Interpreting Operation Selection in Differentiable Architecture Search: A Perspective from Influence-Directed Explanations

PosterID:

Aalborg University, Denmark.
RIKEN AIP, Japan.
East China Normal University, China
Miao Zhang, Wei Huang, Bin Yang

AALBORG UNIVERSITET

RIKEN AIP Center for Advanced Intelligence Project

EAST CHINA NORMAL UNIVERSITY 华东师范大学

NEURAL INFORMATION PROCESSING SYSTEMS

I. Introduction

1. Background:

- DARTS leverages continuous relaxation to convert intractable operation selection problem into a magnitude optimization problem with a bi-level formulation:

$$\min_{\alpha} \; \mathcal{L}_{val}(w^*(\alpha),\alpha) \qquad \text{s.t.} \quad w^*(\alpha)=\operatorname{argmin}_w \; \mathcal{L}_{train}(w,\alpha) \tag{1}$$

- The discretization of DARTS: DARTS considers heuristic methods to derive the final architecture, usually to select the operations with the highest magnitudes, $\hat{\alpha}=\boldsymbol{argmax}(\alpha)$.

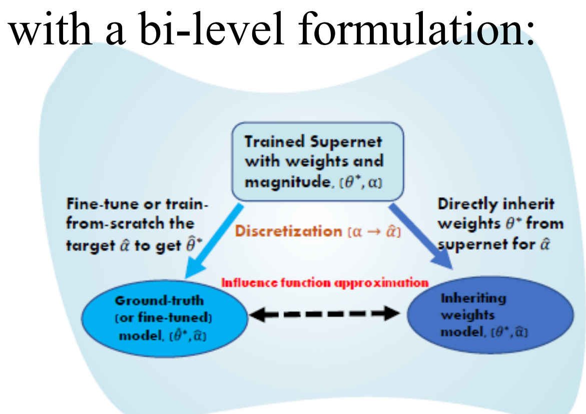


Fig.1 Pictorial depiction of discretization in DARTS.

- Influence functions: is a classic technique from robust statistics that reveals how model parameters change as we upweight or perturb a specific training sample, which has been applied in explaining many modern machine learning applications.

2. Contributions:

- **Reformulate** the operation selection in DARTS by approximating its influence on the supernet with Taylor expansions, interpreting how the validation performance changes when selecting different operations without any additional fine-tuning.
- **Theoretically reveal** the operation strength is not only related to the magnitude but also the second-order information, and accordingly derive a fundamentally new criterion to measure the operation sensitivity, called ***Influential Magnitude***.

II. Interpret Operation Selection with Influence Functions

Rather than deleting a single data point that only brings small changes on the model parameters, we leverage the second-order approximation to reveal the supernet weights change in DARTS. With second-order Taylor expansion on $\hat{\theta}^*$ for $\mathcal{L}(\hat{\theta}^*,\hat{\alpha})$, we have:

$$\Delta\mathcal{L}=\mathcal{L}(\hat{\theta}^*,\hat{\alpha})-\mathcal{L}(\theta^*,\alpha)\approx\mathcal{L}(\theta^*,\hat{\alpha})+\Delta\theta^T\frac{\partial\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta}+1/2\Delta\theta^T\frac{\partial^2\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta\partial\theta}\Delta\theta-\mathcal{L}(\theta^*,\alpha), \tag{2}$$

and based on the implicit function theorem, we have the following theorem.

Theorem 1 *Suppose that DARTS obtains the optimized architecture parameter α with supernet weights θ^* after supernet training, α changes to $\hat{\alpha}$ when conducting architecture discretization, and the train-from-scratch validation loss for $\hat{\alpha}$ is $\mathcal{L}(\hat{\theta}^*,\hat{\alpha})$. If the third and higher derivatives of the loss function $\mathcal{L}$ at optimum is zero or sufficiently small [4], and with $\frac{\partial\mathcal{L}(\hat{\theta}^*,\hat{\alpha})}{\partial\theta}=0$, we have*

$$\Delta\mathcal{L}=\mathcal{L}(\hat{\theta}^*,\hat{\alpha})-\mathcal{L}(\theta^*,\alpha)\approx\mathcal{L}(\theta^*,\hat{\alpha})-\mathcal{L}(\theta^*,\alpha)-1/2\frac{\partial\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta}^T\frac{\partial^2\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta\partial\theta}^{-1}\frac{\partial\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta}. \tag{5}$$

With Theorem 1, we first devised a **DARTS-IF** framework for operation selection.

