6 SUPPLEMENTARY MATERIAL

6.1 Proof of Theorem 2

Let us set some terms,

$$g_{t,i} = -\eta \nabla_{\alpha_{t,i}} \mathcal{L}_{val}(w, \alpha; T_t) \tag{14}$$

Then we have $|g_{t,i}| \leq \eta \cdot \mathbf{L}$ and,

$$\alpha_{t,i} = \alpha_{0,i} - \eta \sum_{s=0}^{t} \nabla_{\alpha_{s,i}} \mathcal{L}_{val}(w, \alpha; T_t) = \sum_{s=0}^{t} g_{s,i} \tag{15}$$

As we initialize $\alpha_{0,i} = 0$ for all $i = 1, \ldots, N$ in 2. Then we prove claim 1,

Proof: The pruning rule can be put as following,

$$\Phi_{o_i}(\alpha_t; T_t) < \theta_t \tag{16}$$

$$\frac{e^{\frac{\alpha_{t,i}}{T_t}}}{\sum_{j=1}^{N} e^{\frac{\alpha_{t,i}}{T_t}}} < \theta_t \tag{17}$$

$$\alpha_{t,i} < T_t \log\left(\sum_{j=1}^{N} e^{\frac{\alpha_{t,i}}{T_t}}\right) - T_t \log\left(\frac{1}{\theta_t}\right) \tag{18}$$

$$\leq T_t \left(\frac{\alpha_t^*}{T_t} + \log(N)\right) - T_t \log\left(\frac{1}{\theta_t}\right) \tag{19}$$

$$< \alpha_t^* - T_t \left(\log\left(\frac{1}{\theta_t}\right) - \log(N)\right) \tag{20}$$

$$< \alpha_t^* - T_t \left(t + \log\left(\frac{1}{N\nu_t}\right)\right) \tag{21}$$

$$< \alpha_t^* - t T_t \rho_t^{-1} \tag{22}$$

$$< \alpha_t^* - 2t\beta_t \tag{23}$$

Where 21 is by setting $\theta_t = \nu_t e^{-t}$ and 19 is since,

$$\log\left(\sum_{i=1}^{N} e^{x_i}\right) \leq \max_i x_i + \log(N) \tag{24}$$

Finally we have,

$$\frac{\alpha_{t,i}}{t} + \beta_t < \frac{\alpha_t^*}{t} - \beta_t, \tag{25}$$

We wish to bound the probability of pruning the best operation, i.e. the operation with the highest expected architecture parameter α. Although involving the empirical values of α, we show that the condition in claim 1 avoids the pruning of the operation with the highest expected α. For this purpose we bound the probability for the deviation of each empirical α from its expected value by the specified margin β_t. For this purpose we introduce the following concentration bound,

Lemma 1 (Hoeffding (Hoeffding, 1963)) *Let $g_1, \ldots, g_t$ be independent bounded random variables with $g_s \in [a_s, b_s]$, where -$\infty < a_s \leq b_s < \infty$ for all $s = 1, \ldots, t$. Then,*

$$\mathbb{P}\left\{\frac{1}{t}\sum_{s=1}^{t}(g_s - \mathbb{E}[g_s]) \geq \quad \beta\right\} \leq e^{-\frac{2\beta^2 t^2}{\sum_{s=1}^{t}(b_s - a_s)^2}} \tag{26}$$

$$\mathbb{P}\left\{\frac{1}{t}\sum_{s=1}^{t}(g_s - \mathbb{E}[g_s]) \leq -\beta\right\} \leq e^{-\frac{2\beta^2 t^2}{\sum_{s=1}^{t}(b_s - a_s)^2}} \tag{27}$$

Our main argument, described in Theorem 3, is that at any time t and for any operation o_i, the empirical value $\frac{\alpha_{t,i}}{t}$ is within β_t of its expected value $\frac{\bar{\alpha}_{t,i}}{t} = \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}[g_{s,i}]$. For the purpose of proving Theorem 3, we first prove the following Corollary 4,

Corollary 4 *For any time t and operations $\{o_i\}_{i=1}^{N} \in \mathcal{O}$ we have,*

$$\mathbb{P}\left\{\frac{1}{t}\left|\alpha_{t,i} - \bar{\alpha}_{t,i}\right| > \beta_t\right\} \leq \frac{6\delta}{\pi^2 N t^2} \tag{28}$$

