UNDERSTANDING AND ROBUSTIFYING DIFFERENTIABLE ARCHITECTURE SEARCH

Arber Zela[1], **Thomas Elsken**[2,1], **Tonmoy Saikia**[1], **Yassine Marrakchi**[1], **Thomas Brox**[1] **& Frank Hutter**[1,2]
[1]Department of Computer Science, University of Freiburg
{zelaa, saikiat, marrakch, brox, fh}@cs.uni-freiburg.de
[2]Bosch Center for Artificial Intelligence
Thomas.Elsken@de.bosch.com

ABSTRACT

Differentiable Architecture Search (DARTS) has attracted a lot of attention due to its simplicity and small search costs achieved by a continuous relaxation and an approximation of the resulting bi-level optimization problem. However, DARTS does not work robustly for new problems: we identify a wide range of search spaces for which DARTS yields degenerate architectures with very poor test performance. We study this failure mode and show that, while DARTS successfully minimizes validation loss, the found solutions generalize poorly when they coincide with high validation loss curvature in the space of architectures. We show that by adding one of various types of regularization we can robustify DARTS to find solutions with smaller Hessian spectrum and with better generalization properties. Based on these observations we propose several simple variations of DARTS that perform substantially more robustly in practice. Our observations are robust across five search spaces on three image classification tasks and also hold for the very different domains of disparity estimation (a dense regression task) and language modelling. We provide our implementation and scripts to facilitate reproducibility[1].

1 INTRODUCTION

Neural Architecture Search (NAS), the process of automatically designing neural network architectures, has recently attracted attention by achieving state-of-the-art performance on a variety of tasks (Zoph & Le, 2017; Real et al., 2019). *Differentiable architecture search* (DARTS) (Liu et al., 2019) significantly improved the efficiency of NAS over prior work, reducing its costs to the same order of magnitude as training a single neural network. This expanded the scope of NAS substantially, allowing it to also apply on more expensive problems, such as semantic segmentation (Chenxi et al., 2019) or disparity estimation (Saikia et al., 2019).

However, several researchers have also reported DARTS to *not* work well, in some cases even no better than random search (Li & Talwalkar, 2019; Sciuto et al., 2019). Why is this? How can these seemingly contradicting results be explained? The overall goal of this paper is to understand and overcome such failure modes of DARTS.

After discussing background and related work in Section 2, we make the following contributions:

1. We propose a set of 12 new NAS benchmarks, based on a wide range of four search spaces in which standard DARTS yields degenerate architectures with poor test performance (Section 3).
2. By computing the eigenspectrum of the Hessian of the validation loss with respect to the architectural parameters, we show that there is a strong correlation between its dominant eigenvalue and the architecture's generalization error. Based on this finding, we propose a simple variation of DARTS with early stopping that performs substantially more robustly (Section 4).

[1] https://github.com/automl/RobustDARTS

3. We show that, related to previous work on sharp/flat local minima, regularizing the inner objective of DARTS more strongly allows it to find solutions with smaller Hessian spectrum and better generalization properties (Section 5).
4. We show that DARTS' discretization of the one-shot model after the search phase substantially worsens performance when the curvature of the validation loss w.r.t. the architectural parameters is large (Section 6).
5. Based on the insights of the previous sections, we propose two practical robustifications of DARTS that overcome its failure modes in all 12 NAS benchmarks we study (Section 7).

Our findings are robust across five different search spaces evaluated on three different image recognition benchmarks each and also hold for the very different domains of language modelling (PTB) and disparity estimation. They consolidate the findings of the various results in the literature and lead to a substantially more robust version of DARTS.

2 Background and Related Work

2.1 Relation between flat/sharp minima and generalization performance

Already Hochreiter & Schmidhuber (1997) observed that flat minima of the training loss yield better generalization performance than sharp minima. Recent work (Keskar et al., 2016; Yao et al., 2018) focuses more on the settings of large/small batch size training, where observations show that small batch training tends to get attracted to flatter minima and generalizes better. Similarly, Nguyen et al. (2018) observed that this phenomenon manifests also in the hyperparameter space. They showed that whenever the hyperparameters overfit the validation data, the minima lie in a sharper region of the space. This motivated us to conduct a similar analysis in the context of differentiable architecture search later in Section 4.1, where we see the same effect in the space of neural network architectures.

2.2 Bi-level Optimization

We start by a short introduction of the bi-level optimization problem (Colson et al., 2007). These are problems which contain two optimization tasks, nested within each other.

Definition 2.1. Given the outer objective function $F : \mathbb{R}^P \times \mathbb{R}^N \to \mathbb{R}$ and the inner objective function $f : \mathbb{R}^P \times \mathbb{R}^N \to \mathbb{R}$, the bi-level optimization problem is given by

$$\min_{y \in \mathbb{R}^P} F(y, \theta^*(y)) \tag{1}$$

$$s.t. \quad \theta^*(y) \in \arg\min_{\theta \in \mathbb{R}^N} f(y, \theta), \tag{2}$$

where $y \in \mathbb{R}^P$ and $\theta \in \mathbb{R}^N$ are the outer and inner variables, respectively. One may also see the bi-level problem as a constrained optimization problem, with the inner problem (and possibly other equality/inequality constraints) as constraints.

In general, even in the case when the inner objective (2) is strongly convex and has an unique minimizer $\theta^*(y) = \arg\min_{\theta \in \mathbb{R}^N} f(y, \theta)$), it is not possible to directly optimize the outer objective (1). A possible method around this issue is to use the implicit function theorem to retrieve the derivative of the solution map (or response map) $\theta^*(y) \in \mathbb{F} \subseteq \mathbb{R}^N$ w.r.t. y (Bengio, 2000; Pedregosa, 2016; Beirami et al., 2017). Another strategy is to approximate the inner problem with a dynamical system (Domke, 2012; Maclaurin et al., 2015; Franceschi et al., 2017; 2018), where the optimization dynamics could, e.g., describe gradient descent. In the case that the minimizer of the inner problem is unique, under some conditions the set of minimizers of this approximate problem will indeed converge to the minimizers of the bilevel problem (1) (see Franceschi et al. (2018)).

However, solving the bi-level optimization problem with non-convex inner objectives is in general a NP-hard problem (Hansen et al., 1992). Resolving to gradient-based algorithms (e.g., gradient descent) for optimizing this objective is on one hand necessary due to the possible high dimensionality of the parameter space, but on the other hand we do not have a global view of the space and we may potentially end in a local minimum.

2.3 NEURAL ARCHITECTURE SEARCH

Neural Architecture Search (NAS) denotes the process of automatically designing neural network architectures in order to overcome the cumbersome trial-and-error process when designing architectures manually. We briefly review NAS here and refer to the recent survey by Elsken et al. (2019b) for a more thorough overview. Prior work mostly employs either reinforcement learning techniques (Baker et al., 2017a; Zoph & Le, 2017; Zhong et al., 2018; Zoph et al., 2018) or evolutionary algorithms (Stanley & Miikkulainen, 2002; Liu et al., 2018b; Miikkulainen et al., 2017; Real et al., 2017; 2019) to optimize the discrete architecture space. As these methods are often very expensive, various works focus on reducing the search costs by, e.g., employing network morphisms (Cai et al., 2018a;b; Elsken et al., 2017; 2019a), weight sharing within one-shot models (Saxena & Verbeek, 2016; Bender et al., 2018; Pham et al., 2018) or multi-fidelity optimization (Baker et al., 2017b; Falkner et al., 2018; Li et al., 2017; Zela et al., 2018), but their applicability still often remains restricted to rather simple tasks and small datasets.

2.4 DIFFERENTIABLE ARCHITECTURE SEARCH (DARTS)

A recent line of work focuses on relaxing the discrete neural architecture search problem to a continuous one that can be solved by gradient descent (Liu et al., 2019; Xie et al., 2019; Casale et al., 2019; Cai et al., 2019). In DARTS (Liu et al., 2019), this is achieved by simply using a weighted sum of possible candidate operations for each layer, whereas the real-valued weights then effectively parametrize the network's architecture. We will now review DARTS in more detail, as our work builds directly upon it.

Continuous relaxation of the search space. In agreement with prior work (Zoph et al., 2018; Real et al., 2019), DARTS optimizes only substructures called cells that are stacked to define the full network architecture. Each cell contains N nodes organized in a directed acyclic graph. The graph contains two inputs nodes (given by the outputs of the previous two cells), a set of intermediate nodes, and one output node (given by concatenating all intermediate nodes). Each intermediate node $x^{(j)}$ represents a feature map. See Figure 1 for an illustration of such a cell. Instead applying a single operation to a specific node during architecture search, Liu et al. (2019) relax the decision which operation to choose by computing the intermediate node as a mixture of candidate operations, applied to predecessor nodes $x^{(i)}, i < j$,

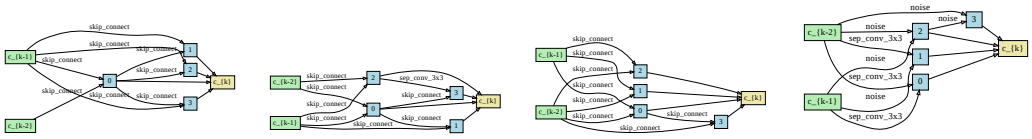
$$x^{(j)} = \sum_{i<j} \sum_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{i,j})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{i,j})} o\left(x^{(i)}\right), \tag{3}$$

where $\mathcal{O}$ denotes the set of all candidate operations (e.g., 3×3 convolution, skip connection, 3×3 max pooling, ...) and $\alpha = (\alpha_o^{i,j})_{i,j,o}$ serves as a real valued parametrization of the architecture.

Gradient-based optimization of the search space. DARTS then optimizes both the weights of the search network (often called *one-shot* model, since the weights of all individual subgraphs/architectures are shared) and architectural parameters by alternating gradient descent. Learning the network weights and the architecture parameters are performed on the training and validation set, respectively. This can be interpreted as solving the bi-level optimization problem (1), (2), where F and f are the validation and training loss, $\mathcal{L}_{valid}$ and $\mathcal{L}_{train}$, respectively, while y and θ denote the architectural parameters α and network weights w, respectively. Note that DARTS only approximates the lower-level solution by a single gradient step (see Appendix A for more details).

At the end of the search phase, a discrete cell is obtained by choosing the k most important incoming operation for each intermediate node while all others are pruned. Importance is measured by the operation weighting factor $\frac{\exp(\alpha_o^{i,j})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{i,j})}$.[2]

[2]One usually searches for two types of cells, a reduction cell (which reduces the spatial dimension), and a normal cell (which preserves spatial resolution).



(a) Space 1 (b) Space 2 (c) Space 3 (d) Space 4

Figure 1: The poor cells standard DARTS finds on spaces S1-S4. For all spaces, DARTS chooses mostly parameter-less operations (skip connection) or even the harmful $Noise$ operation. Here, we show the normal cells; see Figure 17 for the corresponding reduction cells.

3 When DARTS fails

We now describe various search spaces and demonstrate that standard DARTS fails on them. We start with four search spaces similar to the CIFAR-10 search space used in the original DARTS paper (Liu et al., 2018a) but simpler, and evaluated across three different datasets (CIFAR-10, CIFAR-100 and SVHN). We would like to emphasize that these search spaces are in no way special or constructed in an adversarial manner. They use the same macro architecture as the original DARTS paper (Liu et al., 2018a), consisting of normal and reduction cells, but only allow a subset of operators for the cell search space:

- **S1:** This search space uses a different set of two operators per edge, which we identified using an offline process that iteratively dropped the operations from the original DARTS search space with the least importance. This pre-optimized space has the advantage of being quite small while still including many strong architectures. We refer to Appendix B for details on its construction and to Figure 7 in the appendix for an illustration.
- **S2:** The set of candidate operations per edge is $\{3 \times 3$ *SepConv*, *SkipConnect*$\}$. We choose these operations since they are the most frequent in the discovered cells reported by Liu et al. (2019).
- **S3:** The set of candidate operations per edge is $\{3 \times 3$ *SepConv*, *SkipConnect*, *Zero*$\}$, where the *Zero* operation simply replaces every value in the input feature map by zeros.
- **S4:** The set of candidate operations per edge is $\{3 \times 3$ *SepConv*, *Noise*$\}$, where the *Noise* operation simply replaces every value from the input feature map by noise $\epsilon \sim \mathcal{N}(0, 1)$. This is the only space out of S1-S4 that is not a strict subspace of the original DARTS space; we intentionally added the *Noise* operation, which actively harms performance and should therefore not be selected by DARTS.

