IEEE WCCI 2020
IEEE World Congress on Computational Intelligence
Virtual Conference – July 19-24, 2020

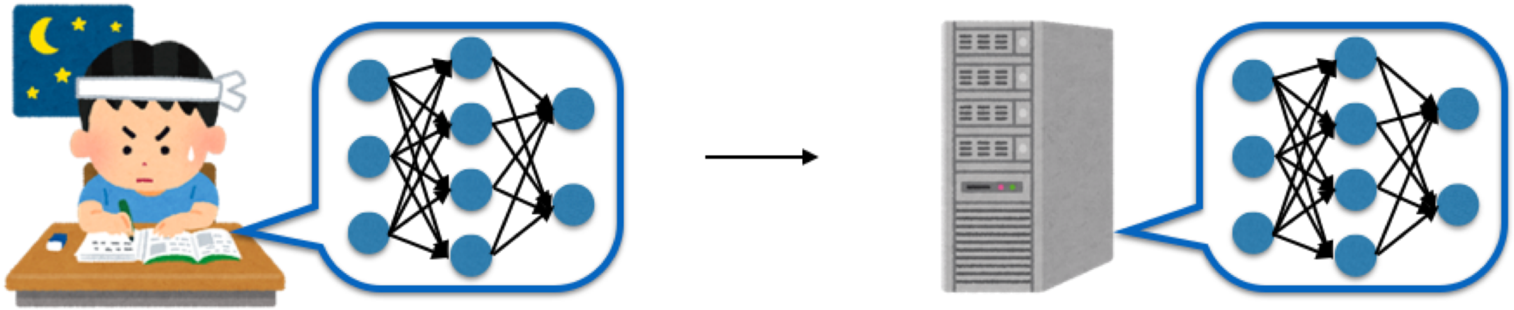
Att-DARTS: Differentiable Neural Architecture Search for Attention

Kohei Nakai, Takashi Matsubara, Kuniaki Uehara
Kobe University, Japan

SPONSORS:
IEEE Advancing Technology for Humanity
IEEE Computational Intelligence Society
IET
The International Neural Network Society (INNS)
EPS Evolutionary Programming Society

Background: Neural Architecture Search

- Deep learning has enabled us to solve various tasks with high performance.
- To achieve high performance, the design of neural networks (NNs) is important.
- Neural Architecture Search (NAS) aims to automate the designing process.



Manually

By NAS

✗ Need much expertise and time

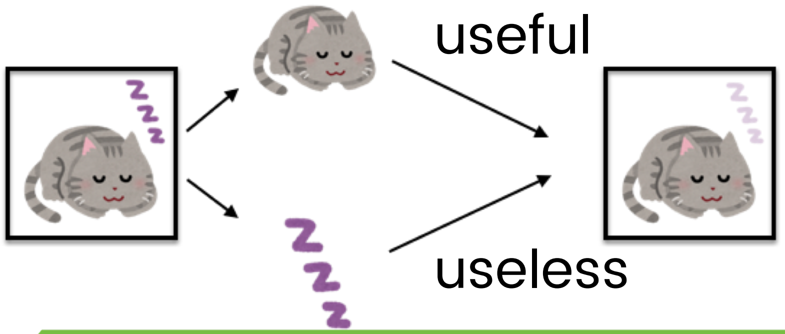
✓ Design NNs fast with high performance

- In this work, we propose a novel NAS for convolutional neural network (CNN).

Background: Attention and CNN

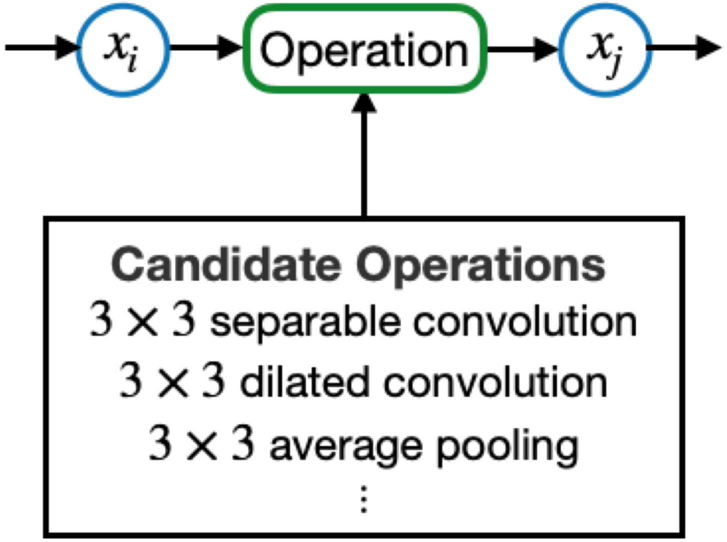
- CNN can capture spatial and channel-wise *features* by repeatedly applying convolutions.
  - ➢ Some of *the features* are useless or harmful to important features.
- Some papers report that inserting attentions to CNN improves its performance.
  - Squeeze-and-Excitation [J. Hu+, CVPR2018] etc.
- **An attention** is the mechanism that focuses on specific parts of the input.
  - ➢ An attention helps CNN discard useless features.

