电子科技大学
University of Electronic Science and Technology of China

Dynamic Neural Networks

Data Mining Lab,
Big Data Research Center, UESTC
Email：weihan@std.uestc.edu.cn

Outline

数据挖掘实验室
Data Mining Lab

- Background & Overview
- Dynamic Architecture & Parameter
- Inference & Training Tricks
- Application & Discussion

Background & Overview

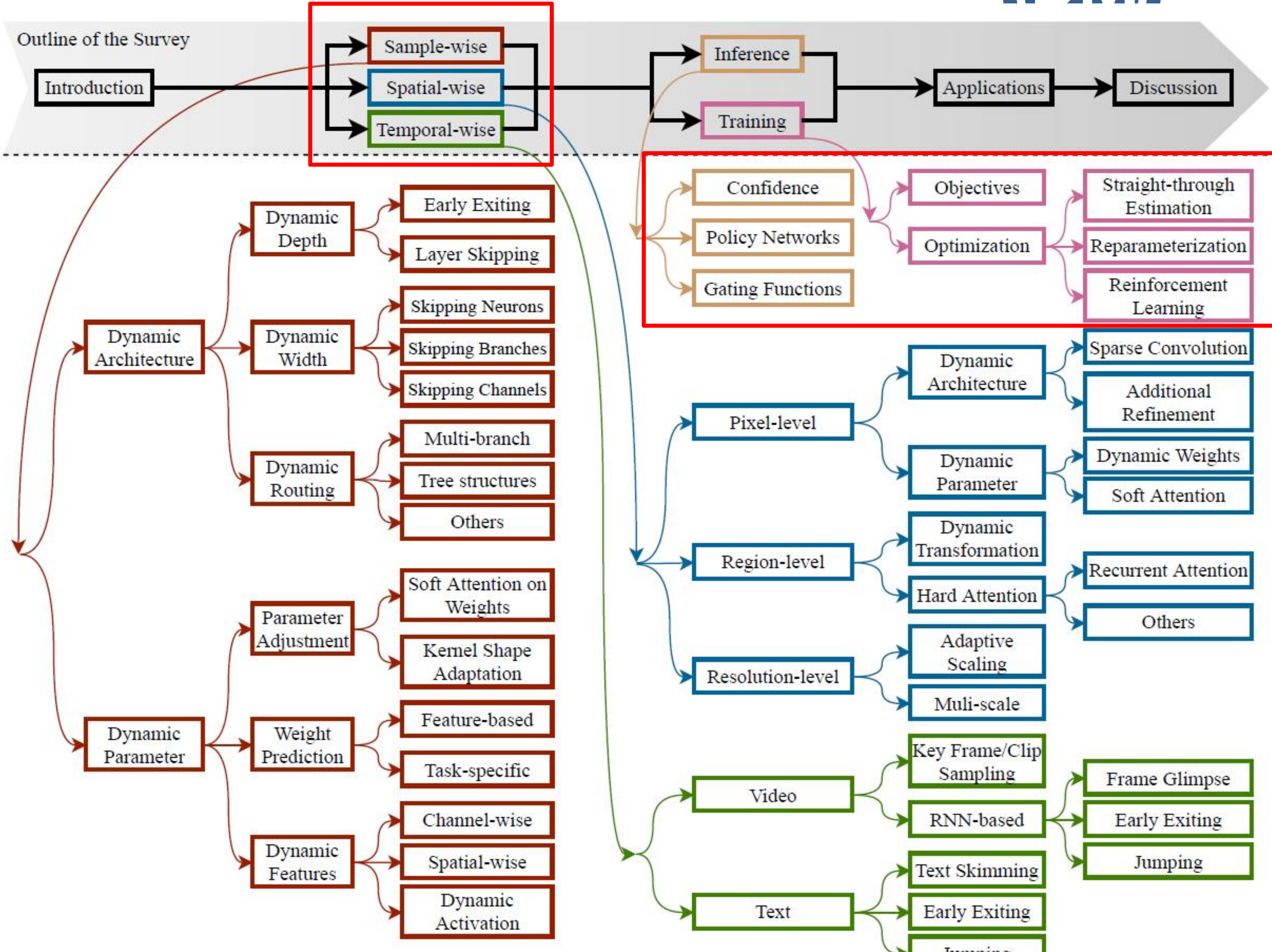
Background

■ Why dynamic?

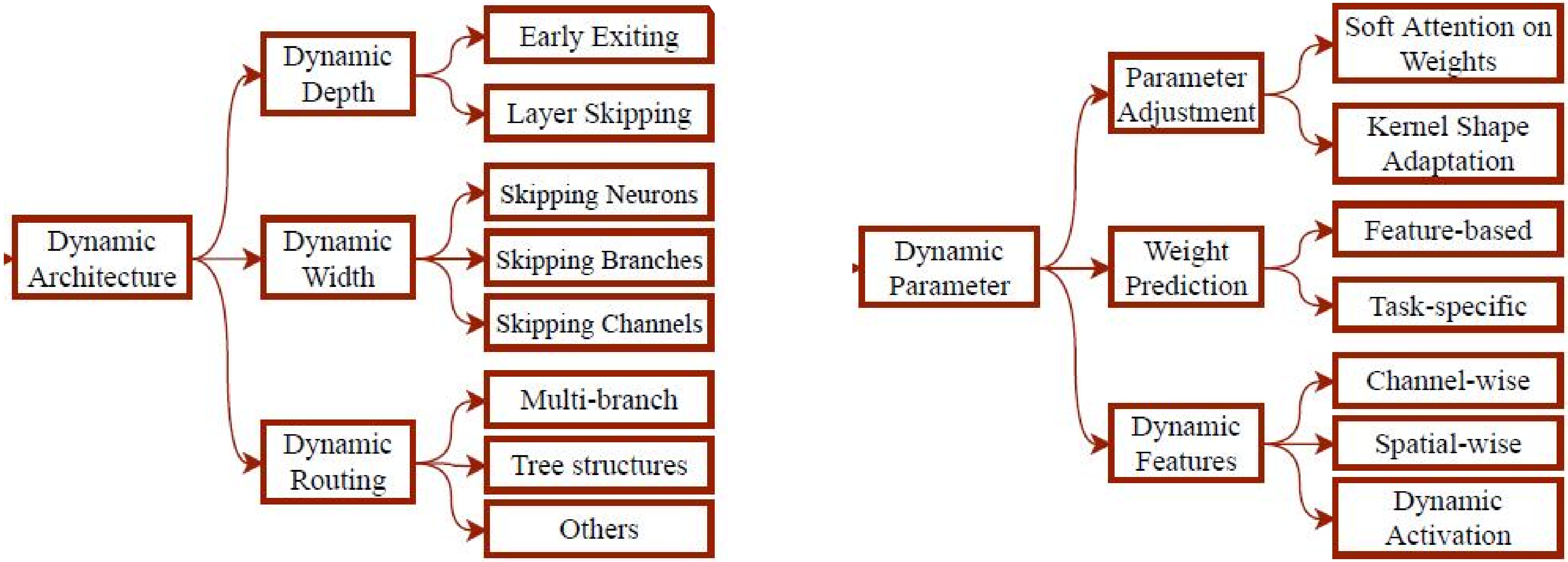
- **Accuracy**: *extra information*
- **Efficiency**: *partial activation*
- **Representation power**: *model capacity*
- **Adaptiveness**: *hardware platforms & environments*
- **Compatibility**: *advanced techniques in deep learning*
- **Generality**: *a wide range of applications*
- **Interpretability**: *process information in a dynamic way*