Algorithm 1 N Differentiable Architecture Search with Influence Functions (DARTS-IF)

1: **Input**: A pretrained supernet after bi-level training process (θ^*, α), candidate operations for each edge $\mathcal{O}$, and set of edges $\mathcal{E}$ from the supernet.
2: **output:** A discrete architecture α^*.
3: **for** $e\in\mathcal{E}$ **do**
4:     **for** $o\in\mathcal{O}$ **do**
5:         Remove candidate operation o from edge e;
6:         Calculate the predictive loss chance $\Delta\mathcal{L}_{o,e}$ based on Eq. (5), that $\Delta\mathcal{L}_{o,e}\approx\mathcal{L}(\theta^*,\hat{\alpha})-\mathcal{L}(\theta^*,\alpha)-1/2\frac{\partial\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta}^T\frac{\partial^2\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta\partial\theta}^{-1}\frac{\partial\mathcal{L}(\theta^*,\hat{\alpha})}{\partial\theta}$, as the operation strength;
7:         Restore o to $\mathcal{O}$;
8:     **end for**
9: **end for**
10: Apply *argmax* on the operation strength $\Delta\mathcal{L}$ and derive the discrete architecture α^* accordingly.

The following **Corollary1** shows a large change on α brings more error in estimation.

Corollary 1 *Based on the Assumption 1-3, we could bound the error between the approximated validation loss $\mathcal{L}(\hat{\theta}^*,\hat{\alpha})=\Delta\mathcal{L}+\mathcal{L}(\theta^*,\alpha)$ and the ground-truth $\tilde{\mathcal{L}}(\hat{\theta}^*,\hat{\alpha})$ in DARTS with $E=\left\|\mathcal{L}(\hat{\theta}^*,\hat{\alpha})-\tilde{\mathcal{L}}(\hat{\theta}^*,\hat{\alpha})\right\|\leqslant\frac{K^3}{6}\max\left|\frac{\partial\mathcal{L}^3}{\partial\theta^3}\right|$, where $K=\frac{C_L}{\lambda}\|\Delta\alpha\|+\frac{C_H C_a^2}{2\sigma_{min}^2\lambda}\|\Delta\alpha\|^2+o(\|\Delta\alpha\|^2)$.*

In this way, we only consider an infinitesimal change on α as **Theorem 2**.

Theorem 2 *Suppose that DARTS obtains the optimized architecture parameter α with supernet weights θ^* after supernet training, and we pose an infinitesimal change on α. Based on implicit function theorem and under the assumption that the third and higher derivatives of the loss function at optimum is zero or sufficiently small [4], the change of validation loss can be estimated as:*

$$\Delta\mathcal{L}=\mathcal{L}(\hat{\theta}^*,\hat{\alpha})-\mathcal{L}(\theta^*,\alpha)\approx-1/2\Delta\alpha^T\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\alpha\partial\theta}H^{-1}\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\theta\partial\alpha}*\Delta\alpha, \tag{7}$$

where $H=\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\theta\partial\theta}$ is the Hessian matrix.

With **Theorem 2**, we observe the relationship between $\Delta\mathcal{L}$ and $\Delta\alpha$, and proposed an ***Influential Magnitude*** to measure operation sensitivity for the operation selection in DARTS.

Definition 1 ***Influential Magnitude*** *($\mathcal{I_M}$): Suppose DARTS obtains the optimized magnitude α with supernet weights θ^* after supernet training, the operation sensitivity can be defined as $\mathcal{I_M}=-\boldsymbol{1}^T\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\alpha\partial\theta}H^{-1}\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\theta\partial\alpha}$.*

With **Definition 1**, we first devised a **DARTS-IF** for operation selection in DARTS.

Algorithm 2 Differentiable Architecture Search with Influence Magnitude (DARTS-IM)

1: **Input**: A pretrained supernet after bi-level training process (θ^*, α), candidate operations for each edge $\mathcal{O}$, and set of edges $\mathcal{E}$ from the supernet.
2: **output:** A discrete architecture α^*.
3: Calculate the influence magnitude $\mathcal{I_M}=-\mathbf{1}^T\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\alpha\partial\theta}H^{-1}\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\theta\partial\alpha}$ based on Definition 1;
4: Apply *argmax* on the influence magnitude $\mathcal{I_M}$ and derive the discrete architecture α^* accordingly.

III. Practical Implementation

For a large neural network, it is impractical to calculate the second-order information, e.g., the Hessian matrix ***H***, let alone the inverse of Hessian. Generally, the core challenge in calculation of Eq.(5) and Eq.(7) is the Inverse-Hessian Vector Products (IHVPs). In this paper, we consider the ***Neumann series*** and ***Sherman-Morrison formula*** to approximate the IHVPs, as shown in **Lemma1 and Lemma 2**.

