DrNAS: Dirichlet Neural Architecture Search

Xiangning Chen*, Ruochen Wang*, Minhao Cheng*, Xiaocheng Tang, Cho-Jui Hsieh

University of California, Los Angeles, DiDI AI Lab

DrNAS - *Effective*, *Robust*, *Efficient* NAS framework

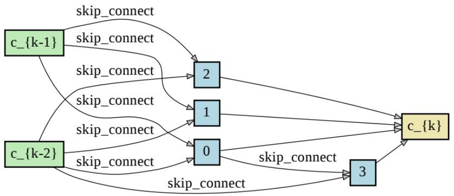
- Effective:
  - Constraint **architecture distribution** learning
  - Strike a balance between **exploration** and **exploitation**

- Robust:
  - **SOTA results** across various spaces and datasets
  - Theoretical benefit to improve **generalization**

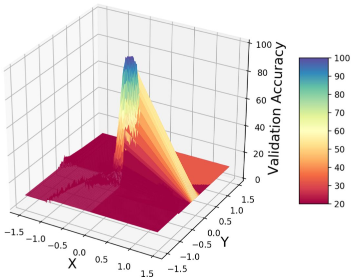
- Efficient:
  - Low GPU memory overhead & **direct search** on large-scale tasks
  - **Gradient-based** optimization

Directly learn an architecture weight doesn't work

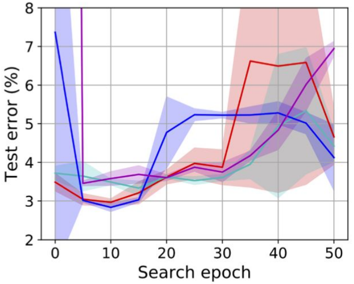
- ➢ Distorted structures
  - All operations are **skip-connection**



- ➢ Sharp landscape
  - **Overfit** the validation set
  - Blowing Hessian norm



- ➢ Significant performance degradation
  - Insufficient **exploration**



Liu et al. "DARTS: Differentiable Architecture Search." *In ICLR, 2019.*
Arber Zela. et al. "Understanding and robustifying differentiable architecture search." *In ICLR, 2020.*
Chen & Hsieh. "Stabilizing Differentiable Architecture Search via Perturbation-based Regularization." *In ICML, 2020.*

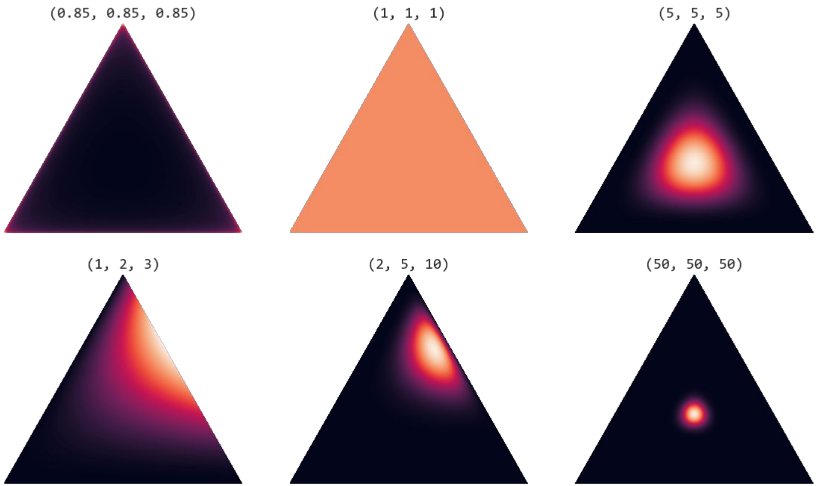
Learn an architecture distribution instead

- Distribution learning naturally encourages exploration compared with point estimation

$$\min_{\beta} E_{q(\theta|\beta)}\left[\mathcal{L}_{val}(w^*,\theta)\right] + \lambda d(\beta,\hat{\beta}) \text{ s.t. } w^* = \arg\min_{w} \mathcal{L}_{train}(w,\theta).$$

$$q(\theta|\beta) \sim Dir(\beta)$$

- Additional distance constraint
  - $\beta \ll 1$ leads to **sparse** samples with high variance (instability)
  - $\beta \gg 1$ leads to **dense** samples with low variance (insufficient exploration)
  - Add a penalty term with the anchor $\hat{\beta} = 1$