When you identify a cat…



Background: Search Space for CNN

- At the search stage, NAS identifies an operation for each CNN layer.
- Candidates include popular convolutions and poolings.
  - Depth-wise separable convolution [F. Chollet+, CVPR2017] and dilated convolution [F. Yu+, ICLR2016].

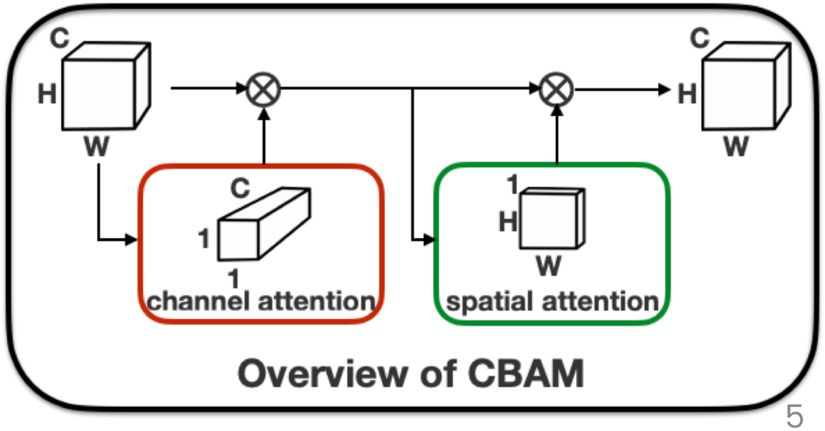

x_i
Operation
x_j
Candidate Operations
3×3 separable convolution
3×3 dilated convolution
3×3 average pooling
⋮

- This search space does **not include any attentions**.
- We propose NAS for CNN **with attentions**.

Related work: Attention for CNN

- There are two types of attentions for images.
  1. Channel attention discards some channels to focus on the remaining channels.
     - Squeeze-and-Excitation [J. Hu+, CVPR2018] etc.
  2. Spatial attention discards some spatial positions to focus on the remaining spatial positions.

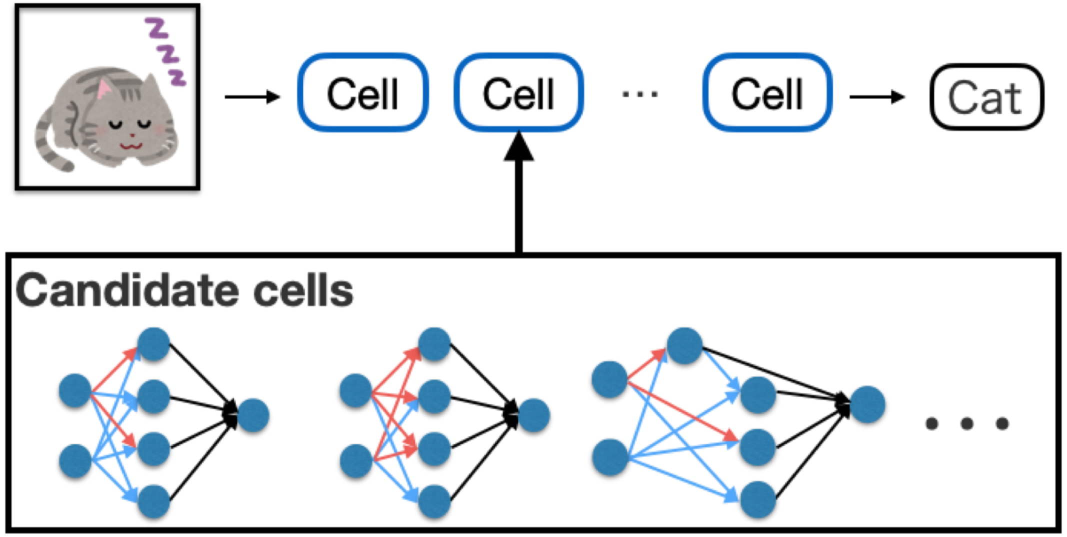
- Some attentions are combinations of channel and spatial attentions.
  - BAM [J. Park+, BMVC2018], CBAM [S. WOO+, ECCV2018]



Overview of CBAM

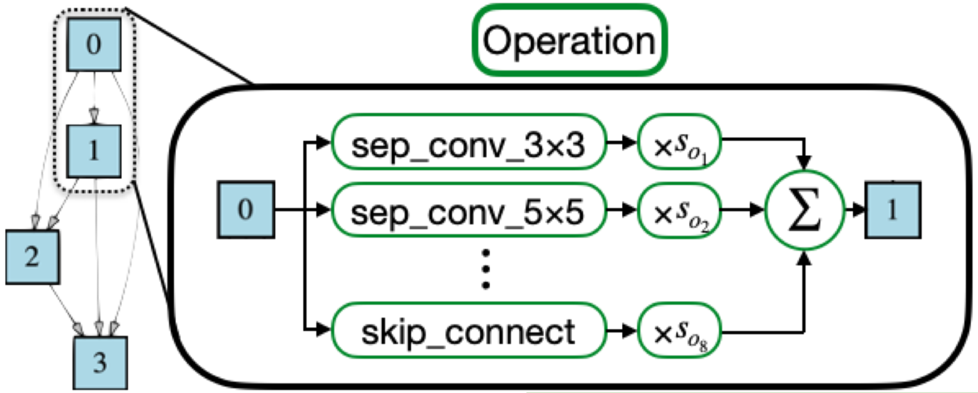
Related work: NAS for CNN

- The entire network is a chain-like structure of repeatedly stacked cells.
- Each cell is expressed as a directed acyclic graph (DAG).
  - Each node expresses an image feature, and each edge expresses an operation.



