# BATS: Binary ArchitecTure Search

Adrian Bulat[1][0000−0002−3185−4979], Brais Martinez[1][0000−0001−7511−8941], and Georgios Tzimiropoulos[1,2][0000−0002−1803−5338]

[1] Samsung AI Center, Cambridge, UK
adrian@adrianbulat.com, brais.mart@gmail.com
[2] Queen Mary University of London, London, UK
g.tzimiropoulos@qmul.ac.uk

**Abstract.** This paper proposes Binary ArchitecTure Search (BATS), a framework that drastically reduces the accuracy gap between binary neural networks and their real-valued counterparts by means of Neural Architecture Search (NAS). We show that directly applying NAS to the binary domain provides very poor results. To alleviate this, we describe, to our knowledge, for the first time, the 3 key ingredients for successfully applying NAS to the binary domain. Specifically, we (1) introduce and design a novel binary-oriented search space, (2) propose a new mechanism for controlling and stabilising the resulting searched topologies, (3) propose and validate a series of new search strategies for binary networks that lead to faster convergence and lower search times. Experimental results demonstrate the effectiveness of the proposed approach and the necessity of searching in the binary space directly. Moreover, (4) we set a new state-of-the-art for binary neural networks on CIFAR10, CIFAR100 and ImageNet datasets. Code will be made available.

**Keywords:** Binary networks, Neural Architecture Search

## 1 Introduction

Network quantization and Network Architecture Search (NAS) have emerged as two important research directions with the goal of designing efficient networks capable of running on mobile devices. Network quantization reduces the size and computational footprint of the models by representing the activations and the weights using $N < 32$ bits. Of particular interest is the extreme case of quantization, binarization, in which the model and the activations are quantized to a single bit [33, 10, 9]. This allows to replace all floating point multiply-add operations inside a convolutional layer with bit-wise operations resulting in a speed-up of up to $57\times$ [33]. Recent years have seen a progressive reduction of the performance gap with real-valued networks. However, research has almost exclusively focused on the ResNet architecture, rather than on efficient ones such as MobileNet. This is widely credited to pointwise convolutions not being amenable to binarization [4, 2].

As an orthogonal direction, NAS attempts to improve the overall performance by automatically searching for optimal network topologies using a various of approaches including evolutionary algorithms [41, 34], reinforcement learning [47,

52] or more recently by taking a differentiable view of the search process via gradient-based approaches [27, 25]. Such methods were shown to perform better than carefully hand-crafted architectures on both classification [25, 27, 32, 52] and fine-grained tasks [24]. However, all of the aforementioned methods are tailored towards searching architectures for real-valued neural networks.

The aim of this work is to propose, for the first time, ways of reducing the accuracy gap between binary and real-valued networks via the state-of-the-art framework of differentiable NAS (DARTS) [27, 25].

We show that direct binarization of existing cells searched in the real domain leads to sub-par results due to the particularities of the binarization process, in which certain network characteristics typically used in real-valued models (e.g. $1 \times 1$ convolutions) may be undesired [4, 2]. Moreover, we show that performing the search in the binary domain comes with a series of challenges of its own: it inherits and amplifies one of the major DARTS [27] drawbacks, that of "cell collapsing", where the resulting architectures, especially when trained for longer, can result in degenerate cells in which most of the connections are skip connections [25] or parameter-free operations. In this work, we propose a novel Binary Architecture Search (BATS) method that successfully tackles the above-mentioned issues and sets a new state-of-the-art for binary neural networks on the selected datasets, yet simultaneously reducing computational resources by a wide margin. In summary, **our contributions** are:

1. We show that directly applying NAS to the binary domain provides very poor results. To alleviate this, we describe, to our knowledge, for the first time, the 3 key ingredients for successfully applying NAS to the binary domain.
2. We devise a novel search space specially tailored to the binary case and incorporate it within the DARTS framework (Section 4.1).
3. We propose a temperature-based mechanism for controlling and stabilising the topology search (Section 4.2).
4. We propose and validate a series of new search strategies for binary networks that lead to faster convergence and lower search times (Section 4.3).
5. We show that our method consistently produces superior architectures for binarized networks within a lower computational budget, setting a new state-of-the-art on CIFAR10/100 and ImageNet datasets (Section 6).

## 2   Related work

**NAS:** While hand-crafted architectures significantly pushed the state-of-the-art in Deep Learning [18, 36, 46, 30, 19, 16], recently, NAS was proposed as an automated alternative, shown to produce architectures that outperform the manually designed ones [35]. NAS methods can be roughly classified in three categories: evolutionary-based [41, 34, 35], reinforcement learning-based [1, 47, 52] and, more recently, one-shot approaches, including differentiable ones [3, 27, 24, 6]. The former two categories require significant computational resources during the search phase (3150 GPU-days for AmoebaNet [34] and 1800 GPU-days for NAS-Net [52]). Hence, of particular interest in our work is the differentiable

search framework (DARTS) of [27] which is efficient. Follow-up work further improves on it by progressively reducing the search space [8] or by constraining the architecture parameters to be one-hot [42]. Our method builds upon the DARTS framework incorporating the progressive training of [8]. In contrast to all the aforementioned methods that search for optimal real-valued topologies, to our knowledge, we are the first to study NAS for binary networks.