Overview

数据挖掘实验室
ning Lab



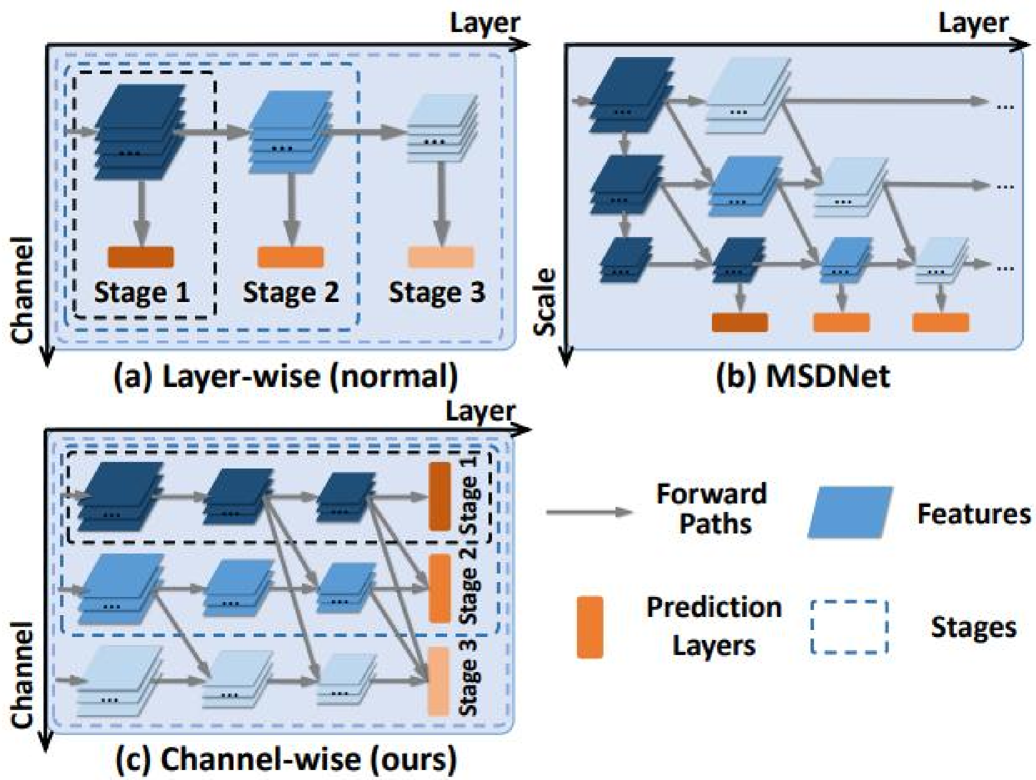
Overview


Dynamic Architecture
Dynamic Depth
Early Exiting
Layer Skipping
Dynamic Width
Skipping Neurons
Skipping Branches
Skipping Channels
Dynamic Routing
Multi-branch
Tree structures
Others
Dynamic Parameter
Parameter Adjustment
Soft Attention on Weights
Kernel Shape Adaptation
Weight Prediction
Feature-based
Task-specific
Dynamic Features
Channel-wise
Spatial-wise
Dynamic Activation

