# Theory-Inspired Path-Regularized Differential Network Architecture Search

**Pan Zhou**, Caiming Xiong, Richard Socher, Steven C.H. Hoi

Salesforce Research

pzhou@salesforce.com

Oct 2020

# Outline

**Background:** what is network architecture search (NAS)

**Theoretical analysis:** why DARTS often select so many skip connections

**Solution:** group-structured sparse gate and path-depth-wise regularization

**Experiments:** higher efficiency and classification accuracy

**Conclusion**

# Outline

**Background:** what is network architecture search (NAS)

**Theoretical analysis:** why DARTS often select so many skip connections

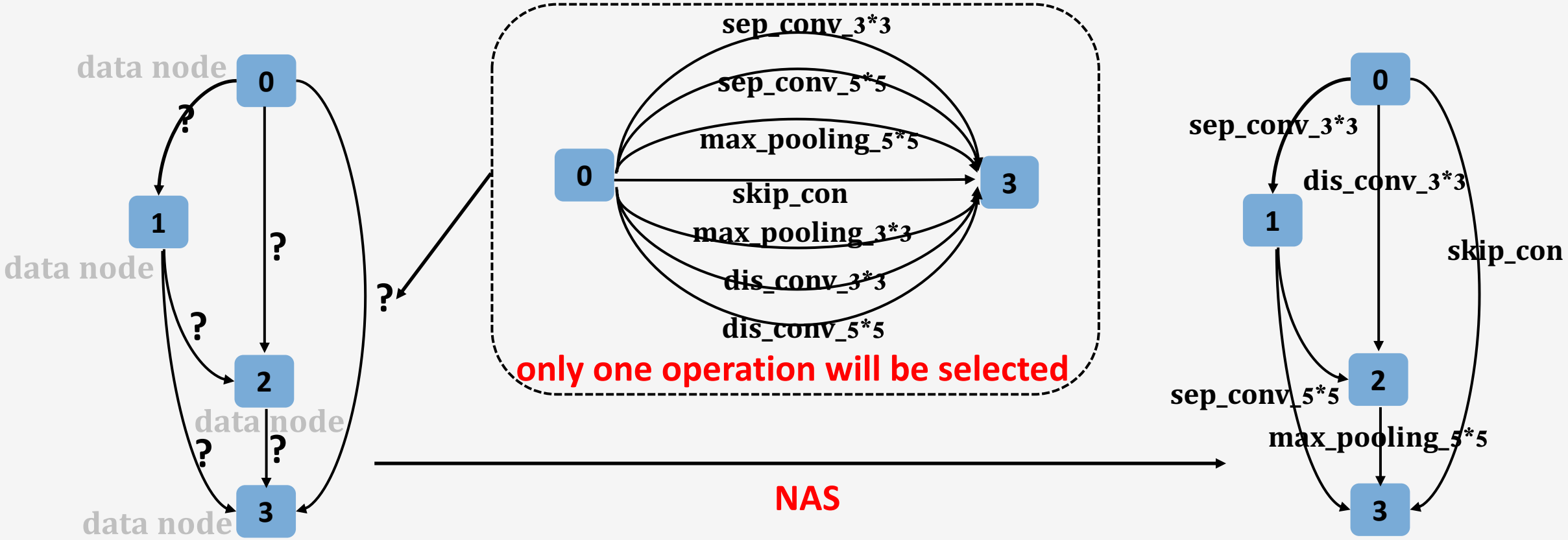
**Solution:** group-structured sparse gate and path-depth-wise regularization

**Experiments:** higher efficiency and classification accuracy

**Conclusion**

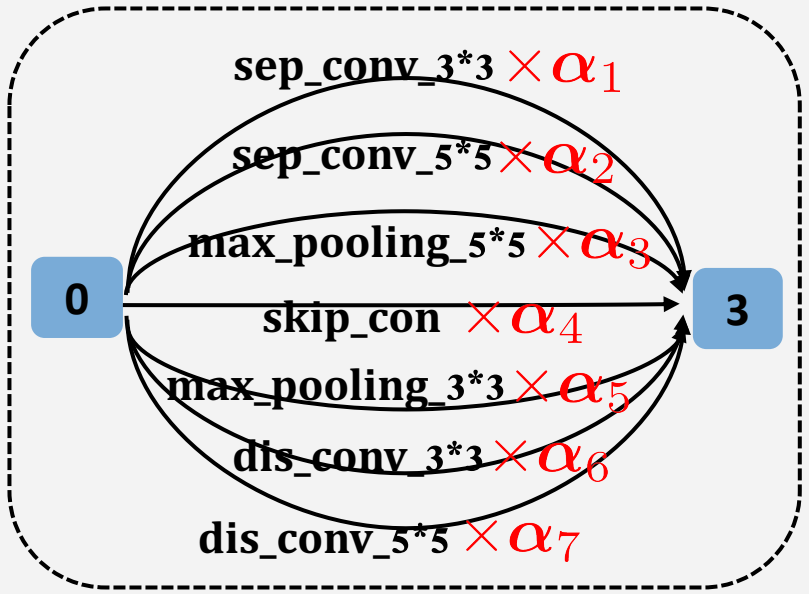
# Background: What Is NAS?

**NAS (network architecture search)** aims to automatically select a proper operation from an operation set for each edge in a dense graph



# Background: What Is NAS?

**Solution:** reinforcement learning (RL) and evolutionary algorithms (EA) are used to solve this discrete operation selection problem.




**Issue: huge search space**

E.g. a graph of 10 nodes has $7^{C_{10}^2} = 7^{45}$ possible operation selections if the operation set is of size 7

**high computational cost** (more than 3000 GPU days)

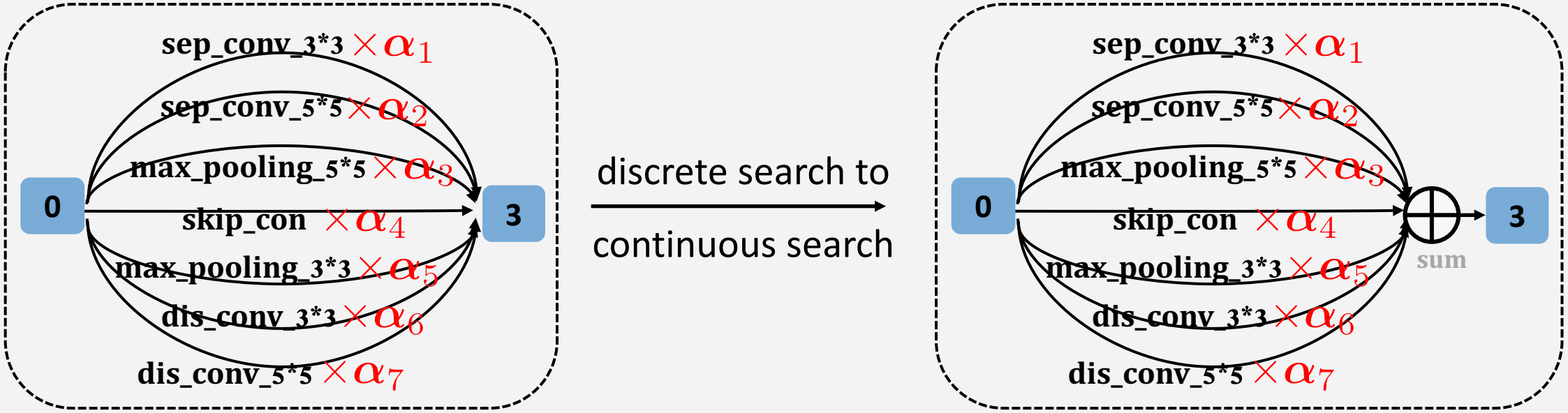
**discrete search space:**

$$\left\{\boldsymbol{\alpha} | \boldsymbol{\alpha}_i \in \{0,1\}, \sum_i \boldsymbol{\alpha}_i = 1\right\}$$

To reduce cost, one often search a small network and then stack several cells to build a large one.

