DARTS: DIFFERENTIABLE ARCHITECTURE SEARCH

Shen Yu 1500012713

2019.5.9

Motivation

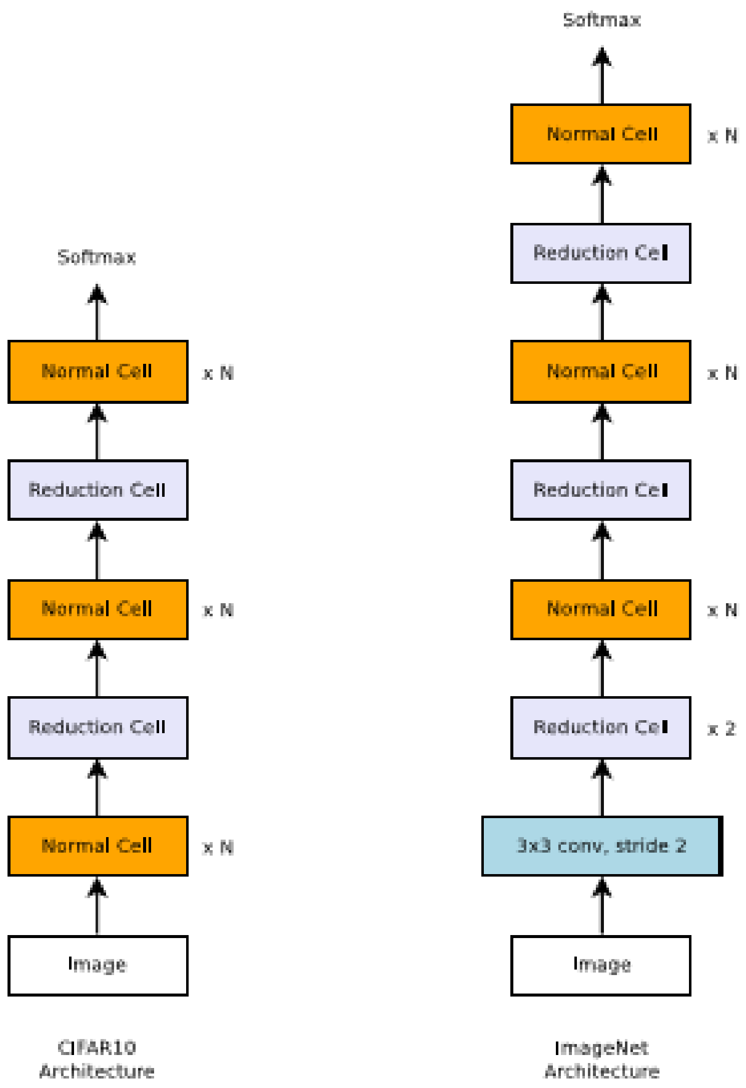
- Substantial effort of human experts on discovering architectures
- Existing NAS algorithms are demanding
  - 2000 GPU days for RL algorithms (NASNet)
  - 3150 GPU days for evolutionary algorithms (AmoebaNet)
  - 225 GPU days for SMBO (PNAS)
- Existing NAS algorithms are non-differentiable

Contributions

- Differentiable network architecture search on both convolutional and recurrent architectures
- Highly competitive results with non-differentiable search techniques
- Remarkable efficiency improvement
- Transferable architectures

Related Work

- Reinforcement learning
  - NASNet
  - ENAS
- Evolutionary algorithm
  - AmoebaNet
- SMBO
  - PNAS