Lemma 1 *With small enough γ, and assuming $\mathcal{L}$ is λ-strongly convex at optimum, $H^{-1}v$ can be formulated as: $H^{-1}v=\gamma\sum_{k=0}^{K}[I-\gamma H]^k V_0=V_0+V_1+...+V_K$, where $H=\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\theta\partial\theta}$, $V_0=\gamma v$, and $V_1=\gamma(I-\gamma H)V_0,...,V_K=\gamma(I-\gamma H)V_{K-1}$.*

Lemma 2 *When assume the empirical Fisher can approximate the Fisher matrix, and H is the Hessian matrix $\frac{\partial^2\mathcal{L}(\theta^*,\alpha)}{\partial\theta\partial\theta}$ in the optimal point, the IHVPs $H^{-1}v$ can be formulated as: $H^{-1}v=F_n^{-1}v=F_{n-1}^{-1}v-\mathrm{r}_n\frac{\mathrm{r}_n{}^T v}{N+\nabla_\theta\mathcal{L}_n{}^T\mathrm{r}_n}=\eta^{-1}v-\sum_{j=1}^{n}\mathrm{r}_j\frac{\mathrm{r}_j{}^T v}{N+\nabla_\theta\mathcal{L}_j{}^T\mathrm{r}_j}$, where $\mathcal{L}=\ell+\eta\mathcal{R}(\theta)$ that ℓ is a cross-entropy loss and $\mathcal{R}$ is the regularization term, $F_n=\frac{1}{n}\sum_{j=1}^{n}\nabla_\theta\mathcal{L}_j\nabla_\theta\mathcal{L}_j^T$ is the empirical Fisher, and $\mathrm{r}_j=F_{j-1}^{-1}\nabla_\theta\mathcal{L}_j$ which can be recurrently calculated through $\mathrm{r}_j=\eta^{-1}\nabla_\theta\mathcal{L}_j-\sum_{i=1}^{j-1}\mathrm{r}_i\frac{\mathrm{r}_i{}^T\nabla_\theta\mathcal{L}_j}{N+\nabla_\theta\mathcal{L}_i{}^T\mathrm{r}_i}$.*

IV. Results

We conducted experiments on **NAS-Bench-201, NAS-Bnech-1shot1**, and **DARTS** space.

Table 2: Best test error (%) on NAS-Bench-1shot1.

Method	Space1	Space2	Space3
DARTS	6.17±0.09	6.30±0.00	6.80±0.00
DARTS-PT	6.25±0.05	**6.28±0.06**	6.69±0.21
DARTS-IM	**6.10±0.24**	6.53±0.05	**6.20±0.00**
PC-DARTS	6.37±0.05	6.30±0.00	6.50±0.00
PC-DARTS-PT	6.14±0.08	6.37±0.12	6.38±0.09
PC-DARTS-IM	**5.90±0.24**	**6.20±0.22**	**6.10±0.08**

Table 3: Search results on DARTS space.

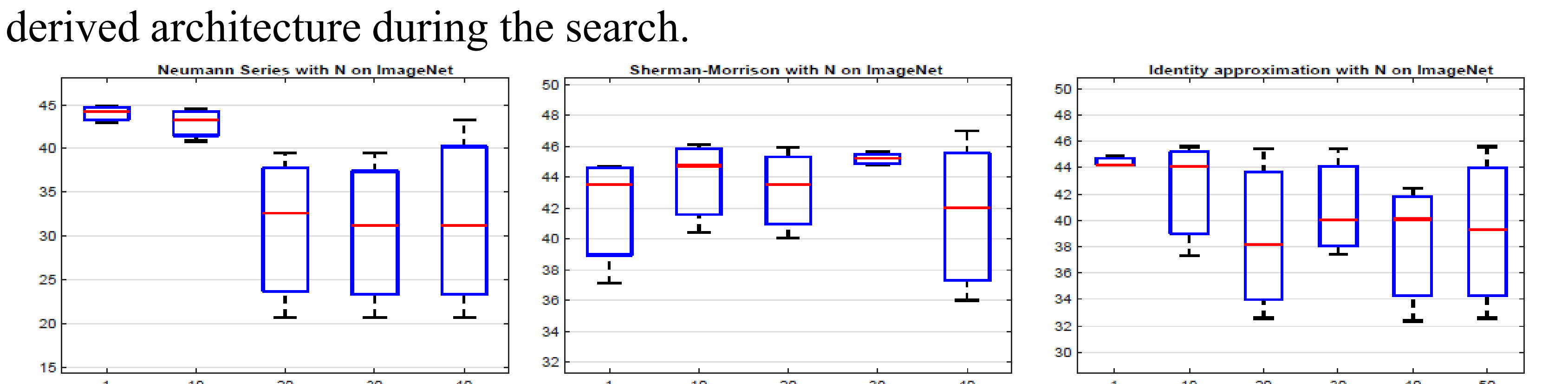
Method	CIFAR-10 Test Error (%) Single	Multi*	ImageNet Best
DARTS	2.76±0.09	3.02±0.45	26.9 / 8.7
PC-DARTS	2.57±0.07	2.92±0.26	25.1 / 7.8
DARTS-PT	**2.61±0.08**	2.89±0.31	26.1 / 8.2
DARTS-IM	**2.50±0.10**	**2.70±0.18**	**25.0 / 7.6**