# Differential Architecture Search (DARTS)

**DARTS** converts discrete operation selection into **continuously weighting a set of operations**




**discrete search space:**

$$\left\{\alpha | \alpha_i \in \{0,1\}, \sum_i \alpha_i = 1\right\}$$

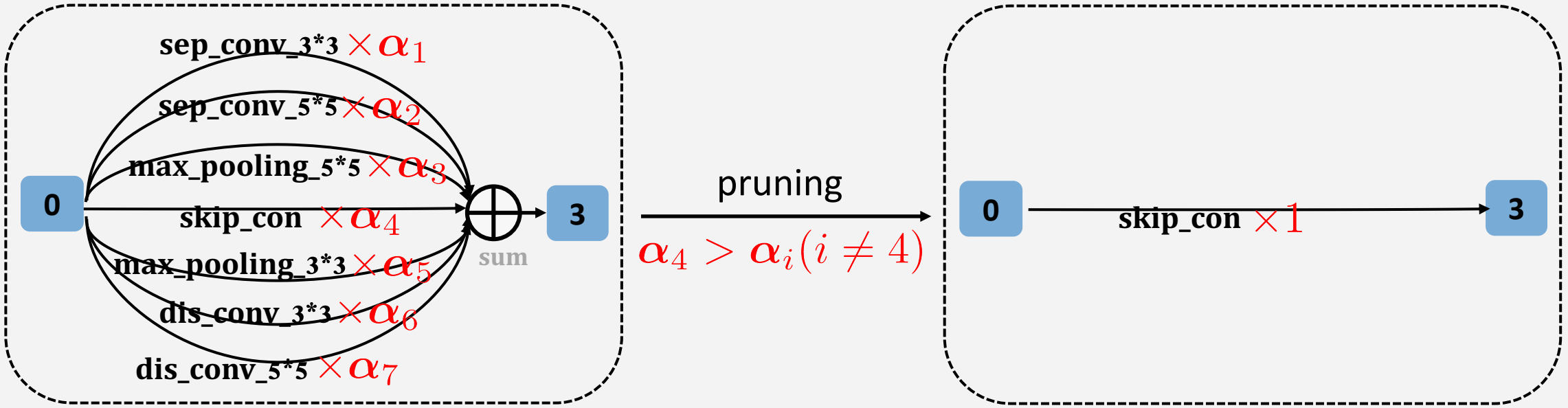
**continuous search space:**

$$\left\{\alpha | \alpha_i \in [0,1], \sum_i \alpha_i = 1\right\}$$

H. Liu et al. DARTS: Differentiable Architecture Search, ICLR'19

# Differential Architecture Search (DARTS)

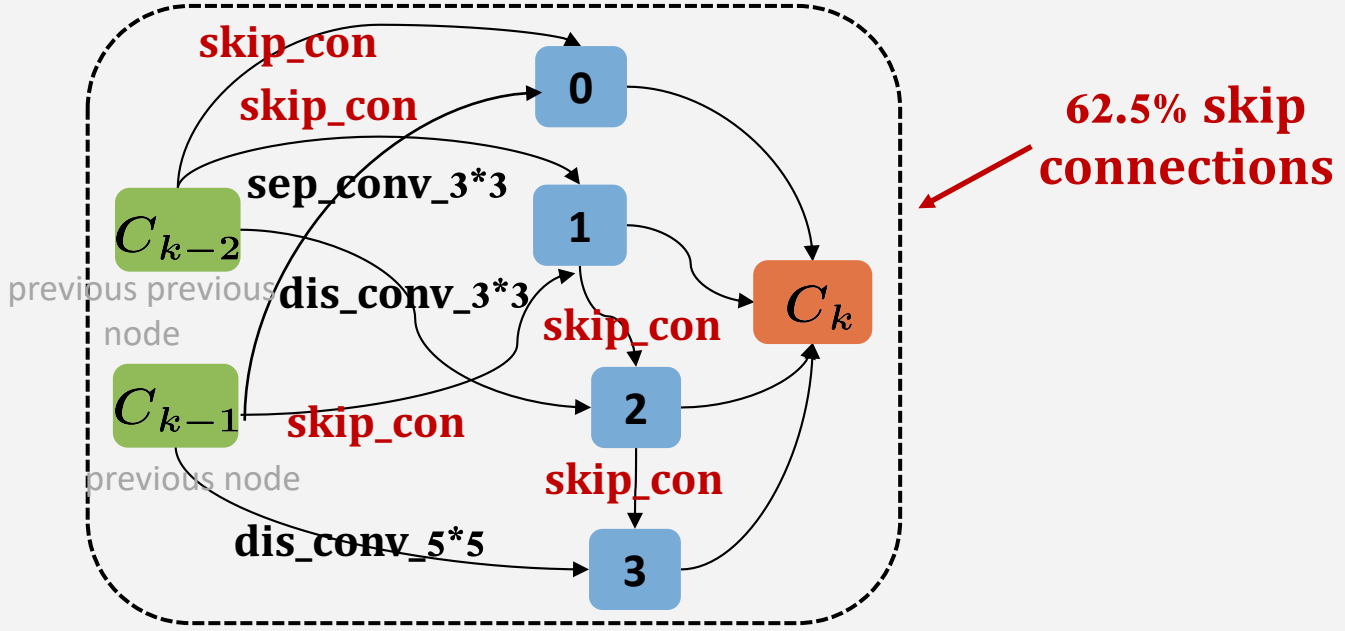
Since the weights of most operations are not exactly zero, one often needs to prune the operations with small weight.




This posteriors pruning often leads to information loss, as it destroys the learnt architecture.