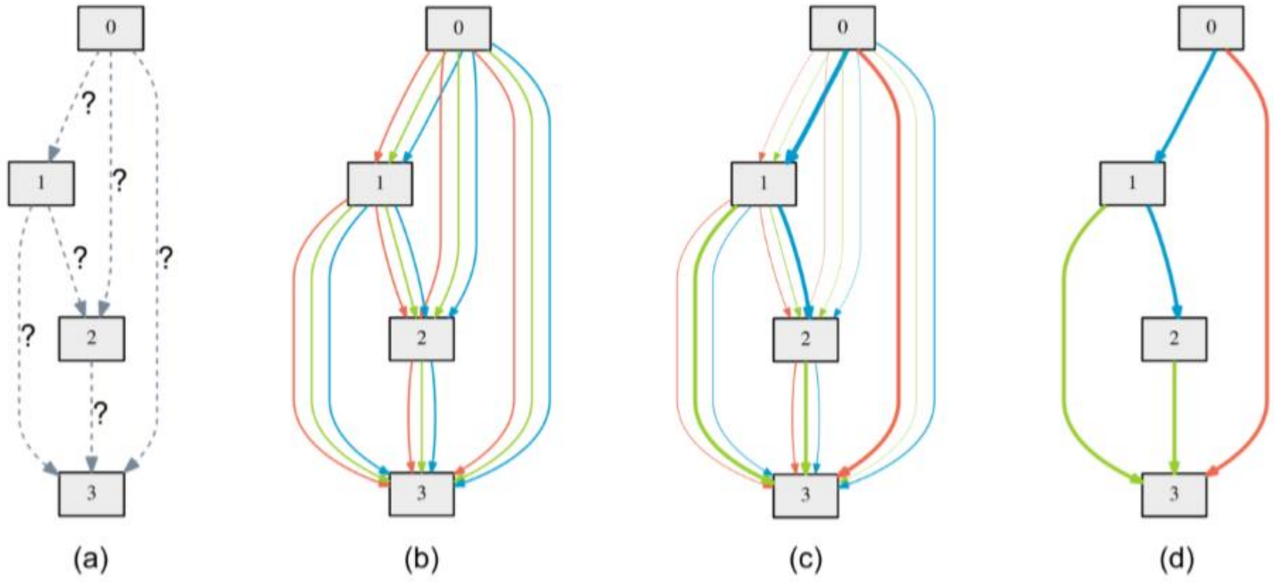


Overview

- DAG of N nodes in each cell

- Softmax over operations

$$\bar{o}^{(i,j)}(x) = \sum_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{(i,j)})} o(x)$$



Approximation

- Bilevel optimization

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

- Approximation 1: Adapt *w* using only a single training step

$$\begin{aligned} & \nabla_\alpha \mathcal{L}_{val}(w^*(\alpha), \alpha) \\ \approx & \nabla_\alpha \mathcal{L}_{val}(w - \xi \nabla_w \mathcal{L}_{train}(w, \alpha), \alpha) \end{aligned}$$

Algorithm

Algorithm 1: DARTS – Differentiable Architecture Search

Create a mixed operation $\bar{o}^{(i,j)}$ parametrized by $\alpha^{(i,j)}$ for each edge (i, j)
while *not converged* **do**
    1. Update architecture α by descending $\nabla_\alpha \mathcal{L}_{val}(w - \xi \nabla_w \mathcal{L}_{train}(w, \alpha), \alpha)$
    ($\xi = 0$ if using first-order approximation)
    2. Update weights w by descending $\nabla_w \mathcal{L}_{train}(w, \alpha)$
Derive the final architecture based on the learned α.

Approximation

- According to chain rule, the gradient is

$$\nabla_\alpha \mathcal{L}_{val}(w', \alpha) - \xi \nabla^2_{\alpha,w} \mathcal{L}_{train}(w, \alpha) \nabla_{w'} \mathcal{L}_{val}(w', \alpha)$$

$$w' = w - \xi \nabla_w \mathcal{L}_{train}(w, \alpha)$$

- Approximation 2: Avoid expensive matrix-vector product

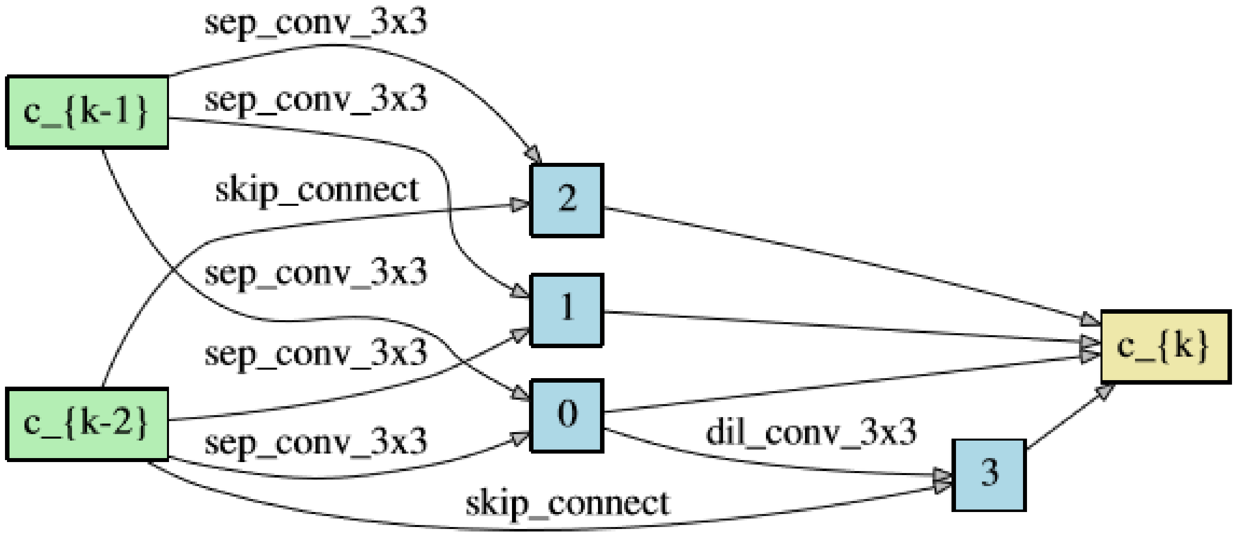
$$\nabla^2_{\alpha,w} \mathcal{L}_{train}(w, \alpha) \nabla_{w'} \mathcal{L}_{val}(w', \alpha) \approx \frac{\nabla_\alpha \mathcal{L}_{train}(w^+, \alpha) - \nabla_\alpha \mathcal{L}_{train}(w^-, \alpha)}{2\epsilon}$$