We ran DARTS on each of these spaces, using exactly the same setup as Liu et al. (2019). Figure 1 shows the poor cells DARTS selected on these search spaces for CIFAR-10 (see Appendix F for analogous results on the other datasets). Already visually, one might suspect that the found cells are suboptimal: the parameter-less skip connections dominate in almost all the edges for spaces S1-S3, and for S4 even the harmful *Noise* operation was selected for five out of eight operations. Table 1 (first column) confirms the very poor performance standard DARTS yields on all of these search spaces for all the aforementioned datasets. We note that Liu et al. (2019) and Xie et al. (2019) argue that the *Zero* operation can help to search for the architecture topology and choice of operators choices jointly, but in our experiments it did not help to reduce the importance weight of the skip connection (compare Figure 1b vs. Figure 1c).

We emphasize that search spaces S1-S3 are very natural, and, as strict subspaces of the original DARTS search space, should merely be easier to search than that. Only S4 was constructed specifically to show-case the failure mode of DARTS selecting the "obviously" suboptimal *Noise* operator.

S5: Very small search space with known global optimum. Knowing the global minimum has the advantage that one can benchmark the performance of algorithms by measuring the regret of chosen points with respect to the known global minimum. Therefore, we created another search space with only one intermediate node for both normal and reduction cells, and 3 operation choices

in each edge, namely 3×3 *SepConv*, *SkipConnection*, and 3×3 *MaxPooling*. The total number of possible architectures in this space is 81, all of which we evaluated a-priori. We dub this space **S5**.

We ran DARTS on this search space three times for each dataset and compared its result to the baseline of Random Search with weight sharing (RS-ws) by Li & Talwalkar (2019). Figure 2 shows the test regret of the architectures selected by DARTS (blue) and RS-ws (green) throughout the search. DARTS manages to find an architecture close to the global minimum, but around epoch 40 the test performance deteriorates. Note that the one-shot (search model) validation error (dashed red line) does not deteriorate but rather converges, indicating that the architectural parameters are overfitting to the validation set. In contrast, RS-ws stays relatively constant throughout the search; when evaluating only the final architecture found, RS-ws indeed outperforms DARTS.

S6: encoder-decoder architecture for disparity estimation. To study whether our findings generalize beyond image recognition, we also analyzed a search space for a very different problem: finding encoder-decoder architectures for the dense regression task of disparity estimation. We base this search space on AutoDispNet (Saikia et al., 2019), which used DARTS for a space containing *normal*, *downsampling* and *upsampling* cells. We again constructed a reduced space, using the following candidate operations for each edge in each of these cells: $\{3 \times 3$ *SepConv*, 3×3 *MaxPool*, *SkipConnect*$\}$. We refer to Appendix D and Saikia et al. (2019) for more details.

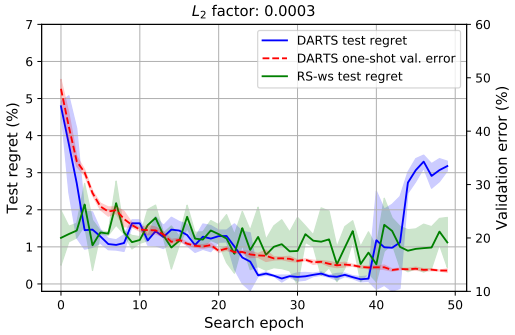


Figure 2: Test regret and validation error of the one-shot model when running DARTS on S5 and CIFAR-10. DARTS finds the global minimum but starts overfitting the architectural parameters to the validation set in the end.

Similarly to the image classification search spaces, we found the normal cell resulting from running standard DARTS on this search space to be mainly composed of parameter-less operations (see Figure 22 in Appendix F). As expected, this causes a large generalization error when AutoDispNet is retrained from scratch (see first row in Table 3 of our later experiments).

4 Architecture Generalization and Hessian Spectrum

4.1 The relationship of large architectural eigenvalues and generalization performance

Why does DARTS perform so poorly in the search spaces described in Section 3? One may hypothesize that DARTS' approximate solution of the bi-level optimization problem by iterative optimization fails, but we actually observe validation errors to nicely progress: Figure 3 (left) shows that the one-shot validation error converges to between 11% and 14% in all cases. This is similar to the one-shot performance of the original one-shot model reported by Liu et al. (2019), even though the cell structures selected by DARTS here are the ones in Figure 1.

Rather, the architectures DARTS finds do not generalize well. This can be seen in Figure 3 (middle). There, we evaluate every 5 epochs the architecture deemed by DARTS to be optimal according to the α values (see Appendix B for the evaluation settings). As one can notice the architectures start to degenerate after a certain number of search epochs, similarly to the results shown in Figure 2. We hypothesized that this might be related to the phenomenon of sharp local minima (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Chaudhari et al., 2017; Yao et al., 2018), which have also been observed in the hyperparameter space (Nguyen et al., 2018). To test this hypothesis, we computed the full Hessian $\nabla^2_\alpha \mathcal{L}_{valid}$ of the validation loss w.r.t. the architectural parameters on a randomly sampled mini-batch from the validation set every two search epochs. Figure 3 (right) shows that the dominant eigenvalue λ^α_{max} increases in standard DARTS for search spaces S1-S4, along with the test error of the final architectures.

Do architectures with large dominant eigenvalues λ^α_{max} also tend to have high test errors? In order to answer this question, we measured these two quantities for 24 different architectures from search

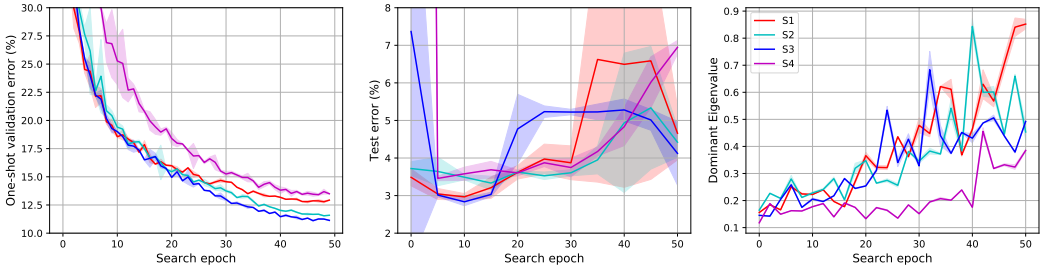


Figure 3: *(Left)* validation error of search model; *(Middle)* Test error of the architectures deemed by DARTS optimal at the end of every 5 epochs; *(Right)* Dominant eigenvalue of $\nabla^2_\alpha \mathcal{L}_{valid}$ throughout DARTS search. The solid line and shaded areas show the mean and standard deviation of 3 independent search runs. All experiments where conducted on CIFAR-10.

space S1 on CIFAR-10, which resulted from running standard DARTS and our regularized versions of it (to be described in the next section) with three different random seeds each. Figure 15 shows that λ^α_{max} indeed correlates with test error (with a Pearson correlation coefficient of 0.867). Having identified this relationship, we now move on to avoid large eigenvalue spectra.

4.2 Early stopping based on large eigenvalues of $\nabla^2_\alpha \mathcal{L}_{valid}$

We just observed that (1) the test error is positively correlated with the largest eigenvalue λ^α_{max} of the Hessian of the validation loss $\nabla^2_\alpha \mathcal{L}_{valid}$, and that (2) λ^α_{max} increases over time. A simple strategy to avoid test errors from increasing is therefore to stop the optimization when λ^α_{max} increases too much.

To implement this idea, we propose a simple heuristic that worked off-the-shelf without the need for any tuning. Let $\overline{\lambda}^\alpha_{max}(i)$ denote the value of λ^α_{max} smoothed over $k = 5$ epochs around i; then, we stop if $\overline{\lambda}^\alpha_{max}(i-k)/\overline{\lambda}^\alpha_{max}(i) < 0.75$ and return the architecture from epoch $i-k$. By this early stopping heuristic, we do not only avoid exploding eigenvalues, which are correlated with poor generalization (see Figure 15), but also shorten the time of the search. Table 1 (DARTS-ES) shows the results for running DARTS with this early stopping criterion across S1-S4 and all three image classification datasets. Early stopping significantly improved DARTS for all settings without ever harming it.

Table 1: Performance of architectures found by DARTS vs. DARTS-ES. For each of the settings we repeat the search 3 times and report the mean $\pm$ std (median) of the 3 found architectures retrained from scratch.

Setting		DARTS	DARTS-ES
C10	S1	4.66 $\pm$ 0.71 (4.57)	**3.05** $\pm$ 0.07 (**3.01**)
	S2	4.42 $\pm$ 0.40 (4.52)	**3.41** $\pm$ 0.14 (**3.39**)
	S3	4.12 $\pm$ 0.85 (3.73)	**3.71** $\pm$ 1.14 (**3.07**)
	S4	6.95 $\pm$ 0.18 (6.86)	**4.17** $\pm$ 0.21 (**4.24**)
C100	S1	29.93 $\pm$ 0.41 (29.88)	**28.90** $\pm$ 0.81 (**28.37**)
	S2	28.75 $\pm$ 0.92 (28.31)	**24.68** $\pm$ 1.43 (**24.03**)
	S3	29.01 $\pm$ 0.24 (28.90)	**26.99** $\pm$ 1.79 (**25.20**)
	S4	24.77 $\pm$ 1.51 (24.92)	**23.90** $\pm$ 2.01 (**23.89**)
SVHN	S1	9.88 $\pm$ 5.50 (7.60)	**2.80** $\pm$ 0.09 (**2.76**)
	S2	3.69 $\pm$ 0.12 (3.73)	**2.68** $\pm$ 0.18 (**2.62**)
	S3	4.00 $\pm$ 1.01 (3.47)	**2.78** $\pm$ 0.29 (**2.65**)
	S4	2.90 $\pm$ 0.02 (2.91)	**2.55** $\pm$ 0.15 (**2.51**)

5 Regularization of inner objective improves generalization of architectures

As we saw in Section 4.1, sharper minima of the validation loss lead to poor generalization performance. In our bi-level optimization setting, the outer variables' trajectory depends on the inner optimization procedure. Therefore, modifying the landscape of the inner objective $\mathcal{L}_{train}$ might potentially redirect the outer variables α to better areas of the architectural space and keep the dominant eigenvalue low. We study two ways of regularization (data augmentation in Section 5.1 and L_2 regularization in Section 5.2) and find that both, along with the early stopping criterion from Section 4.2, make DARTS more robust in practice. Our observations are valid on two different tasks, namely image classification across search spaces S1-S4 and disparity estimation (S6). We highlight that we *do not* alter the regularization of the final training and evaluation phase, but solely that of the DARTS search phase.

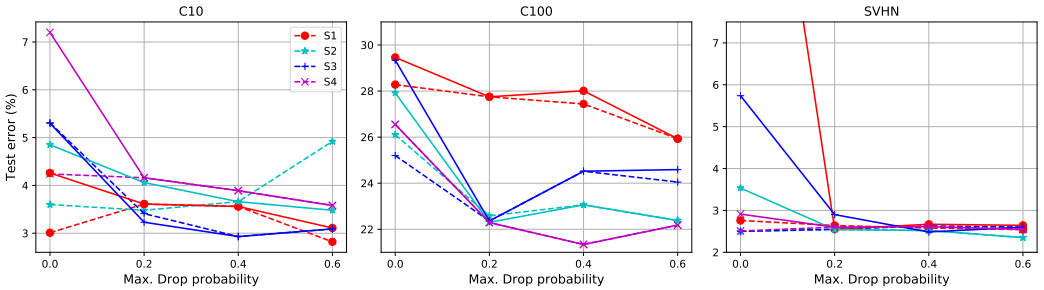


Figure 4: Effect of more regularization via ScheduledDropPath during the DARTS search phase, on the test performance of the architectures discovered by DARTS and DARTS-ES. The results are presented for each of the search spaces S1-S4 and for CIFAR-10, CIFAR-100 and SVHN. The solid lines correspond to DARTS, the dashed lines to DARTS-ES.

5.1 Regularization via data augmentation

We first investigate the effect of regularizing via data augmentation, namely masking out parts of the input and intermediate feature maps via Cutout (CO) (DeVries & Taylor, 2017) and ScheduledDropPath (DP) (Zoph et al., 2018), respectively, during architecture search. We apply DP to randomly zero out mixed operations starting with a drop probability of 0 and linearly increasing it over the course of architecture search until it reaches a maximum value. At the same, we randomly zero out patches in the input images by applying CO; the CO probability also linearly increases.

We ran DARTS plus drop-path (with and without our early stopping criterion, DARTS-ES) with four values of the maximum drop-path probability (0.0, 0.2, 0.4 and 0.6) on all three image classification datasets and search spaces S1-S4. Figure 4 summarizes the results: regularization improves the test performance of DARTS and DARTS-ES in all cases, sometimes very substantially, and at the same time keeps the dominant eigenvalue relatively low. Figure 11 in the Appendix shows the local average of the dominant eigenvalue throughout the DARTS search across all the settings. Table 2 provides additional details, also showing that the one-shot model accuracy consistently drops by increasing the drop-path probability, while the test accuracy improves (up to a certain limit). This demonstrates that overfitting of the architectural parameters is reduced due to an implicit regularization effect (see also Figure 9 in the Appendix).