# Observation: Dominated Skip Connections in DARTS

**Observations:** dominated skip connections in the architectures selected by DARTS




**Problems:**

**1) why DARTS prefers to select so many skip connections?**

**2) how to avoid the dominated skip connections?**

# Outline

**Background:** what is network architecture search

**Theoretical analysis:** why DARTS often select so many skip connections

**Solution:** group-structured sparse gate and path-depth-wise regularization

**Experiments:** higher efficiency and classification accuracy

**Conclusion**

# Formulations of DARTS

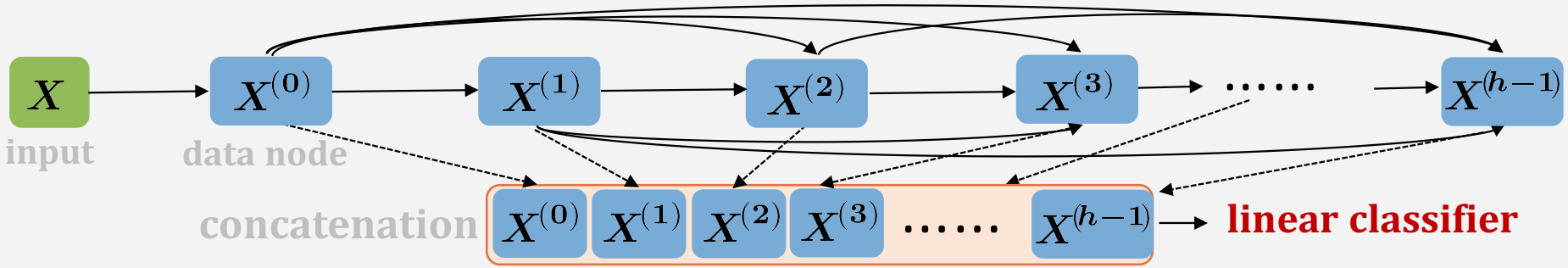
- **Dense directed graph** via connecting current node with all previous nodes

$$\boldsymbol{X}^{(l)} = \sum_{s=0}^{l-1} \left[ \boldsymbol{\alpha}_{s,1}^{(l)} \text{zero}(\boldsymbol{X}^s) + \boldsymbol{\alpha}_{s,2}^{(l)} \text{skip}(\boldsymbol{X}^s) + \boldsymbol{\alpha}_{s,3}^{(l)} \text{conv}(\boldsymbol{W}_s^{(l)}; \boldsymbol{X}^{(s)}) \right] \in \mathbb{R}^{m \times p}$$
$$(l = 1, \cdots, h-1)$$

where $\boldsymbol{\alpha}_{s,1}^{(l)}, \boldsymbol{\alpha}_{s,2}^{(l)}, \boldsymbol{\alpha}_{s,3}^{(l)}$ respectively denote weights of zero, skip and convolution operations.

- **Prediction** by feeding the features in all layers into a linear classifier

$$u_i = \sum_{s=0}^{h-1} \langle \boldsymbol{W}_s, \boldsymbol{X}_i^{(s)} \rangle \in \mathbb{R}$$



# Formulations of DARTS

- **Dense directed graph** via connecting current node with all previous nodes

$$\boldsymbol{X}^{(l)} = \sum_{s=0}^{l-1} \left[ \boldsymbol{\alpha}_{s,1}^{(l)} \text{zero}(\boldsymbol{X}^s) + \boldsymbol{\alpha}_{s,2}^{(l)} \text{skip}(\boldsymbol{X}^s) + \boldsymbol{\alpha}_{s,3}^{(l)} \text{conv}(\boldsymbol{W}_s^{(l)}; \boldsymbol{X}^{(s)}) \right] \in \mathbb{R}^{m \times p} \quad (l = 1, \cdots, h-1)$$

where $\boldsymbol{\alpha}_{s,1}^{(l)}, \boldsymbol{\alpha}_{s,2}^{(l)}, \boldsymbol{\alpha}_{s,3}^{(l)}$ respectively denote weights of zero, skip and convolution operations.

- **Prediction** by feeding the features in all layers into a linear classifier

$$u_i = \sum_{s=0}^{h-1} \langle \boldsymbol{W}_s, \boldsymbol{X}_i^{(s)} \rangle \in \mathbb{R}$$

- **DARTS Model:**

optimize network parameter $\boldsymbol{W}$

$$\min_{\alpha} \; F_{\text{val}}(\boldsymbol{W}^*(\alpha), \alpha), \qquad \text{s.t. } \boldsymbol{W}^*(\alpha) = \text{argmin}_{\boldsymbol{W}} \; F_{\text{train}}(\boldsymbol{W}, \boldsymbol{\alpha}),$$

optimize architecture parameter $\alpha$

where the squared loss $F(\boldsymbol{W}, \boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (u_i - y_i)^2$.

# Theoretical Understanding of Dominated Skip Connections

- **Gradient descent** for optimization:

inner optimization: $$\boldsymbol{W}_s^{(l)}(k+1) = \boldsymbol{W}_s^{(l)}(k) - \eta \nabla_{\boldsymbol{W}_s^{(l)}(k)} F_{\text{train}}(\boldsymbol{W}, \boldsymbol{\alpha}) \ (\forall l, s)$$

outer optimization: $$\boldsymbol{\alpha}_s^{(l)}(k+1) = \boldsymbol{\alpha}_s^{(l)}(k) - \eta \nabla_{\boldsymbol{\alpha}_s^{(l)}(k)} F_{\text{val}}(\boldsymbol{W}_s^{(l)}(k+1), \boldsymbol{\alpha}) \ (\forall l, s)$$

# Theoretical Understanding of Dominated Skip Connections

- **Gradient descent** for optimization:

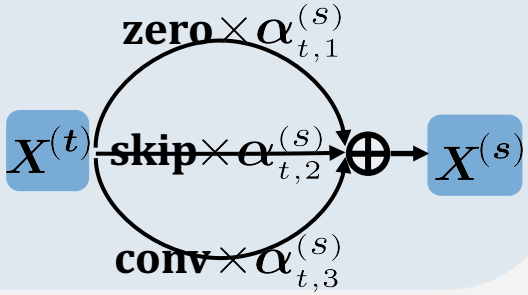
  **inner optimization:** $\boldsymbol{W}_s^{(l)}(k+1) = \boldsymbol{W}_s^{(l)}(k) - \eta \nabla_{\boldsymbol{W}_s^{(l)}(k)} F_{\text{train}}(\boldsymbol{W}, \boldsymbol{\alpha}) \ (\forall l, s)$

**Theorem 1 (Convergence for inner problem, informal).**

Under proper assumptions, for inner problem, the gradient descent algorithm can enjoy linear convergence rate:

$$F_{\text{train}}(\boldsymbol{W}(k+1), \boldsymbol{\alpha}) \leq (1-\lambda) F_{\text{train}}(\boldsymbol{W}(k), \boldsymbol{\alpha}) \quad (\forall k \geq 1),$$

where $\lambda = c\eta \sum_{s=0}^{h-2} (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$ in which $\boldsymbol{\alpha}_{t,2}^{(s)}$ and $\boldsymbol{\alpha}_{t,3}^{(s)}$ respectively denote weights of convolution and skip connections, $c$ is a constant and $\eta$ is learning rate.



# Theoretical Understanding of Dominated Skip Connections

**Theorem 1 (Convergence for inner problem, informal).**
For inner problem, the gradient descent algorithm can enjoy linear convergence rate:

$$F_{\text{train}}(\boldsymbol{W}(k+1), \boldsymbol{\alpha}) \leq (1-\lambda)\, F_{\text{train}}(\boldsymbol{W}(k), \boldsymbol{\alpha}) \quad (\forall k \geq 1),$$

where $\lambda = c\eta \sum_{s=0}^{h-2} (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$ with constant $c$ and learning rate $\eta$ .

