Proceedings of Machine Learning Research 189, 2022 ACML 2022

A Novel Differentiable Mixed-Precision Quantization Search Framework for Alleviating the Matthew Effect and Improving Robustness

Hengyi Zhou z4050037@stu.xjtu.edu.cn
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Hongyi He [*] hongyihe@stu.xjtu.edu.cn
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Wanchen Liu liuwanch79@cetc.com.cn
Science and Technology on Electro-Optical Information Security control Laboratory

Yuhai Li liyuhai.cn@gmail.com
Science and Technology on Electro-Optical Information Security control Laboratory

Haonan Zhang haonanzhang@stu.xjtu.edu.cn
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Longjun Liu [†] liulongjun@xjtu.edu.cn
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Network quantization is an effective and widely-used model compression technique. Recently, several works apply differentiable neural architectural search (NAS) methods to mixed-precision quantization (MPQ) and achieve encouraging results. However, the nature of differentiable architecture search can lead to the Matthew Effect in the mixed-precision. The candidates with higher bit-widths would be trained maturely earlier while the candidates with lower bit-widths may never have the chance to express the desired function. To address this issue, we propose a novel mixed-precision quantization framework. The mixed-precision search is resolved as a distribution learning problem, which alleviates the Matthew effect and improves the generalization ability. Meanwhile, different from generic differentiable NAS methods, search space will grow rapidly as the depth of the network increases in the mixed-precision quantization search. This makes the supernet harder to train and the search process unstable. To this end, we add a skip connection with a gradually decreasing architecture weight between convolutional layers in the supernet to improve robustness. The skip connection will help the optimization of the search process and will not participate in the bit width competition. Extensive experiments on CIFAR-10 and ImageNet demonstrate the effectiveness of the proposed methods. For example, when quantizing ResNet-50 on ImageNet, we achieve a state-of-the-art 156.10x Bitops compression rate while maintaining a 75.87% accuracy.

Keywords: deep neural networks; model compression; neural architecture search; mixed-precision quantization

[†] First Author and Second Author contribute equally to this work.

[†] Longjun Liu is the Corresponding author.

1. Introduction

Deep convolutional neural networks have become the de-facto method and achieved tremendous success in a wide range of tasks. However, high computation complexity impedes the development of DCNNs on edge computing systems, such as mobile phones, drones, autonomous robots, etc.

To this end, many prior arts have been proposed, including, but not limited to, network quantization, network pruning Zhang et al. (2021), knowledge distillation Hinton et al. (2015), tensor decomposition Zhang et al. (2022), structural re-parameterization Zhou et al. (2022) and compact model design Sandler et al. (2018). Network quantization reduces the bit widths of data flowing in a deep neural network, thus allowing the use of lower precision computation units in hardware.

Most early quantization methods usually quantize all (or most of) the layers of a deep model into the same bit widths, which can be categorized into uniform precision quantization. They have shrunk the model size and accelerated the inference process, but still suffer from significant accuracy degradation when performing ultra low-bit quantization. Currently, a wide range of hardware, such as CPUs and FPGAs, have supported mixed-precision computation. This motivates the research of quantizing different filters/channels into different bit widths to further compress deep models and pursue higher accuracy. However, optimizing an appropriate bit-width configuration for a mixed-precision network is computationally expensive. For example, in a neural network with L layers, where each layer has N candidate bit-widths for weights and M candidate bit-widths for activations, the search space will have $(NM)^L$ configurations for mixed-precision search. To reduce the search cost, a series of works Cai and Vasconcelos (2020); Wu et al. (2018); Yu et al. (2020) adopt differentiable architecture search (DARTS) Liu et al. (2019) on mixed-precision quantization and achieve encouraging results. In this paper, we use the *mixed-precision quantization search* to denote optimizing mixed-precision quantization through NAS methods.

However, there are still several issues that have not been fully considered in prior works.

- The direct optimization of architecture weights can naturally lead to the Matthew effect Hong et al. (2020). The Matthew effect in differentiable NAS methods refers to that fewer loss gradients will be propagated to the underperforming operations, which will make them harder to be trained maturely and their performance much worse. The Matthew effect in the mixed-precision quantization search can be more severe since there are natural differences in the expressive abilities of different candidates.

- The DARTS-based optimization used in prior mixed-precision works removes the skip connection from the search space. Apparently, mixed-precision quantization does not need a skip connection as a candidate operation. And in some previous works Chu et al. (2020); Liang et al. (2019), skip connections are even blamed for the performance collapse of DARTS. However, in the mixed-precision quantization search, search space will grow rapidly as the depth of the network increases, which increases the difficulty of supernet training. Hence, if we can utilize the skip connections to help the supernet training without influencing the quantization competition, it will not reduce stability but be helpful for mixed-precision quantization.

In this paper, we propose a novel mixed-precision framework to address the above issues. First, to alleviate the Matthew Effect, we resolve the mixed-precision quantization search as a distribution learning problem, which naturally induces stochasticity and encourages exploration. We treat the architecture weight as random variables sampled from a learnable Dirichlet distribution. To the best of our knowledge, we are the first paper to leverage the Dirichlet distribution on mixed-precision quantization. Furthermore, we propose to add a skip connection between convolutional layers in the supernet. We pre-define the initial architectural weight of the skip connection and gradually reduce it to zero during the search process. This skip connection will alleviate gradient vanishing and help training. Meanwhile, it does not interfere with the search results since it does not participate in the bit widths competition. And when the weights are trained maturely, the skip connections have nearly disappeared, which does not influence the architecture of the quantized model. Moreover, we conduct a series of experiments to evaluate the effectiveness of our framework.