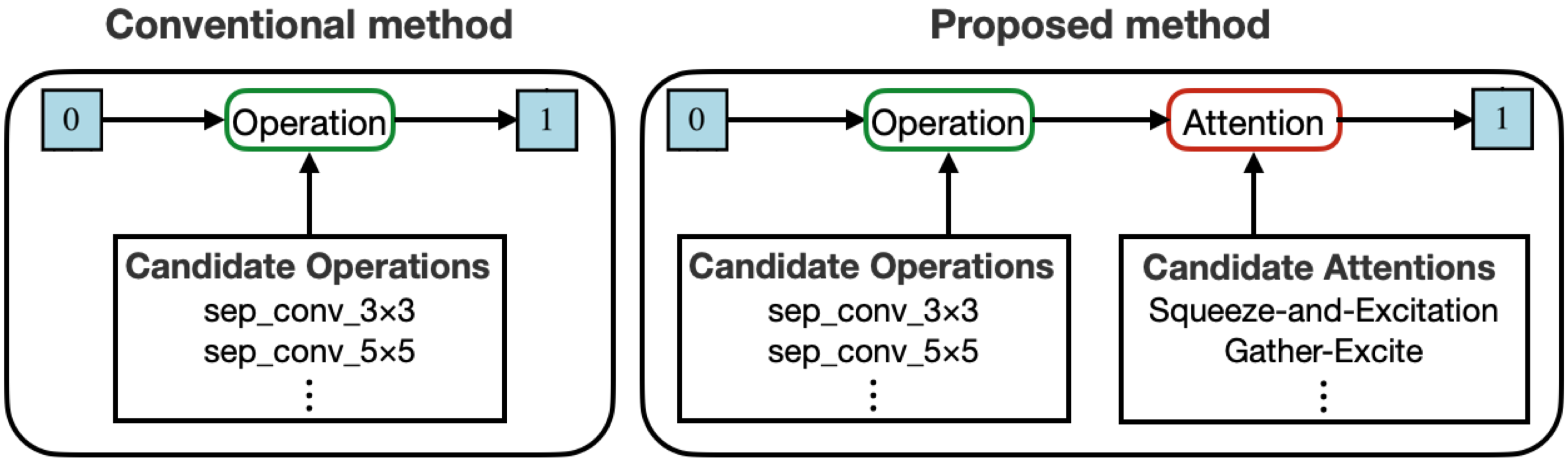
Related work: DARTS [H. Liu+, ICLR2019]

- Each candidate operation has a relative weight s.
- During the search stage, DARTS jointly learns relative weights s as well as weight parameters w (e.g., convolution kernels).
- After the search stage, DARTS chooses
  - an operation with the biggest relative weight s for each edge.
  - two edges with the two biggest relative weights s for each target node.



Proposed Method: Att-DARTS

- Att-DARTS searches for cells including attentions as well as operations.