**Network binarization** is the most extreme case of network quantization in which weights and activations are represented with 1 bit. It allows for up to $57\times$ faster convolutions [33], a direct consequence of replacing all multiplications with bitwise operations [9]. Typically, binarization is achieved by taking the sign of the real-valued weights and features [9, 10]. However, such an approach leads to large drops in accuracy, especially noticeable on large scale datasets. To alleviate this, Rastegari *et al.* [33] introduce analytically computed real-valued scaling factors that scale the input features and weights. This is further improved in [5] that proposes to learn the factors via back-propagation. In Bi-Real Net, Lin *et al.* [28] advocate the use of double-skip connections and of real-valued down-sample layers. Wang *et al.* [40] propose a reinforcement learning-based approach to learn inter-channel correlations to better preserve the sign of convolutions. Ding *et al.* [14] introduce a distribution loss to explicitly regularize the activations and improve the information flow thought the network. While most of these methods operate within the same computational budget, another direction of research attempts to bridge the gap between binary and their real-valued networks by increasing the network size. For example, ABC-Net [23] proposes to use up to $M = N = 5$ branches, equivalent with running $M \times N$ convolutional layers in parallel, while Zhu *et al.* [50] use an ensemble of up to 6 binary models. These methods expand the size of the network while preserving the general architecture of the ResNet [16] or WideResNet [44]. In contrast to all the aforementioned methods in which binary architectures are hand-crafted, we attempt to automatically discover novel binary architectures, without increasing the computational budget. Our discovered architectures set a new state-of-the-art on the most widely used datasets.

**Very recent work on binary NAS** done concurrently with our work include [37] and [38]. As opposed to our work, [37] simple searches for the number of channels in each layer inside a ResNet. [38] uses a completely different search space and training strategy. We note that we outperform significantly both of them in terms of accuracy/efficiency. On ImageNet, using $1.55 \times 10^8$ FLOPs, we obtain a Top-1 accuracy of **66.1%** vs 58.76% using $1.63 \times 10^8$ FLOPS in [38]. On CIFAR-10 we score a top-1 accuracy of 96.1% vs 94.43% in [37]. While [37] achieves an accuracy of 69.65% on ImageNet, they use $6.6 \times 10^8$ FLOPS ($4.2\times$ more than our biggest model).

## 3   Background

**Network binarization:** We binarize the models using the method proposed by Rastegari *et al.* [33] with the modifications introduced in [5]: we learn the weight

scaling factor $\alpha$ via back-propagation instead of computing it analytically while dropping the input feature scaling factor $\beta$. Assume a given convolutional layer $L$ with weights $\mathcal{W} \in \mathbb{R}^{o \times c \times w \times h}$ and input features $\mathcal{I} \in \mathbb{R}^{c \times w_{in} \times h_{in}}$, where $o$ and $c$ represent the number of output and input channels, $(w, h)$ the width and height of the convolutional kernel, $(w_{in} \geq w, h_{in} \geq h)$ the spatial dimensions of $\mathcal{I}$. The binarization is accomplished by taking the sign of $\mathcal{W}$ and $\mathcal{I}$ and then multiplying the output of the convolution by the trainable scaling factor $\alpha \in \mathbb{R}^+$:

$$\mathcal{I} * \mathcal{W} \approx (\operatorname{sign}(\mathcal{I}) \circledast \operatorname{sign}(\mathcal{W})) \odot \alpha, \tag{1}$$

where $\odot$ denotes the element-wise multiplication, $*$ the real-valued convolution and $\circledast$ its binary counterpart. During training, the gradients are used to update the real-valued weights $\mathcal{W}$ while the forward pass is done using $\operatorname{sign}(\mathcal{W})$. Finally, we used the standard arrangement for the binary convolutional layer operations: Batch Norm, Sign, Convolution, Activation.

**DARTS:** Herein, we review DARTS [27] and P-DARTS [8] upon which we build our framework: instead of searching for the whole network architecture, DARTS breaks down the network into a series of $L$ identical cells. Each cell can be seen as a Direct Acyclic Graph (DAG) with $N$ nodes, each representing a feature tensor. Within a cell, the goal is to choose, during the search phase, an operation $o(\cdot)$ from the predefined search space $O$ that will connect a pair of nodes. Given a pair of such nodes $(i, j)$ the information flow from $i$ to $j$ is defined as: $f_{i,j}(\mathbf{x}_i) = \sum_{o \in O} \frac{\exp(\alpha_{i,j}^o)}{\sum_{o' \in O} \exp(\alpha_{i,j}^{o'})} \cdot o(\mathbf{x}_i)$, where $\mathbf{x}_i$ is the output of the $i$-th node and $\alpha_{i,j}^o$ is an architecture parameter used to weight the operation $o(\mathbf{x}_i)$. The output of the node is computed as the sum of all inputs $\mathbf{x}_j = \sum_{i<j} f_{i,j}(\mathbf{x}_i)$ while the output of the cell is a depth-wise concatenation of all the intermediate nodes minus the input ones $(2, 3, \ldots, N-1)$. During the search phase, the networks weights $\mathcal{W}$ and the architectures parameters $\alpha$ are learned using bi-level optimization where the weights and $\alpha$ are optimized on two different splits of the dataset. Because the search is typically done on a much shallower network due to the high computational cost caused by the large search space $O$, DARTS may produce cells that are under-performing when tested using deeper networks. To alleviate this, in [8] the search is done using a series of stages during which the worse performing operations are dropped and the network depth increases. For a complete detailed explanation see [27] and [8].

## 4   Method

This section describes the 3 key components proposed in this work: the binary search space, the temperature regularization and search strategy. As we also show experimentally in Section 5.1, we found that *all 3 components are necessary for successfully applying NAS to the binary domain*[3].

---

[3] When any of the components was not used, the obtained results were very poor.