Dynamic Architecture & Parameter

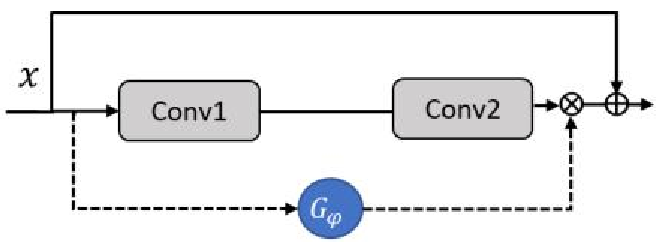
Dynamic Width – Convolutional Channel

- Multi-stage architecture
- Gating functions
- Dynamic Routing

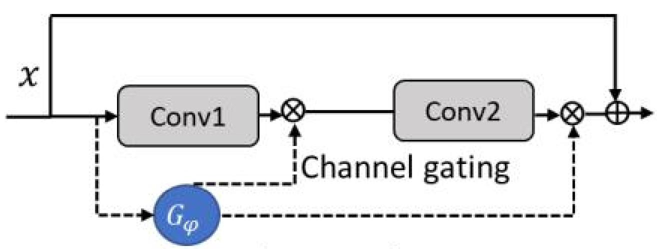


■ Dynamic Width – Convolutional Channel

- Multi-stage architecture
- Gating functions
- Dynamic Routing



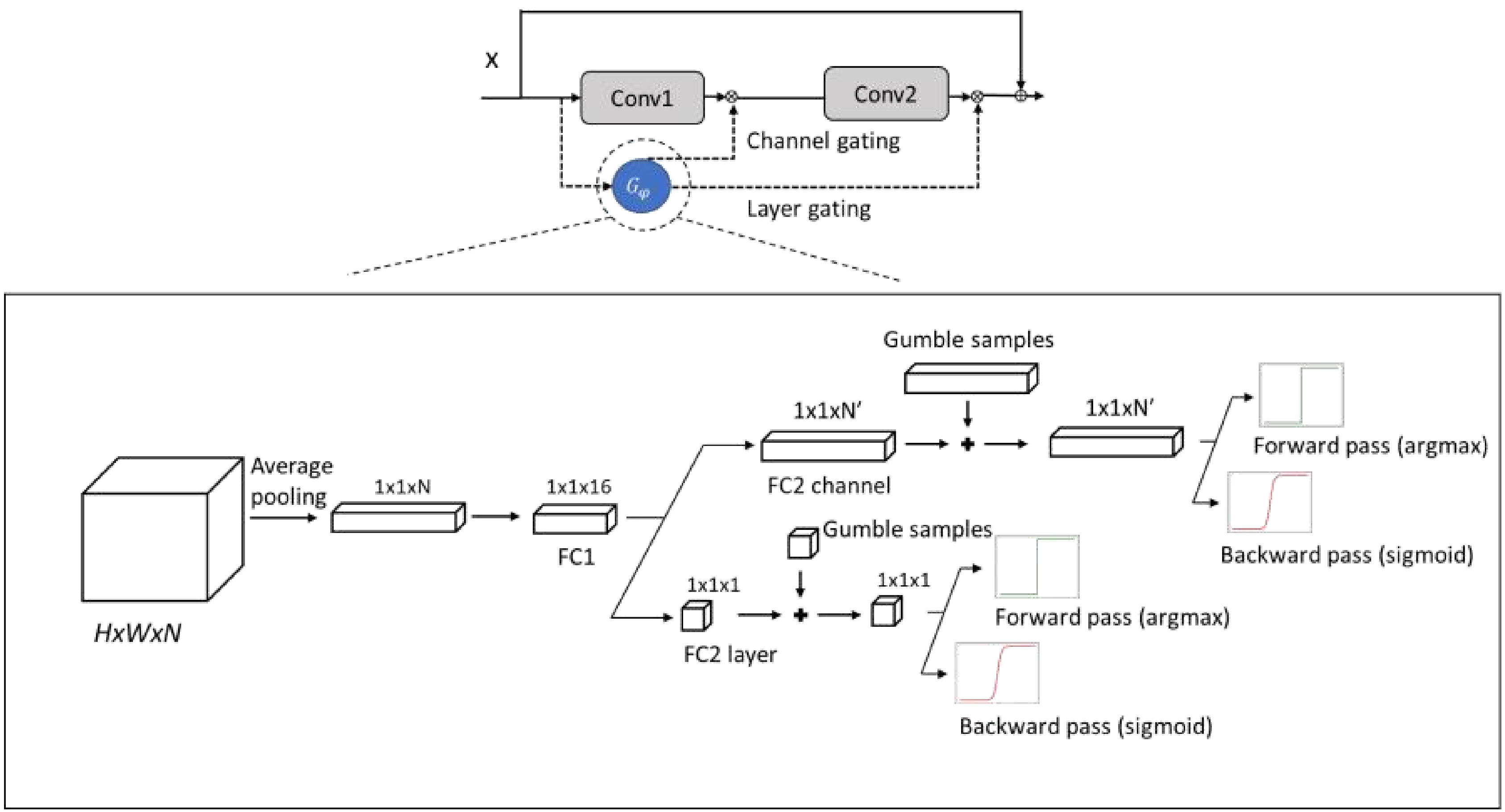
Layer gating



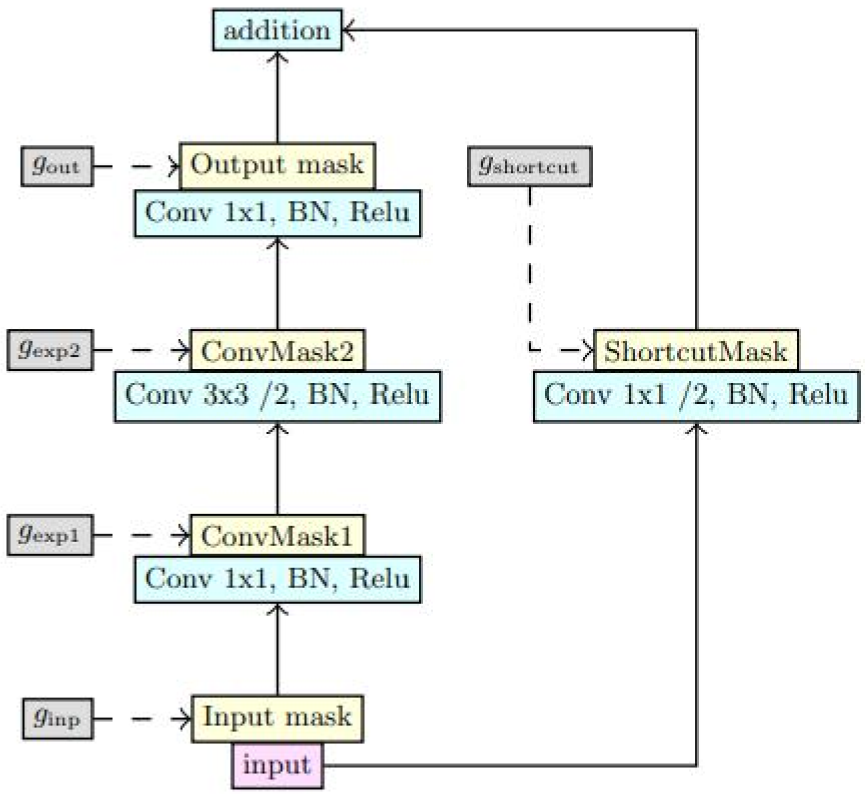
Layer gating

数据挖掘实验室
Data Mining Lab

■ Gating Function – Different Settings

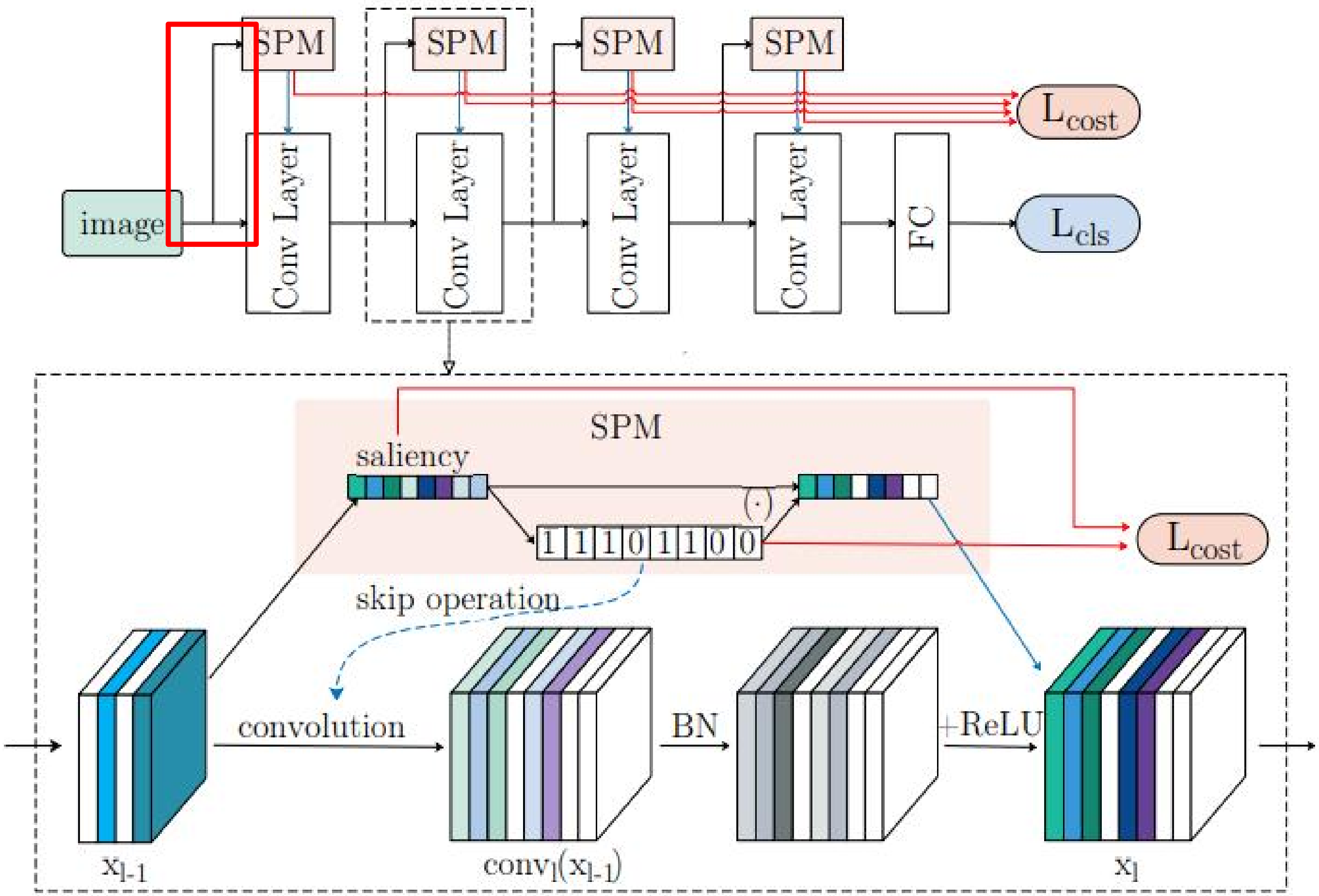


Gating Function – Different Settings



Channel selection using Gumbel Softmax

■ Gating Function – Different Settings



Self-adaptive Network Pruning

■ Gating Function – Different Settings

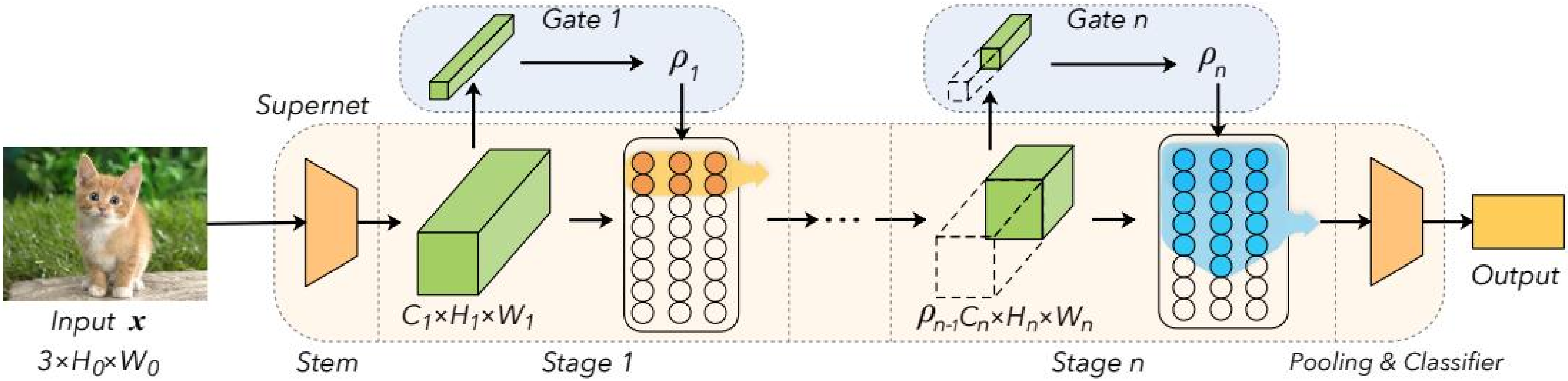


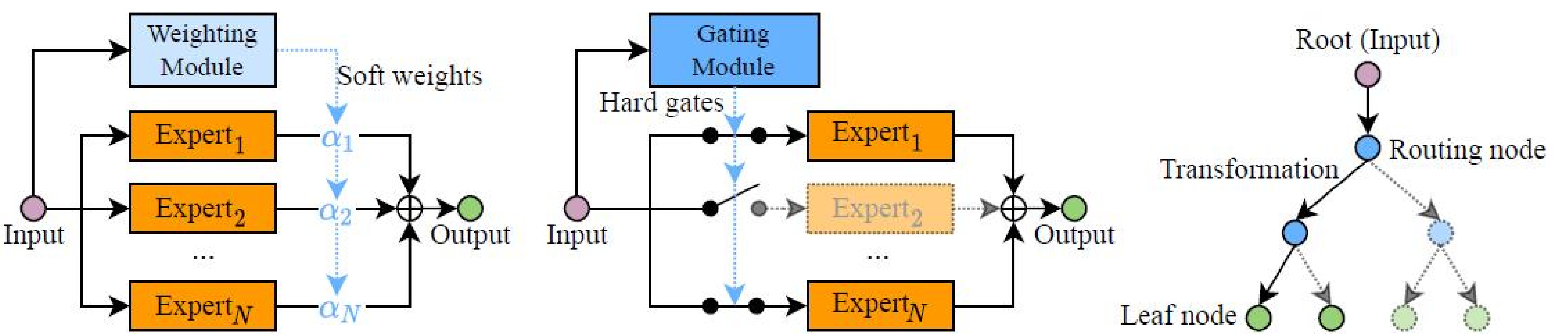
Table 1. Latency comparison of ResNet-50 with 25% channels (on GeForce RTX 2080 Ti). Both *masking* and *indexing* lead to inefficient computation waste, while *slicing* achieves comparable acceleration with *ideal* (the individual ResNet-50 0.25×).

method	full	masking	indexing	slicing (ours)	ideal
latency	12.2 ms	12.4ms	16.6 ms	7.9 ms	7.2 ms

Dynamic Architecture

■ Dynamic Width – Dynamic Routing

- Soft decision tree
- Neural trees & tree-structured
- Controller node / network

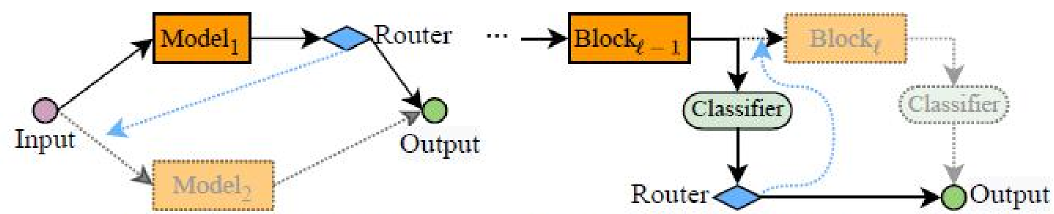


(a) Soft weights for adaptive fusion. (b) Selective execution of MoE branches. (c) Dynamic routing in a tree structure.