- Convergence rate $(1-\lambda)$ depends on the weight $\boldsymbol{\alpha}_{t,2}^{(s)}$ of skip connects more heavily:

$$\lambda = c\eta \sum_{s=0}^{h-2} \lambda_s \quad \text{with} \quad \lambda_s = (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$$

# Theoretical Understanding of Dominated Skip Connections

**Theorem 1 (Convergence for inner problem, informal).**
For inner problem, the gradient descent algorithm can enjoy linear convergence rate:

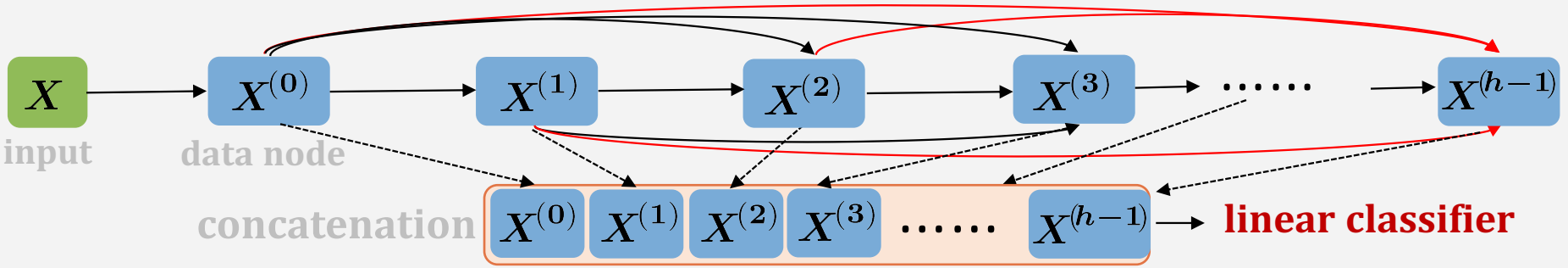
$$F_{\text{train}}(\boldsymbol{W}(k+1), \boldsymbol{\alpha}) \leq (1-\lambda)\, F_{\text{train}}(\boldsymbol{W}(k), \boldsymbol{\alpha}) \quad (\forall k \geq 1),$$

where $\lambda = c\eta \sum_{s=0}^{h-2} (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$ with constant $c$ and learning rate $\eta$ .

- Convergence rate $(1-\lambda)$ depends on the weight $\boldsymbol{\alpha}_{t,2}^{(s)}$ of skip connects more heavily:

$$\lambda = c\eta \sum_{s=0}^{h-2} \lambda_s \quad \text{with} \quad \lambda_s = (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$$

weights of convolutions which connect the last node $\boldsymbol{X}^{(h-1)}$



# Theoretical Understanding of Dominated Skip Connections

**Theorem 1 (Convergence for inner problem, informal).**
For inner problem, the gradient descent algorithm can enjoy linear convergence rate:

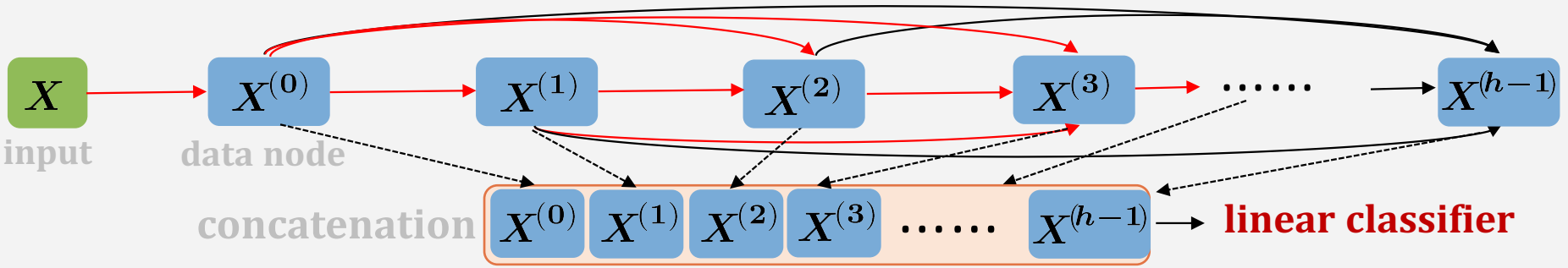
$$F_{\text{train}}(\boldsymbol{W}(k+1), \boldsymbol{\alpha}) \leq (1-\lambda)\, F_{\text{train}}(\boldsymbol{W}(k), \boldsymbol{\alpha}) \quad (\forall k \geq 1),$$

where $\lambda = c\eta \sum_{s=0}^{h-2} (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$ with constant $c$ and learning rate $\eta$ .

- Convergence rate $(1-\lambda)$ depends on the weight $\boldsymbol{\alpha}_{t,2}^{(s)}$ of skip connects more heavily:

$$\lambda = c\eta \sum_{s=0}^{h-2} \lambda_s \quad \text{with} \quad \lambda_s = (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$$

weights of skip connections which do not connect the last node $\boldsymbol{X}^{(h-1)}$



# Theoretical Understanding of Dominated Skip Connections

**Theorem 1 (Convergence for inner problem, informal).**
For inner problem, the gradient descent algorithm can enjoy linear convergence rate:

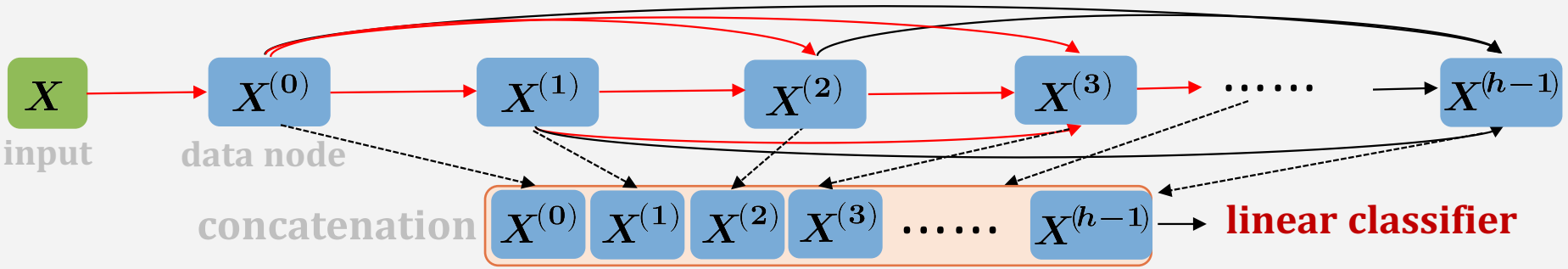
$$F_{\text{train}}(\boldsymbol{W}(k+1), \boldsymbol{\alpha}) \leq (1-\lambda)\, F_{\text{train}}(\boldsymbol{W}(k), \boldsymbol{\alpha}) \quad (\forall k \geq 1),$$

where $\lambda = c\eta \sum_{s=0}^{h-2} (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$ with constant $c$ and learning rate $\eta$ .