### 4.1   Binary Neural Architecture Search Space

Since searching across a large space $O$ is computationally prohibitive, most of the recent NAS methods manually define a series of 8 operations known to produce satisfactory results when used in hand-crafted network topologies: $3 \times 3$ and $5 \times 5$ dilated convolutions, $3 \times 3$ and $5 \times 5$ separable convolutions, $3 \times 3$ max pooling, $3 \times 3$ average pooling, identity (skip) connections plus the zero (no) connection. Table 1 summarizes the operations used for the real-valued case. Having an appropriate search space containing only desirable operations is absolutely essential for obtaining good network architectures: for example, a random search performed on the aforementioned space on CIFAR-10 achieves already 3.29% top-error vs. 2.76% when searched using second-order DARTS [27].

Table 1: Comparison between the commonly used searched space for real valued networks and the proposed one, specially tailed for binary models.

| Real-valued | Ours (proposed) |
|---|---|
| Separable conv. $(3 \times 3)$ | Group convolution $(3 \times 3)$ |
| Separable conv. $(5 \times 5)$ | Group convolution $(5 \times 5)$ |
| Dilated conv. $(3 \times 3)$ | Dilated Group conv. $(3 \times 3)$ |
| Dilated conv. $(5 \times 5)$ | Dilated Group conv. $(5 \times 5)$ |
| Identity (*skip connection*) | |
| Max pooling $(3 \times 3)$ | |
| Average pooling $(3 \times 3)$ | |
| Zero-op | |

However, not all operations used for the real-valued case are suitable for searching binary architectures. In fact, we found that when using the standard DARTS search space, searching in the binary domain *does not converge*. This is due to several reasons: the depth-wise convolutions are notoriously hard to binarize due to what we call the "double approximation problem": the real-valued depth-wise convolution is a "compressed" approximate version of the normal convolution and, in turn, the binary depth-wise in a quantized approximation of the real-valued one. Furthermore, the $1 \times 1$ convolutional layers and the bottleneck block [16] were already shown to be hard to binarize [4] because the features compression that happens in such modules amplifies the high information degradation already caused by binarization. Both $1 \times 1$ and separable convolutions are present in the current search space typically used for search real-valued architectures (see Table 1), making them inappropriate for binarization. Moreover, while the dilated convolution contains 2 serialized convolutions, the separable one contains 4 which (a) causes discrepancies in their convergence speed and (b) can amplify the gradient fading phenomena that often happens during binarization.

To this end, we propose a new search space $O$, shown in Table 1, constructed from a binary-first point of view that avoids or alleviates the aforementioned shortcomings. While we preserve the zeros and identity connections alongside the

$3 \times 3$ max and average pooling layers which do not contain learnable parameters or binary operations, we propose to replace all the convolutional operations with the following new ones: $3 \times 3$ and $5 \times 5$ grouped convolutions and dilated grouped convolutions, also with a kernel size of $3 \times 3$ and $5 \times 5$. This removes all the $1 \times 1$ convolutional layers directly present in the cell's search space while maintaining the efficiency via the usage of grouped convolutions.

We note that there is a clear trade-off here between efficiency and accuracy controlled via the group size and number of channels. In this work, and in contrast to other works that rely on grouped convolutions (e.g. [46]), we propose to use a very high group size which leads to behaviours closer to that of a depthwise convolution (a small one will come at a price of higher computational budgets, while the extreme case #groups=#in_channels will again exacerbate the double-approximation problem). We used a group size of 12 and 3 (only) channels per group (totalling 36 channels) for CIFAR, and a group size of 16 and 5 (only) channels per group (totalling 80 channels) for ImageNet. Note that we also explored the effect of adding a channel shuffle layer [46] after each grouped convolution, typically used to enable cross-group information flow for a set of group convolution layers. However, due to the fact that #groups $\gg$ #channels per group, they were unable to offer accuracy gains. Instead, we found that the $1 \times 1$ convolutions present between different DARTS cells are sufficient for combining information between the different groups.

Furthermore, in the proposed search space, in all convolutional layers, the depth of all operations used is equal to 1 meaning that we used 1 convolution operator per layer, as opposed to 4 convolutions (2 depth-wise separable) used in DARTS, which facilitates learning. Another consequence of using operations with a depth equal to 1 is a latency improvement due to effectively a shallower network. For a visual illustration of differences between DARTS and the proposed BATS please check the supplementary material.

Finally, to improve the gradient and information flow, which is even more critical for the case of binary networks, while also speeding-up the convergence of such operations during the search phase, we explicitly add an identity connection on each convolutional operation such that $f_{i,j}(\mathbf{x}_i) = f_{i,j}(\mathbf{x}_i) + \mathbf{x}_i$.

The efficacy of the proposed search space is confirmed experimentally in Table 5. As the results from Table 5 show, the architectures found by searching within the proposed search space constantly outperform the others.

### 4.2   Search Regularization and Stabilisation

Despite its success and appealing formulation, DARTS accuracy can vary wildly between runs depending on the random seed. In fact, there are cases in which the architectures found perform worse than the ones obtained via random search. Furthermore, especially when trained longer or if the search is performed on larger datasets DARTS can converge towards degenerate solutions dominated by skip connections [8]. While in [8] the authors propose to address this by a) applying a dropout on the skip connections during the architecture search and b) by preserving a maximum of 2 skip connections per cell as a post processing

step that simply promotes the operation with the second highest probability, we found that such mechanism can still result in a large amount of randomness and is not always effective: for example, it can replace skip connections with pooling layers (which have no learning capacity) or the discovered architecture might even already contain too few skip connections.

