more models, we deployed FOX-NAS with TPU based search space on the GPU. Compared with the Once-for-All [1] models, FOX-NAS can achieve almost the same performance while reducing the training cost by 36.5 GPU hours. Under the same accuracy level, FOX-NAS can be 47% faster than MobileNetV3 [8] and 50% faster than FBNetV2-F3 [27].

Latency Comparison of Different Search Space on Different Hardware. It is mentioned in Section 3.2 that different hardware requires different neural network architecture designs to be more efficient. Therefore, we propose two different search spaces: CPU-based and TPU-based. Figure 7 is a comparison diagram of the latency results of running the two backbone subnets on the edge CPU and the edge TPU, respectively. The backbone of the CPU runs more efficiently on the CPU than the backbone of the TPU. On the contrary, the backbone of the TPU runs more efficiently on edge TPU. The reason is that the computing power on the CPU is the bottleneck, and the edge TPU has strong computing power, so the memory access is the bottleneck of the edge TPU.

4.2. Performance Comparison of Different Search Methods

We compare the search methods between FOX-NAS and Once-for-All [1]. In addition, we compare the performance of search using simulated annealing algorithm with or without multivariate regression analysis guidance.

Comparison of the Search Methods Between FOX-NAS and Once-for-All (OFA). Figure 8 shows the comparison between our search and OFA search method. We use multivariate regression to generate performance predictors and then use simulated annealing coupled with guidance from multivariate regression analysis as our search strategy. OFA used predictors based on the neural network and used the evolutionary algorithm as the search strategy.

In this experiment, the target hardware we searched for was the ARM CPU on the Raspberry Pi 4, and OFA was the CPU on the Samsung Note 10 mobile phone. We recorded the performance predicted by FOX-NAS and OFA during the search process. OFA requires 50,000 points to finish the search, so we only selected the first 10,000 points for comparison with FOX-NAS for easy comparison. As shown in Figure 8, FOX-NAS can search for the global optimal solution more quickly. Because the simulated annealing algorithm has the property of accepting inferior search results during the search process, it has more oscillations than the evolutionary algorithm during the convergence process.

Figure 8 also compares the search performance of simulated annealing with or without multivariate regression analysis. After using the guidance of multivariate regression analysis, the simulated annealing can quickly find the global optimal solution near only a few points. In contrast, the simulated annealing search requires more points to find the global optimal solution without multivariate regression analysis.

5. Conclusion

We propose FOX-NAS, a novel neural architecture search method with fast, on-device, and explainable advantages. Unlike the previous approach, we reduced the time required to generate the performance predictors to complete in 3.5 GPU hours. Moreover, we reduced the movement steps required in the search process. Experiments on different hardware showed that the model obtained by our approach has good performance. Thus, we expect our work can reduce the cost of neural architecture search.

References

- [1] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019.
- [2] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.
- [3] Xiangxiang Chu, Bo Zhang, Ruijun Xu, and Jixiang Li. Fairnas: Rethinking evaluation fairness of weight sharing neural architecture search. *arXiv preprint arXiv:1907.01845*, 2019.
- [4] Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, et al. Fbnetv3: Joint architecture-recipe search using neural acquisition function. *arXiv preprint arXiv:2006.02049*, 2020.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection

- and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.
- [9] 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. *arXiv preprint arXiv:1704.04861*, 2017.
- [10] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [11] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [13] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.
- [14] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.
- [15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [16] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104. PMLR, 2018.
- [17] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019.
- [18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.
- [19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [20] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [22] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.
- [23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [24] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.
- [25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [27] Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuan-dong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, et al. Fbnetv2: Differentiable neural architecture search for spatial and channel dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12974, 2020.
- [28] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.
- [29] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.
- [30] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.