Proof:

$$\mathbb{P}\left\{\frac{1}{t}\left|\alpha_{t,i} - \bar{\alpha}_{t,i}\right| > \beta_t\right\} \tag{29}$$

$$= \mathbb{P}\left\{\frac{\alpha_{t,i}}{t} - \frac{\bar{\alpha}_{t,i}}{t} > \beta_t \bigcup \frac{\alpha_{t,i}}{t} - \frac{\bar{\alpha}_{t,i}}{t} < -\beta_t\right\} \tag{30}$$

$$\leq \mathbb{P}\left\{\frac{\alpha_{t,i}}{t} - \frac{\bar{\alpha}_{t,i}}{t} > \beta_t\right\} + \mathbb{P}\left\{\frac{\alpha_{t,i}}{t} - \frac{\bar{\alpha}_{t,i}}{t} < -\beta_t\right\} \tag{31}$$

$$\leq \mathbb{P}\left\{\frac{1}{t}\sum_{s=1}^{t}(g_{s,i} - \mathbb{E}[g_{s,i}]) \geq \beta_t\right\}$$

$$+ \mathbb{P}\left\{\frac{1}{t}\sum_{s=1}^{t}(g_{s,i} - \mathbb{E}[g_{s,i}]) \leq -\beta_t\right\} \tag{32}$$

$$\leq 2e^{-\frac{2\beta_t^2 t}{4\eta^2 \mathbf{L}^2}} \tag{33}$$

$$\leq \frac{6\delta}{\pi^2 N t^2} \tag{34}$$

Where, 31 is by the union bound, 33 is by Lemma 1 with $g_{t,i} \in [-\eta\mathbf{L}, \eta\mathbf{L}]$ and 34 is by setting,

$$\beta_t = \eta\mathbf{L}\sqrt{\frac{2}{t} log\left(\frac{\pi^2 N t^2}{3\delta}\right)} \tag{35}$$

We can now prove Theorem 3,

Proof:

$$\mathbb{P}\left\{\bigcap_{t=1}^{\infty}\frac{1}{t}\left|\alpha_{t,i}-\bar{\alpha}_{t,i}\right|\leq\beta_t\right\} \tag{36}$$

$$=1-\mathbb{P}\left\{\bigcup_{t=1}^{\infty}\frac{1}{t}\left|\alpha_{t,i}-\bar{\alpha}_{t,i}\right|>\beta_t\right\} \tag{37}$$

$$\geq 1-\sum_{t=1}^{\infty}\mathbb{P}\left\{\frac{1}{t}\left|\alpha_{t,i}-\bar{\alpha}_{t,i}\right|>\beta_t\right\} \tag{38}$$

$$\geq 1-\frac{6\delta}{\pi^2 N}\sum_{t=1}^{\infty}\frac{1}{t^2} \tag{39}$$

$$=1-\frac{\delta}{N} \tag{40}$$

where (38) is by the union bound, (39) is by Corollary 4.

Requiring Theorem 3 to hold for all of the operations, we get by the union bound that, the probability of pruning the best operation is less than δ. Furthermore, since $\nu_t \in \Upsilon$, ρ_t goes to 1 as t increases and T_t goes to zero together with β_t. Thus eventually all operations but the best one are pruned. This completes our proof.

6.2 ASAP search and train details

In order to conduct a fair comparison, we follow (Liu et al., 2018b) for all the search and train details. This excludes the new annealing parameters and those related to the training of additional datasets, which are not mentioned in (Liu et al., 2018b). Our code will be made available for future public use.

6.2.1 Search details

Data pre-processing. We apply the following:

- Centrally padding the training images to a size of 40x40.
- Randomly cropping back to the size of 32x32.
- Randomly flipping the training images horizontally.
- Standardizing the train and validation sets to be of a zero-mean and a unit variance.

Operations and cells. We select from the operations mentioned in 4.1, with a stride of 1 for all of the connections within a cell but for the reduction cells' connections to the previous cells, which are with a stride of 2. Convolutional layers are padded so that the spatial resolution is kept. The operations are applied in the order of ReLU-Conv-BN. Following (Zoph and Le, 2017),(Real et al., 2018), depthwise separable convolutions are always applied twice. The cell's output is a 1x1 convolutional layer applied on all of the cells' four intermediate nodes' outputs concatenated, such that the number of channels is preserved. In CIFAR-10, the search lasts up to 0.2 days on NVIDIA GTX 1080Ti GPU.
The annealing schedule. We use the exponential decay annealing schedule as described in Algorithm 1 when setting the annealing schedule as in 9 with $(T_0, \beta, \tau) = (1.6, 0.95, 1)$. It was selected for obtaining a final temperature of 0.1 and 5 epochs of grace-cycles.
Alternate optimization. For a fair comparison, we use the exact training settings from (Liu et al., 2018b). We use the Adam optimizer (Kingma and Ba, 2014) for the architecture weights optimization with momentum parameters of $(0.5, 0.999)$ and a fixed learning rate of 10^{-3}. For the network weights optimization, we use SGD (Robbins and Monro, 1951) with a momentum of 0.9 as the learning rate is following a cosine annealing schedule (Loshchilov and Hutter, 2016) with an initial value of 0.025.