Efficient gradient-based optimization

➢ Pathwise derivative estimator
  ○ Approximate the gradient of Dirichlet samples

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

➢ Alternative updates between network weight and architecture distribution

➢ Determine the operation by the most likely one in expectation (Dirichlet mean) $\frac{\beta_o^{(i,j)}}{\sum_{o'} \beta_{o'}^{(i,j)}}$

$$o^{(i,j)} = \arg\max_{o \in \mathcal{O}} E_{q(\theta_o^{(i,j)}|\beta^{(i,j)})}\left[\theta_o^{(i,j)}\right].$$

➢ The learnt distribution can be beneficial in a post search phase (resource restrictions)

Liu et al. “DARTS: Differentiable Architecture Search.” *In ICLR, 2019.*
Jankowiak & Obermeyer. “Pathwise derivatives beyond the reparameterization trick.” *In ICML, 2018.*

Theoretical benefit of improved generalization

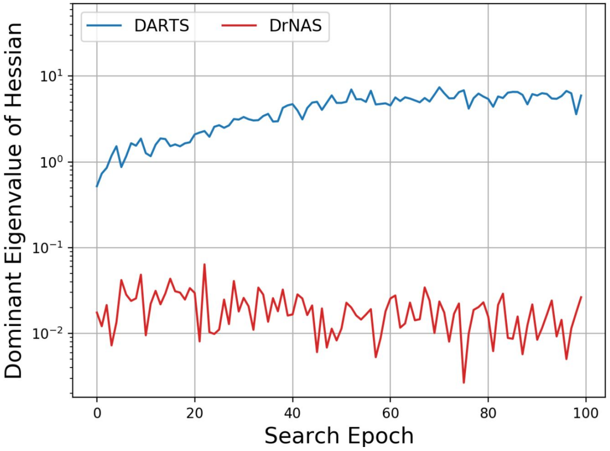
➢ We prove that minimizing the expected validation loss controls the trace norm of the Hessian matrix

Proposition 1 *Let $d(\beta, \hat{\beta}) = \|\beta - \hat{\beta}\|_2 \leq \delta$ and $\hat{\beta} = 1$ in the bi-level formulation. Let μ denote the mean under the Laplacian approximation of Dirichlet. The upper-level objective can be approximated bounded by:*

$$E_{q(\theta|\beta)}(\mathcal{L}_{val}(w, \theta)) \gtrsim \tilde{\mathcal{L}}_{val}(w^*, \mu) + C \cdot tr\big(\nabla^2_{\mu} \tilde{\mathcal{L}}_{val}(w^*, \mu)\big)$$

with:

$$\tilde{\mathcal{L}}_{val}(w^*, \mu) = \mathcal{L}_{val}(w^*, Softmax(\mu)),$$



Progressive learning scheme

- ➢ A direct search on large-scale tasks, no gap between search and evaluation
- ➢ Progressively increase the fraction of channels that are forwarded to the mixed-operation and meanwhile prunes the operation space