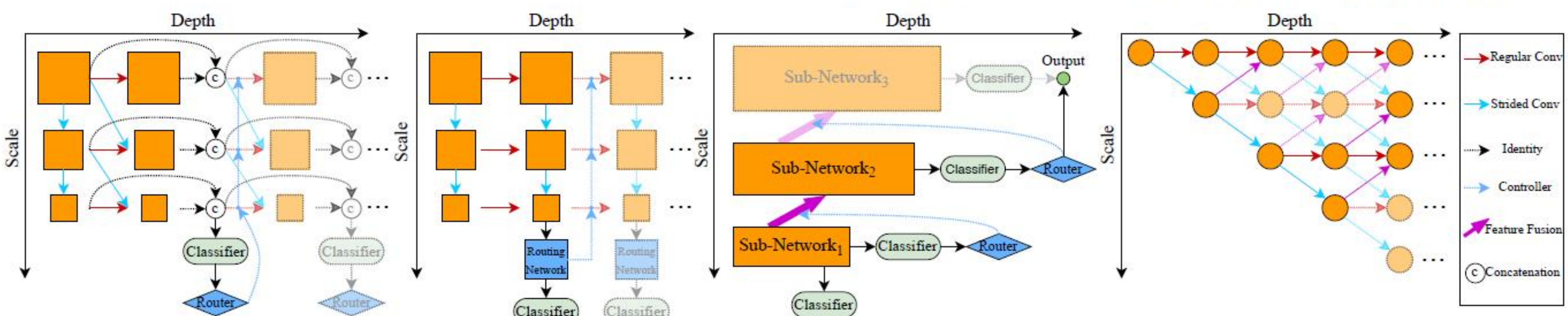
Dynamic Architecture

■ Dynamic Depth – Early exiting

- Cascading of DNNs
- Intermediate classifiers
- Multi-scale architecture with early exits



(a) Cascading of models. (b) Network with intermediate classifiers.

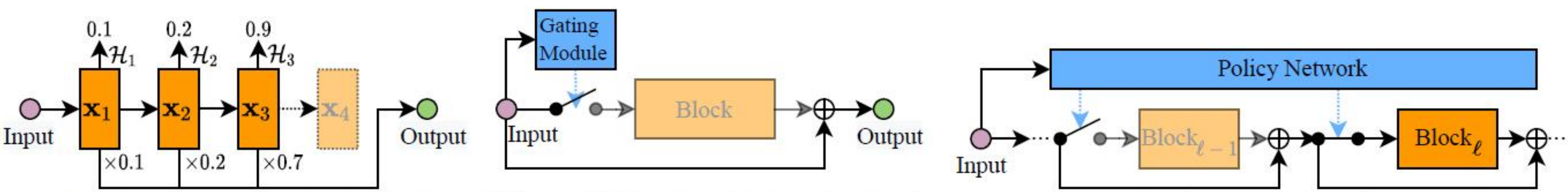


(a) Multi-scale DenseNet. (b) Early exiting with routing networks. (c) Resolution Adaptive Network. (d) Dynamic Routing inside a SuperNet.

Dynamic Architecture

■ Dynamic Depth – Layer skipping

- Halting Score
- Gating Function
- Policy Network



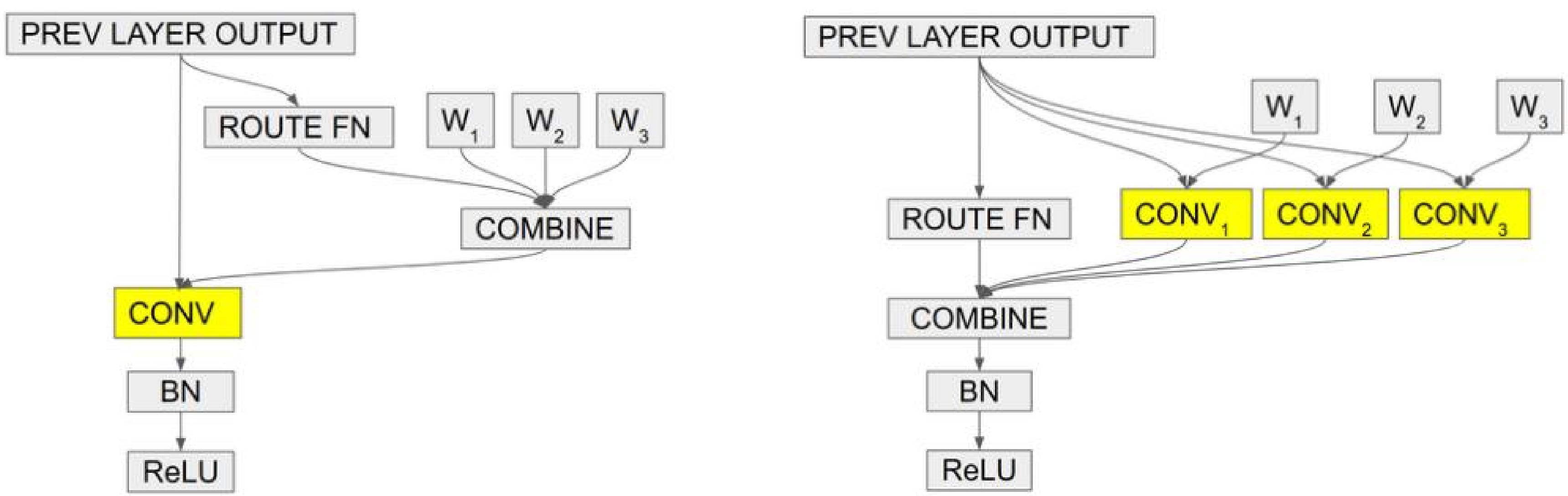
(a) Layer skipping based on halting score. (b) Layer skipping based on a gating function. (c) Layer skipping based on a policy network.

Dynamic Parameter

Parameter Adjustment

- Attention on weight
- Kernel shape adaptation

$$\mathbf{y} = \mathbf{x} \star \tilde{\mathbf{W}} = \mathbf{x} \star \left(\sum_{n=1}^{N} \boxed{\alpha_n} \mathbf{W}_n\right)$$

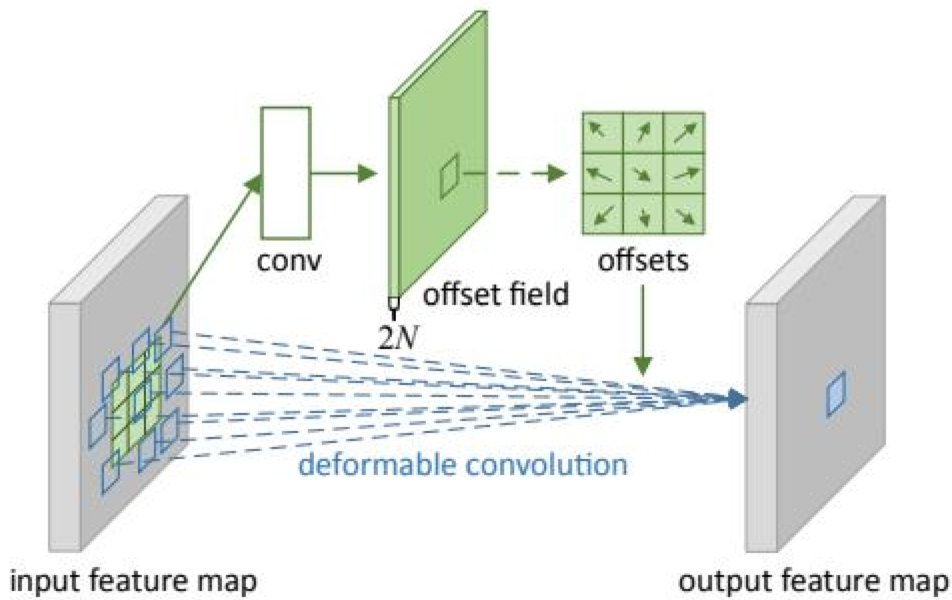


(a) CondConv: $(\alpha_1 W_1 + \ldots + \alpha_n W_n) * x$ (b) Mixture of Experts: $\alpha_1(W_1 * x) + \ldots + \alpha_n(W_n * x)$

Dynamic Parameter

■ Parameter Adjustment

- Attention on weight
- Kernel shape adaptation



Dynamic Parameter

■ Parameter Adjustment

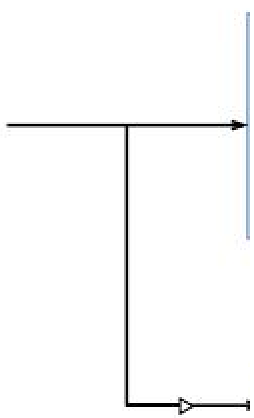
- Attention on weight
- Kernel shape adaptation