* We run the architecture search with multiple times, and average the different derived architecture's test error.

Table 4: Comparison results with NAS baselines on NAS-Bench-201.

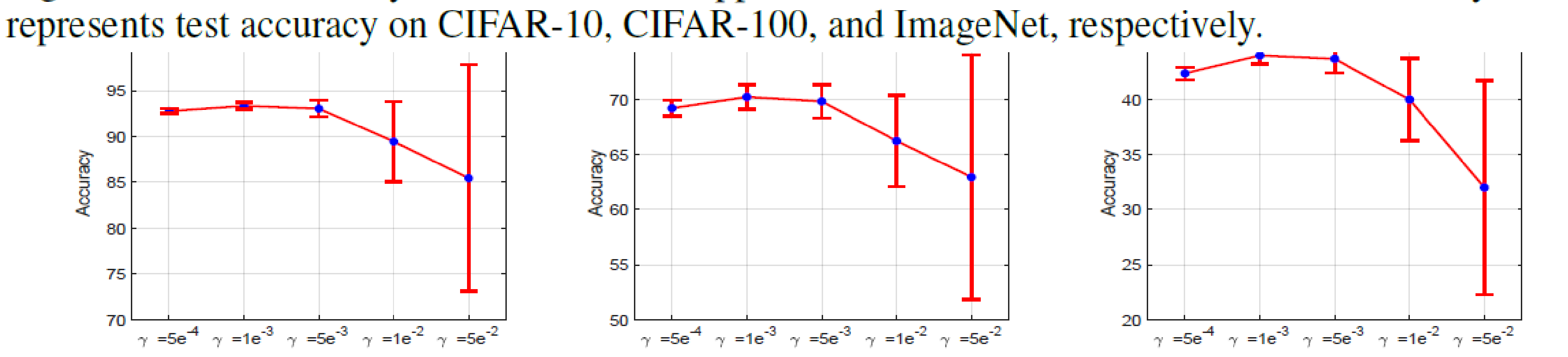
Method	CIFAR-10 Valid(%)	CIFAR-10 Test(%)	CIFAR-100 Valid(%)	CIFAR-100 Test(%)	ImageNet-16-120 Valid(%)	ImageNet-16-120 Test(%)
Random baseline	83.20±13.28	86.61±13.46	60.70±12.55	60.83±12.58	33.34±9.39	33.13±9.66
RandomNAS [26]	80.42±3.58	84.07±3.61	52.12±5.55	52.31±5.77	27.22±3.24	26.28±3.09
ENAS [33]	37.51±3.19	53.89±0.58	13.37±2.35	13.96±2.33	15.06±1.95	14.84±2.10
GDAS [10]	89.88±0.33	93.40±0.49	70.95±0.78	70.33±0.87	41.28±0.46	41.47±0.21
SETN [11]	84.04±0.28	87.64±0.00	58.86±0.06	59.05±0.24	33.06±0.02	32.52±0.21
SNAS [42]	90.10±1.04	92.77±0.84	69.69±2.39	69.35±1.98	42.84±1.79	43.16±2.64
PC-DARTS [43]	89.96±0.15	93.41±0.30	67.12±0.39	67.48±0.89	40.83±0.08	41.31±0.22
DARTS (1st) [27]	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
DARTS (2nd) [27]	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
DARTS-PT [40]	87.34±0.43	89.63±0.19	62.48±2.89	62.35±2.14	36.35±2.76	36.51±2.13
DARTS-IF	90.13±0.54	91.84±0.84	65.47±1.33	67.94±1.23	42.78±3.57	42.50±3.30
DARTS-IM	**90.92±0.34**	**93.61±0.23**	**71.21±0.55**	**71.31±0.40**	**44.70±0.74**	**44.98±0.36**
optimal	91.61	94.37	74.49	73.51	46.77	47.31

Then we analyze the batch size ***N***, hyperparameter γ and **track** performance of the derived architecture during the search.