Such problems are even more noticeable when the search is performed in the binary domain directly. Given that, during search, the input to the node $j$ is obtained by taking a weighted sum of all incoming edges, in order to maximise the flow of information, the architecture parameters $\alpha$ tend to converge to the same value making the selection of the final architecture problematic and susceptible to noise resulting in topologies than may perform worse than a random selection. Furthermore, the search is highly biased towards real-valued operations (pooling and skip connections) that at early stages can offer larger benefits.

To alleviate the aforementioned issues and encourage the search procedure to be more discriminative forcing it to make "harder" decisions, we propose to use a temperature factor $T < 1$ defining the flow from node $i$ to $j$ as follows:
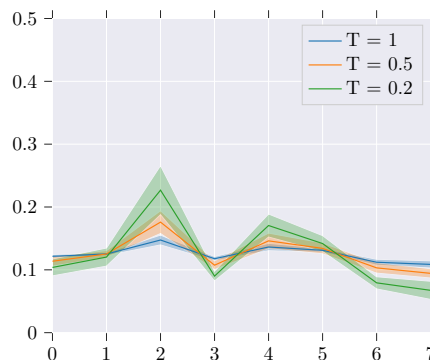


Fig. 1: Distribution of architecture parameters after the first stage for a given cell (data points were sorted by magnitude and do not correspond to the same ops): for low temperatures the network is forced to be more discriminative.

$$f_{i,j}(\mathbf{x}_i) = \sum_{o \in O} \frac{\exp(\alpha_{i,j}^o / T)}{\sum_{o' \in O} \exp(\alpha_{i,j}^{o'} / T)} \cdot o(\mathbf{x}_i). \tag{2}$$

This has the desirable effect of making the distribution of the architecture parameters less uniform and spikier (i.e. more discriminative). Hence, during search, because the information streams are aggregated using a weighted sum, the network cannot equally (or near-equally) rely on all possible operations by pulling information from all of them. Instead, in order to ensure convergence to a satisfactory solution it has to assign the highest probability to a non 0-ops path, enforced by a sub-unitary temperature ($T < 1$). This behaviour also follows closer the evaluation procedure where a single operation will be selected, reducing as such the performance discrepancy between the search, where the network pulls information from all paths, and evaluation.

Fig. 1 depicts the distribution of the architecture parameters for a given cell for different temperatures. For low temperatures, the network is forced to make more discriminative decisions which in turn makes it rely less on identity connections. This is further confirmed by Fig. 2 which depicts the chance of encountering a given op. in a normal cell at the end of the search process for different temperatures.

### 4.3    Binary Search Strategy

Despite their appealing speed-up and space saving, binary networks remain harder to train compared to their real-valued counterparts, with methods typically requiring a pre-training stage [28] or carefully tuning the hyper-parameters and optimizers [33, 9]. For the case of searching binary networks, directly attempting to effectuate an architecture search using binary weights and activations in most of our attempts resulted either in degenerate topologies or the training simply converges to extremely low accuracy values. We also note that, as expected and as our experiments have confirmed, performing the search in the real domain directly and then binarizing the network is sub-optimal.

To alleviate the problem, we propose a two-stage optimization process in which during the search the activations are binarized while the weights
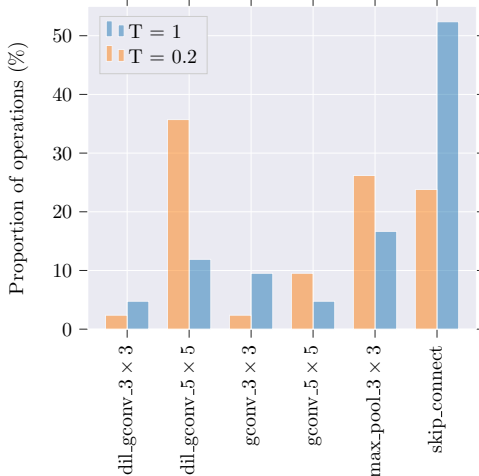


Fig. 2: Probability of a given op. in a cell at the end of the search process for different temperatures: for $T = 0.2$ there is a significant increase in the number of $5 \times 5$ convolutions and a decrease in the number of skip connections.

are kept real, and once the optimal architecture is discovered we proceed with the binarization of the weights too during the evaluation stage[4]. This is motivated by the fact that while the weights of a real-valued network can be typically binarized without a significant drop in accuracy, the same cannot be said about the binarization of the activations where due to the limited number of possible states the information flow inside the network drops dramatically. Hence, we propose to effectively split the problem into two sub-problems: weight and feature binarization and during the search we try to solve the hardest one, that is the binarization of the activations. Once this is accomplished, the binarization of the weights following always results in little drop in accuracy ($\sim 1\%$).

## 5    Ablation studies

### 5.1    Effect of the proposed components

Herein, we analyse the contribution to the overall performance of each proposed component: the newly introduced search space for binary networks (Section 4.1),

---

[4] More specifically, during evaluation, we first train a new network with binary activations and real-valued weights from scratch, and then binarize the weights. Then, the fully binarized network is evaluated on the test set.