2. Related Work

2.1. Uniform Precision Quantization

Network quantization attempts to reduce the bit-width of weight, activation, or other data flowing in a neural network and has been long studied since the very beginning of the blooming era of deep learning. For example, BWN Courbariaux et al. (2015) mainly focuses on weights quantization, while DoReFa Zhou et al. (2016) quantizes not only weight but also activation or even gradient and error. For activation quantization, HWGQ-Net Cai et al. (2017) introduces a clipped ReLU function to avoid the gradient mismatch. PACT Choi et al. (2018) further replaces ReLU with an activation function with a trainable clipping parameter. Most of those early works quantize all or most of the layers of a deep network into the same bit-width, which can be categorized into uniform precision quantization.

2.2. Neural Architecture Search

To the best of our knowledge, Zoph and Le (2017) is the first paper that introduces Neural Architecture Search (NAS). Since then, NAS has achieved state-of-the-art performances in many fields and attracted increasing attention. Most pioneers develop prototypes based on reinforcement learning (RL) Zoph and Le (2017), evolution algorithm Real et al. (2019) and Bayesian optimization Domhan et al. (2015), but suffer from heavy computation costs and challenge of scalability. To this end, Differentiable Architecture Search (DARTS) Liu et al. (2019) resolves the search process as a bilevel optimization problem, which allows the architecture search to be efficiently performed by a gradient-based optimizer.

DARTS achieves encouraging results and inspires a series of other works. However, there are still some flaws in differentiable NAS methods, such as the Co-adaption problem and the Matthew effect Hong et al. (2020), the unfair competition Chu et al. (2020) which leads to performance collapse. And some endeavors Chu et al. (2020); Zela et al. (2020); Liang et al. (2019); Chu et al. (2021) are dedicated to solving those problems. Since differentiable NAS methods are very efficient and easy to implement, several works utilize them to perform mixed-precision quantization.

2.3. Mixed Precision Quantization

Even though uniform quantization has achieved success in compressing deep models, it can be suboptimal since different layers can have different quantization sensitivities. To address this issue, a series of mixed-precision quantization works have been proposed. Some works define a dominant metric of quantization sensitivity. HAWQ Dong et al. (2019) and HAWQv2 Dong et al. (2020) are representatives, which use the eigenvalue/trace of a Hessian matrix as a metric of quantization sensitivity and then determine the relative quantization levels of layers based on the metric. ZeroQ Cai et al. (2020) also designs a sensitivity metric through the KL divergence and uses a Pareto frontier approach to perform mixed-precision quantization. Liu et al. (2021) adopts the difference of the output of a channel before and after quantization as the sensitivity metric and constructs the multipoint quantization with a greedy selection procedure.

Different from those metric-based methods, many other studies do not define a dominant metric of quantization sensitivity but seek to automatically determine the exact bit precision configurations. ALQ Qu et al. (2020) proposes a special quantization framework and incrementally trains an adaptive bit width without gradient approximation. HAQ Wang et al. (2019) and ADMM Kingma and Ba (2015) leverage the reinforcement learning and the Alternating Direction Method of Multipliers (ADMM), respectively, to optimize the bit widths of both weights and activations for each layer on different hardware platforms. DNAS Wu et al. (2018) applies differentiable neural architecture search on mixed-precision quantization and optimizes with gradient descent. BP-NAS Yu et al. (2020), EDMIPS Cai and Vasconcelos (2020) and GMPQ Wang et al. (2021) futher improve the differentiable nerual architecture search framework on MPQ and achieve state-of-the-art results. HMQ Habi et al. (2020) utilizes the Gumbel-Softmax estimator to simultaneously search bit widths and threshold of each quantizer. Uhlich et al. (2020) uses stochastic gradient descent to search the stepsize and dynamic range for each layer, and then infers the bit width through them. And FleXOR Lee et al. (2020a) achieves fractionally mixed-precision quantization through designing an encryption algorithm.

However, those previous works do not investigate the Matthew Effect and the robustness problem in the mixed-precision quantization. In this paper, we propose a novel mixed-precision quantization search framework to address those problems.

3. Methodology

3.1. Mixed Precision Quantization Search

In this section, we formulate the mixed-precision quantization search problem and define the model complexity regularizer. Aussming we have a neural network $\mathcal{N}$ with L layers. Let $\mathcal{O}_w = \{b_1^w, b_2^w, ..., b_n^w\}$, $\mathcal{O}_a = \{b_1^a, b_2^a, ..., b_m^a\}$ $(n, m = |\mathcal{O}_w|, |\mathcal{O}_a|)$ denote the set of candidate bit widths (i.e. the search space) for weight and activation, respectively. In this paper, we perform a differentiable architecture search to achieve mixed-precision quantization. To this end, we first build a supernet $\mathcal{SN}$ based on the original architecture. Let Q denote the quantizer (quantization technique) for uniformly quantizing each layer, the forward process of the l^{th} convolutional layer in the supernet will be:

$$\begin{aligned} F_l &= \sum_{i=1}^{n} \theta_{l,i}^{w} [Q(b_i^w, w_i) * \sum_{j=1}^{m} \theta_{l,j}^{a} Q(b_j^a, X_l)] \\ &= \sum_{i=1}^{n} \theta_{l,i}^{w} \cdot (w_i^q * \sum_{j=1}^{m} \theta_{l,j}^{a} \cdot a_j^q) \end{aligned} \tag{1}$$

where X_l, F_l are the input and output of this layer, $*$ denotes the convolution operator, $\theta_{l,i}^w$, $\theta_{l,j}^a$ are the architecture weight of weight and activation, respectively. And $w_i^q = Q(b_i^w, w_i)$, $a_j^q = Q(b_j^a, X_l)$ are the b_i^w bit quantized weight and b_j^a bit quantized activation. For simplicity, we use $\theta = \theta^w \cup \theta^a$ to denote the set of architecture weights.