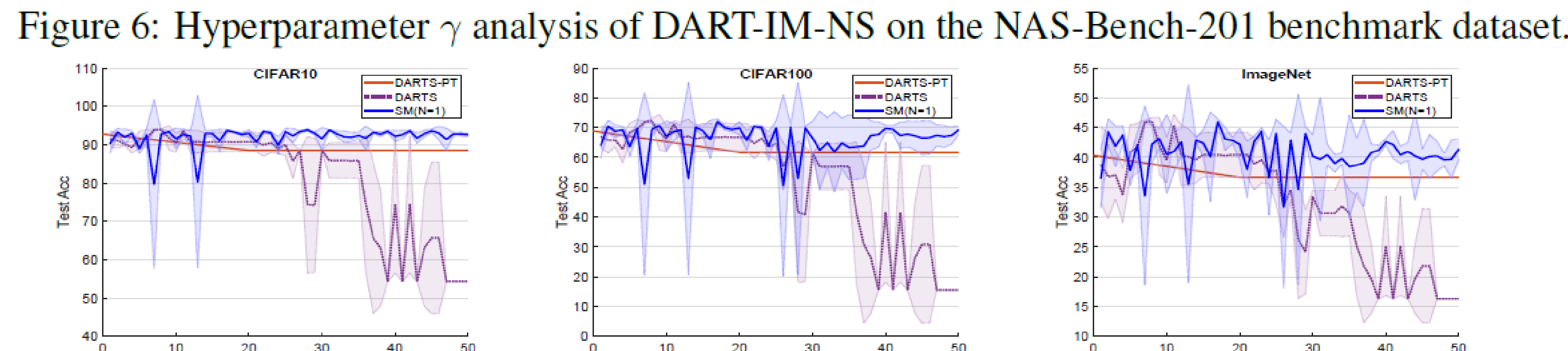
(g) Neumann series. (h) Sherman-Morrison formula. (i) Identity

Figure 5: Ablation study on N under two approximation methods, where x-axis is N and y-axis represents test accuracy on CIFAR-10, CIFAR-100, and ImageNet, respectively.

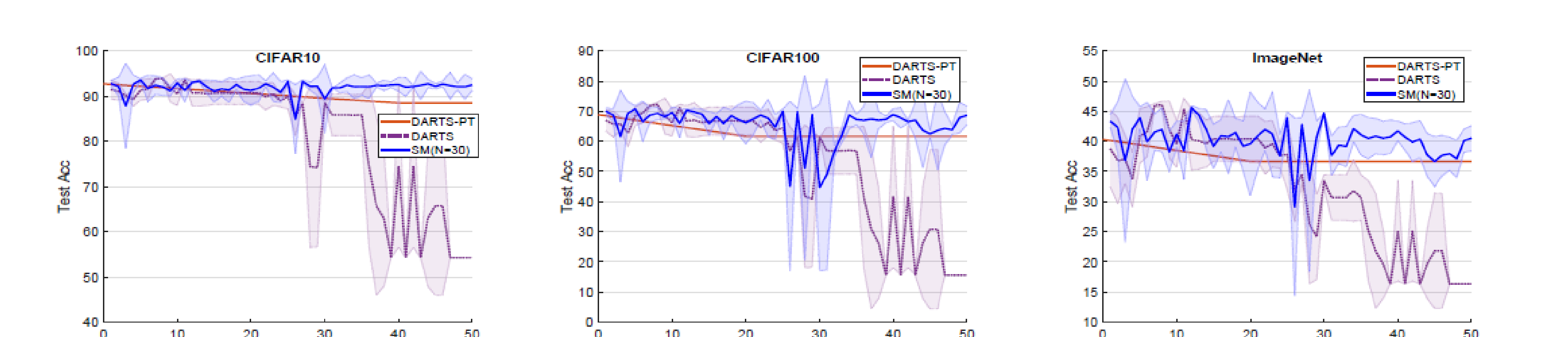


(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet-16-120

Figure 6: Hyperparameter γ analysis of DART-IM-NS on the NAS-Bench-201 benchmark dataset.



(a) Neumann series. (b) Sherman-Morrison formula. (c) Identity



(d) Neumann series. (e) Sherman-Morrison formula. (f) Identity

Figure 7: Track performance of the derived architectures during the search on NAS-Bench-201 with Sherman-Morrison formula under different N for CIFAR-10, CIFAR-100, and ImageNet, respectively.