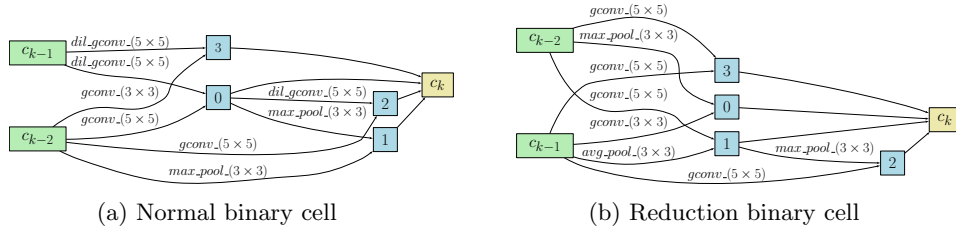(a) Normal binary cell                (b) Reduction binary cell

Fig. 3: Binary normal (a) and reduction (b) cells discovered by our method. Notice the prevalence of large kernels $(5 \times 5)$ and of the wider reduction cell, both contributing to increasing the filters' diversity and the flow of information.

the effect of the temperature adjustment (Section 4.2) and finally, of performing the search using real-valued weights and binary activations (Section 4.3).

We emphasize that *all 3 ingredients are needed in order to stabilize training and obtain good accuracy.* This suggests that it is not straightforward to show the effect of each component in isolation. To this end, we always *keep fixed* 2 of the proposed improvements and vary the other component to understand its impact. We also note that in all cases the final evaluated networks are fully binary (i.e. both the weights and the activations are binarized).

Table 2: Effect of search space on Top-1 accuracy on CIFAR-10.

| Search space | Temperature | Accuracy |
|---|---|---|
| Ours | Ours | **93.7 ± 0.6** |
| [27] | Ours | 51.0 ± 7.5 |
| Ours | [27] | **86.0 ± 2.0** |

**Importance of new search space:** We keep the temperature from Section 4.2 and search strategy from Section 4.3 fixed and then compare the proposed search space with that used in [27]. As the results from Table 2 show, searching for binary networks using the search space of [27] leads to much worse results ($\sim 51\%$ vs.$\sim 94\%$)

Table 3: Effect of temperature on Top-1 accuracy on CIFAR-10. T=1 was used in DARTS [27]. *Often leads to degenerate solution (i.e. all skip connections).

| Temperature | 1 | 0.5 | 0.2 | 0.05 |
|---|---|---|---|---|
| Accuracy | 86.0* | 92.1 | 93.7 | 93.3 |

and a high variation in the performance of the networks (a std. of 7.5% across 5 runs). This clearly shows that the previously used search space is not suitable for binary networks.

**Impact of temperature:** We keep the search space from Section 4.1 and search strategy from Section 4.3 fixed and evaluate the use of the temperature proposed in Section 4.2. As Table 3 shows, a decrease in temperature has a direct impact on discovering of higher performing models. At the opposite spectrum, a low temperature often leads to degenerate solutions. Note

that decreasing the temperature further has a negative impact on the accuracy, because, as the temperature goes to 0, the distribution of the architecture parameters resembles a delta function, which hinders the training process.

**Importance of search strategy:**
We keep the search space from Section 4.1 and temperature from Section 4.2 fixed and evaluate the effect of the proposed search strategy from Section 4.3. Table 4 summarizes the different strategies for searching binary architectures and their properties. The first row represents the case where the search is done in the real domain and then binarization follows. As the second row shows, if the search is done in the binary domain for both weights and activations, the search does not converge. As the third row shows, searching in the binary domain for the weights while keeping the activations real converges but results in low accuracy networks. Finally, the last row shows the proposed search strategy where the search is done in the binary domain for the activations and in the real domain for the weights. This configuration is the only one that yields stable search procedure and high accuracy networks (see std). Note that for all cases, the architectures are fully binarized (i.e. both activations and weights) prior to evaluation on the test set.
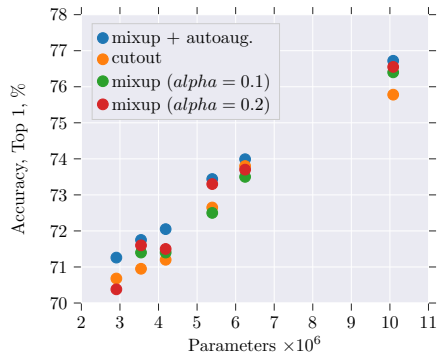


Fig. 4: Effect of data augmentation: Top-1 accuracy on CIFAR-100 of the network constructed using the discovered cell for different number of network parameters.

Table 4: Different strategies for searching binary architectures and their properties. The last row represents the proposed strategy. **All resulting topologies were trained using binary weights and activations during evaluation.**

| Search domain | | Search cost | Stable | Accuracy |
| --- | --- | --- | --- | --- |
| W | A | (GPU-days) | | |
| Real | Real | 0.2 | Yes | $88.4 \pm 0.8$ |
| Binary | Binary | 0.3 | No | $84.6 \pm 1.7$ |
| Binary | Real | 0.25 | Yes | $91.1 \pm 1.7$ |
| Real | Binary | 0.25 | Yes | $\mathbf{93.7 \pm 0.6}$ |

### 5.2   Impact of augmentation

While the positive impact of data augmentation has been explored for the case of real-valued networks, to the best of our knowledge, there is little to no work on this for the case of binary networks. Despite the fact that binarization is considered to be an extreme case of regularization [9], we found that augmentation

is equally important, and confirm that most of the augmentations applied to real-valued networks, can also improve their binarized counterparts. See Fig. 4.

## 6    Experiments

We conducted experiments on three different image classification datasets, namely CIFAR10 [21], CIFAR100 [21] and ImageNet [12]. Note that the architecture search is carried out on CIFAR-10 and tested on all three.