Conventional method

0 → Operation → 1

Candidate Operations
sep_conv_3×3
sep_conv_5×5
⋮

Proposed method

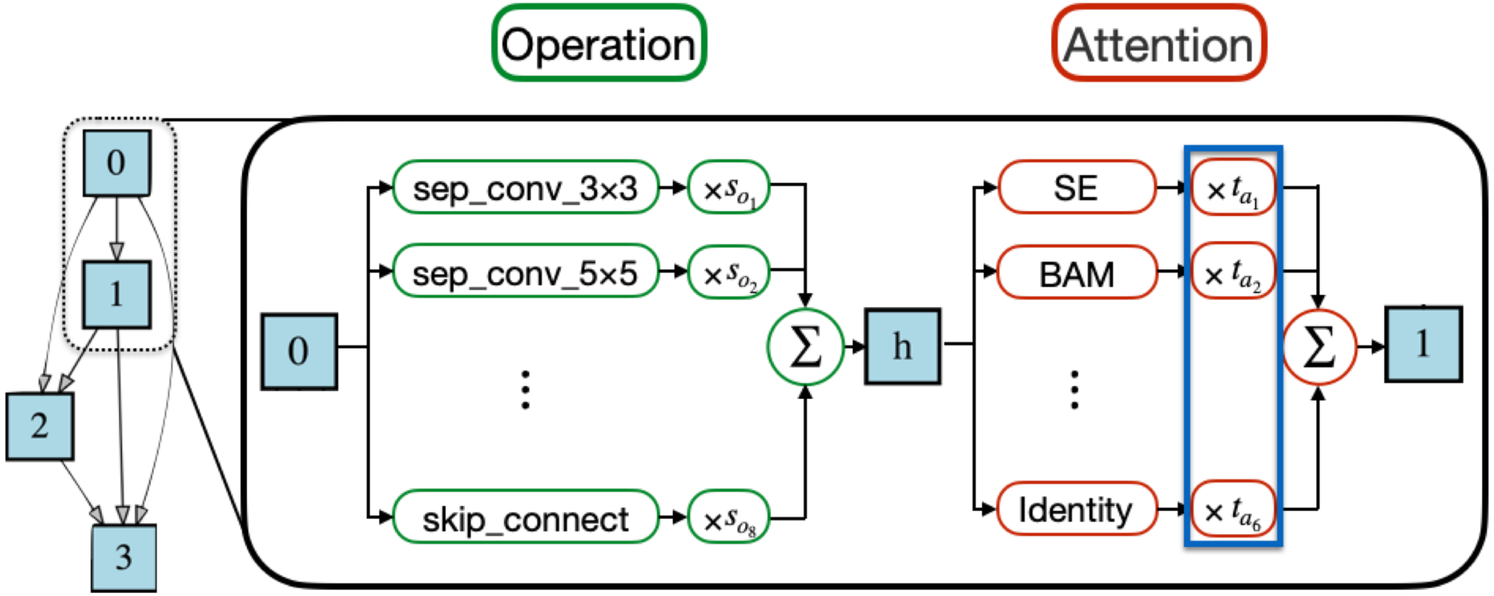
0 → Operation → Attention → 1

Candidate Operations
sep_conv_3×3
sep_conv_5×5
⋮

Candidate Attentions
Squeeze-and-Excitation
Gather-Excite
⋮

Proposed Method: Att-DARTS

- Each candidate attention has a relative weight t.
- Att-DARTS learns relative weights t and chooses an attention with the biggest relative weight t for each edge.



Experiment

- We evaluated Att-DARTS using CIFAR-10, CIFAR-100, and ImageNet (ILSVRC2012).

- The experimental procedure was as follows:
  1. We initialized Att-DARTS using CIFAR-10 to obtain candidate cells.
  2. We built a CNN composed of the cells and trained it from scratch.
  3. We repeated 1. and 2. for four times with different random seeds and chose the best cells based on the best validation accuracy.
  4. We retrained the best cells from scratch using CIFAR-10, CIFAR-100, and ImageNet.
     This allows us to check the transferability to CIFAR-100 and ImageNet.

Results

Dataset for architecture search : CIFAR-10

Dataset for architecture evaluation: CIFAR-10 and CIFAR-100

Number of cells : 20

- Compared with DARTS, Att-DARTS reduced both the classification error and the number of parameters.
- Att-DARTS also reduced the classification error for CIFAR-100.

Architecture	**Test Error (%)**		**Params (M)**
	CIFAR-10	**CIFAR-100**	
DARTS + cutout	2.76 ± 0.09	16.69 ± 0.28	3.3
Att-DARTS + cutout	**2.54** ± 0.10	**16.54** ± 0.40	**3.2**

Results

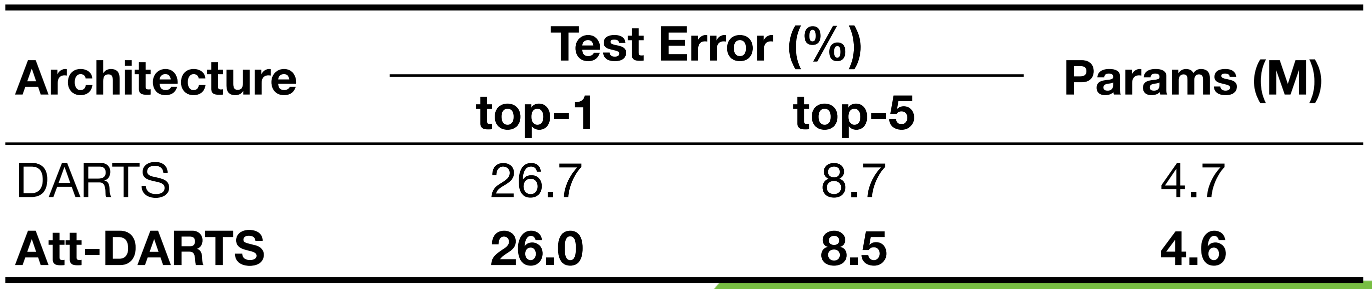
Dataset for architecture search : CIFAR-10

Dataset for architecture evaluation : ImageNet

Number of cells : 14

- Att-DARTS found a CNN using CIFAR-10 with lower classification error for not only CIFAR-10 but also CIFAR-100 and ImageNet.