We now show a similar experiment for disparity estimation on S6. In this case, we vary the strength of standard data augmentation methods, such as shearing or brightness change, rather then masking parts of features, which is unreasonable for this task. The augmentation strength is described by a "augmentation scaling factor" (see Appendix D for details). Table 3 summarizes the results. The best test performance is obtained for the network with maximum augmentation.[3]

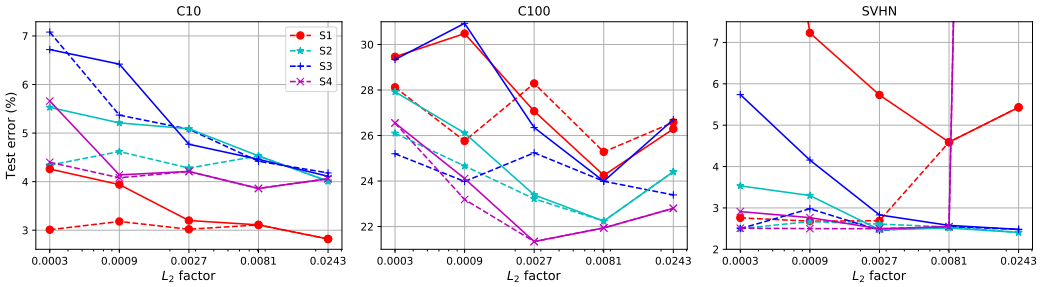


Figure 5: Effect of increasing L_2 regularization of the inner objective. The figure is analogous to Figure 4.

[3]We note that we could not test the early stopping method on AutoDispNet since AutoDispNet relies on custom operations to compute feature map correlation (Dosovitskiy et al., 2015) and resampling, for which second order derivatives are currently not available (which are required to compute the Hessian).

Table 2: Validation (train) and test accuracy on CIFAR-10 of the one-shot and final evaluation model, respectively. The values in the last column show the maximum eigenvalue λ^{α}_{max} (computed on a random sampled mini-batch) of the Hessian, at the end of search for different maximum drop path probability). The four blocks in the table state results for the search spaces S1-S4, respectively.

Space	Drop Prob.	Valid (train) acc. one-shot (%)	Test acc. final (%)	Params (M)	λ^{α}_{max}
S1	0.0	87.22 (97.95)	96.16	2.24	1.023
	0.2	84.24 (83.61)	96.39	2.63	0.148
	0.4	82.28 (75.69)	96.44	2.63	0.192
	0.6	79.17 (63.01)	**96.89**	3.38	0.300
S2	0.0	88.49 (98.69)	95.15	0.93	0.684
	0.2	85.29 (84.49)	95.94	1.28	0.270
	0.4	82.92 (78.00)	96.34	1.28	0.304
	0.6	79.68 (66.15)	**96.52**	1.21	0.292
S3	0.0	88.78 (99.26)	94.70	2.21	0.496
	0.2	85.61 (85.14)	96.78	3.62	0.179
	0.4	83.03 (78.59)	**97.07**	4.10	0.156
	0.6	79.86 (66.94)	96.91	4.46	0.239
S4	0.0	86.33 (95.63)	92.80	1.05	0.400
	0.2	81.01 (79.44)	95.84	1.44	0.070
	0.4	79.49 (73.86)	96.11	1.44	0.064
	0.6	74.54 (61.24)	**96.42**	1.44	0.057

Table 3: Effect of more augmentation on the architecture generalization found by AutoDispNet. The search was conducted on FlyingThings3D (FT) and the final architecture was evaluated on both FT and Sintel. Lower is better.

Aug. Scale	One-shot valid EPE	FT test EPE	Sintel test EPE	Params (M)
0.0	4.49	3.83	5.69	9.65
0.1	3.53	3.75	5.97	9.65
0.5	3.28	3.37	5.22	9.43
1.0	4.61	3.12	5.47	12.46
1.5	5.23	2.60	4.15	12.57
2.0	7.45	**2.33**	**3.76**	12.25

Table 4: Effect of more L_2 regularization on the architecture generalization found by AutoDispNet. Lower is better.

L_2 reg. factor	One-shot valid EPE	FT test EPE	Sintel test EPE	Params (M)
3×10^{-4}	3.95	3.25	6.13	11.00
9×10^{-4}	5.97	**2.30**	4.12	13.92
27×10^{-4}	4.25	2.72	4.83	10.29
81×10^{-4}	4.61	**2.34**	**3.85**	12.16

5.2 Increased L_2 regularization

We now test different L_2 regularization factors $3i \cdot 10^{-4}$ for $i \in \{1, 3, 9, 27, 81\}$ for image classification and S1-S4. Standard DARTS in fact does already include a small amount of L_2 regularization; $i = 1$ yields its default. In Figure 5 we can see that the test performance of standard DARTS (solid lines) can be significantly improved for other L_2 factors than the default $3 \cdot 10^{-4}$ across all datasets and search spaces, while keeping the dominant eigenvalue low (see Figure 12 in the Appendix). DARTS with early stopping (dashed lines) also benefits from additional regularization. Again, we observe the implicit regularization effect on the outer objective which reduces the overfitting of the architectural parameters. The suboptimal cells composed mostly of parameter-less operations perform poorly, even though the one-shot validation error surprisingly converges to a relatively low value (see also Figure 10 in the Appendix).

We conduct the same experiment for disparity estimation on S6. The search is done on the FlyingThings3D (FT) dataset (Mayer et al., 2016) and the extracted architectures are afterwards retrained from scratch and evaluated on the FT and Sintel (Butler et al., 2012a) datasets. We report the average end point error (EPE), which is the Euclidean distance between the predicted and ground truth disparity maps. Table 4 summarizes the results. We observe similar trends here: when the regularization strength increases, DARTS finds networks with better test performance.

In Appendix E we also show that increasing L_2 regularization improves generalization performance for a search space for language modelling (Penn TreeBank).

6 Discussion and further analysis

In this section we discuss one potential hypothesis that may explain why DARTS fails in several scenarios as shown in this paper. Afterwards we provide empirical justification on why the curvature of the validation objective is a good indicator for the generalization performance of the architectures found by DARTS.

As it is well known in the settings of large vs. small batch training (Yao et al., 2018; Keskar et al., 2016; Hochreiter & Schmidhuber, 1997), one potential hypothesis explaining the relationship between the sharpness of minimas and generalization properties of a neural network, is based on the fact that the training function is much more sensitive at a sharp minimizer, e.g. to the variations in the input data. This may lead to relatively large accuracy drop even from small discrepancies

between training and test data. Analogously, we conjectured that this also holds for α rather than w and therefore we investigated the relation between large eigenvalues (w.r.t. α) - as a proxy for sharp minima - and generalization performance. Similarly, sharp minima would also be much more sensitive to variations in the architecture, which is relevant since DARTS discretizes (by taking the *argmax* over operations in each edge) the optimal α^* after search, resulting in α^d somewhere in the neighbourhood of α^*. In the case of a sharp minimum α^*, α^d might already have a significantly larger objective function value, while in the case of a flat minimum, α^d is expected to have an objective value similar to α^*.

To make this more crisp, we conducted the following experiment: after the DARTS search has finished, we discretize the architecture and evaluate it with the search model's weights, rather than retraining, and compare the performance to the one-shot model's performance. Figure 6 (*Bottom*) shows the drop in performance due to this discretization step for some of the settings: this drop is much larger when there is little regularization (drop prob = 0), resulting in large eigenvalues (see Figures 13 and 14 Appendix), corresponding to a sharp minmum. Figure 6 (*Top*) shows the relationship between the dominant eigenvalues at the end of search and the drop in accuracy when doing the discretization step as described above.

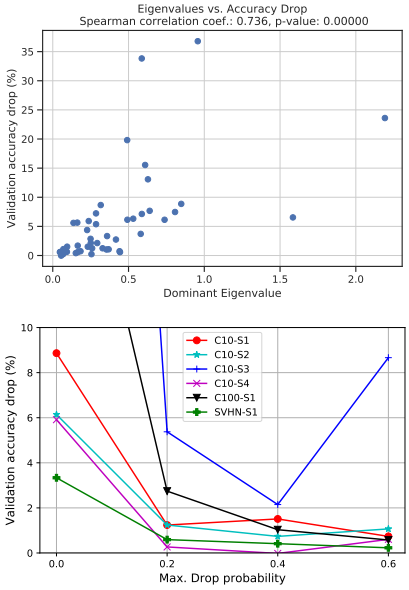


Figure 6: Drop in accuracy after discretizing the search model. (*Top*) Correlation with the landscape curvature. (*Bottom*) Example of some of the settings from Section 5.

Why does DARTS get attracted to these bad regions in the architecture space? This might arise potentially from some premature convergence in the weights' space, supposing that the landscape of the training loss $\mathcal{L}_{train}$ does not change significantly after each α update. In a non-convex landscape, this minimizer $w^{(1)}$ of $\mathcal{L}_{train}$ might not necessarily be the one that also minimizes $\mathcal{L}_{valid}$ (see Franceschi et al. (2018)), i.e. there might exist another $w^{(2)} \in \arg\min_w \mathcal{L}_{train}(\alpha, w)$ such that $\mathcal{L}_{valid}(\alpha, w^{(2)}) < \mathcal{L}_{valid}(\alpha, w^{(1)})$. More concretely, the parameterless operations such as *skip connections* might get higher weight especially in the beginning of search due to the easiness of gradients to flow through these paths during training. Regularizing the inner objective by adding a convex term or perturbing the inputs will eventually redirect also the gradients flowing backwards, and consecutively also the attention DARTS focuses on different architectural operations.

7 Practical Robustification of DARTS

While we demonstrated in Section 5 that additional regularization serves to robustify DARTS, the regularization by L2 penalty and data augmentation introduces additional hyperparameters. While one could of course run an extensive optimization of these for each dataset at a time, this is not useful for practical applications since the runtime for this routine would have to be counted as part of the runtime of the resulting method (Lindauer & Hutter, 2019). Rather, in this section, we propose practical robustification methods that come with minimal computational overhead.

7.1 DARTS with adaptive regularization

Based on the insights from the aforementioned analysis and empirical results, one may think of a way to adapt DARTS' hyperparameters (L_2 regularization or max. drop path probability value) in an automated way, in order to keep the architectural weights in areas of the validation loss objective with smaller curvature. The simplest off-the-shelf procedure towards this desiderata would be to increase the regularization strength whenever the dominant eigenvalue starts increasing rapidly.

Algorithm 1 (DARTS-ADA, see the Appendix) shows such a procedure. We use the same criterion (`stop_criter`) used in the DARTS-ES (Section 4.2), roll back DARTS to the architectural and

Table 5: Empirical evaluation of practical robustified versions of DARTS. Each entry is the test error after retraining the selected architecture as described in Section 7.2.

Setting		RS-ws	DARTS	R-DARTS(DP)	R-DARTS(L2)	DARTS-ES	DARTS-ADA
C10	S1	3.23	3.84	3.11	**2.78**	**3.01**	3.10
	S2	3.66	4.85	3.48	**3.31**	**3.26**	3.35
	S3	2.95	3.34	2.93	**2.51**	2.74	**2.59**
	S4	8.07	7.20	**3.58**	**3.56**	3.71	4.84
C100	S1	**23.30**	29.46	25.93	24.25	28.37	**24.03**
	S2	**21.21**	26.05	22.30	**22.24**	23.25	23.52
	S3	23.75	28.90	**22.36**	23.99	23.73	**23.37**
	S4	28.19	22.85	22.18	**21.94**	**21.26**	23.20
SVHN	S1	2.59	4.58	**2.55**	4.79	2.72	**2.53**
	S2	2.72	3.53	**2.52**	**2.51**	2.60	2.54
	S3	2.87	3.41	**2.49**	**2.48**	2.50	2.50
	S4	3.46	3.05	2.61	**2.50**	2.51	**2.46**

search model parameters at that point (`stop_epoch`) whenever this criterion is met, and continue the search with a larger regularization value R for the remaining epochs (larger by a factor of η). This procedure is repeated whenever the criterion is met, unless the regularization value exceeds some maximum predefined value R_{max}.

7.2 A Simple Robustification of DARTS

In order to hedge against brittle optimization runs, the original DARTS paper (Liu et al., 2018a) already suggested to run the search phase of DARTS four times, resulting in four architectures, and to return the best of these four architectures w.r.t. validation performance when retrained from scratch for a limited number of epochs. Then the returned architecture is retrained from scratch as usual, for 600 epochs using the evaluation settings (see Appendix B.1). We propose to use the same procedure, with the only difference that the four runs of the search phase use different amounts of regularization. The resulting RobustDARTS (R-DARTS) method is conceptually very simple, trivial to implement, and likely to work well if any of the tried regularization strengths works well.