The goal of mixed-precision quantization is to determine the weight bit width and activation bit width for each layer to achieve the optimal trade-off between accuracy and complexity. Let $C(\mathcal{SN}, \theta)$ be the model complexity and C_0 be the pre-defined model complexity constraint. The mix-precision quantization problem can be formulated as follows:

$$\begin{aligned} \min_{\theta} \ & \mathcal{L}_{val}(w^*(\theta), \theta) \\ \text{s.t.} \ & w^*(\theta) = arg \min_{w} \mathcal{L}_{train}(w, \theta) \\ & C(\mathcal{SN}, \theta) \leqslant C_0 \end{aligned} \tag{2}$$

where $\mathcal{L}_{train}$ and $\mathcal{L}_{val}$ are the training and the validation loss, respectively. Equation 2 implies a typical bilevel constrained optimization problem, which can be solved by the method of Lagrange multiplier.

$$\mathcal{L} = \mathcal{L}_{cls} + \mu \mathcal{L}_{mc} \tag{3}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{mc}$ denote the classification loss and the model complexity loss, respectively. μ is a hyperparameter to balance the importance of different losses. In this paper, we utilize the Bit-operations (BOPs) as the metric of model complexity, for it can constrain both computation and storage cost Cai and Vasconcelos (2020); Wang et al. (2020).

$$\mathcal{L}_{mc} = \sum_{l=1}^{L} \sum_{i=1}^{n} \theta_{l,i}^{w} b_i^w \cdot \sum_{j=1}^{m} \theta_{l,j}^{a} b_j^a \cdot \text{FLOPs}^l \tag{4}$$

where FLOPs^l is the floating-point operations (FLOPs) of the l^{th} layer.

For comparison with prior works, we also define the average weight bit $\bar{b}^w$, the average activation bit $\bar{b}^a$ and the average network bit $\bar{b}$ for the quantized networks.

$$\begin{aligned} \bar{b}^w &= \frac{\text{Quantized Weights Memory}}{\text{Float Weights Memory}} * 32 = \text{Weights Compression rate} * 32 \\ \bar{b}^a &= \frac{\text{Quantized Activations Memory}}{\text{Float Activations Memory}} * 32 = \text{Activations Compression rate} * 32 \\ \bar{b} &= \sqrt{\frac{\text{Quantized Bitops}}{\text{Float Bitops}}} * 32 = \sqrt{\text{Bitops compression rate}} * 32 \end{aligned} \tag{5}$$

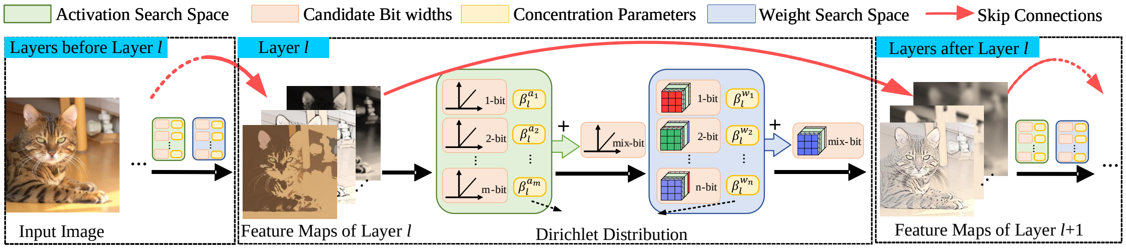

Figure 1: An overview of the supernet in our framework. The Dirichlet distribution is used to re-parameterize the architecture weights, which encourages exploration. The skip connections with gradually decaying architecture weights will be added between convolutional layers, which helps to stabilize the searching process.

3.2. Distribution Learning for MPQ

The optimization procedure in DARTS will lead to the Co-adaption problem and the Matthew effect. The same problems will also exist in the mixed-precision quantization search if it follows a similar search process. In this paper, distinguishing from the generic NAS, we redefine the Co-adaption problem and the Matthew effect in mixed-precision quantization search: 1) Candidates with some special bit widths would have natural advantages in the competition and perform better in the early stage of supernet training. While the architecture parameters of underperforming candidates will be lowered and fewer loss gradients will be propagated, which will make them harder to be trained maturely and cause the Matthew Effect. 2) Those candidates with natural advantages would be trained maturely with fewer epochs and express the desired function earlier than the others. Hence, those underperforming candidates may rarely have the chance to express the desired function. This makes the mixed-precision quantization prefer candidate bit widths which have the natural advantages but not the optimal results.

The Co-adaption problem and the Matthew effect in the generic DARTS can be attributed to the performance gap between different candidate operations over various application scenarios. Hence, without the complexity regularizer, those problems can be even worse in the mixed-precision quantization since there are natural differences in the expressive abilities of candidate bit widths. The higher bit-width candidates will almost always occupy the advantaged positions. And the performance gap can not be ignored, especially in pursuing ultra low-bit quantization. Whereas, under the complexity regularizer, the Co-adaption performance will not be eliminated. Reverse discrimination could be introduced, and the training of some special bit widths is encouraged while the others are inhibited.