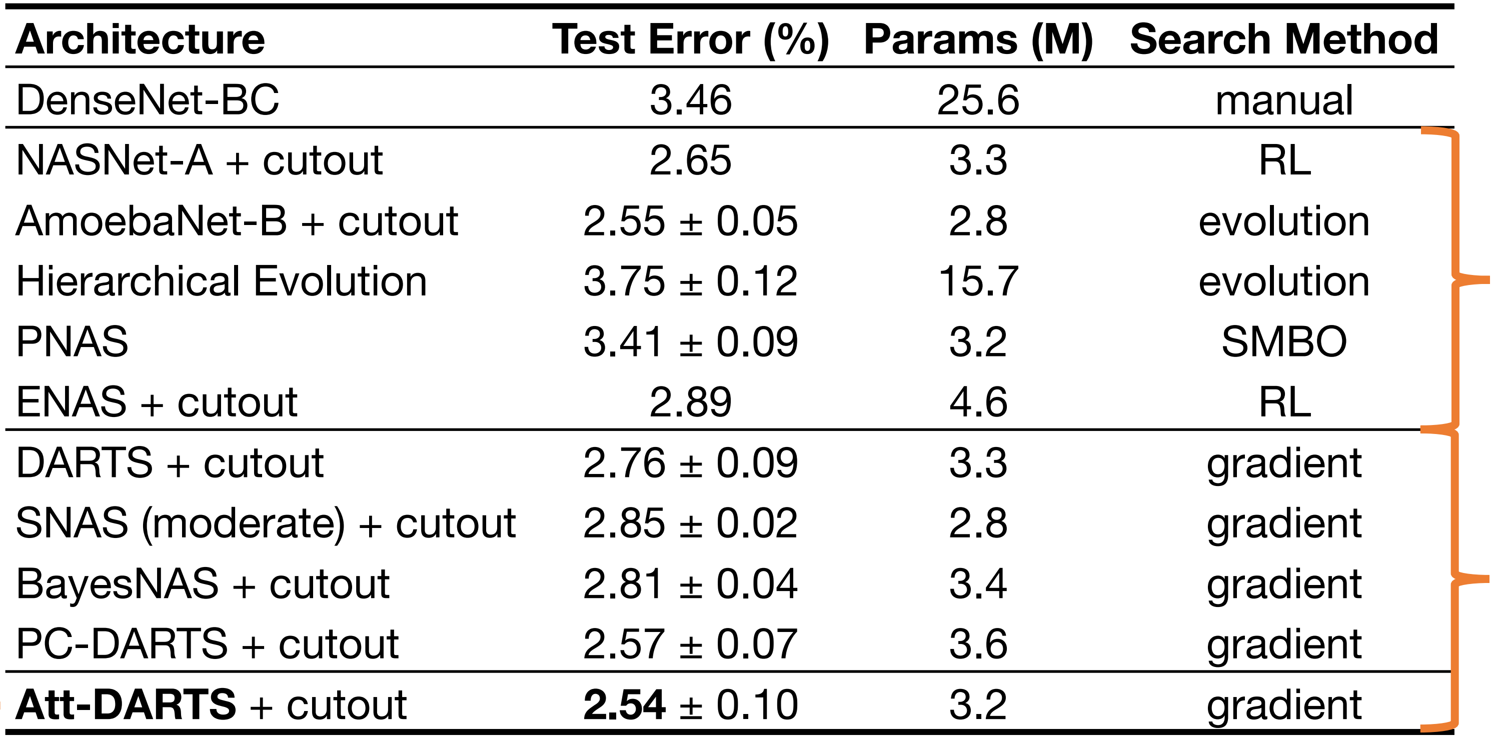
- The best CNN found by Att-DARTS is transferable to CIFAR-100 and ImageNet.



Architecture	Test Error (%)		Params (M)
	top-1	top-5	
DARTS	26.7	8.7	4.7
Att-DARTS	**26.0**	**8.5**	**4.6**

Results

Comparison with state-of-the-art architectures on CIFAR-10



Architecture	Test Error (%)	Params (M)	Search Method
DenseNet-BC	3.46	25.6	manual
NASNet-A + cutout	2.65	3.3	RL
AmoebaNet-B + cutout	2.55 ± 0.05	2.8	evolution
Hierarchical Evolution	3.75 ± 0.12	15.7	evolution
PNAS	3.41 ± 0.09	3.2	SMBO
ENAS + cutout	2.89	4.6	RL
DARTS + cutout	2.76 ± 0.09	3.3	gradient
SNAS (moderate) + cutout	2.85 ± 0.02	2.8	gradient
BayesNAS + cutout	2.81 ± 0.04	3.4	gradient
PC-DARTS + cutout	2.57 ± 0.07	3.6	gradient
Att-DARTS + cutout	**2.54** ± 0.10	3.2	gradient

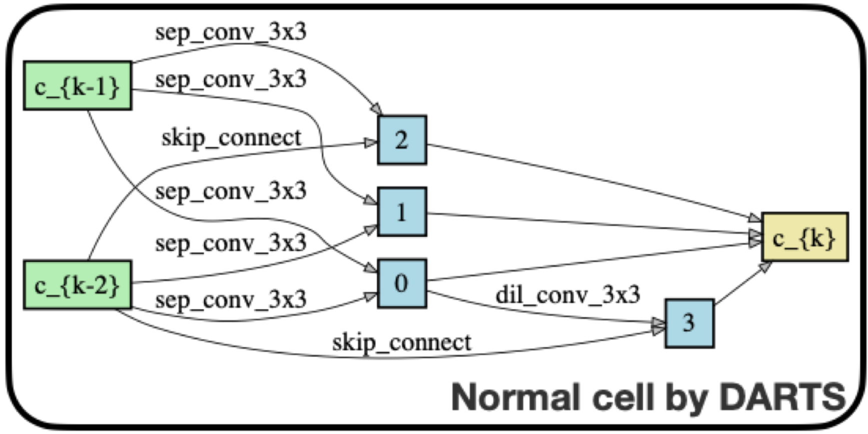
Att-DARTS's approach to search attentions can be combined with gradient-based methods.