- Convergence rate $(1-\lambda)$ depends on the weight $\boldsymbol{\alpha}_{t,2}^{(s)}$ of skip connects more heavily:

$$\lambda = c\eta \sum_{s=0}^{h-2} \lambda_s \quad \text{with} \quad \lambda_s = (\boldsymbol{\alpha}_{s,3}^{(h-1)})^2 \prod_{t=0}^{s-1} (\boldsymbol{\alpha}_{t,2}^{(s)})^2$$

weight product gives heavier dependence



# Theoretical Understanding of Dominated Skip Connections

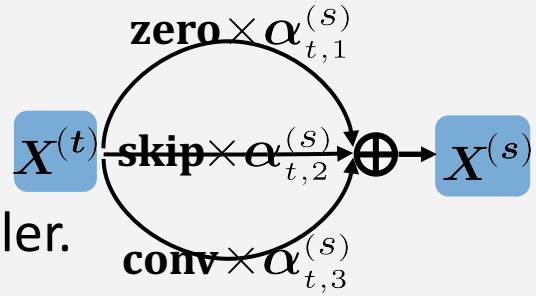
- Since training and validation data are drawn from a same distribution, we have

$$\mathbb{E}[F_{\text{train}}(\boldsymbol{W}), \boldsymbol{\alpha})] = \mathbb{E}[F_{\text{val}}(\boldsymbol{W}), \boldsymbol{\alpha})]$$

- **When skip connections have larger weights, the validation loss can decrease faster**

- Since all types of operations between two nodes share a softmax distribution

$$\boldsymbol{\alpha}_{t,i}^{(s)} = \frac{\exp(\boldsymbol{\beta}_{t,i}^{(s)})}{\sum_{i=1}^{3} \exp(\boldsymbol{\beta}_{t,i}^{(s)})} \longrightarrow \sum_i \boldsymbol{\alpha}_{t,i}^{(s)} = 1$$



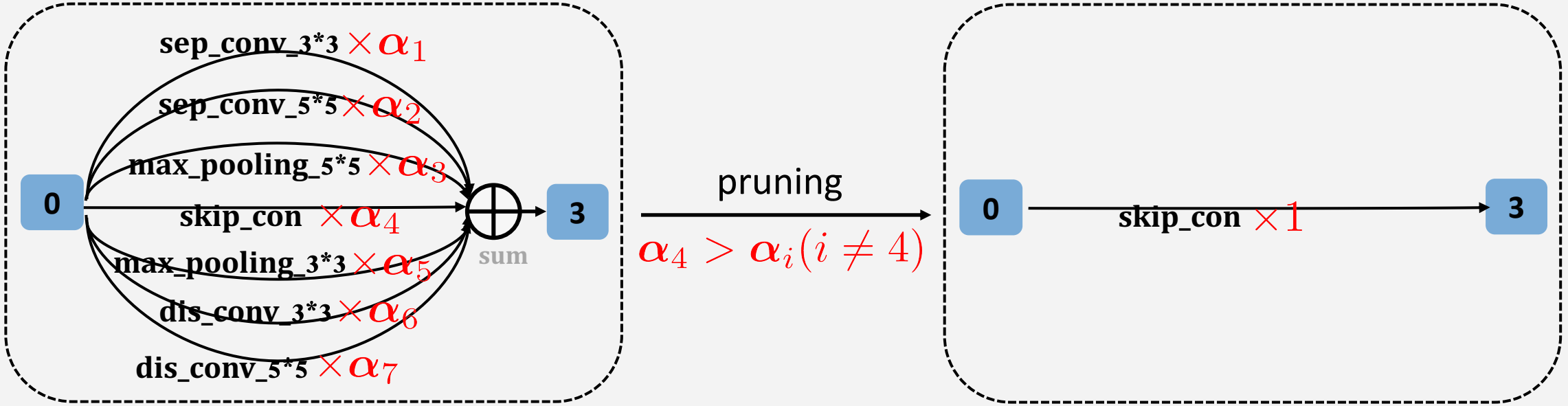

if weight of skip connection becomes larger, other weights become smaller.

- In the outer level, DARTS will increase the weights of skip connections and reduce the weights of other operations.

outer optimization: $\boldsymbol{\alpha}_s^{(l)}(k+1) = \boldsymbol{\alpha}_s^{(l)}(k) - \eta \nabla_{\boldsymbol{\alpha}_s^{(l)}(k)} F_{\text{val}}(\boldsymbol{W}_s^{(l)}(k+1), \boldsymbol{\alpha}) \ (\forall l, s)$

# Theoretical Understanding of Dominated Skip Connections

- After searching, the **posterior pruning will preserve** most of **skip connections** and **prune** most of **other operations**.




- Our theoretical result can answer the first question:

**why DARTS prefers to select so many skip connections?**

# Outline

**Background:** what is network architecture search

**Theoretical analysis:** why DARTS often select so many skip connections

**Solution:** group-structured sparse gate and path-depth-wise regularization

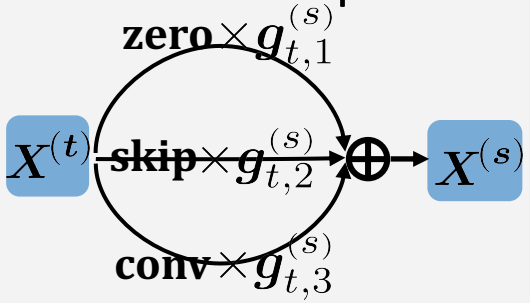
**Experiments:** higher efficiency and classification accuracy

**Conclusion**

# Solution to Reduce Skip Connections

**One solution:** independent gate implemented by Bernoulli distribution for each operation

$$g_{t,i}^{(s)} \sim \text{Bernoulli}\left(\frac{\exp(\beta_{t,i}^{(s)})}{1+\exp(\beta_{t,i}^{(s)})}\right)$$




Theorem 2 (Convergence for inner problem, informal).
When we replace the weights from a softmax distribution with the independent gate, then **increasing $g_{t,i}^{(s)}$ of any operations can reduce or maintain the validation loss**.

**Issue:**

**independent gate leads to a dense network and thus performance degradation,**

since posterior pruning for a compact network prunes operations with non-zero weights.

# Solution to Reduce Skip Connections

**Solution: group-structured sparsity regularization on the stochastic gates**

- Step 1. use Gumbel trick to produce an approximate Bernoulli variable $u$

$$u = \frac{\exp\left(\frac{v_1 + \ln v_2}{\tau}\right)}{1+\exp\left(\frac{v_1 + \ln v_2}{\tau}\right)} \approx \text{Bernoulli}(v_2), \quad v_1 \sim \text{Uniform}(0,1), \quad v_2 = \frac{\exp(\boldsymbol{\beta}_{t,i}^{(s)})}{1+\exp(\boldsymbol{\beta}_{s,i}^{(t)})}$$

- Step 2. rescale u from [0,1] to [a,b] (a<0, b>1), and feed $\boldsymbol{g}_{t,i}^{(s)}$ into a hard threshold gate

$$\boldsymbol{g}_{t,i}^{(s)} = a + (b-a)u, \qquad \boldsymbol{g}_{t,i}^{(s)} = \min(1, \max(0, \boldsymbol{g}_{t,i}^{(s)}))$$

The gate $\boldsymbol{g}_{t,i}^{(s)}$ can become sparse and its activation probability is computable.

$$\boldsymbol{g}_{t,i}^{(s)} = \begin{cases} 0, & \text{if } u \in (0, -\frac{a}{b-a}], \text{ sparse} \\ \boldsymbol{g}_{t,i}^{(s)}, & \text{if } u \in (-\frac{a}{b-a}, \frac{1-a}{b-a}], \\ 1, & \text{if } u \in (\frac{1-a}{b-a}, 1], \end{cases} \qquad \mathbb{P}(\boldsymbol{g}_{t,i}^{(s)} \neq 0) = \Theta\left(\boldsymbol{\beta}_{t,i}^{(s)} - \tau \ln \frac{-a}{b}\right),$$

gate activation probability

where $\Theta$ denotes the sigmoid function.