In this paper, we seek to alleviate the Co-adaption problem and the Matthew effect in mixed-precision quantization without expensive computational cost. Unlike DropNAS Hong et al. (2020) which utilizes the dropout, we propose to formulate the mixed-precision problem as a distribution learning, which naturally induces stochasticity and encourages exploration. Inspired by previous works Blei et al. (2003); Lee et al. (2020b); Joo et al. (2020); Chen et al. (2021), we leverage Dirichlet distribution to the continuously relaxed model architecture weights. Dirichlet distribution is a widely used distribution over the probability simplex and enjoys a nice property that it enables gradient-based training.

Specifically, we will convert the supernet to a stochastic supernet whose edges are executed stochastically and treat the architecture weight θ as random variables sampled from a Dirichlet distribution. The formulation of the bilevel optimization problem in equation 2 will transfer to:

$$\begin{aligned} &\min_{\beta} E_{dt(\theta|\beta)}[\mathcal{L}_{val}(w^*(\theta),\theta)] \\ &\text{s.t.} \quad w^*(\theta) = \arg\min_{w} \mathcal{L}_{train}(w,\theta) \end{aligned} \tag{6}$$

where $dt(\theta|\beta)$ is the distribution of θ parameterized by the Dirichlet concentration parameter β, i.e. $dt(\theta|\beta) \sim Dir(\beta)$. The gradient of β can not be directly calculation, but could be approximated via pathwise derivative estimators Jankowiak and Obermeyer (2018).

$$\frac{d\theta_i}{d\beta_j} = -\frac{\frac{\partial F_{Beta}}{\partial \beta_j}(\theta_j|\beta_j,\beta_{tot}-\beta_j)}{\beta_{Beta}(\theta_j|\beta_j,\beta_{tot}-\beta_j)} \times \frac{\delta_{i,j}-\theta_i}{1-\theta_j} \quad i,j = 1,...,|\mathcal{O}| \tag{7}$$

where F_{Beta} and f_{Beta} denote the CDF and PDF of beta distribution respectively, $\delta_{i,j}$ is the indicator function, and β_{tot} is the sum of concentrations. F_{Beta} is the irregularized incomplete beta function, for which the gradient can be computed by simple numerical approximation. When the search phase finishes, we select the bit width for the l^{th} layer by Equation 8.

$$\begin{aligned} b_l^w &= \arg\min_{\theta_i^w \in \mathcal{O}_w} E_{dt(\theta_i^w|\beta_l^w)}[\theta_i^w] \\ b_l^a &= \arg\min_{\theta_i^a \in \mathcal{O}_a} E_{dt(\theta_i^a|\beta_l^a)}[\theta_i^a] \end{aligned} \tag{8}$$

It has been proved that formulating the differentiable architecture search as a distribution learning problem can improve the generalization ability Chen et al. (2021). And through this transformation, we encourage the exploration of the mixed-precision quantization search, which alleviates the Matthew effect and the Co-adaption problem.

3.3. Skip Connection to Stabilize Search

A big difference between the mixed-precision quantization and the generic NAS is that the size of search space in mixed-precision quantization will grow rapidly as the depth of the network increases. For example, if $n = m = 4$ and $L = 20$, the size of search space will be $4^{20} \times 4^{20} = 2^{80}$. But if $L = 50$, the size of search space will be $4^{50} \times 4^{50} = 2^{200}$, which is far greater than $8^{14} \times 8^{14} = 2^{112}$ in DARTS. The rapidly growing search space size increases the difficulty of supernet training and makes the search unstable, which motivates us to pursue a novel component for stabilizing the search process in our framework.

Since ResNet He et al. (2016) constructs a residual block with skip connections and significantly improves training stability, the skip connection has been widely used. But in DARTS, superfluous skip connections can cause unfair competition and even lead to the performance collapse. However, a recent study Chu et al. (2021) claims that the skip connection in DARTS actually plays twofold roles: an auxiliary connection to stabilize the

```
Algorithm 1: The whole pipeline of our framework
Input: Network $\mathcal{N}$ with $L$ layers; Set of candidate bit widths for weight and activation
       $\mathcal{O}_w$, $\mathcal{O}_a$; The quantizer $Q$; Search epochs $E_s$; Quantization epochs $E_q$.
Output: The quantized network $\mathcal{QN}$.
Build a supernet $\mathcal{SN}$ based on $\mathcal{O}_w$, $\mathcal{O}_a$, and $\mathcal{N}$.
Add a skip connection between every two layers in the supernet $\mathcal{SN}$.
The search process:
for $e = 1 \rightarrow E_s$ do
    Update weights $w$ by descending $\nabla_w \mathcal{L}_{train}(w, \theta(\beta))$
     Update concentration parameters $\beta$ by descending $\nabla_\beta \mathcal{L}_{val}(w, \beta)$
end
Build the quantization network $\mathcal{QN}$ with the searched bit width configurations.
Train $\mathcal{QN}$ from scratch with $E_q$ and get the quantization network.
```

supernet training and a candidate operation to build the final network. Motivated by this conclusion, we propose to add a skip connection between each two layers in the supernet (Fig. 1) with a special architecture weight η. η is a pre-defined value, which does not participate in the bit-widths competition but alleviates the gradient vanishing problem. For simplicity, we use θ to denote $\theta(\beta)$ in this section. When the skip connection is added to the supernet, the forward process of the k^{th} convolutional layer will be:

$$\begin{aligned} F_k^s &= \sum_{i=1}^{n} \theta_{k,i}^w \cdot (w_i^q * \sum_{j=1}^{m} \theta_{k,j}^a \cdot a_j^q) + \eta \cdot X_k \\ \eta &= \eta_0 - g(epoch) \end{aligned} \tag{9}$$