$$w^\pm = w \pm \epsilon \nabla_{w'} \mathcal{L}_{val}(w', \alpha)$$

- First-order approximation: ξ = 0, 2 forward, 2 backward
- Second-order approximation: ξ ≠ 0, 4 forward, 4 backward

Deriving architectures

- Retain the top-k strongest operations from all previous nodes
- k=2 for conv cell, k=1 for recurrent cell



Experiments

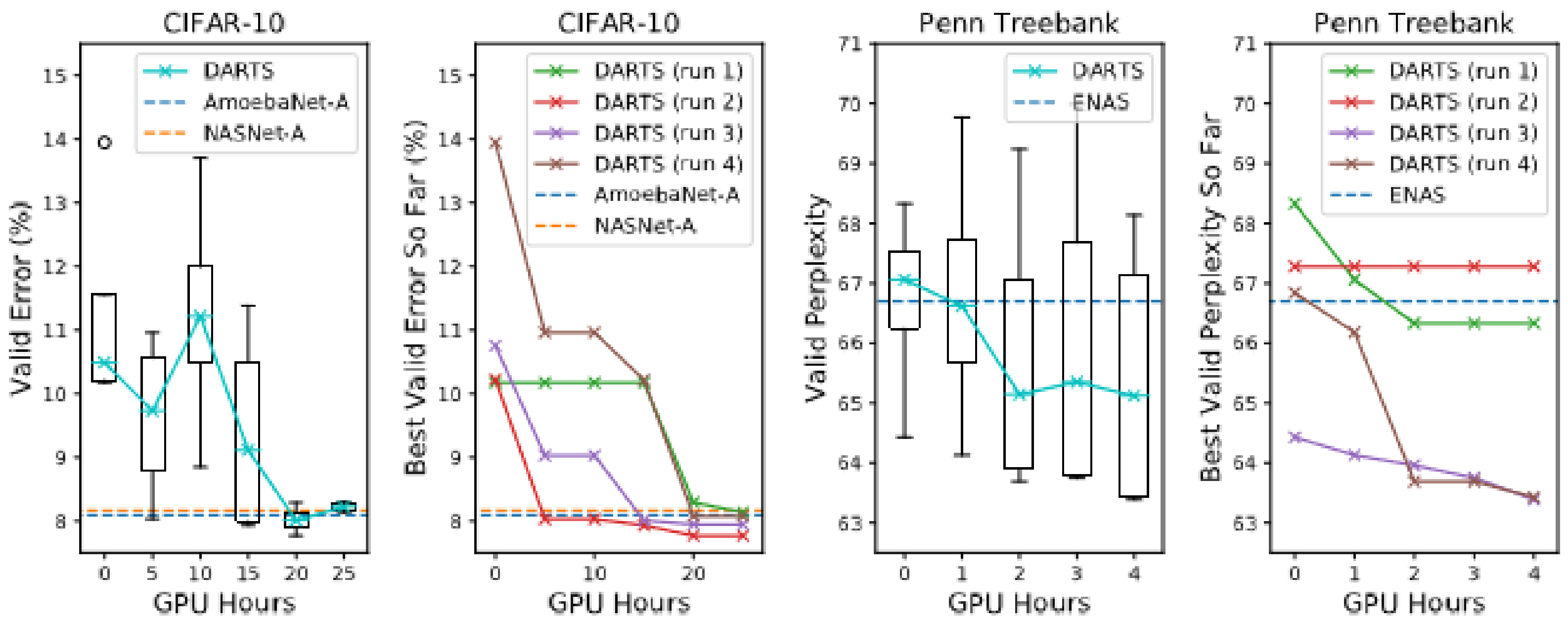
- Search space for CIFAR-10

TABLE 1. Selection of POs in current NAS methods.