# Solution to Reduce Skip Connections

**Solution: group-structured sparsity regularization on the stochastic gates**

- Step 3. divide the operations in the cell into two groups, skip connection group and non-skip connection group, and compute their average gate activation probabilities:

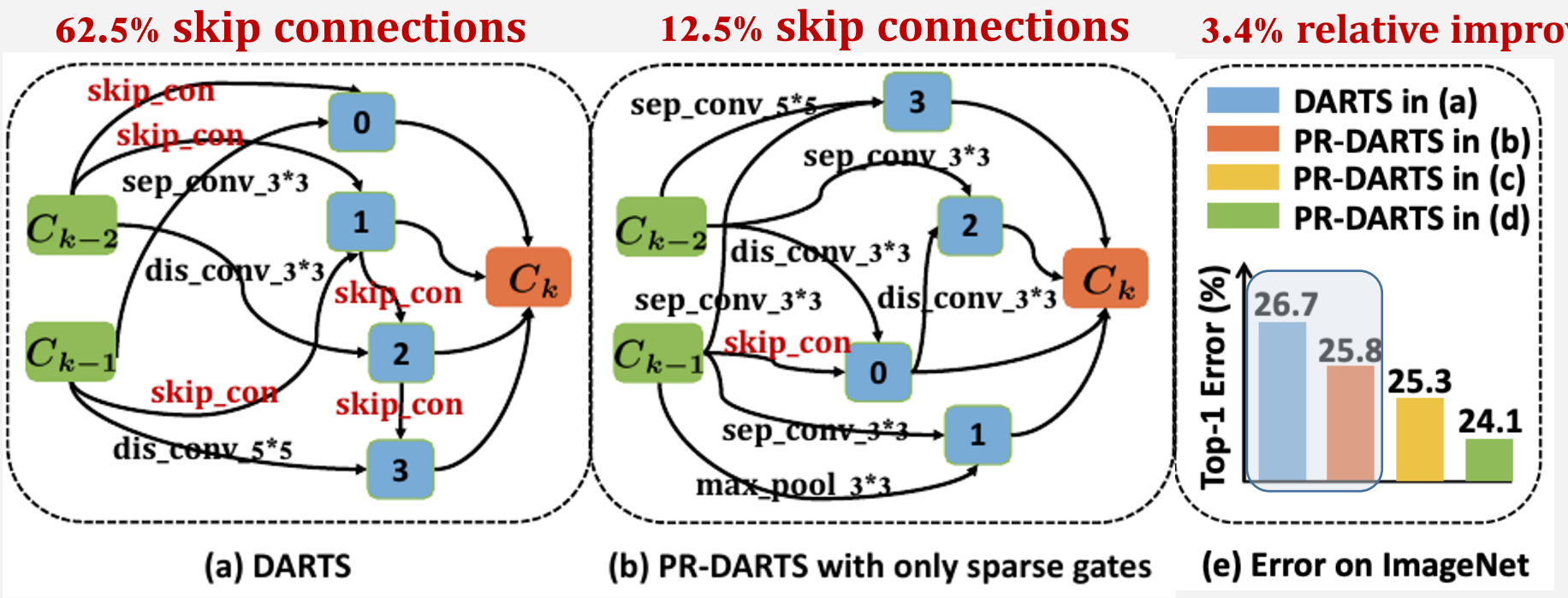
$$\mathcal{L}_{\text{skip}}(\boldsymbol{\beta}) = \zeta \sum_{l=1}^{h-1} \sum_{s=0}^{l-1} \Theta\left(\boldsymbol{\beta}_{s,t_{\text{skip}}}^{(l)} - \tau \ln \frac{-a}{b}\right), \quad \mathcal{L}_{\text{non-skip}}(\boldsymbol{\beta}) = \frac{\zeta}{r-1} \sum_{l=1}^{h-1} \sum_{s=0}^{l-1} \sum_{1 \le t \le r, t \ne t_{\text{skip}}} \Theta\left(\boldsymbol{\beta}_{s,t}^{(l)} - \tau \ln \frac{-a}{b}\right),$$

- Step 4. we penalize these two terms independently to avoid competition between skip connection and other type operations.

# Solution to Reduce Skip Connections

**Solution:** **group-structured sparsity regularization on the stochastic gates**

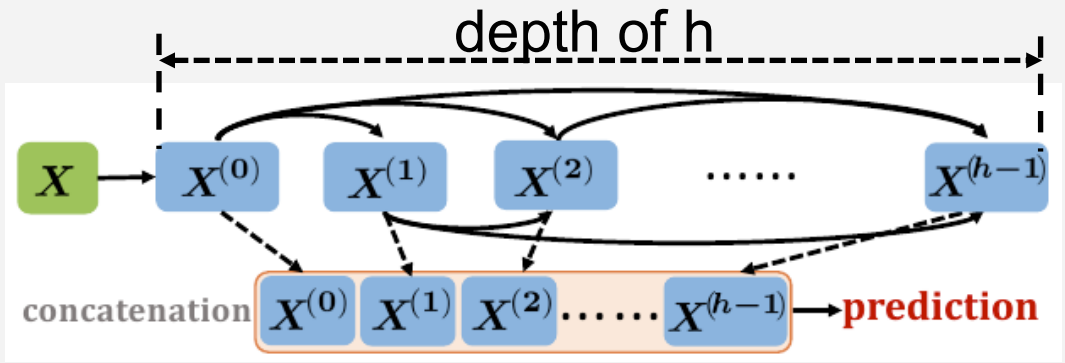
**Advantages**: this solution greatly reduce the skip connections in the selected network.



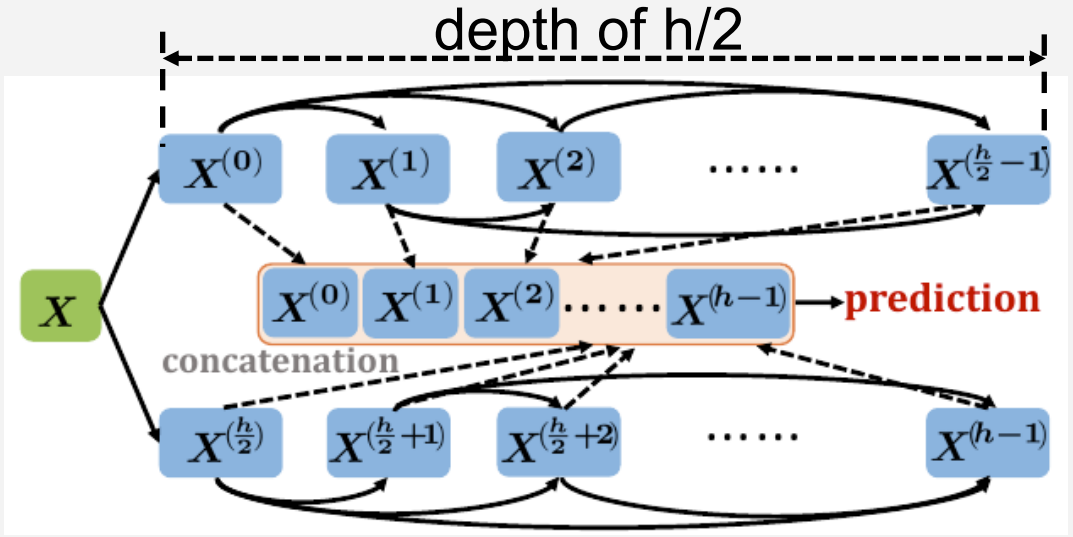
# Solution to Avoid Shallow Networks

**Issues of independent gates**: searching algorithm prefers to select shallow networks due to their faster convergence rate over deep ones.