These methods tend to take a longer time.

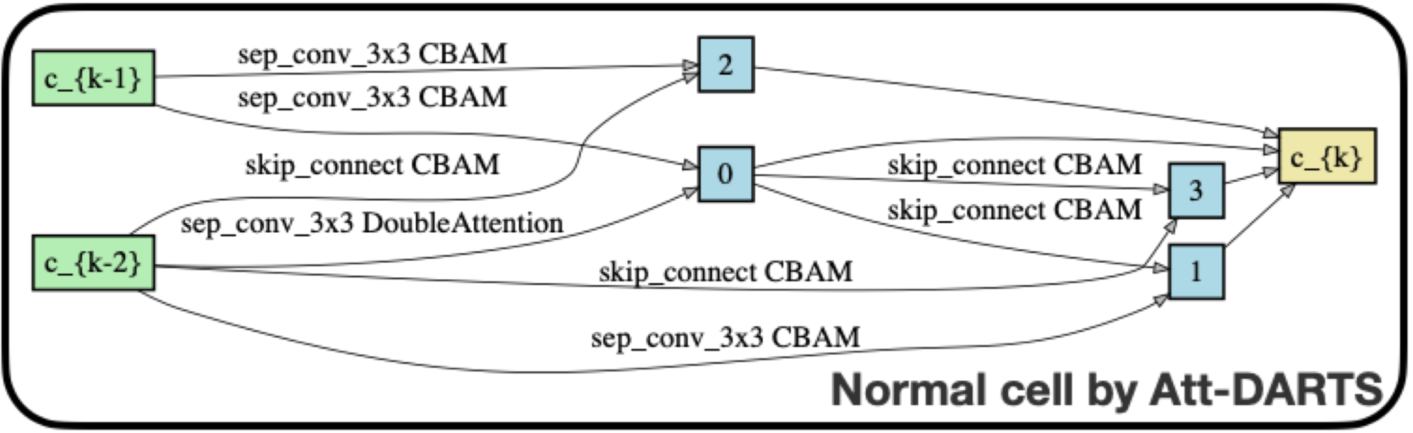
Att-DARTS achieved the best among gradient methods.

Results

- Only operations are arranged in cells by the conventional method, DARTS.
- An attention is inserted after each operation by Att-DARTS.