POs	NAS-Net	ENAS	NAO	DARTS	PNASNet	Evolutionary search
Identity	✓	✓	✓	✓	✓	✓
1×1 **Conv**	✓	×	✓	×	×	✓
1×1 **SepConv**	✓	×	✓	×	×	×
2×2 **Conv**	×	×	✓	×	×	×
2×2 **SepConv**	×	×	✓	×	×	×
3×3 **Conv**	✓	×	×	×	×	×
3×3 **SepConv**	×	✓	✓	✓	✓	✓
3×3 **DilConv**	✓	×	×	✓	✓	×
1×3 *then* 3×1 **Conv**	✓	×	×	×	×	✓
5×5 **SepConv**	✓	✓	×	✓	✓	×
5×5 **DilConv**	×	×	×	✓	×	×
7×7 **SepConv**	✓	×	×	×	✓	×
1×7 *then* 7×1 **Conv**	✓	×	×	×	×	×
2×2 **MP**	×	×	✓	×	×	×
3×3 **MP**	✓	✓	✓	✓	✓	✓
5×5 **MP**	✓	×	×	×	×	×
7×7 **MP**	✓	×	×	×	×	×
2×2 **AP**	×	×	✓	×	×	×
3×3 **AP**	✓	✓	✓	✓	✓	✓
None	×	×	×	✓	×	×

Search progress

- Run 4 times with different random seeds



Architectures

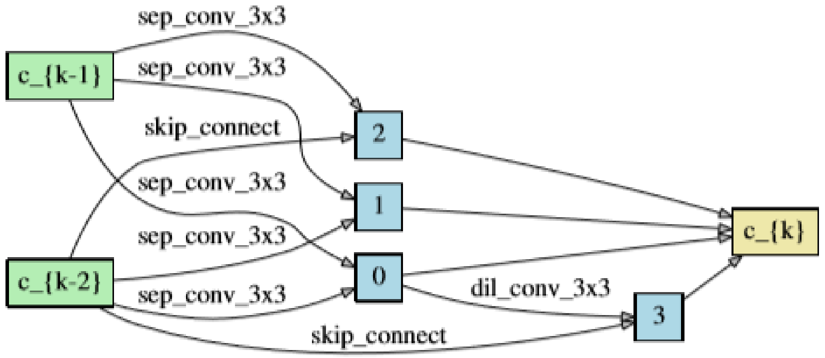


Figure 4: Normal cell learned on CIFAR-10.

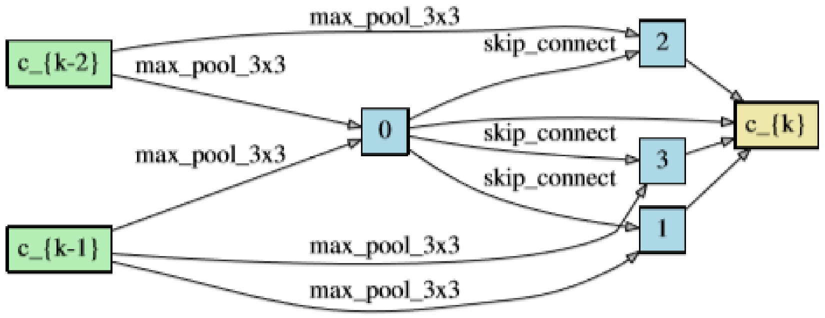


Figure 5: Reduction cell learned on CIFAR-10.

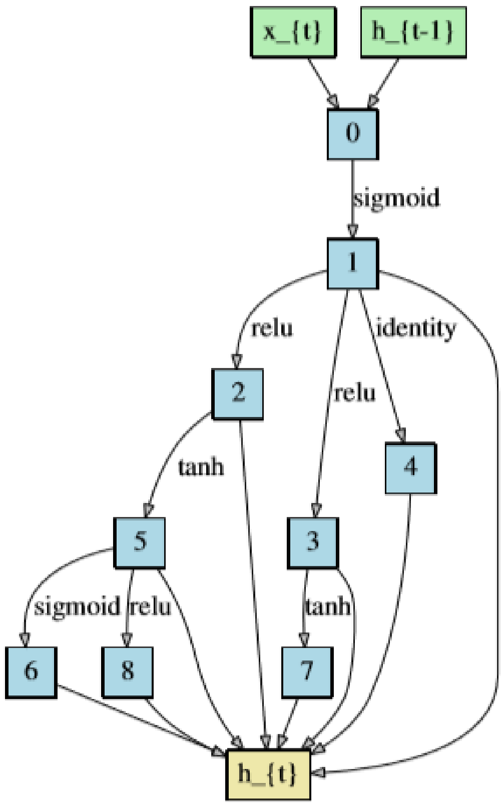


Figure 6: Recurrent cell learned on PTB.