7.3 Empirical Evaluation

Table 5 evaluates the performance of our practical robustifications of DARTS, DARTS-ADA and R-DARTS (based on either L2 or ScheduledDropPath regularization), by comparing them to the original DARTS, DARTS-ES and Random Search with weight sharing (RS-ws). For each of these methods, as proposed in the DARTS paper (Liu et al., 2018a), we ran the search four independent times with different random seeds and selected the architecture used for the final evaluation based on a validation run as described above. As the table shows, in accordance with Li & Talwalkar (2019), RS-ws often outperforms the original DARTS; however, with our robustifications, DARTS typically performs substantially better than RS-ws. The improvements of DARTS-ADA are consistent across all settings compared to the default DARTS, indicating that a gradual increase of regularization during search prevents ending up in the bad regions of the architectural space. Finally, RobustDARTS yielded the best performance and since it is also easier to implement than DARTS-ES and DARTS-ADA, it is the method that we recommend to be used in practice.

8 Conclusions

We showed that the generalization performance of architectures found by DARTS is related to the eigenvalues of the Hessian matrix of the validation loss w.r.t. the architectural parameters. Standard DARTS often results in degenerate architectures with large eigenvalues and poor generalization. Our empirical results show that properly regularizing the inner objective helps controlling the eigenvalue and therefore improves generalization. We also proposed a simple early stopping criterion for DARTS based on tracking the largest eigenvalue. Our findings substantially improve our understanding of DARTS' failure modes and lead to much more robust versions. They are consistent across many different search spaces on image recognition tasks and also for the very different domains of language modelling and disparity estimation. Our code is available for reproducibility.

ACKNOWLEDGMENTS

The authors acknowledge funding by the Robert Bosch GmbH, support by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme through grant no. 716721, and by BMBF grant DeToL.

REFERENCES

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *International Conference on Learning Representations*, 2017a.

Bowen Baker, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Accelerating Neural Architecture Search using Performance Prediction. In *NIPS Workshop on Meta-Learning*, 2017b.

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On optimal generalizability in parametric learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3455–3465. Curran Associates, Inc., 2017.

Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and simplifying one-shot architecture search. In *International Conference on Machine Learning*, 2018.

Y. Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8):1889–1900, Aug 2000. ISSN 0899-7667. doi: 10.1162/089976600300015187.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012a.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012b.

Han Cai, Tianyao Chen, Weinan Zhang, Yong Yu, and Jun Wang. Efficient architecture search by network transformation. In *AAAI*, 2018a.

Han Cai, Jiacheng Yang, Weinan Zhang, Song Han, and Yong Yu. Path-Level Network Transformation for Efficient Architecture Search. In *International Conference on Machine Learning*, June 2018b.

Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations*, 2019.

Francesco Casale, Jonathan Gordon, and Nicolo Fusi. Probabilistic neural architecture search. *arXiv preprint*, 2019.

P. Chaudhari, Anna Choromanska, S. Soatto, Yann LeCun, C. Baldassi, C. Borgs, J. Chayes, Levent Sagun, and R. Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*, 2017.

Liu Chenxi, Chen Liang Chieh, Schroff Florian, Adam Hartwig, Hua Wei, Yuille Alan L., and Fei Fei Li. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Benot Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization, 2007.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Justin Domke. Generic methods for optimization-based modeling. In Neil D. Lawrence and Mark Girolami (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 318–326, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Simple And Efficient Architecture Search for Convolutional Neural Networks. In *NIPS Workshop on Meta-Learning*, 2017.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Efficient multi-objective neural architecture search via lamarckian evolution. In *International Conference on Learning Representations*, 2019a.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019b.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1437–1446, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/falkner18a.html`.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for linear bilevel programming. *SIAM J. Sci. Stat. Comput.*, 13(5):1194–1217, September 1992. ISSN 0196-5204. doi: 10.1137/0913069.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, January 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*, 2017. Published online: `iclr.cc`.

Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search. *CoRR*, abs/1902.07638, 2019.

Marius Lindauer and Frank Hutter. Best practices for scientific research on neural architecture search. *arXiv preprint arXiv:1909.02453*, 2019.

H. Liu, K. Simonyan, O. Vinyals, C.Fernando, and K. Kavukcuoglu. Hierarchical representations for efficient architecture search. In *International Conference on Learning Representations (ICLR) 2018 Conference Track*, April 2018a.

Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical Representations for Efficient Architecture Search. In *International Conference on Learning Representations*, 2018b.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122, Lille, France, 07–09 Jul 2015. PMLR.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Dan Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, and Babak Hodjat. Evolving Deep Neural Networks. In *arXiv:1703.00548*, March 2017.

Thanh Dai Nguyen, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Stable bayesian optimization. *International Journal of Data Science and Analytics*, 6(4):327–339, Dec 2018. ISSN 2364-4168. doi: 10.1007/s41060-018-0119-9.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 737–746, New York, New York, USA, 20–22 Jun 2016. PMLR.

Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In *International Conference on Machine Learning*, 2018.

Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V. Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2902–2911, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Aging Evolution for Image Classifier Architecture Search. In *AAAI*, 2019.

Tonmoy Saikia, Yassine Marrakchi, Arber Zela, Frank Hutter, and Thomas Brox. Autodispnet: Improving disparity estimation with automl, 2019.

Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4053–4061. Curran Associates, Inc., 2016.

Christian Sciuto, Kaicheng Yu, Martin Jaggi, Claudiu Musat, and Mathieu Salzmann. Evaluating the search phase of neural architecture search. *arXiv preprint*, 2019.

Kenneth O Stanley and Risto Miikkulainen. Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10:99–127, 2002.

Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. SNAS: stochastic neural architecture search. In *International Conference on Learning Representations*, 2019.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4949–4959. Curran Associates, Inc., 2018.

Arber Zela, Aaron Klein, Stefan Falkner, and Frank Hutter. Towards automated deep learning: Efficient joint neural architecture and hyperparameter search. In *ICML 2018 Workshop on AutoML (AutoML 2018)*, July 2018.

Zhao Zhong, Jingchen Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Practical block-wise neural network architecture generation. In *CVPR*. IEEE Computer Society, 2018.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2018.

A MORE DETAIL ON DARTS

Here we present a detailed description of DARTS architectural update steps. We firstly provide the general formalism which computes the gradient of the outer level problem in (1) by means of the implicit function theorem. Afterwards, we present how DARTS computes the gradient used to update the architectural parameters α.

A.1 DERIVATIVE WITH SMOOTHED NON-QUADRATIC LOWER LEVEL PROBLEM

Consider the general definition of the bi-level optimization problem as given by (1) and (2). Given that f is twice continuously differentiable and that all stationary points are local minimas, one can make use of the implicit function theorem to find the derivative of the solution map $\theta^*(y)$ w.r.t. y (Bengio, 2000). Under the smoothness assumption, the optimality condition of the lower level (2) is $\nabla_\theta f(y, \theta) = \mathbf{0}$, which defines an implicit function $\theta^*(y)$. With the assumption that $\min_\theta f(y, \theta)$ has a solution, there exists a (y, θ^*) such that $\nabla_\theta f(y, \theta^*) = \mathbf{0}$. Under the condition that $\nabla_\theta f(y, \theta^*) = \mathbf{0}$ is continuously differentiable and that $\theta^*(y)$ is continuously differentiable at y, implicitly differentiating the last equality from both sides w.r.t. y and applying the chain rule, yields:

$$\frac{\partial(\nabla_\theta f)}{\partial \theta}(y, \theta^*) \cdot \frac{\partial \theta^*}{\partial y}(y) + \frac{\partial(\nabla_\theta f)}{\partial y}(y, \theta^*) = \mathbf{0}. \tag{4}$$

Assuming that the Hessian $\nabla^2_\theta f(y, \theta^*)$ is invertible, we can rewrite (4) as follows:

$$\frac{\partial \theta^*}{\partial y}(y) = -\Big(\nabla^2_\theta f(y, \theta^*)\Big)^{-1} \cdot \frac{\partial(\nabla_\theta f)}{\partial y}(y, \theta^*). \tag{5}$$

Applying the chain rule to (1) for computing the total derivative of *F* with respect to y yields:

$$\frac{dF}{dy} = \frac{\partial F}{\partial \theta} \cdot \frac{\partial \theta^*}{\partial y} + \frac{\partial F}{\partial y}, \tag{6}$$

where we have omitted the evaluation at (y, θ^*). Substituting (5) into (6) and reordering yields:

$$\frac{dF}{dy} = \frac{\partial F}{\partial y} - \frac{\partial F}{\partial \theta} \cdot \Big(\nabla^2_\theta f\Big)^{-1} \cdot \frac{\partial^2 f}{\partial \theta \partial y}. \tag{7}$$

equation 7 computes the gradient of *F*, given the function $\theta^*(y)$, which maps outer variables to the inner variables minimizing the inner problem. However, in most of the cases obtaining such a mapping is computationally expensive, therefore different heuristics have been proposed to approximate dF/dy (Maclaurin et al., 2015; Pedregosa, 2016; Franceschi et al., 2017; 2018).

A.2 DARTS ARCHITECTURAL GRADIENT COMPUTATION

DARTS optimization procedure is defined as a bi-level optimization problem where $\mathcal{L}_{valid}$ is the outer objective (1) and $\mathcal{L}_{train}$ is the inner objective (2):

$$\min_\alpha \mathcal{L}_{valid}(\alpha, w^*(\alpha)) \tag{8}$$

$$s.t. \quad w^*(\alpha) = \arg\min_w \mathcal{L}_{train}(\alpha, w), \tag{9}$$

where both losses are determined by both the architecture parameters α (outer variables) and the network weights w (inner variables). Based on Appendix A.1, under some conditions, the total derivative of $\mathcal{L}_{valid}$ w.r.t. α evaluated on $(\alpha, w^*(\alpha))$ would be:

$$\frac{d\mathcal{L}_{valid}}{d\alpha} = \nabla_\alpha \mathcal{L}_{valid} - \nabla_w \mathcal{L}_{valid} \big(\nabla^2_w \mathcal{L}_{train}\big)^{-1} \nabla^2_{\alpha,w} \mathcal{L}_{train}, \tag{10}$$

where $\nabla_\alpha = \frac{\partial}{\partial \alpha}$, $\nabla_w = \frac{\partial}{\partial w}$ and $\nabla^2_{\alpha,w} = \frac{\partial^2}{\partial \alpha \partial w}$. Computing the inverse of the Hessian is in general not possible considering the high dimensionality of the model parameters w, therefore resolving to gradient-based iterative algorithms for finding w^* is necessary. However, this would also require to

optimize the model parameters w till convergence each time α is updated. If our model is a deep neural network it is clear that this computation is expensive, therefore Liu et al. (2019) propose to approximate $w^*(\alpha)$ by updating the current model parameters w using a single gradient descent step:

$$w^*(\alpha) \approx w - \xi \nabla_w \mathcal{L}_{train}(\alpha, w), \tag{11}$$

where ξ is the learning rate for the virtual gradient step DARTS takes with respect to the model weights w. From equation 11 the gradient of $w^*(\alpha)$ with respect to α is

$$\frac{\partial w^*}{\partial \alpha}(\alpha) = -\xi \nabla^2_{\alpha,w} \mathcal{L}_{train}(\alpha, w), \tag{12}$$

By setting the evaluation point $w^* = w - \xi \nabla_w \mathcal{L}_{train}(\alpha, w)$ and following the same derivation as in Appendix A.1, we obtain the DARTS architectural gradient approximation:

$$\frac{d\mathcal{L}_{valid}}{d\alpha}(\alpha) = \nabla_\alpha \mathcal{L}_{valid}(\alpha, w^*) - \xi \nabla_w \mathcal{L}_{valid}(\alpha, w^*) \nabla^2_{\alpha,w} \mathcal{L}_{train}(\alpha, w^*), \tag{13}$$

where the inverse Hessian $\nabla^2_w \mathcal{L}^{-1}_{train}$ in (10) is replaced by the learning rate ξ. This expression however contains again an expensive vector-matrix product. Liu et al. (2019) reduce the complexity by using the finite difference approximation around $w^{\pm} = w \pm \epsilon \nabla_w \mathcal{L}_{valid}(\alpha, w^*)$ for some small $\epsilon = 0.01 / \left\| \nabla_w \mathcal{L}_{valid}(\alpha, w^*) \right\|_2$ to compute the gradient of $\nabla_\alpha \mathcal{L}_{train}(\alpha, w^*)$ with respect to w as

$$\begin{aligned} \nabla^2_{\alpha,w} \mathcal{L}_{train}(\alpha, w^*) &\approx \frac{\nabla_\alpha \mathcal{L}_{train}(\alpha, w^+) - \nabla_\alpha \mathcal{L}_{train}(\alpha, w^-)}{2\epsilon \nabla_w \mathcal{L}_{valid}(\alpha, w^*)} \qquad \Leftrightarrow \\ \nabla_w \mathcal{L}_{valid}(\alpha, w^*) \nabla^2_{\alpha,w} \mathcal{L}_{train}(\alpha, w^*) &\approx \frac{\nabla_\alpha \mathcal{L}_{train}(\alpha, w^+) - \nabla_\alpha \mathcal{L}_{train}(\alpha, w^-)}{2\epsilon}. \end{aligned} \tag{14}$$

In the end, combining equation 13 and equation 14 gives the gradient to compute the architectural updates in DARTS:

$$\frac{d\mathcal{L}_{valid}}{d\alpha}(\alpha) = \nabla_\alpha \mathcal{L}_{valid}(\alpha, w^*) - \frac{\xi}{2\epsilon} \big(\nabla_\alpha \mathcal{L}_{train}(\alpha, w^+) - \nabla_\alpha \mathcal{L}_{train}(\alpha, w^-) \big) \tag{15}$$

In all our experiments we always use $\xi = \eta$ (also called second order approximation in Liu et al. (2019)), where η is the learning rate used in SGD for updating the parameters w.