6.2.2 Training details

CIFAR-10. The training architecture consists of 20 cells stacking up: 18 normal cells and 2 reduction cells, located at the $1/3$ and $2/3$ of the total network depth respectively. We double the number of channels after each reduction cell. We train the network for 1500 epochs with a batch size of 128. We use the SGD nesterov-momentum optimizer (Robbins and Monro, 1951) with a momentum of 0.9, following a cycles cosine annealing learning rate (Loshchilov and Hutter, 2016) with an initial value of 0.025. We apply a weight decay of $3 \cdot 10^{-4}$ and a norm gradient clipping at 5. We add an auxiliary loss (Szegedy et al., 2015) after the last reduction cell with a weight of 0.4. In addition to data pre-processing, we use cutout augmentations (DeVries and Taylor, 2017) with a length of 16 and a drop-path regularization (Larsson et al., 2016) with a probability of 0.2. WE also use the AutoAugment augmentation scheme.
ImageNet. Our training architecture starts with stem cells that reduce the input image resolution from 224 to 56 (3 reductions), similar to MobileNet-V1 (Howard et al., 2017). We then stack 14 cells: 12 normal cells and 2 reduction cells. The reduction cells are placed after the fourth and eighth normal cells. The normal cells start with 50 channels, as the number of channels is doubled after each reduction cell. We also added SE layer(Hu et al., 2018) at the end of each cell. In total, the network contains 5.7 million parameters. We train the network on 2 GPUs for 250 epochs with a batch size of 256. We use the SGD nesterov-momentum optimizer (Robbins and Monro, 1951) with a momentum of 0.9, following a cosine learning rate with an initial learning rate value of 0.2. We apply a weight decay of $1 \cdot 10^{-4}$ and a norm gradient clipping at 5. We add an auxiliary loss after the last reduction cell with the weight of 0.4 and a label smoothing (Reed et al., 2014) of 0.1. During training, we normalize the input image and crop it with a random cropping factor in the range of 0.08 to 1. In addition, autoaugment augmentations and randomly horizontal

flipping are applied. During testing, we resize the input image to the size of 256x256 and applying a fixed central crop to the size of 224x224.

Additional datasets. Our additional classification datasets consist of the Following:

CINIC-10: (Darlow et al., 2018) is an extension of CIFAR-10 by ImageNet images, down-sampled to match the image size of CIFAR-10. It has $270,000$ images of 10 classes, i.e. it has larger train and test sets than those of CIFAR-10.

CIFAR-100: (Torralba et al., 2008) A natural image classification dataset, containing 100 classes with 600 images per class. The image size is 32x32 and the train-test split is $50,000$:$10,000$ images respectively.

FREIBURG: (Jund et al., 2016) A groceries classification dataset consisting of 5000 images of size 256x256, divided into 25 categories. It has imbalanced class sizes ranging from 97 to 370 images per class. Images were taken in various aspect ratios and padded to squares.

SVHN: (Netzer et al., 2011) A dataset containing real-world images of digits and numbers in natural scenes. It consists of $600,000$ images of size 32x32, divided into 10 classes. The dataset can be thought of as a real-world alternative to MNIST, with an order of magnitude more images and significantly harder real-world scenarios.

FMNIST: (Xiao et al., 2017) A clothes classification dataset with a $60,000$:$10,000$ train-test split. Each example is a grayscale image of size 28x28, associated with a label from 10 classes of clothes. It is intended to serve as a direct drop-in replacement for the original MNIST dataset as a benchmark for machine learning algorithms.

The training scheme use for those was similar to the one used for CIFAR-10, with some minor adjustments and modifications - mainly the use of standard color augmentations instead of autoaugment regimes, and default training length of 600 epochs instead of 1500 epochs. For the FREIBURG dataset, we resized the original images from 256x256 to 64x64. For CINIC-10, we were training for 400 epochs instead of 600, since this dataset is quite large. Note that for each of those datasets, all of the cells were trained with exactly the same network architecture and hyper-parameters, unlike our ImageNet comparison at 4.3, where each cell was embedded into a different architecture and trained with a completely different scheme.