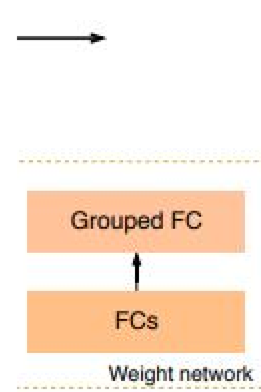
Dynamic Parameter

■ Weight

- Gen
- Task

Model	# Params	FLOPs	Top-1 err.
ShuffleNetV2 [22] (0.5×)	1.4M	41M	39.7
+ SE [11]	1.4M	41M	37.5
+ SK [16]	1.5M	42M	37.5
+ CondConv [38] (2×)	1.5M	41M	37.3
+ WeightNet (1×)	1.5M	41M	**36.7**
+ CondConv [38] (4×)	1.8M	41M	35.9
+ WeightNet (2×)	1.8M	41M	**35.5**
ShuffleNetV2 [22] (1.5×)	3.5M	299M	27.4
+ SE [11]	3.9M	299M	26.4
+ SK [16]	3.9M	306M	26.1
+ CondConv [38] (2×)	5.2M	303M	26.3
+ WeightNet (1×)	**3.9M**	301M	**25.6**
+ CondConv [38] (4×)	8.7M	306M	26.1
+ WeightNet (2×)	**5.9M**	**303M**	**25.2**
ShuffleNetV2 [22] (2.0×)	5.5M	557M	25.5
+ WeightNet (2×)	10.1M	565M	**23.7**
ResNet50 [7]	25.5M	3.86G	24.0
+ SE [11]	26.7M	3.86G	22.8
+ CondConv [38] (2×)	72.4M	3.90G	23.4
+ WeightNet (1×)	31.1M	3.89G	**22.5**

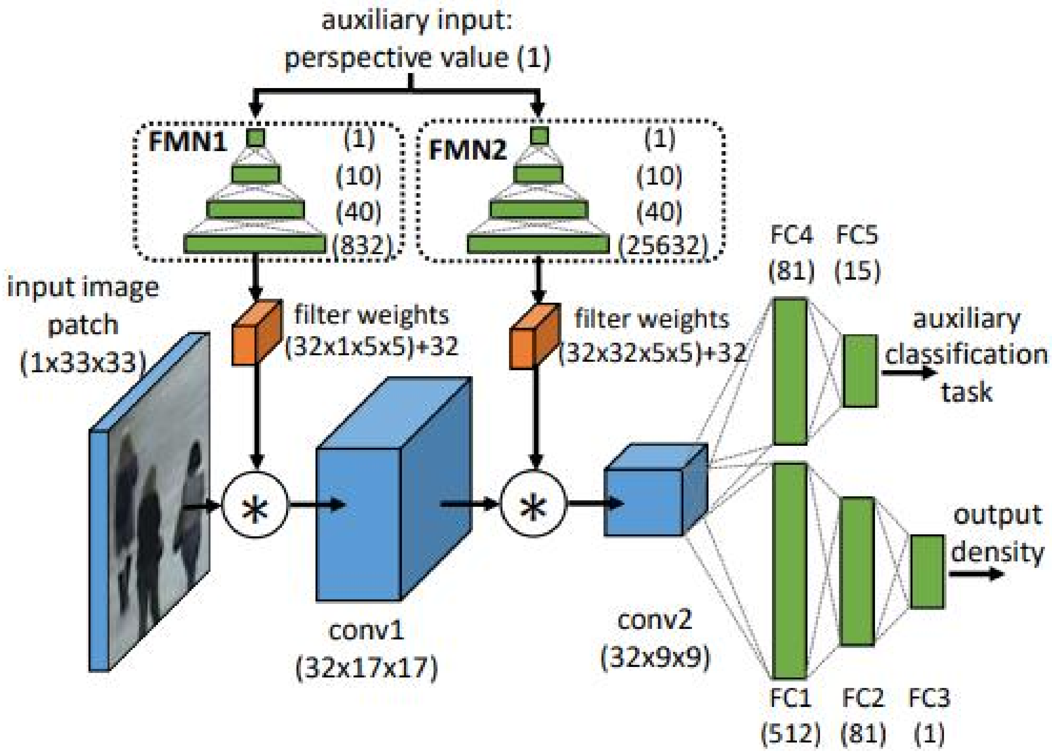




(b)

■ Weight Prediction

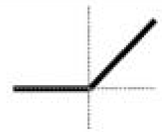
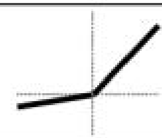
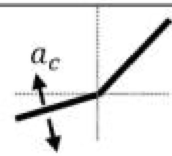
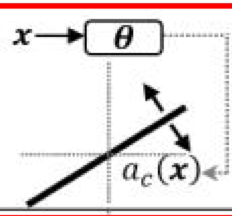
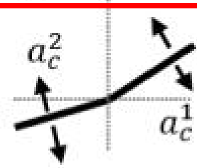
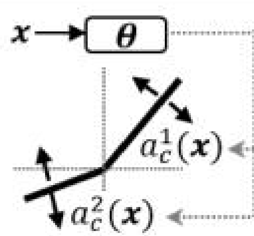
- General architectures
- Task-specific information



Dynamic Parameter

数据挖掘实验室
Data Mining Lab

- Dyn

		Type	K	relation to DY-ReLU
ReLU [27,17]		static	2	special case $a_c^1(x) = 1,\ b_c^1(x) = 0$ $a_c^2(x) = 0,\ b_c^2(x) = 0$
LeakyReLU [25]		static	2	special case $a_c^1(x) = 1,\ b_c^1(x) = 0$ $a_c^2(x) = \alpha,\ b_c^2(x) = 0$
PReLU [10]		static	2	special case $a_c^1(x) = 1,\ b_c^1(x) = 0$ $a_c^2(x) = a_c,\ b_c^2(x) = 0$
SE [14]		dynamic	1	special case $a_c^1(x) = a_c(x),\ b_c^1(x) = 0$ $0 \leq a_c(x) \leq 1$
Maxout [7]		static	1,2,3,...	DY-ReLU is a dynamic and efficient Maxout.
DY-ReLU		dynamic	1,2,3,...	identical

Dynamic Parameter

数据挖掘实验室
Data Mining Lab

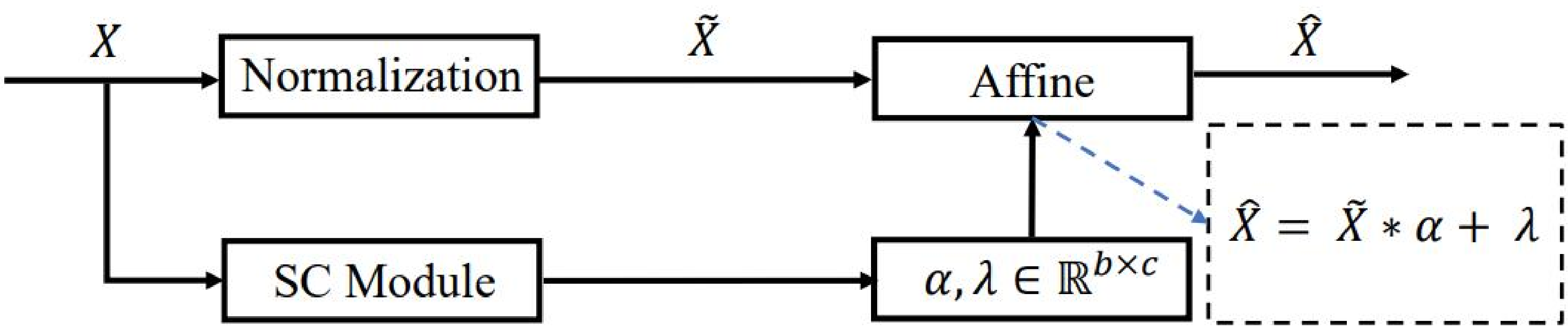
■ Dynamic Features

- Channel-wise attention
- Spatial-wise attention
- Dynamic activation functions

Activation	K	MobileNetV2 ×0.35			MobileNetV2 ×1.0		
		#Param	MAdds	Top-1	#Param	MAdds	Top-1
ReLU	2	1.7M	59.2M	60.3	3.5M	300.0M	72.0
RReLU [40]	2	1.7M	59.2M	$60.0_{(-0.3)}$	3.5M	300.0M	$72.5_{(+0.5)}$
LeakyReLU [25]	2	1.7M	59.2M	$60.9_{(+0.6)}$	3.5M	300.0M	$72.7_{(+0.7)}$
PReLU [10]	2	1.7M	59.2M	$63.1_{(+2.8)}$	3.5M	300.0M	$73.3_{(+1.3)}$
SE[14]+ReLU	2	2.1M	62.0M	$62.8_{(+2.5)}$	5.1M	307.5M	$74.2_{(+2.2)}$
Maxout [7]	2	2.1M	106.6M	$64.9_{(+4.6)}$	5.7M	575.8M	$75.1_{(+3.1)}$
Maxout [7]	3	2.4M	157.6M	$65.4_{(+5.1)}$	7.8M	860.2M	$75.8_{(+3.8)}$
DY-ReLU-B	2	2.7M	65.0M	$66.4_{(+6.1)}$	7.5M	315.5M	$\mathbf{76.2}_{(+4.2)}$
DY-ReLU-B	3	3.1M	67.8M	$\mathbf{66.6}_{(+6.3)}$	9.2M	322.8M	$\mathbf{76.2}_{(+4.2)}$