B Details on search spaces in Section 3 and final architecture evaluations

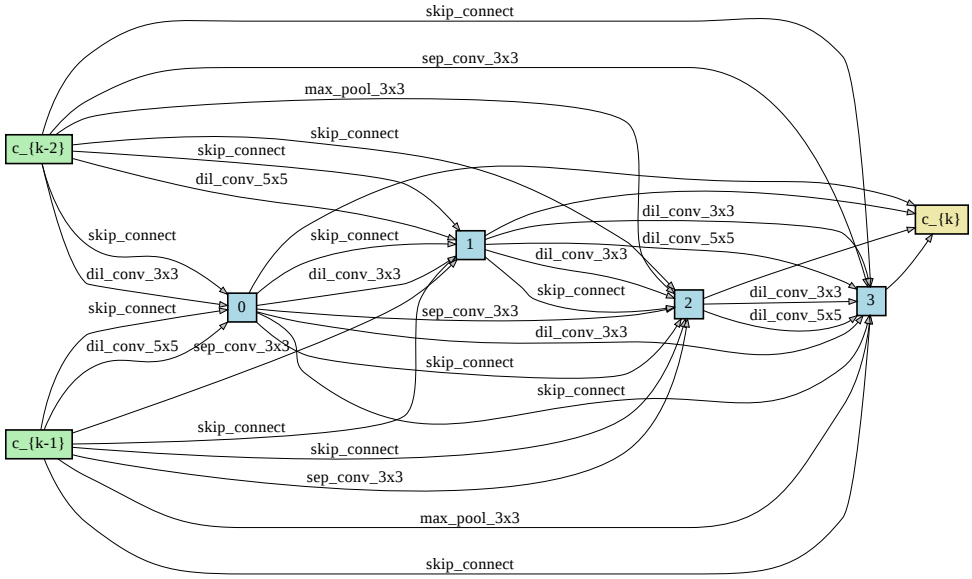
We ran DARTS on the default search space to find the two most important operations per edge. S1 is then defined to contain only these most important operations per edge. Refer to Figure 7 for an illustration of this pre-optimized space.

B.1 Architecture Evaluation

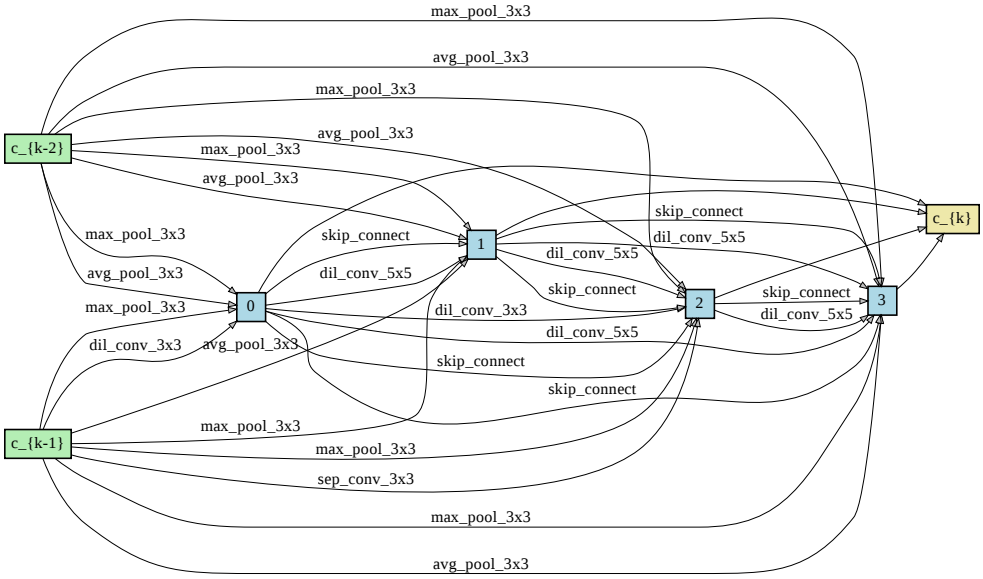
For CIFAR-100 and SVHN we use 16 number of initial filters and 8 cells when training architectures from scratch for all the experiments we conduct. The rest of the settings is the same as in Liu et al. (2019).

On CIFAR-10, when scaling the ScheduledDropPath drop probability, we use the same settings for training from scratch the found architectures as in the original DARTS paper, i.e. 36 initial filters and 20 stacked cells. However, for search space S2 and S4 we reduce the number of initial filters to 16 in order to avoid memory issues, since the cells found with more regularization usually are composed only with separable convolutions. When scaling the L_2 factor on CIFAR-10 experiments we use 16 initial filters and 8 stacked cells, except the experiments on S1, where the settings are the same as in Liu et al. (2019), i.e. 36 initial filters and 20 stacked cells.

Note that although altering the regularization factors during DARTS search, when training the final architectures from scratch we always use the same values for them as in Liu et al. (2019), i.e. ScheduledDropPath maximum drop probability linearly increases from 0 towards 0.2 throughout training, Cutout is always enabled with cutout probability 1.0, and the L_2 regularization factor is set to $3 \cdot 10^{-4}$.



(a) Normal cell space



(b) Reduction cell space

Figure 7: Search space **S1**.

C ADDITIONAL EMPIRICAL RESULTS

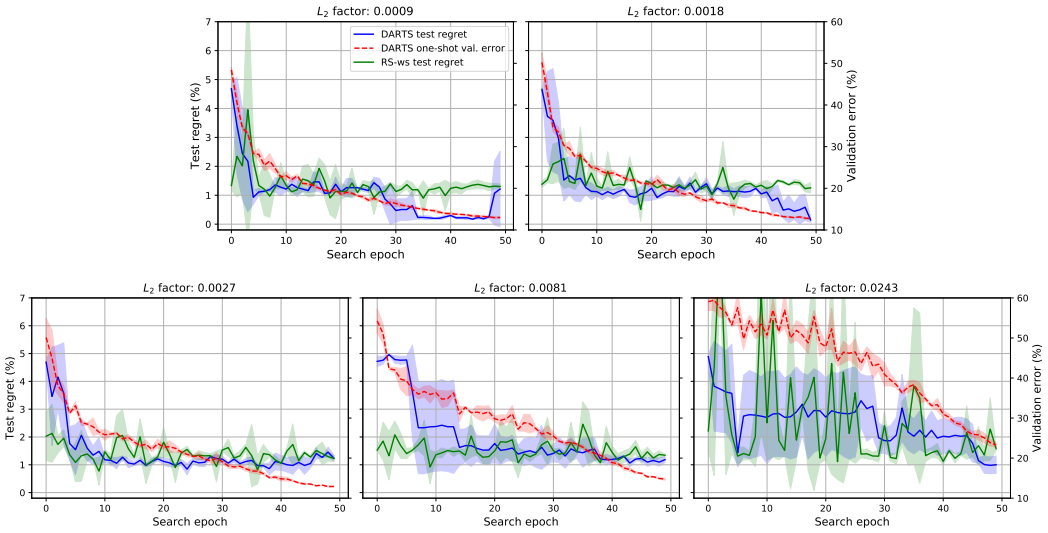


Figure 8: Test regret and validation error of the one-shot model when running DARTS on S5 and CIFAR-10 with different L_2 regularization values. The architectural parameters' overfit reduces as we increase the L_2 factor and successfully finds the global minimum. However, we notice that the architectural parameters start underfitting as we increase to much the L_2 factor, i.e. both validation and test error increase.

Table 6: Validation (train) and test accuracy on CIFAR-10 of the one-shot and final evaluation model, respectively. The values in the last column show the maximum eigenvalue λ^{α}_{max} (computed on a random sampled mini-batch) of the Hessian, at the end of search for different maximum drop path probability). The four blocks in the table state results for the search spaces S1-S4, respectively.

	Drop	**Valid acc.**			**Test acc.**			**Params**			λ^{α}_{max}		
	Prob.	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
S1	0.0	87.22	87.01	86.98	96.16	94.43	95.43	2.24	1.93	2.03	1.023	0.835	0.698
	0.2	84.24	84.32	84.22	96.39	96.66	96.20	2.63	2.84	2.48	0.148	0.264	0.228
	0.4	82.28	82.18	82.79	96.44	**96.94**	96.76	2.63	2.99	3.17	0.192	0.199	0.149
	0.6	79.17	79.18	78.84	**96.89**	96.93	**96.96**	3.38	3.02	3.17	0.300	0.255	0.256
S2	0.0	88.49	88.40	88.35	95.15	95.48	96.11	0.93	0.86	0.97	0.684	0.409	0.268
	0.2	85.29	84.81	85.36	95.15	95.40	96.14	1.28	1.44	1.36	0.270	0.217	0.145
	0.4	82.03	82.66	83.20	96.34	**96.50**	**96.44**	1.28	1.28	1.36	0.304	0.411	0.282
	0.6	79.86	80.19	79.70	**96.52**	96.35	96.29	1.21	1.28	1.36	0.292	0.295	0.281
S3	0.0	88.78	89.15	88.67	94.70	96.27	96.66	2.21	2.43	2.85	0.496	0.535	0.446
	0.2	85.61	85.60	85.50	96.78	96.84	**96.74**	3.62	4.04	2.99	0.179	0.185	0.202
	0.4	83.03	83.24	83.43	**97.07**	**96.85**	96.48	4.10	3.74	3.38	0.156	0.370	0.184
	0.6	79.86	80.03	79.68	96.91	94.56	96.44	4.46	2.30	2.66	0.239	0.275	0.280
S4	0.0	86.33	86.72	86.46	92.80	93.22	93.14	1.05	1.13	1.05	0.400	0.442	0.314
	0.2	81.01	82.43	82.03	95.84	96.08	96.15	1.44	1.44	1.44	0.070	0.054	0.079
	0.4	79.49	79.67	78.96	96.11	96.30	96.28	1.44	1.44	1.44	0.064	0.057	0.049
	0.6	74.54	74.74	74.37	**96.42**	**96.36**	**96.64**	1.44	1.44	1.44	0.057	0.060	0.066

C.1 ADAPTIVE DARTS DETAILS

We evaluated DARTS-ADA (Section 7.1) with $R = 3 \cdot 10^{-4}$ (DARTS default), $R_{max} = 3 \cdot 10^{-2}$ and $\eta = 10$ on all the search spaces and datasets we use for image classification. The results are shown in Table 5 (DARTS-ADA). The function `train_and_eval` conducts the normal DARTS search for one epoch and returns the architecture at the end of that epoch's updates and the `stop` value if a decision was made to stop the search and rollback to `stop_epoch`.