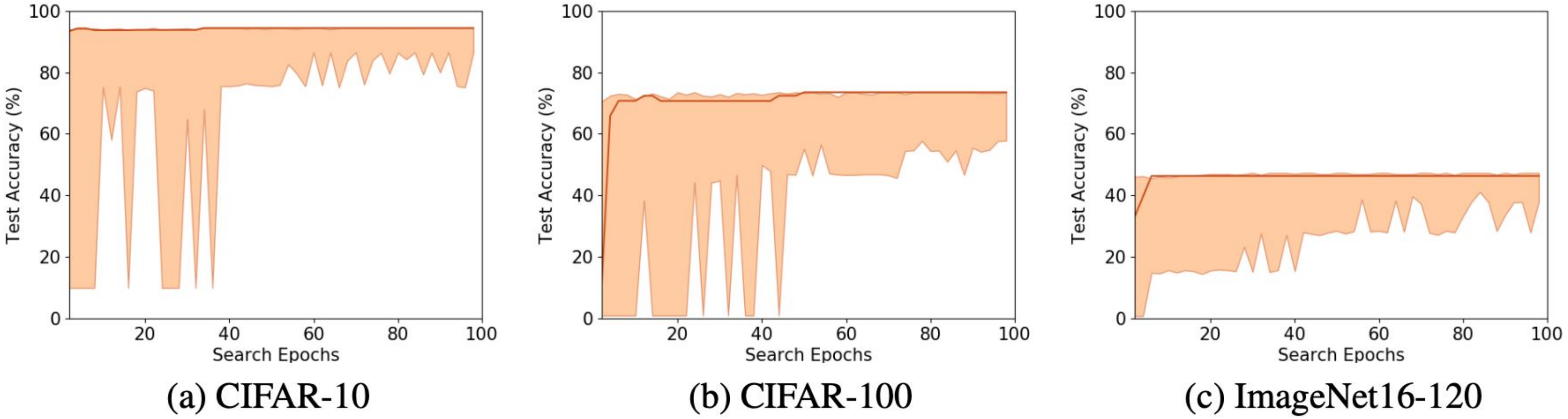
Strong Empirical Results

- ➢ On NAS-Bench-201, we achieve the best accuracy on all 3 datasets
- ➢ **Oracle** on CIFAR-100 with 0 variance

Method	**CIFAR-10**		**CIFAR-100**		**ImageNet-16-120**	
	validation	test	validation	test	validation	test
ResNet (He et al., 2016)	90.83	93.97	70.42	70.86	44.53	43.63
Random (baseline)	90.93 ± 0.36	93.70 ± 0.36	70.60 ± 1.37	70.65 ± 1.38	42.92 ± 2.00	42.96 ± 2.15
RSPS (Li & Talwalkar, 2019)	84.16 ± 1.69	87.66 ± 1.69	45.78 ± 6.33	46.60 ± 6.57	31.09 ± 5.65	30.78 ± 6.12
Reinforce (Zoph et al., 2018)	91.09 ± 0.37	93.85 ± 0.37	70.05 ± 1.67	70.17 ± 1.61	43.04 ± 2.18	43.16 ± 2.28
ENAS (Pham et al., 2018)	39.77 ± 0.00	54.30 ± 0.00	10.23 ± 0.12	10.62 ± 0.27	16.43 ± 0.00	16.32 ± 0.00
DARTS (1st) (Liu et al., 2019)	39.77 ± 0.00	54.30 ± 0.00	38.57 ± 0.00	38.97 ± 0.00	18.87 ± 0.00	18.41 ± 0.00
DARTS (2nd) (Liu et al., 2019)	39.77 ± 0.00	54.30 ± 0.00	38.57 ± 0.00	38.97 ± 0.00	18.87 ± 0.00	18.41 ± 0.00
GDAS (Dong & Yang, 2019)	90.01 ± 0.46	93.23 ± 0.23	24.05 ± 8.12	24.20 ± 8.08	40.66 ± 0.00	41.02 ± 0.00
SNAS (Xie et al., 2019)	90.10 ± 1.04	92.77 ± 0.83	69.69 ± 2.39	69.34 ± 1.98	42.84 ± 1.79	43.16 ± 2.64
DSNAS (Hu et al., 2020)	89.66 ± 0.29	93.08 ± 0.13	30.87 ± 16.40	31.01 ± 16.38	40.61 ± 0.09	41.07 ± 0.09
PC-DARTS (Xu et al., 2020)	89.96 ± 0.15	93.41 ± 0.30	67.12 ± 0.39	67.48 ± 0.89	40.83 ± 0.08	41.31 ± 0.22
DrNAS	$\mathbf{91.55 \pm 0.00}$	$\mathbf{94.36 \pm 0.00}$	$\mathbf{73.49 \pm 0.00}$	$\mathbf{73.51 \pm 0.00}$	$\mathbf{46.37 \pm 0.00}$	$\mathbf{46.34 \pm 0.00}$
optimal	91.61	94.37	73.49	73.51	46.77	47.31

Exploration vs. Exploitation

➢ Accuracy range of 100 sampled architectures vs. Dirichlet mean



➢ DrNAS learns to encourage exploration at the early stages and then gradually reduces it towards the end