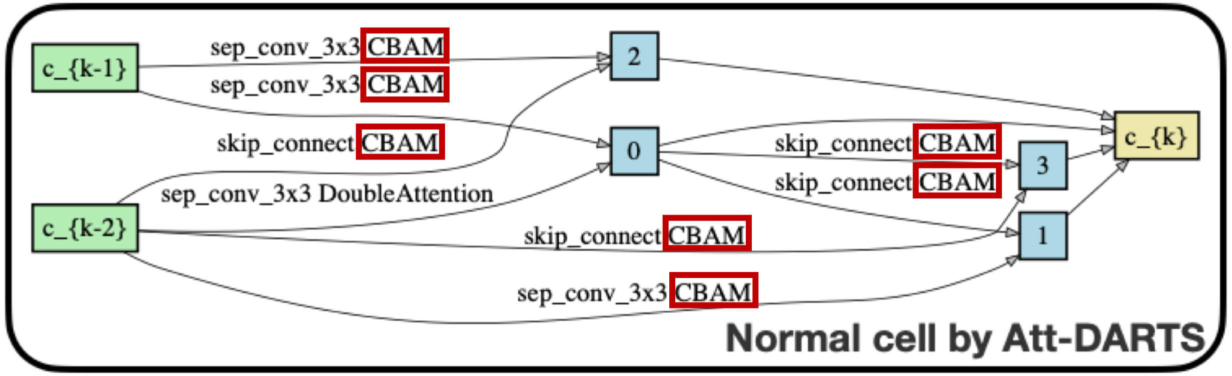
Normal cell by DARTS



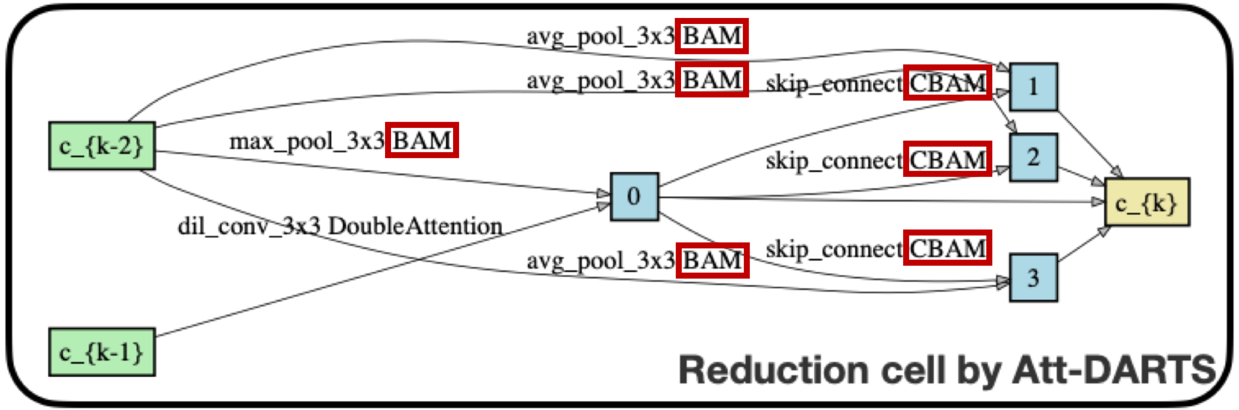
Normal cell by Att-DARTS

Results

- **BAM** and **CBAM** are mainly chosen as attentions.



Normal cell by Att-DARTS

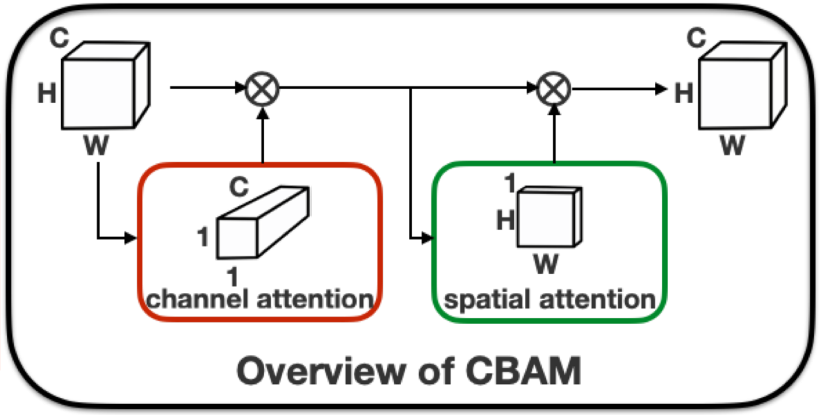


Reduction cell by Att-DARTS

Related work: Attention for CNN (reshown)

- There are two types of attentions for images.
  1. Channel attention masks channels and enables CNN to focus on important channels.
     - Squeeze-and-Excitation [J. Hu+, CVPR2018] etc.
  2. Spatial attention masks spatial positions and enables CNN to focus on important spatial positions.

- Some attentions are **combinations of channel and spatial attentions**.
  - **BAM** [J. Park+, BMVC2018], **CBAM** [S. WOO+, ECCV2018]



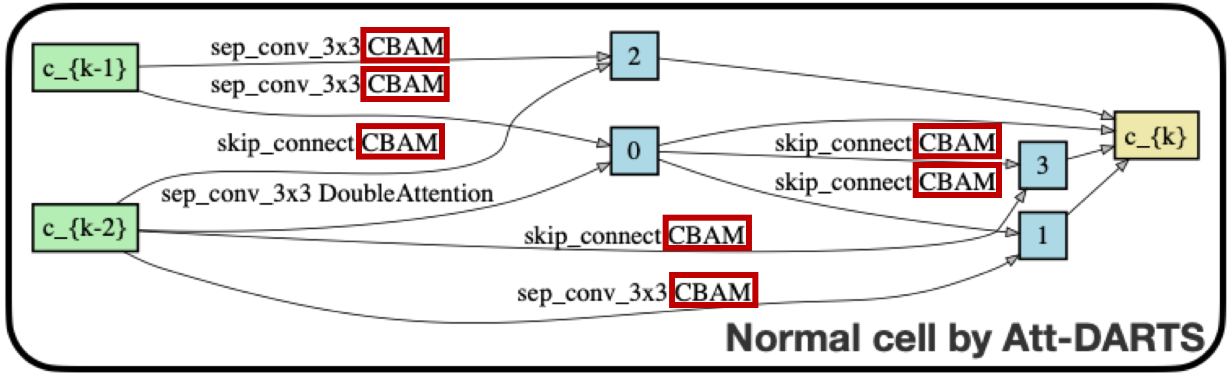
Overview of CBAM

Results

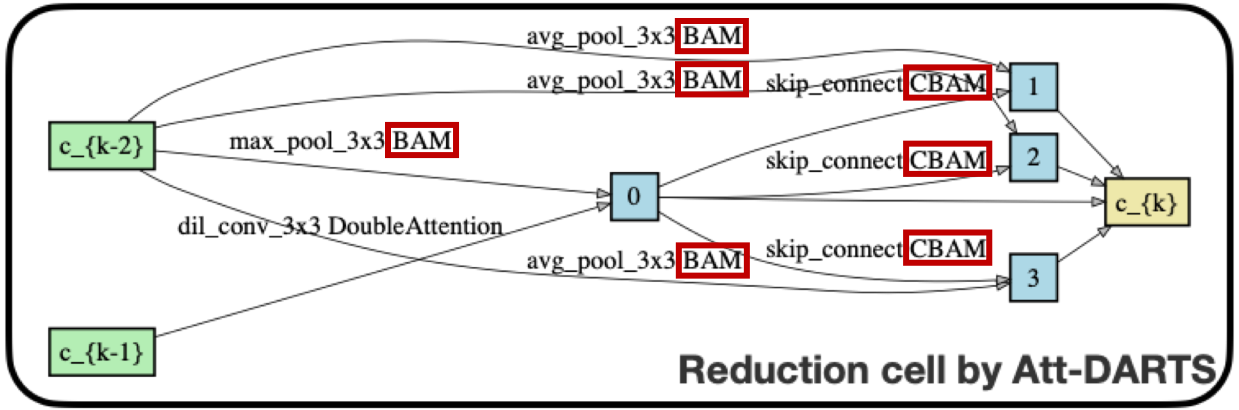
- BAM and CBAM are mainly chosen as attentions.