Performance

Architecture	Test Error (%)	Params (M)	Search Cost (GPU days)	#ops	Search Method
DenseNet-BC (Huang et al., 2017)	3.46	25.6	–	–	manual
NASNet-A + cutout (Zoph et al., 2018)	2.65	3.3	2000	13	RL
NASNet-A + cutout (Zoph et al., 2018)†	2.83	3.1	2000	13	RL
BlockQNN (Zhong et al., 2018)	3.54	39.8	96	8	RL
AmoebaNet-A (Real et al., 2018)	3.34 ± 0.06	3.2	3150	19	evolution
AmoebaNet-A + cutout (Real et al., 2018)†	3.12	3.1	3150	19	evolution
AmoebaNet-B + cutout (Real et al., 2018)	2.55 ± 0.05	2.8	3150	19	evolution
Hierarchical evolution (Liu et al., 2018b)	3.75 ± 0.12	15.7	300	6	evolution
PNAS (Liu et al., 2018a)	3.41 ± 0.09	3.2	225	8	SMBO
ENAS + cutout (Pham et al., 2018b)	2.89	4.6	0.5	6	RL
ENAS + cutout (Pham et al., 2018b)*	2.91	4.2	4	6	RL
Random search baseline‡ + cutout	3.29 ± 0.15	3.2	4	7	random
DARTS (first order) + cutout	3.00 ± 0.14	3.3	1.5	7	gradient-based
DARTS (second order) + cutout	2.76 ± 0.09	3.3	4	7	gradient-based

* Obtained by repeating ENAS for 8 times using the code publicly released by the authors. The cell for final evaluation is chosen according to the same selection protocol as for DARTS.

† Obtained by training the corresponding architectures using our setup.

‡ Best architecture among 24 samples according to the validation error after 100 training epochs.

Performance

Architecture	Perplexity valid	Perplexity test	Params (M)	Search Cost (GPU days)	#ops	Search Method
Variational RHN (Zilly et al., 2016)	67.9	65.4	23	–	–	manual
LSTM (Merity et al., 2018)	60.7	58.8	24	–	–	manual
LSTM + skip connections (Melis et al., 2018)	60.9	58.3	24	–	–	manual
LSTM + 15 softmax experts (Yang et al., 2018)	58.1	56.0	22	–	–	manual
NAS (Zoph & Le, 2017)	–	64.0	25	1e4 CPU days	4	RL
ENAS (Pham et al., 2018b)*	68.3	63.1	24	0.5	4	RL
ENAS (Pham et al., 2018b)†	60.8	58.6	24	0.5	4	RL
Random search baseline‡	61.8	59.4	23	2	4	random
DARTS (first order)	60.2	57.6	23	0.5	4	gradient-based
DARTS (second order)	58.1	55.7	23	1	4	gradient-based

* Obtained using the code (Pham et al., 2018a) publicly released by the authors.
† Obtained by training the corresponding architecture using our setup.
‡ Best architecture among 8 samples according to the validation perplexity after 300 training epochs.

Performance

Architecture	Test Error (%) top-1	Test Error (%) top-5	Params (M)	$+\times$ (M)	Search Cost (GPU days)	Search Method
Inception-v1 (Szegedy et al., 2015)	30.2	10.1	6.6	1448	–	manual
MobileNet (Howard et al., 2017)	29.4	10.5	4.2	569	–	manual
ShuffleNet $2\times$ ($g = 3$) (Zhang et al., 2017)	26.3	–	$\sim$5	524	–	manual
NASNet-A (Zoph et al., 2018)	26.0	8.4	5.3	564	2000	RL
NASNet-B (Zoph et al., 2018)	27.2	8.7	5.3	488	2000	RL
NASNet-C (Zoph et al., 2018)	27.5	9.0	4.9	558	2000	RL
AmoebaNet-A (Real et al., 2018)	25.5	8.0	5.1	555	3150	evolution
AmoebaNet-B (Real et al., 2018)	26.0	8.5	5.3	555	3150	evolution
AmoebaNet-C (Real et al., 2018)	24.3	7.6	6.4	570	3150	evolution
PNAS (Liu et al., 2018a)	25.8	8.1	5.1	588	$\sim$225	SMBO
DARTS (searched on CIFAR-10)	26.7	8.7	4.7	574	4	gradient-based

Q&A