where F_k^s is the output feature, which is used to distinguish with F_k. η_0, g are pre-defined initial value and the decay strategy of η, respectively. $epoch$ is the current search epoch and we decrease η to 0 in the search phase. Let $\mathcal{L}$ stand for the loss fuction, the backward process of the k-th convolution layer will be:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial X_k} &= \frac{\partial \mathcal{L}}{\partial F_k^s} \cdot \frac{\partial F_k^s}{\partial X_k} = \frac{\partial \mathcal{L}}{\partial F_k^s} \cdot (\sum_{i=1}^{n} \theta_{k,i}^w \cdot (w_i^q * \sum_{j=1}^{m} \theta_{k,j}^a \cdot \frac{\partial a_j^q}{\partial X_k}) + \eta \cdot \mathbb{1}) \\ &= \frac{\partial \mathcal{L}}{\partial F_k^s} \cdot (\frac{\partial F_k}{\partial X_k} + \eta \cdot \mathbb{1}) = \frac{\partial \mathcal{L}}{\partial F_L^s} \cdot \prod_{r=1}^{L-k} (\frac{\partial F_{k+r}}{\partial X_{k+r}} + \eta \cdot \mathbb{1}) \end{aligned} \tag{10}$$

where $\mathbb{1}$ denotes a tensor whose items are all ones. In this paper, we set the initial value of η as η_0. According to equation 10, the skip connection can alleviate the gradient vanishing and stabilize the supernet training when the weights of candidates have not been trained maturely. And we gradually decrease η in the search phase. When the search is near to finish, η will tend to be zero, which will not disturb bit-width competition. In conclusion, the skip connection mainly works in the early stage of search.

The whole pipeline of our quantization algorithm is illustrated in Algorithm 1. We first search the bit width configurations and then quantize the network from scratch.

Table 1: Comparison with prior arts on CIFAR-10. We abbreviate accuracy as ’Acc’, compression rate as ’Comp’, weights as ’W’, activations as ’A’, bitops as ’B’ and mixed-precision as ’MP’, respectively. $\overline{b}^w$, $\overline{b}^a$, $\overline{b}$ are calculated following Equation 5.

Network	Method	Acc (%)	Δ Acc (%)	$\overline{b}^a$	W Comp	$\overline{b}^w$	B Comp	$\overline{b}$
ResNet-20	Baseline	92.1	+0.0	32	1	32	1	32
	TTQ Zhu et al. (2017)	91.13	-0.97	32	16	2	16	8
	HAWQ Dong et al. (2019)	92.22	+0.12	2.43_{MP}	13.11	4	64.49	4
	BP-NAS Yu et al. (2020)	92.3	+0.2	3.14_{MP}	10.19	MP	81.53	3.5
	Ours	92.62	+0.52	3.59_{MP}	8.91	3.39_{MP}	83.52	3.50
	Dorefa Zhou et al. (2016)	89.9	-2.2	3	10.67	3	113.78	3
	PACT Choi et al. (2018)	91.1	-1.0	3	10.67	3	113.78	3
	LQ-Nets Zhang et al. (2018)	91.6	-0.5	3	10.67	3	113.78	3
	BP-NAS Yu et al. (2020)	92.12	+0.02	2.86_{MP}	10.74	MP	95.61	3.3
	BP-NAS Yu et al. (2020)	92.04	-0.06	2.65_{MP}	12.08	MP	116.89	2.9
	Ours	92.36	+0.26	2.85_{MP}	11.2	3.33_{MP}	113.2	3.00
	Dorefa Zhou et al. (2016)	88.2	-3.9	2	16	2	256	2
	PACT Choi et al. (2018)	89.7	-2.4	2	16	2	256	2
	LQ-Nets Zhang et al. (2018)	90.2	-1.9	2	16	2	256	2
	Ours	91.08	-1.02	1.81_{MP}	17.64	2.62_{MP}	258.52	1.99
ResNet-56	Baseline	93.47	+0.0	32	1	32	1	32
	TTQ Zhu et al. (2017)	93.56	+0.09	32	16	2	16	8
	Ours	93.95	+0.48	3.44_{MP}	9.28	3.63_{MP}	83.57	3.50
	Ours	93.40	-0.07	2.60_{MP}	12.21	3.32_{MP}	113.59	3.00
	Ours	92.65	-0.82	1.42_{MP}	22.43	2.69_{MP}	258.02	1.99
MobileNetv2	Baseline	95.22	+0.0	32	1	32	1	32
	Ours	94.81	-0.41	7.44_{MP}	4.36	7.33_{MP}	29.37	5.90
	Ours	94.41	-0.81	6.73_{MP}	5.43	5.88_{MP}	60.28	4.12
	Ours	93.47	-1.75	5.96_{MP}	5.96	5.37_{MP}	101.18	3.18

4. Experiments

In this section, we evaluate our framework on classification tasks. First, we provide the implementation details and compare our method with the prior quantization methods. Besides, we conduct a series of ablation studies to investigate the impact of components in our framework. All the experiments are accomplished with PyTorch Paszke et al. (2019).

4.1. Network Quantization

We present the experiment results on two widely-used image classification datasets, CIFAR-10 Krizhevsky (2009) and ImageNet Russakovsky et al. (2015) to illustrate the effectiveness of the proposed quantization method. We directly quote some results of the existing methods from previous works.