↓

- Combining channel and spatial attentions is promising.



Normal cell by Att-DARTS



Reduction cell by Att-DARTS

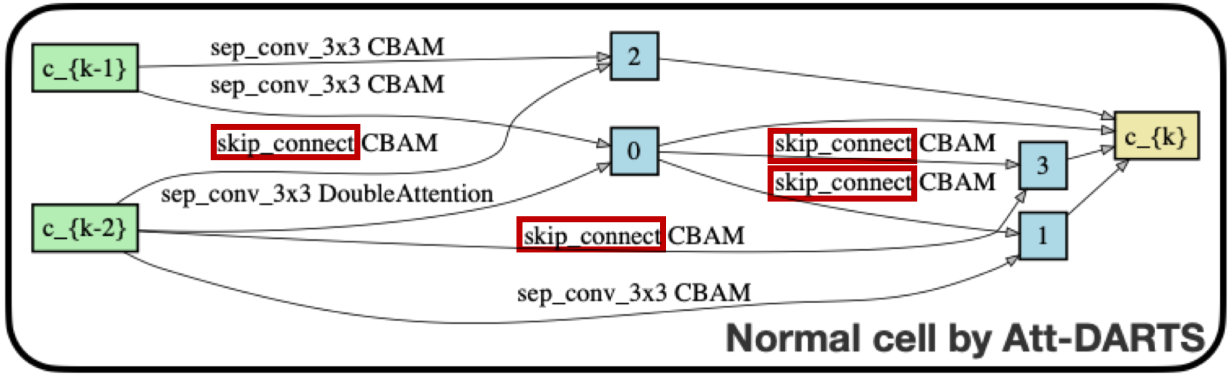
Results

- The skip connection is chosen as an operation the most frequently.

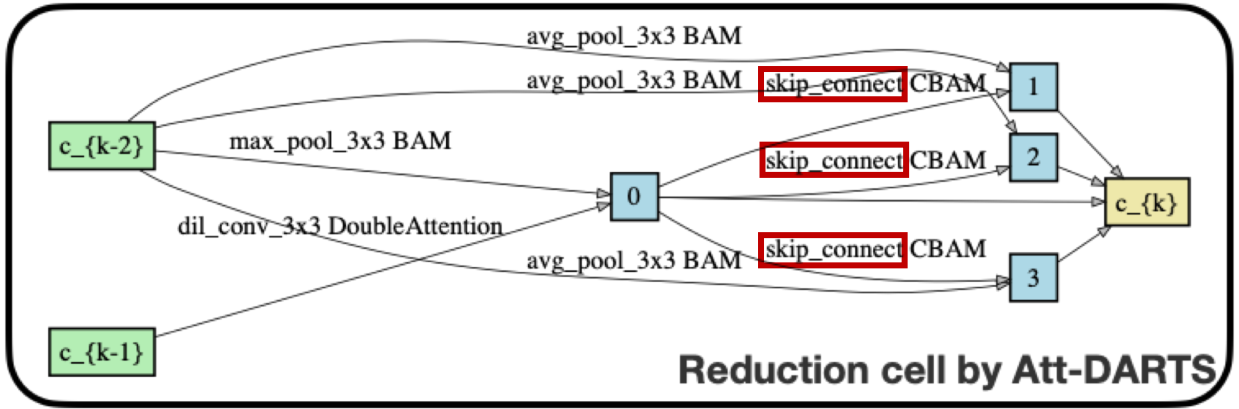
- Attentions are repeatedly applied without convolution operations.
    - From node c_{k-1} to node 3 via node 0 in the normal cell.

⬇

- Repeating attentions is promising.



Normal cell by Att-DARTS



Reduction cell by Att-DARTS

Conclusion

- We proposed Att-DARTS, a novel NAS method that searches over attentions as well as operations.
  - Att-DARTS thus automatically generates CNN with attentions.
- We applied Att-DARTS using CIFAR-10 for CNN discovery.
  Our results suggested that Att-DARTS found the CNN
  - that achieved **lower** classification error.
  - that had **fewer** number of parameters.
  - that was **transferable** from CIFAR-10 to CIFAR-100 and ImageNet, in other words, whose performance did **not depend** on the dataset used during the search stage.
- Regarding the cells found by Att-DARTS, we found
  - combining channel and spatial attentions is promising.
  - repeating attentions is promising.