### 6.1    Architecture search

**Implementation details:** The training process consists of 3 stages during which the network depth is increased from 5 to 11 and finally 17 cells. Concomitantly, at each stage, the search space is gradually reduced to 8, 5 and respectively 3 ops. As opposed to [8], we did not use *on-skip* dropout since we found that it leads to unstable solutions. Instead, we used the mechanism proposed in Section 4.2, setting the temperature for the normal and reduction cells to 0.2 and 0.15 respectively. The latter is lower to encourage wider cells. During each stage, the network is trained using a batch size of 96 for 25 epochs. For the first 10 epochs only the network weights are learned. The remaining 15 epochs update both the architecture and network parameters. All the parameters are optimized using Adam [20]. The learning rate was set to $\eta = 0.0006$, weight decay to $w_d = 0.001$ and momentum to $\beta = (0.5, 0.999)$ for the architecture parameters and $\eta = 0.001$, $w_d = 0.0003$, $\beta = (0.9, 0.999)$ for the network parameters. To keep the search time low, we used the first-order optimization approach of DARTS. All methods were implemented in PyTorch [31].

**Discovered topologies:** Overall, our approach is capable of discovering high-performing binary cells, offering new or validating existing insights about optimal binary topologies. As depicted in Fig. 3, the cells found tend to prefer convolutional layers with larger kernel sizes ($5 \times 5$) which help alleviating the limited representational power found in such networks (a $3 \times 3$ kernel has maximum number of $2^9$ unique binary filters while a $5 \times 5$ has $2^{25}$). In addition, to preserve the information flow, a real-valued path that connects one of the input nodes to the output is always present.

Furthermore, since the down-sampling operation compresses the information across the spatial dimension, to compensate for this, the reduction cell tends to be wider (i.e. more information can flow through) as opposed to the normal cell which generally is deeper. As Table 5 shows, our searched architecture outperforms all prior ones searched in the real-valued domain and then binarized.

### 6.2    Comparison against state-of-the-art

In this section, we compare the performance of the searched architecture against state-of-the-art network binarization methods on CIFAR10/100 and ImageNet.

Table 5: Comparison with state-of-the-art NAS methods on CIFAR10 and CI-FAR100. For all methods we apply the binarization process described in Section 3 on the best cell provided by the authors.

| Architecture | Test Acc. (%) | | Params | Search Cost | Search Method |
|---|---|---|---|---|---|
| | C10 | C100 | (M) | (GPU-days) | |
| ResNet-18 [16] | 91.0 | 66.3 | 11.2 | - | manual |
| Random search | 92.9 | 70.2 | - | - | random |
| NASNet-A [52] | 90.3 | 66.1 | 3.9 | 1800 | RL |
| AmoebaNet-A [34] | 91.5 | 65.6 | 3.2 | 3150 | evolution |
| DARTS (first order) [27] | 89.4 | 63.3 | 3.3 | 1.5 | grad-based |
| DARTS (second order) [27] | 91.1 | 64.0 | 3.3 | 4.0 | grad-based |
| P-DARTS [8] | 89.5 | 63.9 | 3.6 | 0.3 | grad-based |
| Ours | 93.7 | 70.7 | 2.8 | 0.25 | grad-based |
| Ours + autoaug. [11] | 94.1 | 71.3 | 2.8 | 0.25 | grad-based |
| Ours (medium) | 94.6 | 73.5 | 5.4 | 0.25 | grad-based |
| Ours (medium) + autoaug. [11] | 94.9 | 74.0 | 5.4 | 0.25 | grad-based |
| Ours (large) | 95.5 | 75.7 | 10.0 | 0.25 | grad-based |
| Ours (large) + autoaug. [11] | 96.1 | 76.8 | 10.0 | 0.25 | grad-based |

**Implementation details:** This section refers to training the discovered architectures from scratch for evaluation purposes. For all models we use Adam with a starting learning rate of 0.001, $\beta = (0.9, 0.999)$ and no weight decay. The learning rate was reduced following a cosine scheduler [29]. In line with other works [27], on CIFAR10 and CIFA100, the models were trained for 600 epochs while on ImageNet for 75. The batch size was set to 256. The drop-path [22] was set to 0 for all experiments since binary networks are less prone to overfitting and already tend to use all parallel paths equally well. Following [27], we add an auxiliary tower [39] with a weight of 0.4. Unless otherwise stated, the network trained consisted of 20 cells and 36 initial channels with a group size of 12 channels for CIFAR10/100 and respectively 80 initial channels and 14 cells for ImageNet. For data augmentation, similarly to [27, 8], we padded the images appropriately on CIFAR10/100 and applied CutOut [13] with a single hole and a length of 16px for CIFAR-10 and Mix-up ($\alpha = 0.2$) for CIFAR-100 respectively. For ImageNet, we simply resized the images to $256 \times 256$px, and then randomly cropped them to $224 \times 224$px. During testing, the images were center-cropped.

**CIFAR10:** When compared against network topologies discovered by existing state-of-the-art NAS approaches, as the results from Table 5 show, our method significantly outperforms all of them on both CIFAR-10 and CIFAR-100 datasets. When using 20% the parameters than the competing methods, our approach offers gains of 5 to 7% on CIFAR-100 and $2.5 - 4\%$ on CIFAR-10. This clearly confirms the importance of the proposed search space and strategy. Fur-

thermore, we test how the discovered architecture performs when scaled-up. In order to increase the capacity of the models we concomitantly adjust the width and #groups of the cells. By doing so, we obtain a 2% improvement of CIFAR-10 and 5% on CIFAR-100 almost matching the performance of a real-valued ResNet-18 model. We note that for fairness all other topologies we compared against where modified using the structure described in Section 3.

For an exhaustive comparison on CIFAR-10 see the supplementary material.