**Theorem 2 (Convergence Comparison between shallow and deep networks, informal)**
With proper assumptions, shallow network B can converge faster than the deep network A.




(a) Deep network A




(b) Shallow network B

# Solution to Avoid Shallow Networks

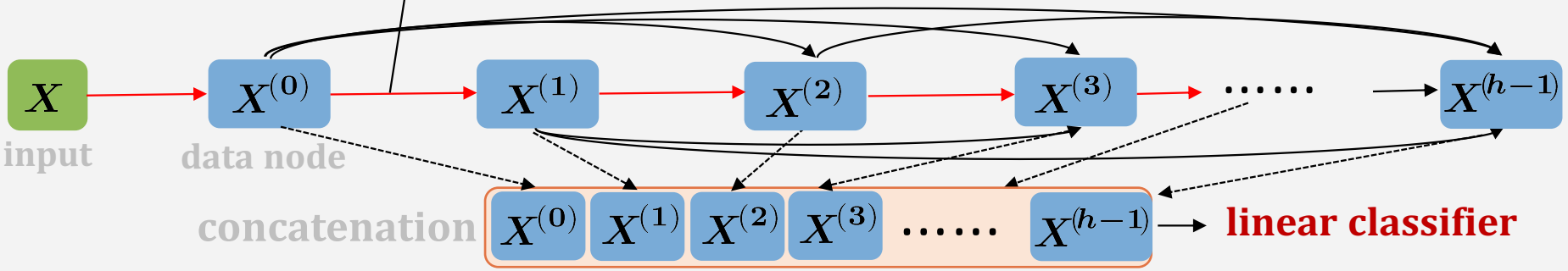
**Solution:** path-depth-wise regularization

- Step 1. probability that all neighboring nodes are connected via parameterized operations

$$\mathcal{L}_{\text{path}}(\boldsymbol{\beta}) = \prod_{l=1}^{h-1} \mathbb{P}_{l,l+1}(\boldsymbol{\beta}) = \prod_{l=1}^{h-1} \sum_{O_t \in \mathcal{O}_p} \Theta\left(\beta_{l,t}^{(l+1)} - \tau \ln \frac{-a}{b}\right).$$

corresponding activation probability of parameterized operations

connected by learnable parameterized operations, e.g. various types of convolutions



# Solution to Avoid Shallow Networks

**Solution:** path-depth-wise regularization

- Step 1. probability that all neighboring nodes are connected via parameterized operations

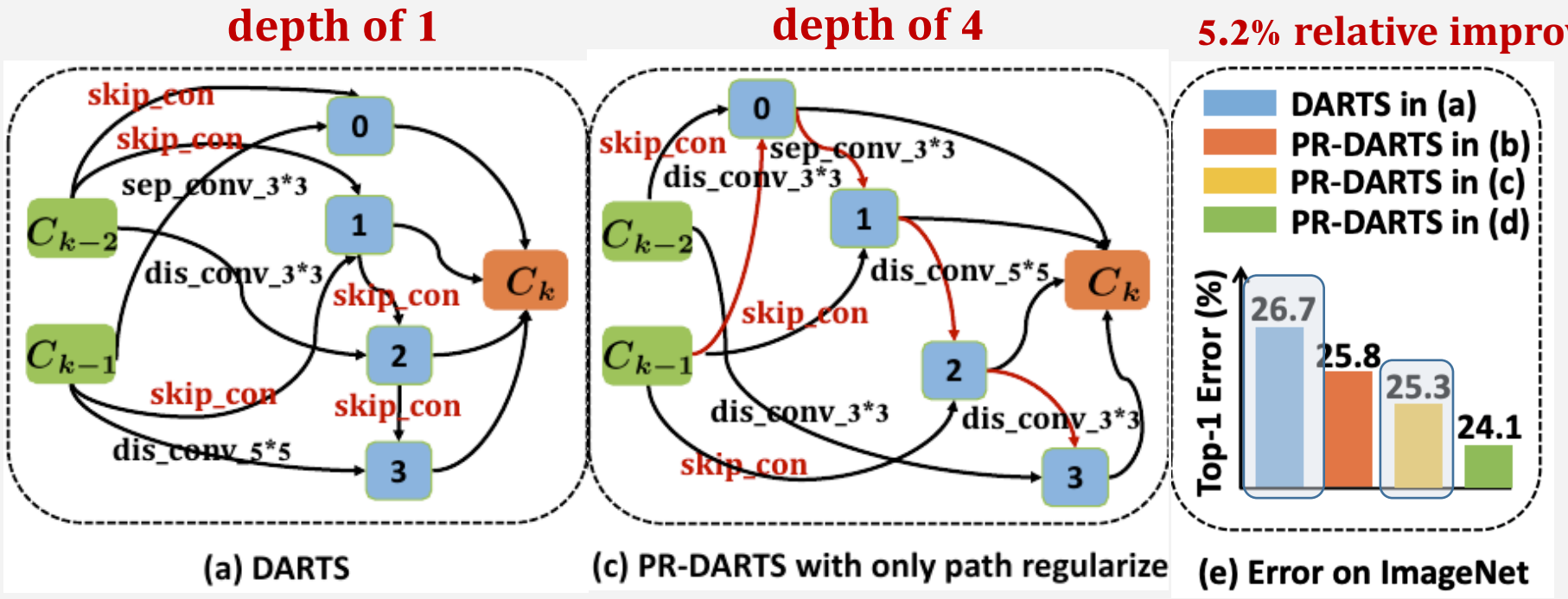
$$\mathcal{L}_{\text{path}}(\boldsymbol{\beta}) = \prod_{l=1}^{h-1} \mathbb{P}_{l,l+1}(\boldsymbol{\beta}) = \prod_{l=1}^{h-1} \sum_{O_t \in \mathcal{O}_p} \Theta\left(\boldsymbol{\beta}_{l,t}^{(l+1)} - \tau \ln \frac{-a}{b}\right).$$

- Step 2. we encourage the selected model to be deep via penalizing small $\mathcal{L}_{\text{path}}(\boldsymbol{\beta})$