4.1.1. CIFAR-10

For CIFAR-10, we choose ResNet-20, ResNet-56 He et al. (2016) and MobileNetv2 Sandler et al. (2018), which are the most popular DNN models in compression works, to demonstrate the effectiveness of our methods. We utilize HWGQ-Net Cai et al. (2017) for uniformly quantizing each layer. First, we search the bit-width configurations under the model complexity regularizer for 50 epochs, with the same hyper-parameter setting of a batch size of 256, a weight decay of 5e-4, and the SGD optimizer with a momentum of 0.9. The learning rate for network parameters is set to 0.1 initially and divided by 10 at every 20

Table 2: Comparison with Top-1 accuracy of prior arts on ImageNet. We abbreviate accuracy as ”Acc”, compression rate as ”Comp”, weights as ”W”, activations as ”A”, bitops as ”B”, mixed-precision as ”MP”, respectively. $\overline{b}^w$, $\overline{b}^a$, $\overline{b}$ are calculated following Equation 5.

Network	Method	Acc (%)	Δ Acc (%)	$\overline{b}^a$	W Comp	$\overline{b}^w$	B Comp	$\overline{b}$
ResNet-18	Baseline	69.54	+0.0	32	1	32	1	32
	TTQ Zhu et al. (2017)	66.6	-2.94	32	16	2	16	8
	Dorefa Zhou et al. (2016)	68.1	-1.44	4	8	4	64	4
	PACT Choi et al. (2018)	69.2	-0.34	4	8	4	64	4
	LQ-Nets Zhang et al. (2018)	69.3	-0.24	4	8	4	64	4
	DSQ Gong et al. (2019)	69.56	+0.02	4	8	4	64	4
	EdMIPS Cai and Vasconcelos (2020)	≈ 67.5	-2.04	MP	-	MP	-	≈ 3.5
	Ours	69.56	+0.02	3.40_{MP}	8.75	3.66_{MP}	85.98	3.45
	Dorefa Zhou et al. (2016)	67.5	-2.04	3	10.67	3	113.78	3
	PACT Choi et al. (2018)	68.1	-1.44	3	10.67	3	113.78	3
	LQ-Nets Zhang et al. (2018)	68.2	-1.34	3	10.67	3	113.78	3
	DSQ Gong et al. (2019)	68.66	-0.88	3	10.67	3	113.78	3
	EdMIPS Cai and Vasconcelos (2020)	≈ 66.5	-3.04	MP	-	MP	-	≈ 2.8
	Ours	68.72	-0.82	2.99_{MP}	10.66	3.00_{MP}	126.03	2.85
	Ours	68.25	-1.29	2.79_{MP}	12.23	2.62_{MP}	174.18	2.42
ResNet-50	Baseline	76.14	+0.0	32	1	32	1	32
	Dorefa Zhou et al. (2016)	71.4	-4.74	4	8	4	64	4
	LQ-Nets Zhang et al. (2018)	75.1	-1.04	4	8	4	64	4
	PACT Choi et al. (2018)	76.5	+0.36	4	8	4	64	4
	BP-NAS Yu et al. (2020)	76.67	+0.53	MP	-	MP	71.65	3.8
	EdMIPS Cai and Vasconcelos (2020)	≈ 73.0	-3.14	MP	-	MP	-	≈ 3.5
	Ours	76.42	+0.28	3.84_{MP}	8.83	3.62_{MP}	90.96	3.36
	Dorefa Zhou et al. (2016)	69.9	-6.24	3	10.67	3	113.78	3
	PACT Choi et al. (2018)	75.3	-0.84	3	10.67	3	113.78	3
	LQ-Nets Zhang et al. (2018)	74.2	-1.94	3	10.67	3	113.78	3
	HAQ Wang et al. (2019)	75.3	-0.84	MP	10.57	MP	-	-
	HAWQ Dong et al. (2019)	75.48	-0.66	2_{MP}	12.28	4_{MP}	-	-
	HAWQv2 Dong et al. (2020)	75.76	-	4_{MP}	2_{MP}	12.24	-	-
	EdMIPS Cai and Vasconcelos (2020)	≈ 72.5	-3.64	MP	-	MP	-	≈ 2.9
	BP-NAS Yu et al. (2020)	75.71	-0.43	MP	-	MP	118.98	2.9
	Ours	75.87	-0.27	3.29_{MP}	12.22	2.62_{MP}	156.10	2.56

epochs, while the learning rate for architecture weights θ is set to 0.01. All concentration parameters β are initialized to 0.01 to ensure fairness. When the search phase finishes, we select the bit width for weights and activations following Equation 8. Then, we quantize the network from scratch for 600 epochs with an initial learning rate of 0.2 and the cosine decay strategy. For MobileNetv2, we utilize DoReFa Zhou et al. (2016) for uniformly quantizing each layer and change the initial learning rate of the quantizing phase to 0.025.

The experimental results are summarized in Table 1. Our results compare favorably to other quantization methods. For example, On ResNet-20, compared with uniform quantization methods, Dorefa Zhou et al. (2016), PACT Choi et al. (2018) and LQ-Nets Zhang et al. (2018), our method achieves higher accuracy when quantizing the network to 3 average bit. Compared with the mixed-precision method HAWQ Dong et al. (2019), which achieves 92.22% accuracy with a 64.49x Bitops compression rate, our method achieves a little higher accuracy of 92.36% with an almost double Bitops compression rate 113.2x.

4.1.2. ImageNet

For ImageNet, we use ResNet-18 and ResNet-50 He et al. (2016) as the baseline models. In the searching phase, the search epoch is set to 25, and the learning rate for network

Table 3: Comparison of Top-1 accuracies of ResNet-56 on CIFAR-10 with/without Dirichlet distribution learning and skip connections.