**ImageNet:** Herein, we compare our approach against related state-of-the-art binarization and quantization methods. As the results from Table 7 show, our discovered architecture outperforms all existing binarization methods while using a lower computational budget for the normal thin model and offers a gain of 5% for the 2x-wider one. Furthermore, our methods compared favorable even against methods that either significantly increase the network computational requirements or use more bits for quantization.

Table 6: Comparison with selected SOTA binary methods on ImageNet in terms of computational cost. Notice, that within a similar budget our method achieves the highest accuracy. For a full comparison see Table 7.

| Architecture | Accuracy (%) | | # Operations | | # bits |
|---|---|---|---|---|---|
| | Top-1 | Top-5 | $\mathbf{FLOPS}\times10^8$ | $\mathbf{BOPS}\times10^9$ | (W/A) |
| BNN [10] | 42.2 | 69.2 | 1.314 | 1.695 | 1/1 |
| XNOR-Net [33] | 51.2 | 73.2 | 1.333 | 1.695 | 1/1 |
| CCNN [43] | 54.2 | 77.9 | 1.333 | 1.695 | 1/1 |
| Bi-Real Net [28] | 56.4 | 79.5 | 1.544 | 1.676 | 1/1 |
| XNOR-Net++ [5] | 57.1 | 79.9 | 1.333 | 1.695 | 1/1 |
| CI-Net [40] | 59.9 | 84.2 | − | − | 1/1 |
| **BATS (Ours)** | 60.4 | 83.0 | 0.805 | 1.149 | 1/1 |
| **BATS [2x-wider] (Ours)** | **66.1** | **87.0** | 1.210 | 2.157 | 1/1 |

### 6.3   Network efficiency

The current most popular settings of binarizing neural networks preserve the first and last layer real-valued. However, for the currently most popular binarized architecture, i.e. ResNet-18 [16], the first convolutional layer accounts for approx. 6.5% of the total computational budget ($1.2 \times 10^8$ FLOPs out of the total of $1.8 \times 10^9$ FLOPs), a direct consequence of the large kernel size ($7 \times 7$) and the high input resolution. In order to alleviate this, beyond using the ImageNet stem cell used in DARTS [27] that replaces the $7 \times 7$ layer with two $3 \times 3$, we replaced the second convolution with a grouped convolution ($g = 4$). This alone allows us to more than halve the number of real-valued operations (see Table 6). Additionally, in order to transparently show the computational cost of the tested models, we separate the binary operations (BOPs) and FLOPs into

two distinctive categories. Furthermore, as Table 6 shows, our method surpasses sophisticated approaches while having a significantly lower number of ops.

Table 7: Comparison with state-of-the-art binarization methods on ImageNet, including against approaches that use low-bit quantization (upper section) and ones that increase the capacity of the network (middle section). For the latter case the last column also indicates the capacity scaling used. Models marked with ** use real-valued downsampling convolutional layers.

| Architecture | Accuracy (%) | | # bits |
| | Top-1 | Top-5 | (W/A) |
|---|---|---|---|
| BWN [10] | 60.8 | 83.0 | 1/32 |
| TTQ [49] | 66.6 | 87.2 | 2/32 |
| HWGQ [7] | 59.6 | 82.2 | 1/2 |
| LQ-Net [45] | 59.6 | 82.2 | 1/2 |
| SYQ [15] | 55.4 | 78.6 | 1/2 |
| DOREFA-Net [48] | 62.6 | 84.4 | 1/2 |
| ABC-Net $(M, N = 1)$ [23] | 42.2 | 67.6 | 1/1 |
| ABC-Net $(M, N = 5)$ [23] | 65.0 | 85.9 | $(1/1) \times 5^2$ |
| Struct. Approx. [51] | 64.2 | 85.6 | $(1/1) \times 4$ |
| Struct. Approx.** [51] | 66.3 | 86.6 | $(1/1) \times 4$ |
| CBCN [26] | 61.4 | 82.8 | $(1/1) \times 4$ |
| Ensamble [50] | 61.0 | - | $(1/1) \times 6$ |
| BNN [10] | 42.2 | 69.2 | 1/1 |
| XNOR-Net [33] | 51.2 | 73.2 | 1/1 |
| CCNN [43] | 54.2 | 77.9 | 1/1 |
| Bi-Real Net** [28] | 56.4 | 79.5 | 1/1 |
| Rethink. BNN** [17] | 56.6 | 79.4 | 1/1 |
| XNOR-Net++ [5] | 57.1 | 79.9 | 1/1 |
| CI-Net** [40] | 59.9 | 84.2 | 1/1 |
| **BATS (Ours)** | 60.4 | 83.0 | 1/1 |
| **BATS [2x-wider] (Ours)** | **66.1** | **87.0** | 1/1 |

## 7   Conclusion

In this work we introduce a novel Binary Architecture Search (BATS) that drastically reduces the accuracy gap between binary models and their real-valued counterparts, by searching for the first time directly for binary architectures. To this end we, (a) designed a novel search space specially tailored to binary networks, (b) proposed a new regularization method that helps stabilizing the search process, and (c) introduced an adapted search strategy that speed-ups the overall network search. Experimental results conducted on CIFAR10, CIFAR100 and ImageNet demonstrate the effectiveness of the proposed approach and the need of doing the search in the binary space directly.