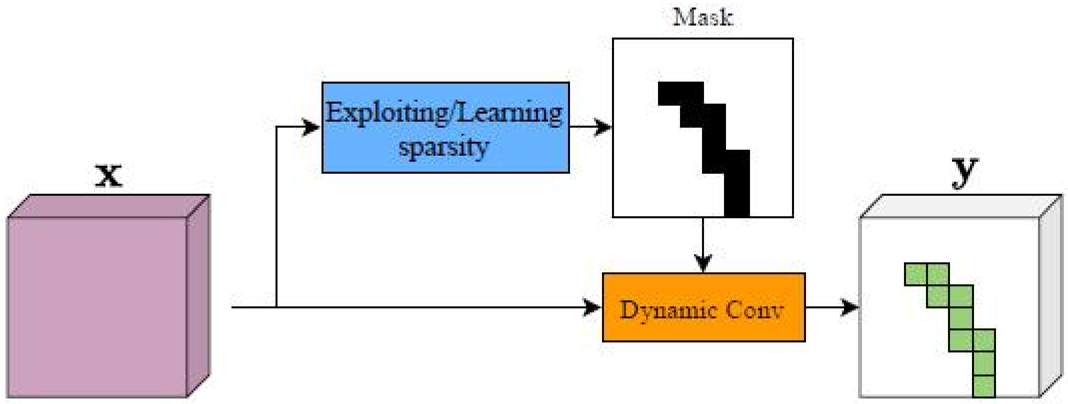
Dynamic Features

- Channel-wise attention
- Spatial-wise attention
- Dynamic activation functions



■ Pixel-level Dynamic Networks

- Dynamic sparse convolution
- Dynamic reception fields



Other Dynamic Networks

■ Region-level Dynamic Networks

- Dynamic transformations
- Hard attention on selected patches

■ Resolution-level Dynamic Networks

- Adaptive scaling ratios
- Dynamic resolution in multi-scale architectures

■ Temporal-wise Dynamic Networks

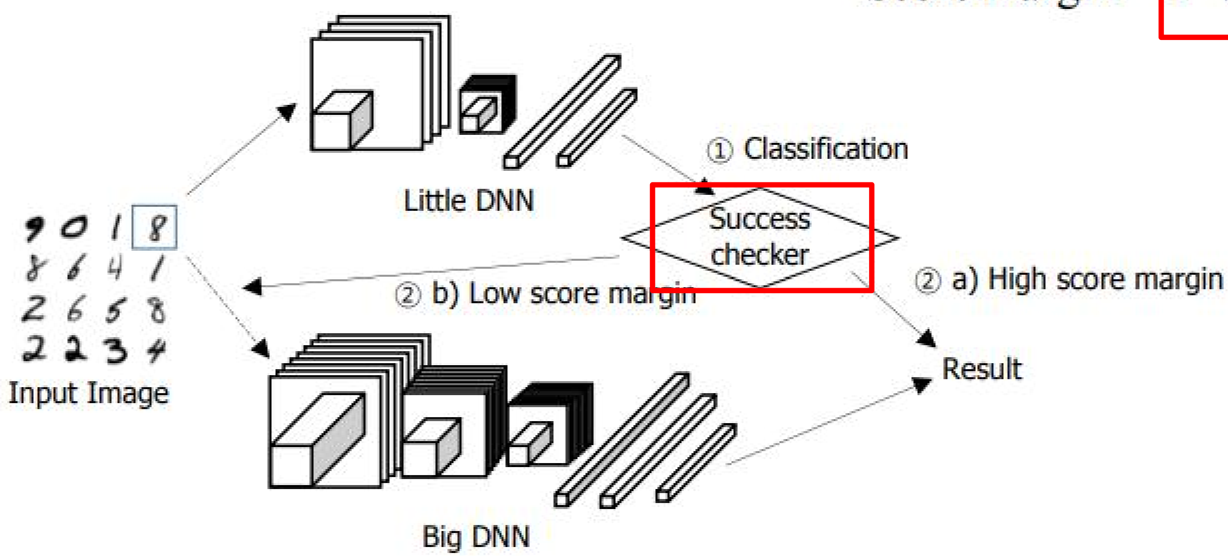
Inference & Training Tricks

Inference

■ Confidence-based Criteria

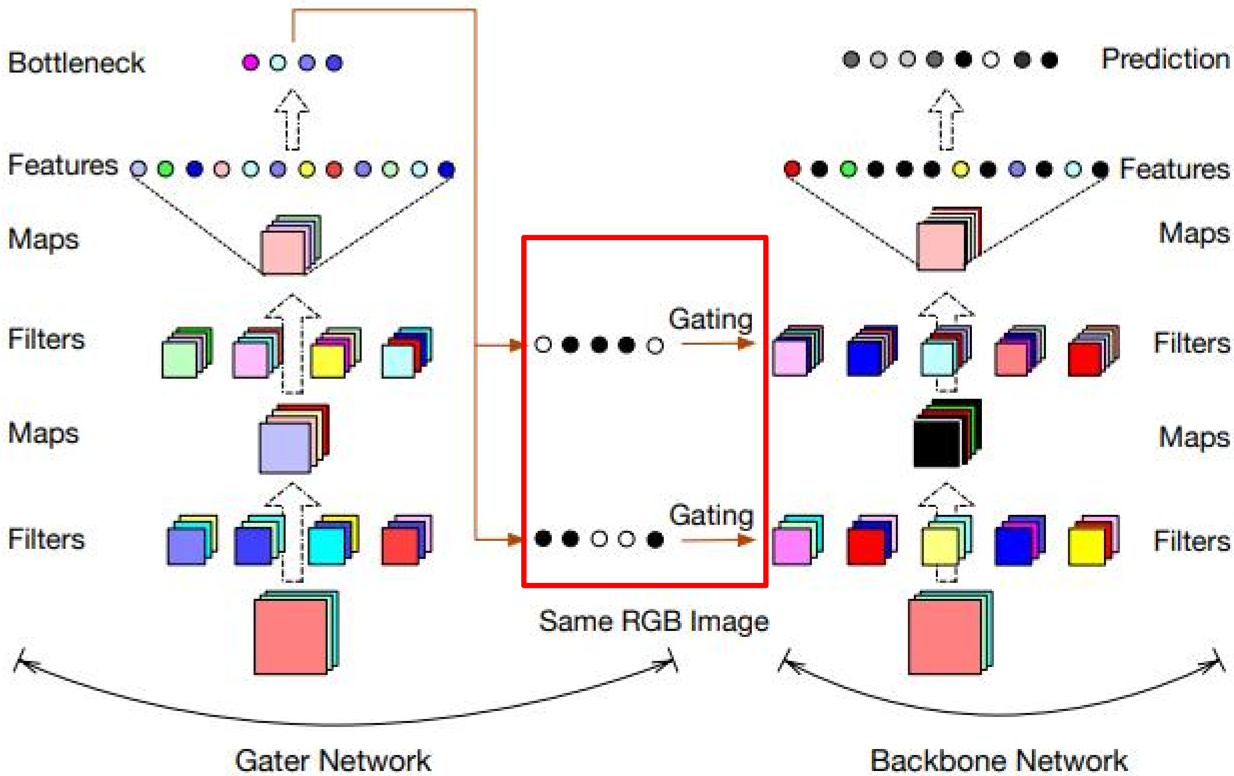
$$\text{entropy}(\boldsymbol{y}) = \sum_{c \in C} y_c \log y_c,$$

$$\text{Score margin} = \boxed{1^{\text{st}}\ \text{score}} - 2^{\text{nd}}\ \text{score}$$