Dirichlet	Skip	Acc (%)	Average network bit
✓	✓	93.40	3.00
✗	✓	93.38	3.01
✓	✗	93.25	3.04
✗	✗	93.08	3.01

parameters is set to 0.1 initially and divided by 10 at every 10 epochs. And in the quantizing phase, the training epoch and weight decay are set to 100 and 1e-4, respectively. The rest of the hyperparameters are set following CIFAR-10 experiments.

The experimental results are summarized in Table 2. Our method achieves a better accuracy-complexity trade-off under various resource constraints. On ResNet-18, our method achieves an 85.98× Bitops compression rate with a 69.59% accuracy, which is even higher than the full-precision baseline. On ResNet-50, our method achieves the state-of-the-art Bitops compression rate 156.10x with maintaining 75.87% accuracy.

4.2. Ablation Study

4.2.1. Skip Connection and Distribution Learning

In this section, we investigate the performance of Dirichlet distribution learning and the skip connections of our framework when they work alone. We conduct experiments with the following settings: 1) replacing the Dirichlet distribution with softmax operation, 2) removing the skip connection in the supernet; 3) enforcing both of the above. The experimental results of quantizing ResNet-56 to about 3 network average bit is shown in Table 3. The accuracies of the quantized networks drop by 0.02% and 0.15% without Dirichlet distribution and skip connection, respectively, while their average network bits both increase. Both of them fail to find the optimal bit allocation. But they still perform better than the baseline, which proves the effectiveness of Dirichlet distribution and the skip connections.

4.2.2. Stability

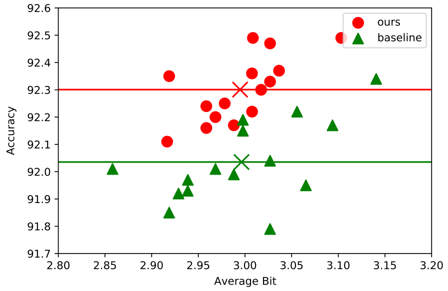

Figure 2: 15 sets of our method and baseline. Line denotes the average accuracies.

Table 4: Comparison of Top-1 accuracies of ResNet-2 on CIFAR-10 with different initial architecture weight η_0

initial η	Acc (%)	Average network bit
2	90.94	2.02
1	91.08	1.99
0.8	91.10	2.01
0.5	91.17	1.99
0	90.54	2.01

In our framework, we utilize skip connections to stabilize the supernet training. To verify the robustness, we search several times for quantizing ResNet-20 under the same complexity decay. The network bit of the baseline (without skip connections) is 2.9963 averagely and scatterd over a -0.1382 to 0.1443 error range with a 5.18e-3 variance, and the average quantized accuracy is 92.03%. Compared with the baseline, our method achieves 2.9947 average bit networks, whose error only ranges from -0.0781 to 0.0984 with a 2.13e-3 variance. And our method improves the average performance of quantized ResNet-20 to 92.30%. Those experimental results illustrated in Figure 2 demonstrate that our method can improve the stability of the search.

4.2.3. The initial value of η

The initial value η_0 of the architecture weight η on skip connections is an important hyperparameter, which influences the power of the auxiliary skip connections in the supernet. To investigate the influence of different η_0, we quantize ResNet-20 to 2 average bit with $\eta_0 \in \{2, 1, 0.8, 0.5, 0\}$. Table 4 illustrates that when η_0 is within a reasonable range, our method maintains good performance. However, overlarge or too small η_0 leads to worse performance. In conclusion, setting $\eta_0 = 1$ is a suitable strategy, which avoids the complicated hyper-parameter tuning process while gaining significant accuracy boost.

4.3. Visualization

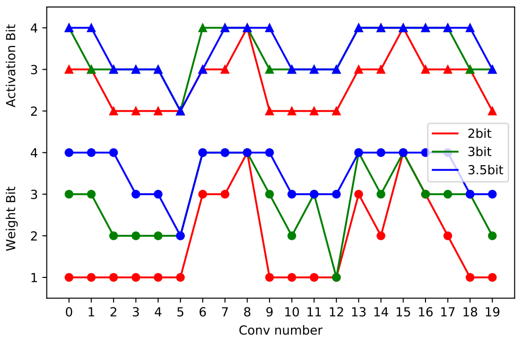

Figure 3: Bit allocation of ResNet-20.

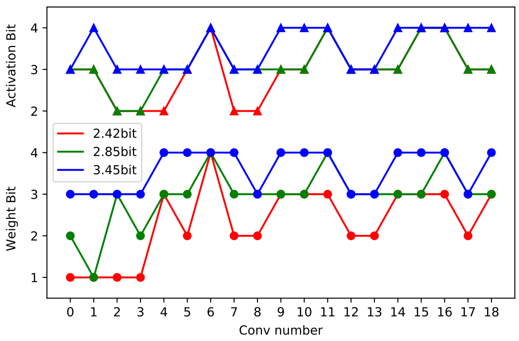

Figure 4: Bit allocation of ResNet-18.

In this section, we present the searched bit allocation for ResNet-20 on CIFAR-10 and ResNet-18 on ImageNet. The bit width curves shown in Fig 3 and 4 illustrates that the quantization bit for activations or weights of the same layer mostly increases or stays the same when the average network bit increases. It can be interpreted as that, when the supernet gets more computational resources, it will reweigh the importance of all layers and allocate those extra resources to them. This phenomenon demonstrates convincingly that our method is consistent with different datasets and stable enough to find the inner pattern of the mixed-precision quantization.