Algorithm 1: DARTS_ADA

/* E: epochs to search; R: initial regularization value; R_{max}: maximal regularization value; stop_criter: stopping criterion; η: regularization increase factor */
Input : E, R, R_{max}, stop_criter, η

/* start search for E epochs */
for *epoch* ***in*** *E* **do**
    /* run DARTS for one epoch and return stop=True together with the stop_epoch */
    /* and the architecture at stop_epoch if the criterion is met */
    stop, stop_epoch, arch ← train_and_eval(stop_criter);
    if *stop* & $R \leq R_{max}$ **then**
        /* start DARTS from stop_epoch with a larger R */
        arch ← DARTS_ADA(E - stop_epoch, $\eta \cdot R$, R_{max}, stop_criter, η);
        break
    end
end

Output: arch

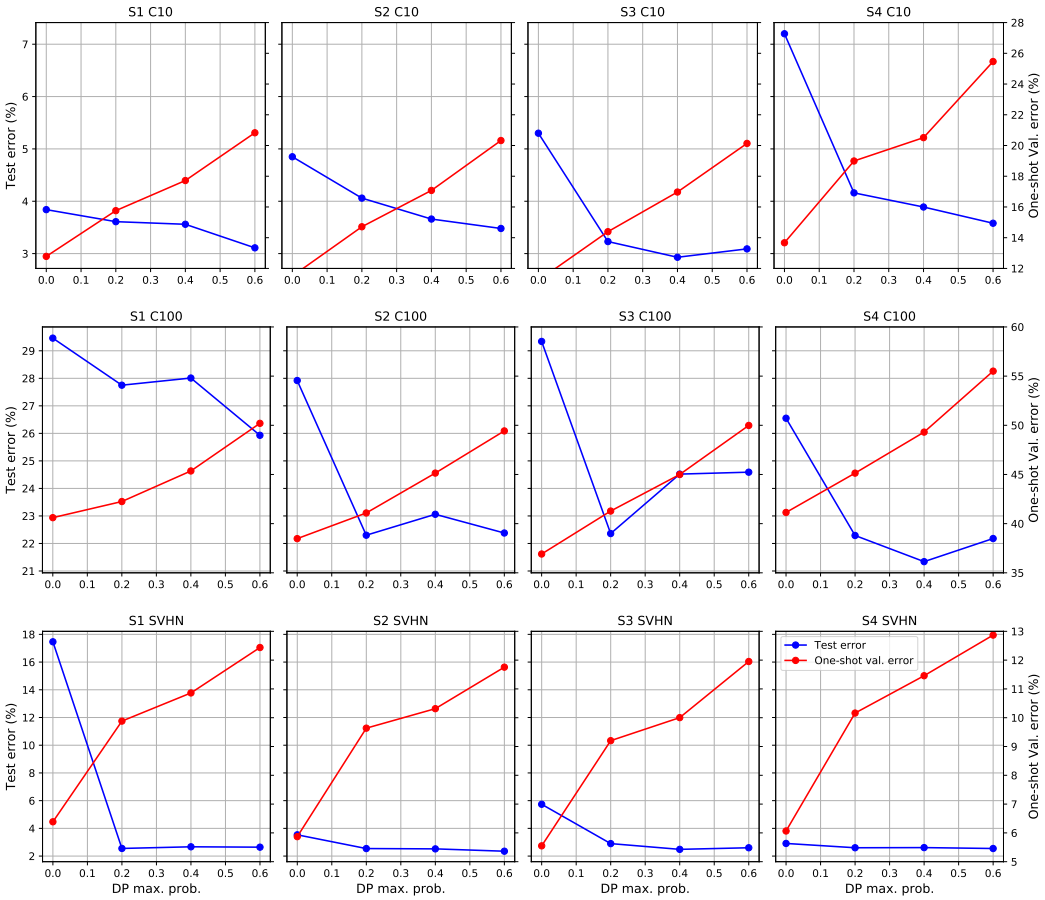


Figure 9: Test errors of architectures along with the validation error of the one-shot model for each dataset and space when scaling the ScheduledDropPath drop probability. Note that these results (blue lines) are the same as the ones in Figure 5.

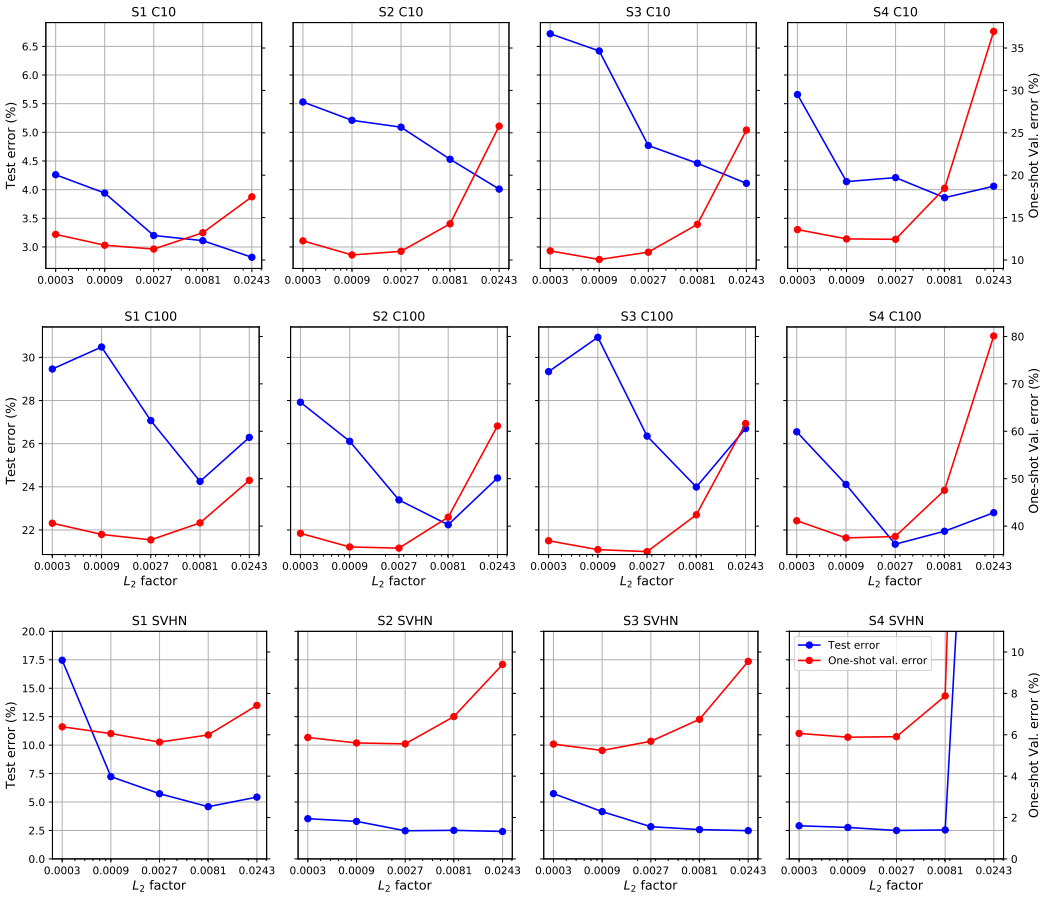


Figure 10: Test errors of architectures along with the validation error of the one-shot model for each dataset and space when scaling the L_2 factor. Note that these results (blue lines) are the same as the ones in Figure 4.

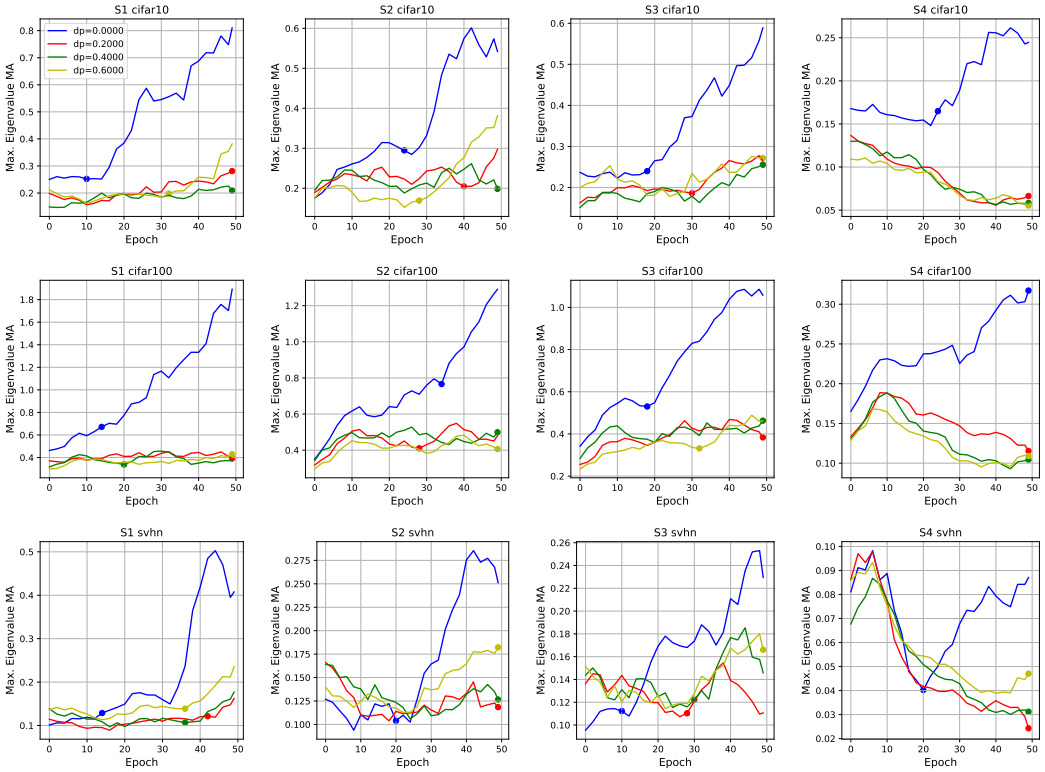


Figure 11: Local average of the dominant EV λ^{α}_{max} throughout DARTS search (for different drop path prob. values). Markers denote the early stopping point based on the criterion in Section 4.2.

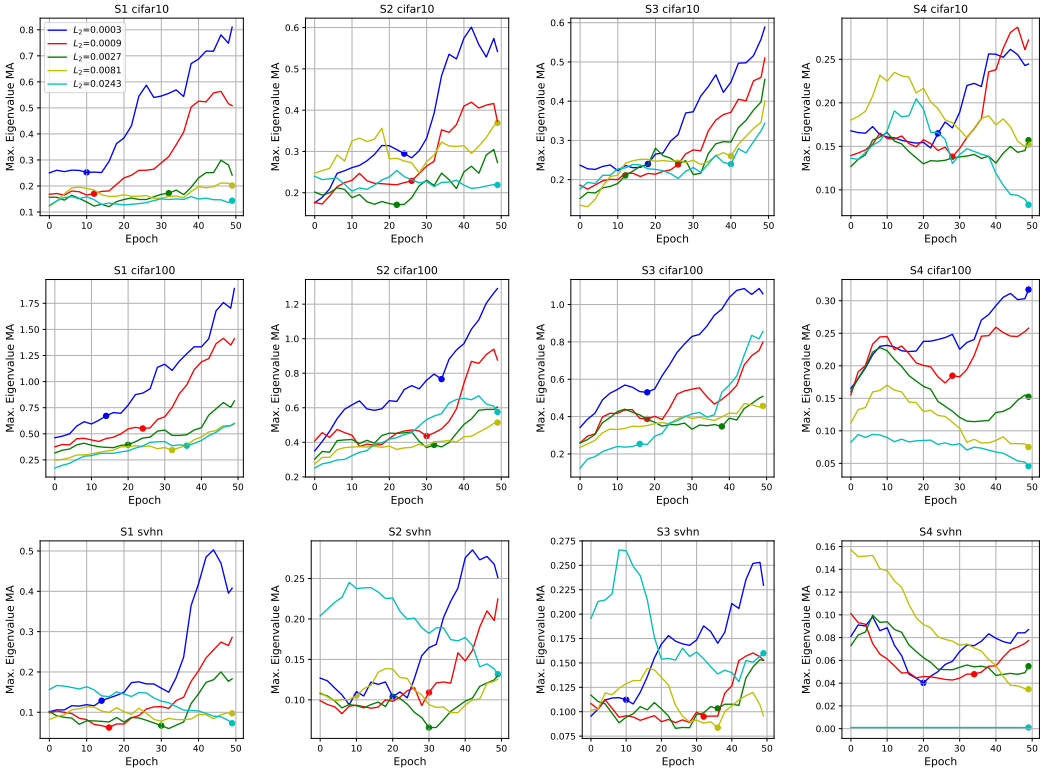


Figure 12: Effect of L_2 regularization no the EV trajectory. The figure is analogous to Figure 11.