数据挖掘实验室
Data Mining Lab

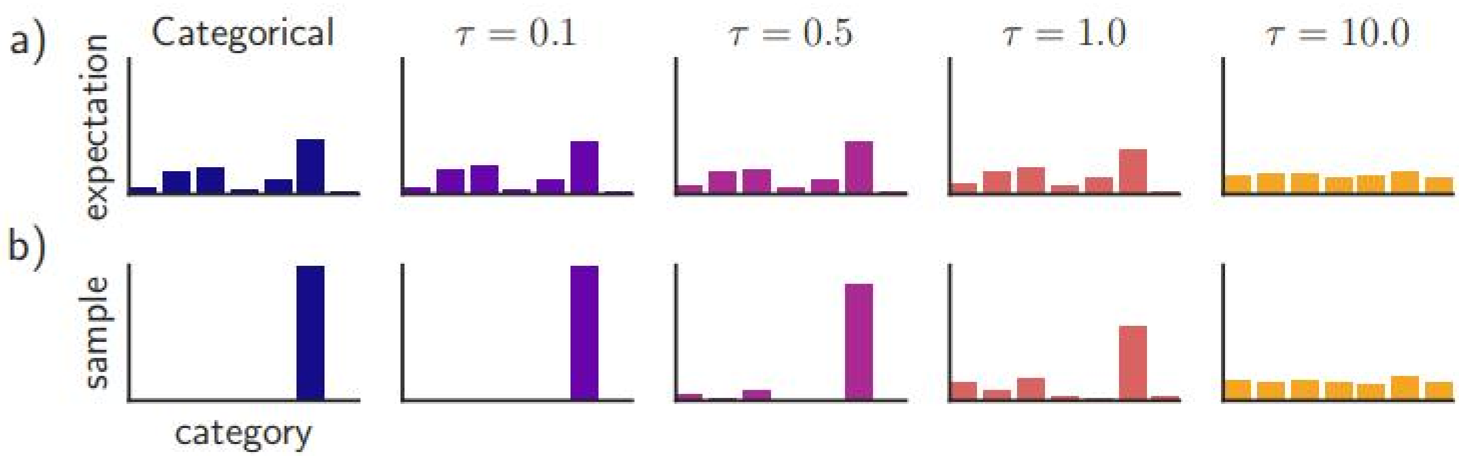
■ Policy Networks



■ Gating Functions – Gumbel Softmax

$$z = \text{one\_hot}(\underset{i}{\text{argmax}}[g_i + \log \pi_i])$$

$$y_i = \frac{\exp(\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp(\log(\pi_j) + g_j)/\tau)} \quad \text{for} \quad i = 1, \ldots, k$$



■ Gating Functions – Sigmoid

$$f(b_i, \beta_1, \beta_2) = \text{Sigmoid} \circ \text{Log}(b_i) = \frac{1}{1+(\frac{b_i}{\beta_1})^{-\beta_2}}$$

$$\mathbf{a}_l^t = \sigma\left(s\mathbf{e}_l^t\right)$$

■ Gating Functions – Hash

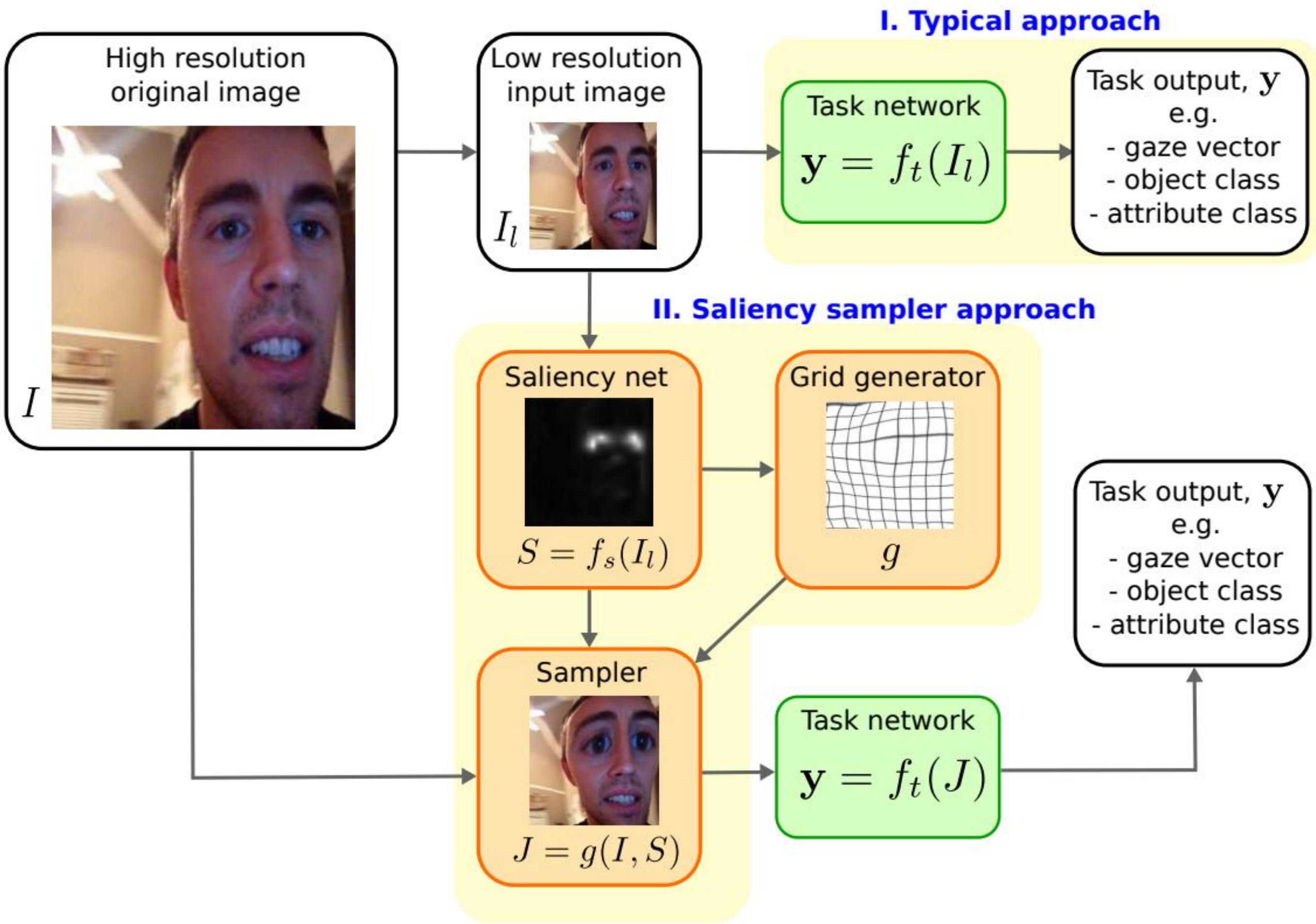
$$s_1 = \max(0, \min(1, a \cdot \sigma(s^l(x^{l-1}) + \xi) - b))$$

Training

- Training multi-exit networks
- Gradient estimation
- Reparameterization
- Reinforcement learning
- Encouraging sparsity
- Clustering hypothesis

Application & Discussion

- Fine-grained classification



■ Few-shot learning

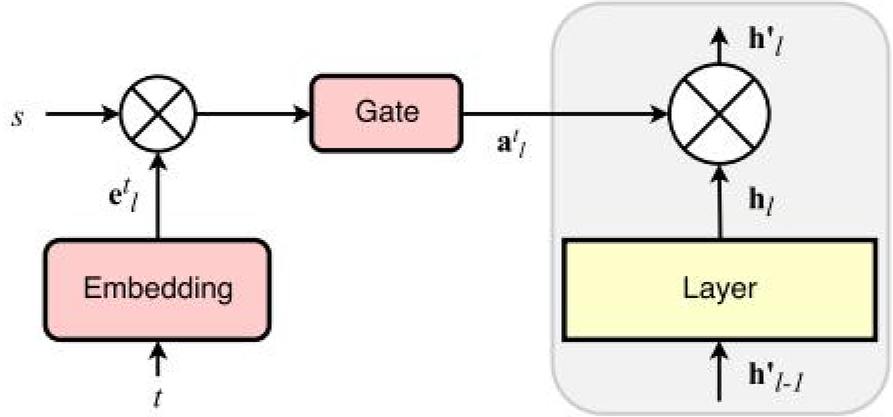
Table 6: Few-shot ImageNet Classification on ImageNet. Our model is competitive compared to the state-of-the-art meta learning model without hallucinator.

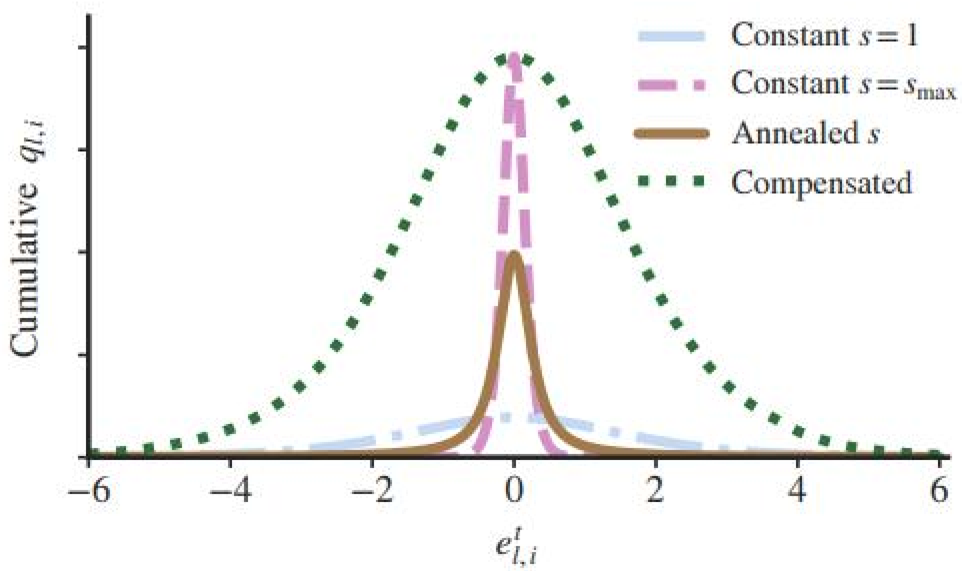
Method	Novel Top-5 Acc n=1	Novel Top-5 Acc n=2	All Top-5 Acc n=1	All Top-5 Acc n=2
LogReg [17]	38.4	51.1	40.8	49.9
PN [38]	39.3	54.4	49.5	61.0
MN [42]	43.6	54.0	54.4	61.0
TAFE-Net	43.0	53.9	**55.7**	**61.9**
LogReg w/ Analogies [17]	40.7	50.8	52.2	59.4
PN w/ G [45]	45.0	55.9	56.9	63.2