On CIFAR-10 and ImageNet

Architecture	Test Error (%)	Params (M)	Search Cost (GPU days)	Search Method
DenseNet-BC (Huang et al., 2017)*	3.46	25.6	-	manual
NASNet-A (Zoph et al., 2018)	2.65	3.3	2000	RL
AmoebaNet-A (Real et al., 2019)	3.34 ± 0.06	3.2	3150	evolution
AmoebaNet-B (Real et al., 2019)	2.55 ± 0.05	2.8	3150	evolution
PNAS (Liu et al., 2018)*	3.41 ± 0.09	3.2	225	SMBO
ENAS (Pham et al., 2018)	2.89	4.6	0.5	RL
DARTS (1st) (Liu et al., 2019)	3.00 ± 0.14	3.3	0.4	gradient
DARTS (2nd) (Liu et al., 2019)	2.76 ± 0.09	3.3	1.0	gradient
SNAS (moderate) (Xie et al., 2019)	2.85 ± 0.02	2.8	1.5	gradient
GDAS (Dong & Yang, 2019)	2.93	3.4	0.3	gradient
BayesNAS (Zhou et al., 2019)	2.81 ± 0.04	3.4	0.2	gradient
ProxylessNAS (Cai et al., 2019)†	2.08	5.7	4.0	gradient
PARSEC (Casale et al., 2019)	2.81 ± 0.03	3.7	1	gradient
P-DARTS (Chen et al., 2019)	2.50	3.4	0.3	gradient
PC-DARTS (Xu et al., 2020)	2.57 ± 0.07	3.6	0.1	gradient
SDARTS-ADV (Chen & Hsieh, 2020)	2.61 ± 0.02	3.3	1.3	gradient
GAEA + PC-DARTS (Li et al., 2020)	2.50 ± 0.06	3.7	0.1	gradient
DrNAS (without progressive learning)	2.54 ± 0.03	4.0	0.4‡	gradient
DrNAS	2.46 ± 0.03	4.1	0.6‡	gradient

2.46% test error on CIFAR-10

Architecture	Test Error(%)		Params (M)	Search Cost (GPU days)	Search Method
	top-1	top-5			
Inception-v1 (Szegedy et al., 2015)	30.1	10.1	6.6	-	manual
MobileNet (Howard et al., 2017)	29.4	10.5	4.2	-	manual
ShuffleNet 2× (v1) (Zhang et al., 2018)	26.4	10.2	~ 5	-	manual
ShuffleNet 2× (v2) (Ma et al., 2018)	25.1	-	~ 5	-	manual
NASNet-A (Zoph et al., 2018)	26.0	8.4	5.3	2000	RL
AmoebaNet-C (Real et al., 2019)	24.3	7.6	6.4	3150	evolution
PNAS (Liu et al., 2018)	25.8	8.1	5.1	225	SMBO
MnasNet-92 (Tan et al., 2019)	25.2	8.0	4.4	-	RL
DARTS (2nd) (Liu et al., 2019)	26.7	8.7	4.7	1.0	gradient
SNAS (mild) (Xie et al., 2019)	27.3	9.2	4.3	1.5	gradient
GDAS (Dong & Yang, 2019)	26.0	8.5	5.3	0.3	gradient
BayesNAS (Zhou et al., 2019)	26.5	8.9	3.9	0.2	gradient
DSNAS (Hu et al., 2020)†	25.7	8.1	-	-	gradient
ProxylessNAS (GPU) (Cai et al., 2019)†	24.9	7.5	7.1	8.3	gradient
PARSEC (Casale et al., 2019)	26.0	8.4	5.6	1	gradient
P-DARTS (CIFAR-10) (Chen et al., 2019)	24.4	7.4	4.9	0.3	gradient
P-DARTS (CIFAR-100) (Chen et al., 2019)	24.7	7.5	5.1	0.3	gradient
PC-DARTS (CIFAR-10) (Xu et al., 2020)	25.1	7.8	5.3	0.1	gradient
PC-DARTS (ImageNet) (Xu et al., 2020)†	24.2	7.3	5.3	3.8	gradient
GAEA + PC-DARTS (Li et al., 2020)†	24.0	7.3	5.6	3.8	gradient
DrNAS (without progressive learning)†	24.2	7.3	5.2	3.9	gradient
DrNAS†	23.7	7.1	5.7	4.6	gradient

† The architecture is searched on ImageNet, otherwise it is searched on CIFAR-10 or CIFAR-100.

23.7% top-1 test error on ImageNet

DrNAS

Effective, *Robust*, *Efficient* NAS framework

Code: https://github.com/xiangning-chen/DrNAS