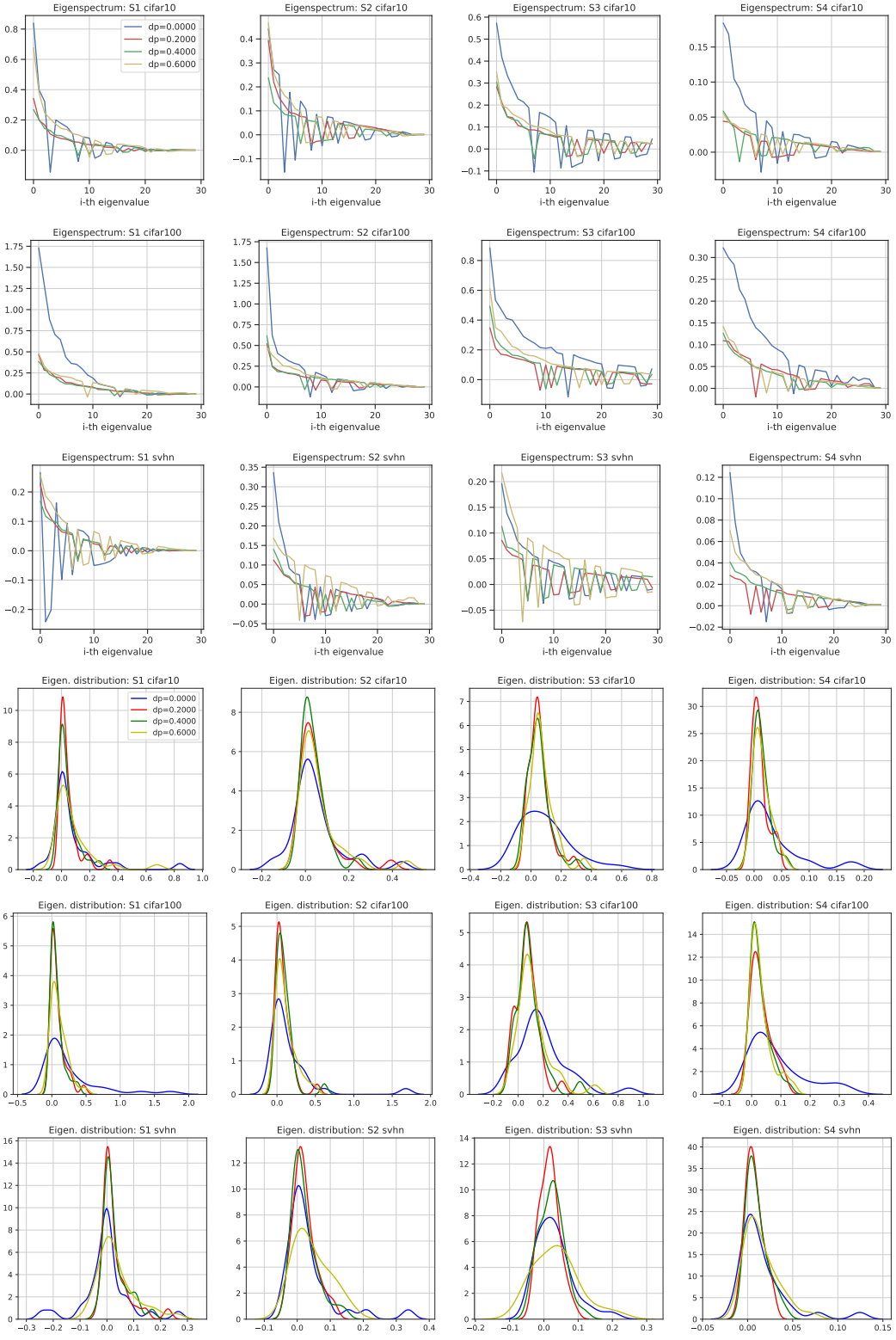


Figure 13: Effect of ScheduledDropPath and Cutout on the full eigenspectrum of the Hessian at the end of architecture search for each of the search spaces. Since most of the eigenvalues after the 30-th largest one are almost zero, we plot only the largest (based on magnitude) 30 eigenvalues here. We also provide the eigenvalue distribution for these 30 eigenvalues. Notice that not only the dominant eigenvalue is larger when $dp = 0$ but in general also the others.

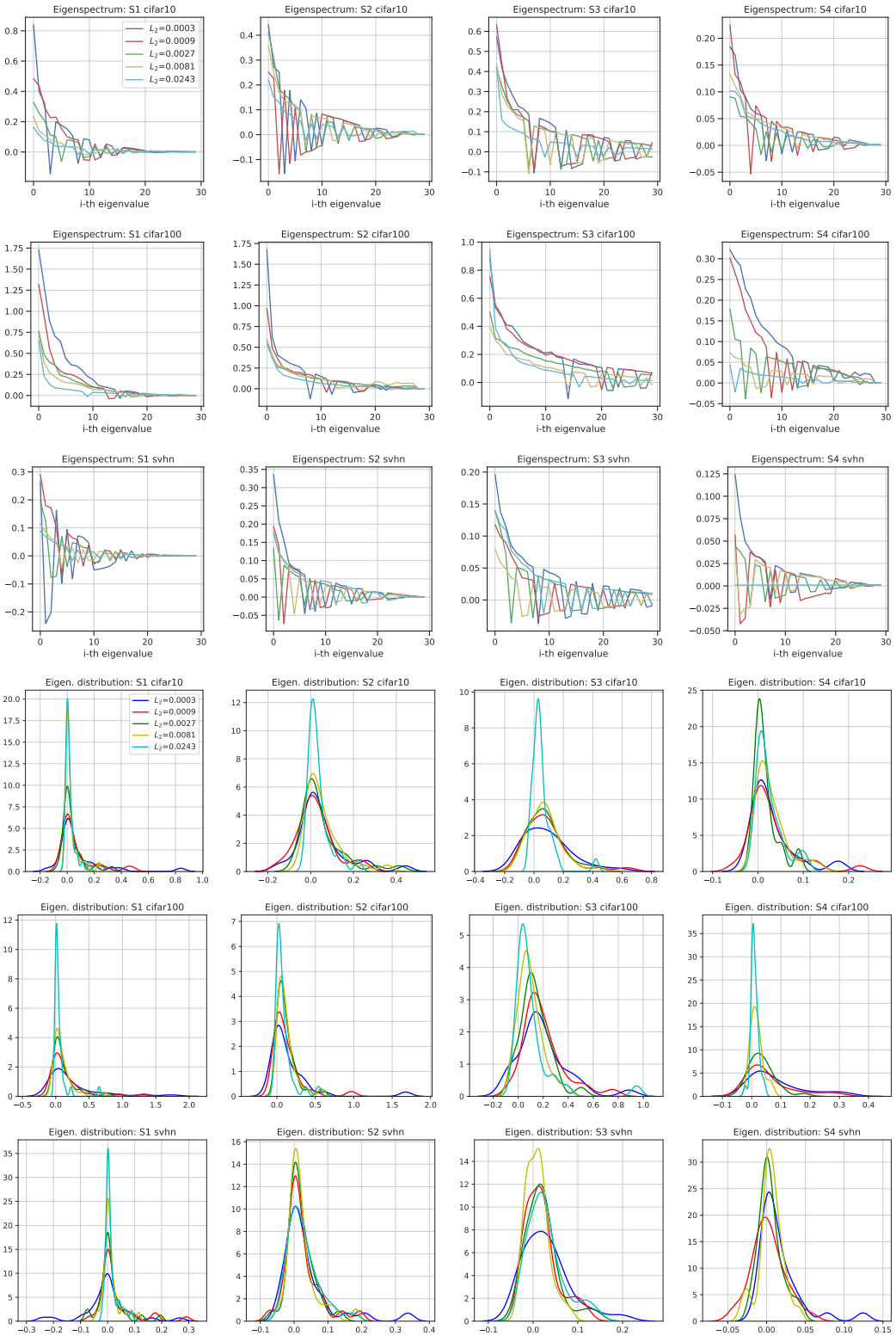


Figure 14: Effect of L_2 regularization on the full eigenspectrum of the Hessian at the end of architecture search for each of the search spaces. Since most of the eigenvalues after the 30-th largest one are almost zero, we plot only the largest (based on magnitude) 30 eigenvalues here. We also provide the eigenvalue distribution for these 30 eigenvalues. Notice that not only the dominant eigenvalue is larger when $L_2 = 3 \cdot 10^{-4}$ but in general also the others.

C.2 A CLOSER LOOK AT THE EIGENVALUES

Over the course of all experiments from the paper, we tracked the largest eigenvalue across all configuration and datasets to see how they evolve during the search. Figures 11 and 12 shows the results across all the settings for image classification. It can be clearly seen that increasing the inner objective regularization, both in terms of L_2 or data augmentation, helps controlling the largest eigenvalue and keeping it to a small value, which again helps explaining why the architectures found with stronger regularization generalize better. The markers on each line highlight the epochs where DARTS is early stopped. As one can see from Figure 15, there is indeed some correlation between the average dominant eigenvalue throughout the search and the test performance of the found architectures by DARTS.

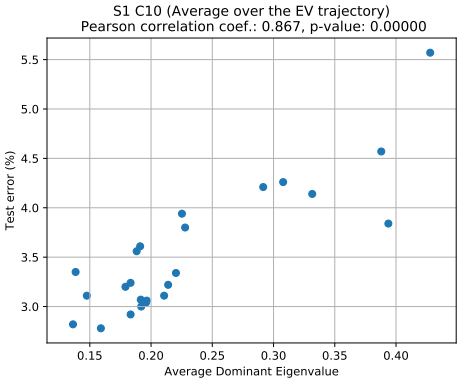


Figure 15: Correlation between the average (across search epochs) dominant eigenvalue of $\nabla^2_\alpha \mathcal{L}_{valid}$ and the test error of architectures.

Figures 13 and 14 (top 3 rows) show the full spectrum (sorted based on eigenvalue absolute values) at the end of search, whilst bottom 3 rows plot the distribution of eigenvalues in the eigenspectrum. As one can see, not only the dominant eigenvalue is larger compared to the cases when the regularization is stronger and the generalization of architectures is better, but also the other eigenvalues in the spectrum have larger absolute value, indicating a sharper objective landscape towards many dimensions. Furthermore, from the distribution plots note the presence of more negative eigenvalues whenever the architectures are degenerate (lower regularization value) indicating that DARTS gets stuck in a point with larger positive and negative curvature of the validation loss objective, associated with a more degenerate Hessian matrix.

D DISPARITY ESTIMATION

D.1 DATASETS

We use the FlyingThings3D dataset (Mayer et al., 2016) for training AutoDispNet. It consists of rendered stereo image pairs and their ground truth disparity maps. The dataset provides a training and testing split consisting of $21,818$ and 4248 samples respectively with an image resolution of 960×540. We use the Sintel dataset (Butler et al. (2012b)) for testing our networks. Sintel is another synthetic dataset from derived from an animated movie which also provides ground truth disparity maps (1064 samples) with a resolution of 1024×436.

D.2 TRAINING

We use the AutoDispNet-C architecture as described in Saikia et al. (2019). However, we use the smaller search which consists of three operations: $MaxPool3 \times 3$, $SepConv3 \times 3$, and $SkipConnect$. For training the search network, images are downsampled by a factor of two and trained for $300k$ mini-batch iterations. During search, we use SGD and ADAM to optimize the inner and outer objectives respectively. Differently from the original AutoDispNet we do not warm-start the search model weights before starting the architectural parameter updates. The extracted network is also trained for $300k$ mini-batch iterations but full resolution images are used. Here, ADAM is used for optimization and the learning rate is annealed to 0 from $1e-4$, using a cosine decay schedule.

D.3 EFFECT OF REGULARIZATION ON THE INNER OBJECTIVE

To study the effect of regularization on the inner objective for AutoDispNet-C we use experiment with two types of regularization: data augmentation and of $L2$ regularization on network weights.

Data augmentation. Inspite of fairly large number of training samples in FlyingThings3D, data augmentation is crucial for good generalization performance. Disparity estimation networks employ spatial transformations such as translation, cropping, shearing and scaling. Additionally, appearance transformations such as additive Gaussian noise, changes in brightness, contrast, gamma and color are also applied. Parameters for such transformations are sampled from a uniform or Gaussian distribution (parameterized by a mean and variance). In our experiments, we vary the data augmentation strength by multiplying the variance of these parameter distributions by a fixed factor, which we dub the *augmentation scaling factor*. The extracted networks are evaluated with the same augmentation parameters. The results of increasing the augmentation strength of the inner objective can be seen in Table 3. We observe that as augmentation strength increases DARTS finds networks with more number of parameters and better test performance. The best test performance is obtained for the network with maximum augmentation for the inner objective. At the same time the one-shot validation error increases when scaling up the augmentation factor, which again enforces the argument that the overfitting of architectural parameters is reduced by this implicit regularizer.
L2 regularization. We study the effect of increasing regularization strength on the weights of the network. The results are shown in Table 4. Also in this case best test performance is obtained with the maximum regularization strength.

E Results on Penn Treebank

Here we investigate the effect of more L_2 regularization on the inner objective for searching recurrent cells on Penn Treebank (PTB). We again used a reduced search space with only *ReLU* and *identity* mapping as possible operations. The rest of the settings is the same as in (Liu et al., 2019).

We run DARTS search four independent times with different random seeds, each with four L_2 regularization factors, namely 5×10^{-7} (DARTS default), 15×10^{-7}, 45×10^{-7} and 135×10^{-7}. Figure 16 shows the test perplexity of the architectures found by DARTS with the aforementioned L_2 regularization values. As we can see, a stronger regularization factor on the inner objective makes the search procedure more robust. The median perplexity of the discovered architectures gets better as we increase the L_2 factor from 5×10^{-7} to 45×10^{-7}, while the one-shot validation mean perplexity increases. This observation is similar to the ones on image classification shown in Figure 8, showing again that properly regularizing the inner objective helps reduce overfitting the architectural parameters.