# Solution to Avoid Shallow Networks

**Solution:** path-depth-wise regularization

**Advantages**: this solution can search much deeper networks than DARTS.



# Network Architecture Search Model

- **Architecture search model (PR-DARTS):**

optimize network parameter $W$

$$\min_{\beta} F_{\text{val}}(\boldsymbol{W}^*(\boldsymbol{\beta}), \boldsymbol{\beta}) + \lambda_1 \mathcal{L}_{\text{skip}}(\boldsymbol{\beta}) + \lambda_2 \mathcal{L}_{\text{non-skip}}(\boldsymbol{\beta}) - \lambda_3 \mathcal{L}_{\text{path}}(\boldsymbol{\beta}), \ \text{s.t.} \boldsymbol{W}^*(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{W}} F_{\text{train}}(\boldsymbol{W}, \boldsymbol{\beta})$$

optimize architecture parameter $\alpha$

- **Advantages:**

(1) it avoids unfair competition between skip and non-skip connections

$$\min_{\beta} F_{\text{val}}(\boldsymbol{W}^*(\boldsymbol{\beta}), \boldsymbol{\beta}) + \lambda_1 \mathcal{L}_{\text{skip}}(\boldsymbol{\beta}) + \lambda_2 \mathcal{L}_{\text{non-skip}}(\boldsymbol{\beta}) - \lambda_3 \mathcal{L}_{\text{path}}(\boldsymbol{\beta}), \ \text{s.t.} \boldsymbol{W}^*(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{W}} F_{\text{train}}(\boldsymbol{W}, \boldsymbol{\beta})$$

(2) it avoids unfair competition between shallow and deep networks

$$\min_{\beta} F_{\text{val}}(\boldsymbol{W}^*(\boldsymbol{\beta}), \boldsymbol{\beta}) + \lambda_1 \mathcal{L}_{\text{skip}}(\boldsymbol{\beta}) + \lambda_2 \mathcal{L}_{\text{non-skip}}(\boldsymbol{\beta}) - \lambda_3 \mathcal{L}_{\text{path}}(\boldsymbol{\beta}), \ \text{s.t.} \boldsymbol{W}^*(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{W}} F_{\text{train}}(\boldsymbol{W}, \boldsymbol{\beta})$$

# Outline

**Background:** what is network architecture search

**Theoretical analysis:** why DARTS often select so many skip connections

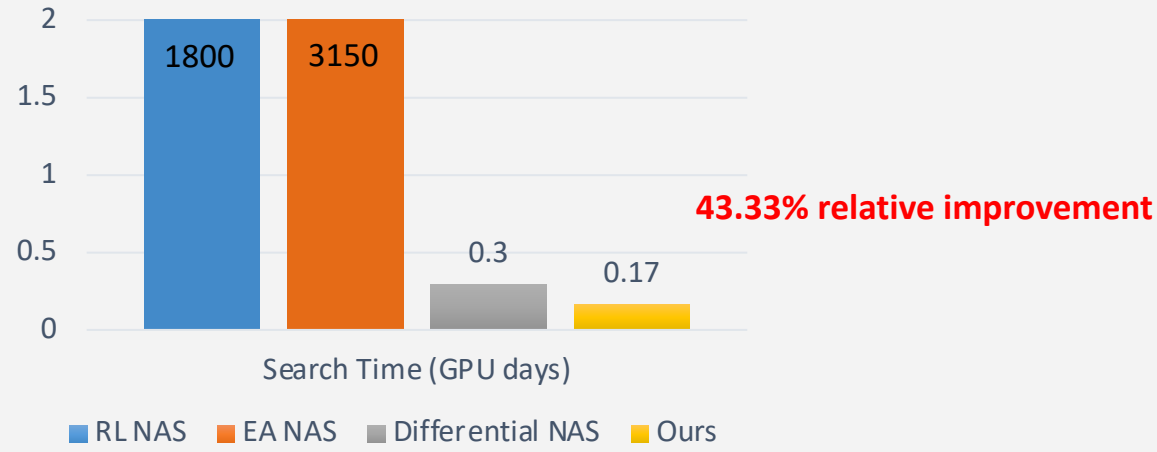
**Solution:** group-structured sparse gate and path-depth-wise regularization

**Experiments:** higher efficiency and classification accuracy

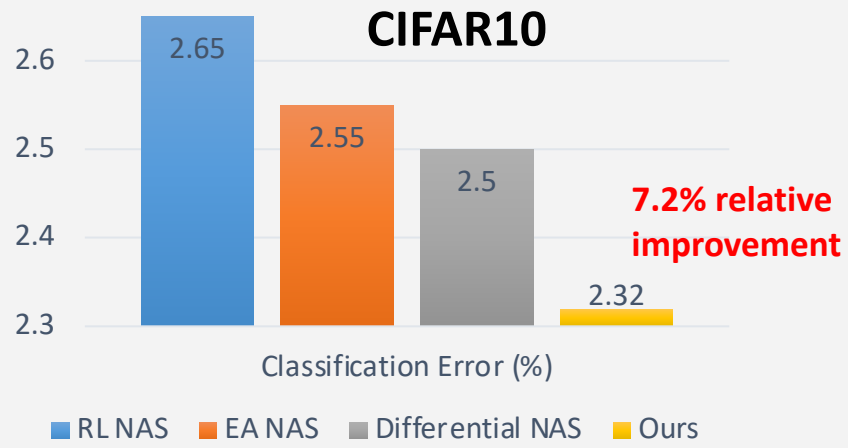
**Conclusion**

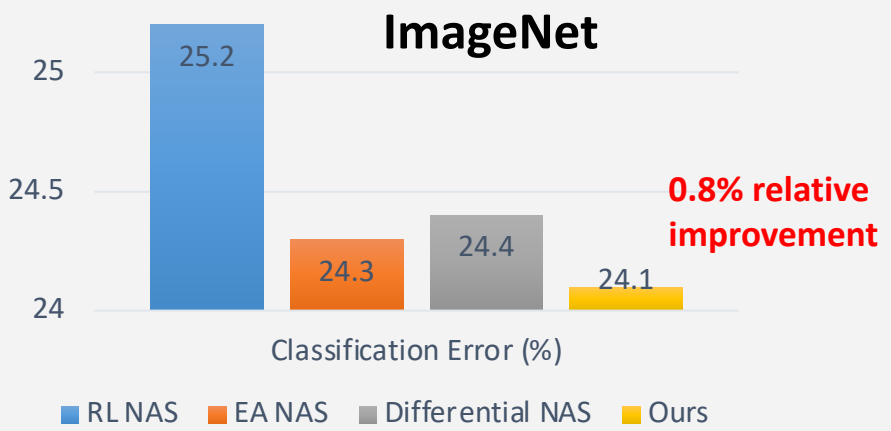
# Experimental Results

## Search Time on CIFAR10: much higher search efficiency




## Accuracy on CIFAR10 and ImageNet: much smaller classification error



# Outline

**Background:** what is network architecture search

**Theoretical analysis:** why DARTS often select so many skip connections

**Solution:** group-structured sparse gate and path-depth-wise regularization

**Experiments:** higher efficiency and classification accuracy

**Conclusion**

# Conclusion

- **Problems:**

(1) why DARTS prefers to select so many skip connections?

**more skip connections lead to faster convergence speed and thus are selected.**

(2) how to avoid the dominated skip connections?

**we propose a group-structured sparsity regularization and a path-depth-wise regularization**


# Thanks !