5. Conclusion

In this paper, we propose a novel mixed-precision quantization framework for alleviating the Matthew effect and improving robustness. We study and redefine the Matthew effect and the Co-adaption problem in the mixed-precision quantization. Then, we resolve the mixed-precision quantization search as a distribution learning problem, which encourages the exploration and alleviates the Matthew effect and the Co-adaption problem. Furthermore, we investigate the lack of skip connections in the search space of mixed-precision quantization. We propose to add a skip connection between every two convolutional layers in the supernet, which improves robustness and does not influence the bit-widths competition. Extensive experiments depict the effectiveness of our framework compared with the state-of-the-art methods. Of note is that our method is simple to implement using mainstream frameworks.

Acknowledgments

This work was supported in part by National Defense Basic Scientific Research program of China under Grant JCKY202210C005, in part by the Natural Science Foundation of Shaanxi Province under Grant 2022JM-366, and in part by the China Postdoctoral Science Foundation under Grant 227556, and Grant 2020T130514.

References

David M. Blei, A. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13166–13175, 2020.

Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2346–2355, 2020.

Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5406–5414, 2017.

Xiangning Chen, Ruochen Wang, Minhao Cheng, Xiaocheng Tang, and Cho-Jui Hsieh. Drnas: Dirichlet neural architecture search. *ArXiv*, abs/2006.10355, 2021.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and K. Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *ArXiv*, abs/1805.06085, 2018.

Xiangxiang Chu, Tianbao Zhou, Bo Zhang, and Jixiang Li. Fair darts: Eliminating unfair advantages in differentiable architecture search. *ArXiv*, abs/1911.12126, 2020.

Xiangxiang Chu, Xiaoxing Wang, Bo Zhang, Shun Lu, Xiaolin Wei, and Junchi Yan. Darts-: Robustly stepping out of performance collapse without indicators. *ArXiv*, abs/2009.01027, 2021.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, 2015.

Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *IJCAI*, 2015.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 293–302, 2019.

Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *ArXiv*, abs/1911.03852, 2020.

Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tian-Hao Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4851–4860, 2019.

Hai Victor Habi, Roy H. Jennings, and Arnon Netzer. Hmq: Hardware friendly mixed precision quantization block for cnns. *ArXiv*, abs/2007.09952, 2020.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Weijun Hong, Guilin Li, Weinan Zhang, Ruiming Tang, Yunhe Wang, Zhenguo Li, and Yong Yu. Dropnas: Grouped operation dropout for differentiable architecture search. *ArXiv*, abs/2201.11679, 2020.

Martin Jankowiak and Fritz Obermeyer. Pathwise derivatives beyond the reparameterization trick. *ArXiv*, abs/1806.01851, 2018.

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognit.*, 107:107514, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Dongsoo Lee, Se Jung Kwon, Byeongwook Kim, Yongkweon Jeon, Baeseong Park, and Jeongin Yun. Flexor: Trainable fractional quantization. *ArXiv*, abs/2009.04126, 2020a.

Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. *ArXiv*, abs/2001.00689, 2020b.

Hanwen Liang, Shifeng Zhang, Jiacheng Sun, Xingqiu He, Weiran Huang, Kechen Zhuang, and Zhenguo Li. Darts+: Improved differentiable architecture search with early stopping. *ArXiv*, abs/1909.06035, 2019.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *ArXiv*, abs/1806.09055, 2019.

Xingchao Liu, Mao Ye, Dengyong Zhou, and Qiang Liu. Post-training quantization with multiple points: Mixed precision without mixed precision. In *AAAI*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Zhongnan Qu, Zimu Zhou, Yun Cheng, and Lothar Thiele. Adaptive loss-aware quantization for multi-bit networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7985–7994, 2020.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):4780–4789, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso García, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Mixed precision dnns: All you need is a good parametrization. In *ICLR*, 2020.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8604–8612, 2019.

Ying Wang, Yadong Lu, and Tijmen Blankevoort. Differentiable joint pruning and quantization for hardware efficiency. *ArXiv*, abs/2007.10463, 2020.

Ziwei Wang, Han Xiao, Jiwen Lu, and Jie Zhou. Generalizable mixed-precision quantization via attribution rank preservation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5271–5280, 2021.

Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Péter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *ArXiv*, abs/1812.00090, 2018.

Haibao Yu, Qi Han, Jianbo Li, Jianping Shi, Guangliang Cheng, and Bin Fan. Search what you want: Barrier panelty nas for mixed precision quantization. In *ECCV*, 2020.

Arber Zela, Thomas Elsken, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, and Frank Hutter. Understanding and robustifying differentiable architecture search. *ArXiv*, abs/1909.09656, 2020.

Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. *ArXiv*, abs/1807.10029, 2018.

Haonan Zhang, Longjun Liu, Hengyi Zhou, Wenxuan Hou, Hongbin Sun, and Nanning Zheng. Akecp: Adaptive knowledge extraction from feature maps for fast and efficient channel pruning. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

Haonan Zhang, Longjun Liu, Hengyi Zhou, Hongbin Sun, and Nanning Zheng. Cmd: controllable matrix decomposition with global optimization for deep neural network compression. *Machine Learning*, 2022.

Hengyi Zhou, Longjun Liu, Haonan Zhang, Hongyi He, and Nanning Zheng. Cmb: A novel structural re-parameterization block without extra training parameters. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2022. doi: 10.1109/IJCNN55064.2022.9892874.

Shuchang Zhou, Zekun Ni, Xinyu Zhou, He Wen, Yuxin Wu, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *ArXiv*, abs/1606.06160, 2016.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J. Dally. Trained ternary quantization. *ArXiv*, abs/1612.01064, 2017.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *ArXiv*, abs/1611.01578, 2017.