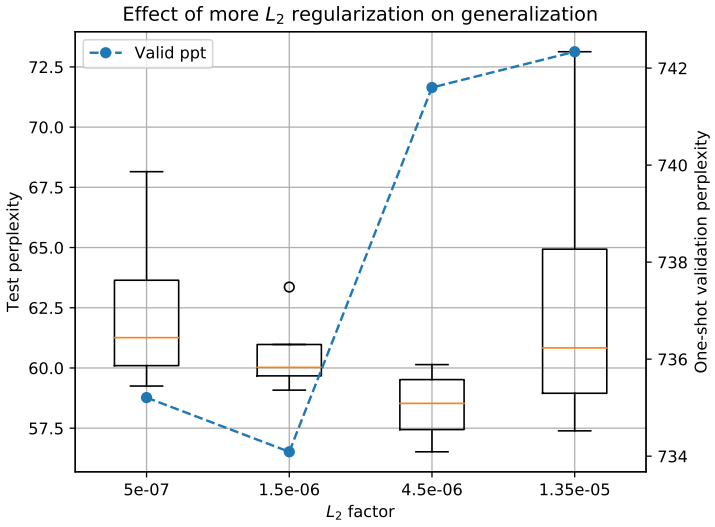
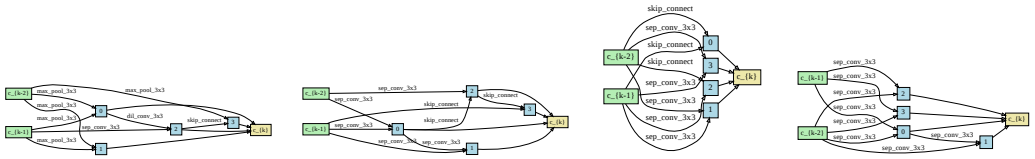


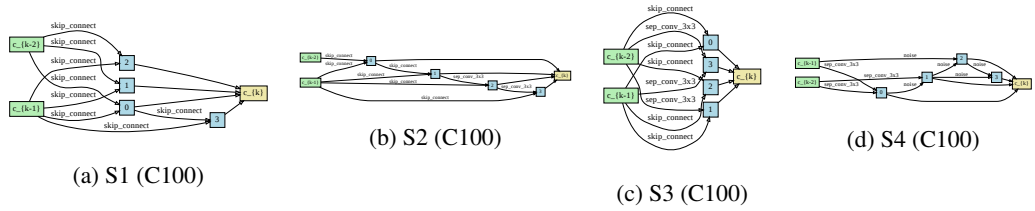
Figure 16: Performance of recurrent cells found with different L_2 regularization factors on the inner objective on PTB. We run DARTS 4 independent times with different random seeds, train each of them from scratch with the evaluation settings for 1600 epochs and report the median test perplexity. The blue dashed line denotes the validation perplexity of the search model.

F DISCOVERED CELLS ON SEARCH SPACES S1-S4 FROM SECTION 3 ON OTHER DATASETS

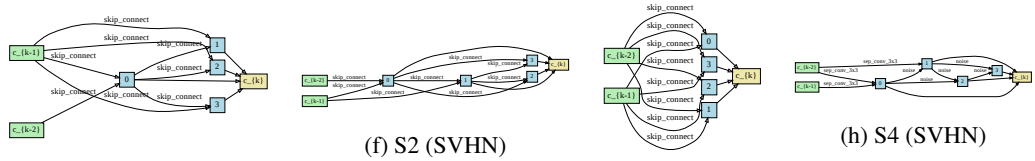


(a) S1 (b) S2 (c) S3 (d) S4

Figure 17: Reduction cells found by DARTS when ran on CIFAR-10 with its default hyperparameters on spaces S1-S4. These cells correspond with the normal ones in Figure 1.

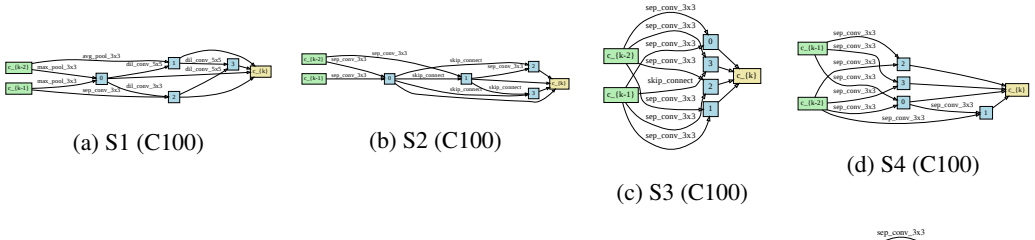


(a) S1 (C100) (b) S2 (C100) (c) S3 (C100) (d) S4 (C100)

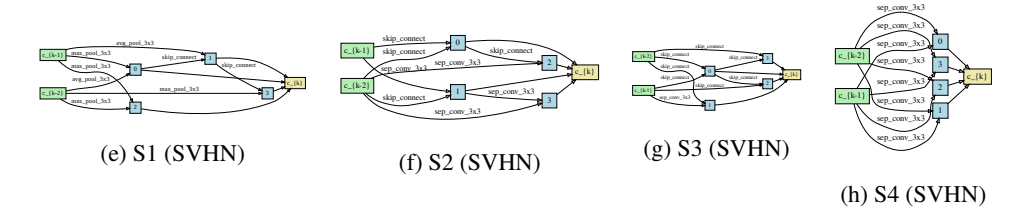


(e) S1 (SVHN) (f) S2 (SVHN) (g) S3 (SVHN) (h) S4 (SVHN)

Figure 18: Normal cells found by DARTS on CIFAR-100 and SVHN when ran with its default hyperparameters on spaces S1-S4. Notice the dominance of parameter-less operations such as *skip connection* and *pooling* ops.

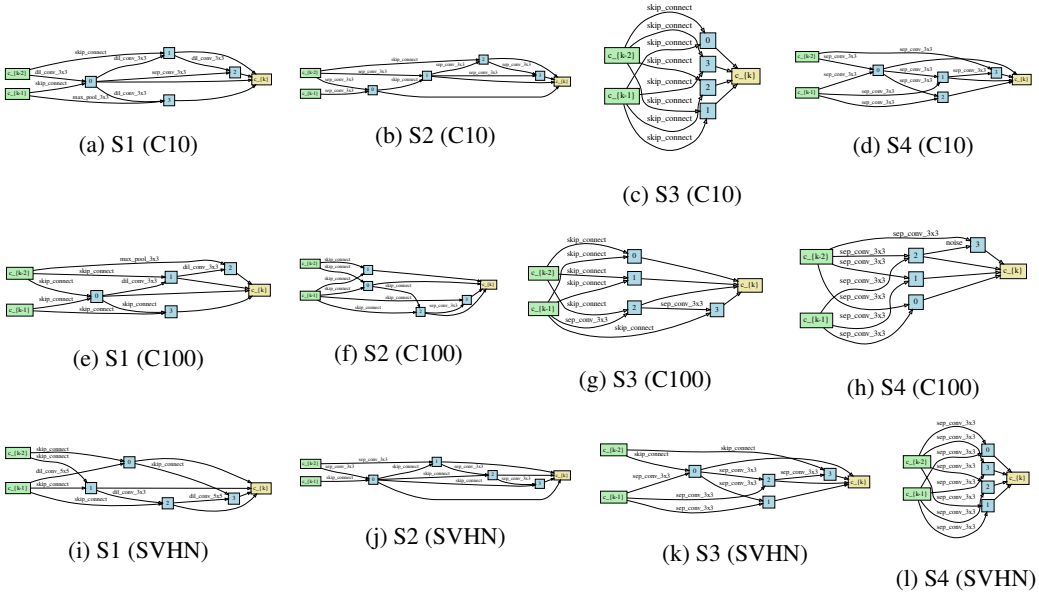


(a) S1 (C100) (b) S2 (C100) (c) S3 (C100) (d) S4 (C100)



(e) S1 (SVHN) (f) S2 (SVHN) (g) S3 (SVHN) (h) S4 (SVHN)

Figure 19: Reduction cells found by DARTS on CIFAR-100 and SVHN when ran with its default hyperparameters on spaces S1-S4.

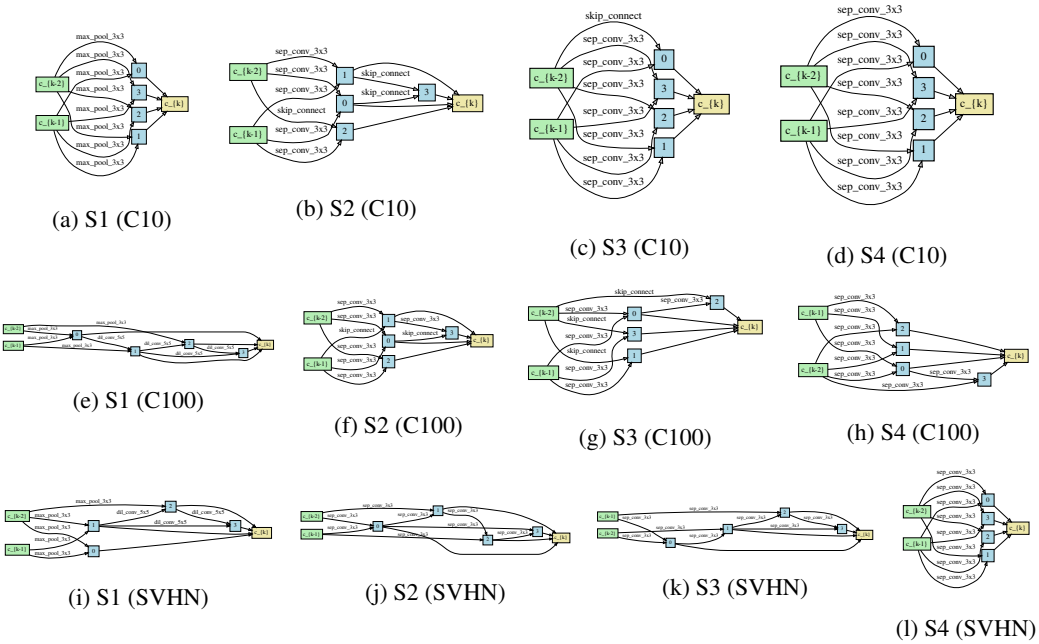


(a) S1 (C10) (b) S2 (C10) (c) S3 (C10) (d) S4 (C10)

(e) S1 (C100) (f) S2 (C100) (g) S3 (C100) (h) S4 (C100)

(i) S1 (SVHN) (j) S2 (SVHN) (k) S3 (SVHN) (l) S4 (SVHN)

Figure 20: Normal cells found by DARTS-ES when ran with DARTS default hyperparameters on spaces S1-S4.



(a) S1 (C10) (b) S2 (C10) (c) S3 (C10) (d) S4 (C10)

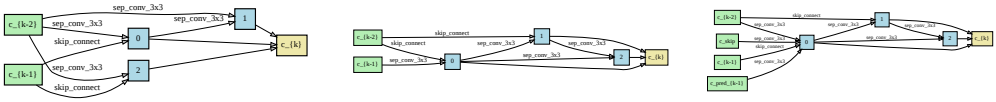
(e) S1 (C100) (f) S2 (C100) (g) S3 (C100) (h) S4 (C100)

(i) S1 (SVHN) (j) S2 (SVHN) (k) S3 (SVHN) (l) S4 (SVHN)

Figure 21: Reduction cells found by DARTS-ES when ran with DARTS default hyperparameters on spaces S1-S4.

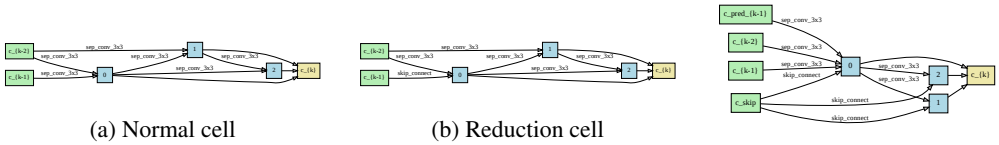
(a) Normal cell (b) Reduction cell (c) Upsampling cell

Figure 22: Cells found by AutoDispNet when ran on S6-d. These cells correspond to the results on the first row of Table 3.



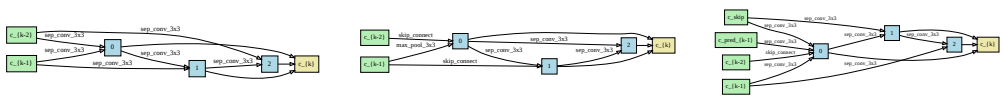
(a) Normal cell (b) Reduction cell (c) Upsampling cell

Figure 23: Cells found by AutoDispNet when ran on S6-d. These cells correspond to the results on the last row of Table 3.



(a) Normal cell (b) Reduction cell (c) Upsampling cell

Figure 24: Cells found by AutoDispNet when ran on S6-d. These cells correspond to the results on the first row of Table 4.



(a) Normal cell (b) Reduction cell (c) Upsampling cell

Figure 25: Cells found by AutoDispNet when ran on S6-d. These cells correspond to the results on the last row of Table 4.