# References

1. Baker, B., Gupta, O., Naik, N., Raskar, R.: Designing neural network architectures using reinforcement learning. In: International Conference on Learning Representations (2017)
2. Bethge, J., Yang, H., Bornstein, M., Meinel, C.: Back to simplicity: How to train accurate BNNs from scratch? arXiv preprint arXiv:1906.08637 (2019)
3. Brock, A., Lim, T., Ritchie, J., Weston, N.: SMASH: One-shot model architecture search through hypernetworks. In: International Conference on Learning Representations (2018)
4. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: IEEE International Conference on Computer Vision. pp. 3706–3714 (2017)
5. Bulat, A., Tzimiropoulos, G.: Xnor-net++: Improved binary neural networks. In: British Machine Vision Conference (2019)
6. Cai, H., Zhu, L., Han, S.: ProxylessNAS: Direct neural architecture search on target task and hardware. International Conference on Learning Representations (2019)
7. Cai, Z., He, X., Sun, J., Vasconcelos, N.: Deep learning with low precision by half-wave gaussian quantization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5918–5926 (2017)
8. Chen, X., Xie, L., Wu, J., Tian, Q.: Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. IEEE International Conference on Computer Vision (2019)
9. Courbariaux, M., Bengio, Y., David, J.P.: Binaryconnect: Training deep neural networks with binary weights during propagations. In: Advances on Neural Information Processing Systems (2015)
10. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv preprint arXiv:1602.02830 (2016)
11. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: AutoAugment: Learning augmentation policies from data. IEEE Conference on Computer Vision and Pattern Recognition (2019)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
13. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
14. Ding, R., Chin, T.W., Liu, Z., Marculescu, D.: Regularizing activation distribution for training binarized deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11408–11417 (2019)
15. Faraone, J., Fraser, N., Blott, M., Leong, P.H.: Syq: Learning symmetric quantization for efficient deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4300–4309 (2018)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Helwegen, K., Widdicombe, J., Geiger, L., Liu, Z., Cheng, K.T., Nusselder, R.: Latent weights do not exist: Rethinking binarized neural network optimization. In: Advances in neural information processing systems. pp. 7533–7544 (2019)

18. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
22. Larsson, G., Maire, M., Shakhnarovich, G.: Fractalnet: Ultra-deep neural networks without residuals. arXiv preprint arXiv:1605.07648 (2016)
23. Lin, X., Zhao, C., Pan, W.: Towards accurate binary convolutional neural network. In: Advances on Neural Information Processing Systems. pp. 345–353 (2017)
24. Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L.: Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 82–92 (2019)
25. Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.J., Fei-Fei, L., Yuille, A., Huang, J., Murphy, K.: Progressive neural architecture search. In: European Conference on Computer Vision. pp. 19–34 (2018)
26. Liu, C., Ding, W., Xia, X., Zhang, B., Gu, J., Liu, J., Ji, R., Doermann, D.: Circulant binary convolutional networks: Enhancing the performance of 1-bit dcnns with circulant back propagation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2691–2699 (2019)
27. Liu, H., Simonyan, K., Yang, Y.: DARTS: Differentiable architecture search. International Conference on Learning Representations (2019)
28. Liu, Z., Wu, B., Luo, W., Yang, X., Liu, W., Cheng, K.T.: Bi-Real Net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm. In: European Conference on Computer Vision. pp. 747–763 (2018)
29. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
30. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: European Conference on Computer Vision. pp. 122–138 (2018)
31. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035 (2019)
32. Pham, H., Guan, M., Zoph, B., Le, Q., Dean, J.: Efficient neural architecture search via parameters sharing. In: International Conference on Machine Learning. pp. 4095–4104 (2018)
33. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: XNOR-Net: ImageNet classification using binary convolutional neural networks. In: European Conference on Computer Vision. pp. 525–542 (2016)
34. Real, E., Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: AAAI Conf. on Artificial Intelligence. vol. 33, pp. 4780–4789 (2019)

35. Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y.L., Tan, J., Le, Q.V., Kurakin, A.: Large-scale evolution of image classifiers. In: International Conference on Machine Learning. pp. 2902–2911 (2017)
36. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)
37. Shen, M., Han, K., Xu, C., Wang, Y.: Searching for accurate binary neural architectures. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
38. Singh, K.P., Kim, D., Choi, J.: Learning architectures for binary networks. arXiv preprint arXiv:2002.06963 (2020)
39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
40. Wang, Z., Lu, J., Tao, C., Zhou, J., Tian, Q.: Learning channel-wise interactions for binary convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 568–577 (2019)
41. Xie, L., Yuille, A.: Genetic CNN. In: IEEE International Conference on Computer Vision. pp. 1379–1388 (2017)
42. Xie, S., Zheng, H., Liu, C., Lin, L.: Snas: stochastic neural architecture search. In: International Conference on Learning Representations (2018)
43. Xu, Z., Cheung, R.C.: Accurate and compact convolutional neural networks with trained binarization. arXiv preprint arXiv:1909.11366 (2019)
44. Zagoruyko, S., Komodakis, N.: Wide residual networks. British Machine Vision Conference (2016)
45. Zhang, D., Yang, J., Ye, D., Hua, G.: Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: European Conference on Computer Vision. pp. 365–382 (2018)
46. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6848–6856 (2018)
47. Zhong, Z., Yan, J., Wu, W., Shao, J., Liu, C.L.: Practical block-wise neural network architecture generation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2423–2432 (2018)
48. Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv (2016)
49. Zhu, C., Han, S., Mao, H., Dally, W.J.: Trained ternary quantization. arXiv preprint arXiv:1612.01064 (2016)
50. Zhu, S., Dong, X., Su, H.: Binary ensemble neural network: More bits per network or more networks per bit? In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4923–4932 (2019)
51. Zhuang, B., Shen, C., Tan, M., Liu, L., Reid, I.: Structured binary neural networks for accurate image classification and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 413–422 (2019)
52. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8697–8710 (2018)