Method	aPY s	aPY H
LATEM	73.0	0.2
ALE [1	73.7	8.7
DeViSE	76.9	9.2
SJE [2]	55.7	6.9
ESZSL	70.1	4.6
SYNC	66.3	13.3
Relation	-	-
DEM [5	75.1	**19.4**
f-CLSW	-	-
SE† [41	-	-
SP-AEN	63.4	22.6
TAFE-N	75.4	**36.8**

数据挖掘实验室
Data Mining Lab

■ Continual learning



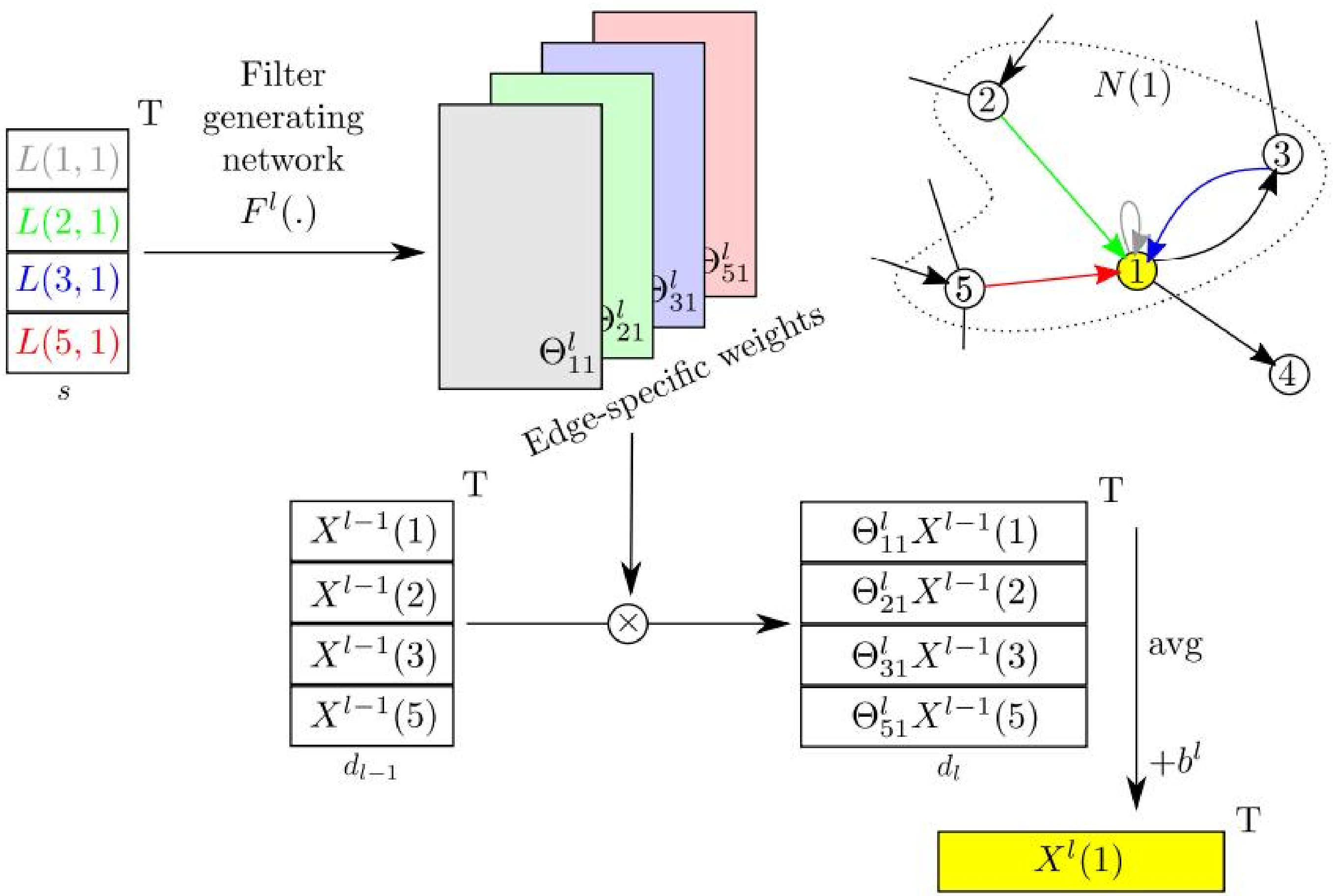


Application

■ Adversarial attack

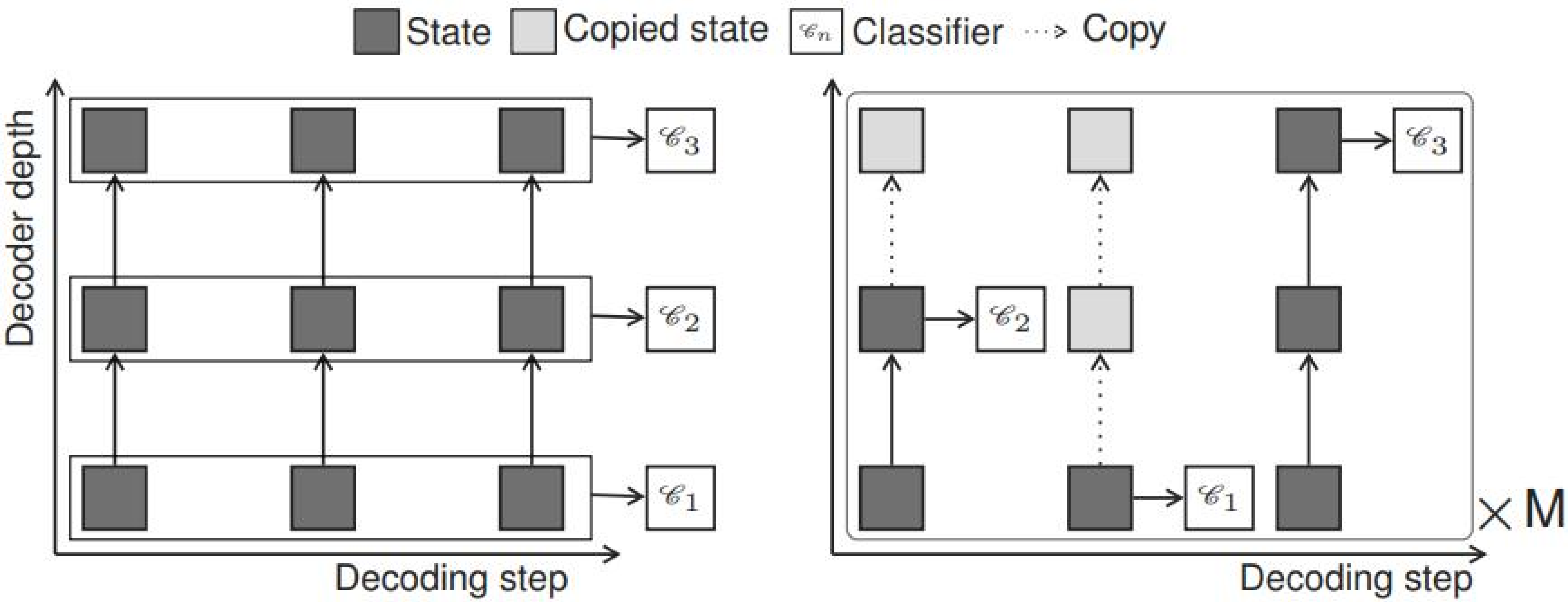
ADV. TRAINING	NO ATTACK	PGD-20	PGD-20 (AVG.)	PGD-20 (MAX.)	DEEPSLOTH
UNDEFENDED	0.77 / 89%	0.79 / 29%	0.85 / 10%	0.81 / 27%	**0.01** / 13%
PGD-10	0.61 / 72%	0.55 / 38%	0.64 / 23%	0.58 / 29%	**0.33** / 70%
PGD-10 (AVG.)	0.53 / 72%	0.47 / 36%	0.47 / 35%	0.47 / 35%	**0.32** / 70%
PGD-10 (MAX.)	0.57 / 72%	0.51 / 37%	0.54 / 30%	0.52 / 34%	**0.32** / 70%
OURS	0.74 / 72%	0.71 / 38%	0.82 / 14%	0.77 / 21%	**0.44** / 67%
OURS + PGD-10	0.61 / 73%	0.55 / 38%	0.63 / 23%	0.58 / 28%	**0.33** / 70%

- Graph learning



Application

■ Time serial model



(a) Aligned training

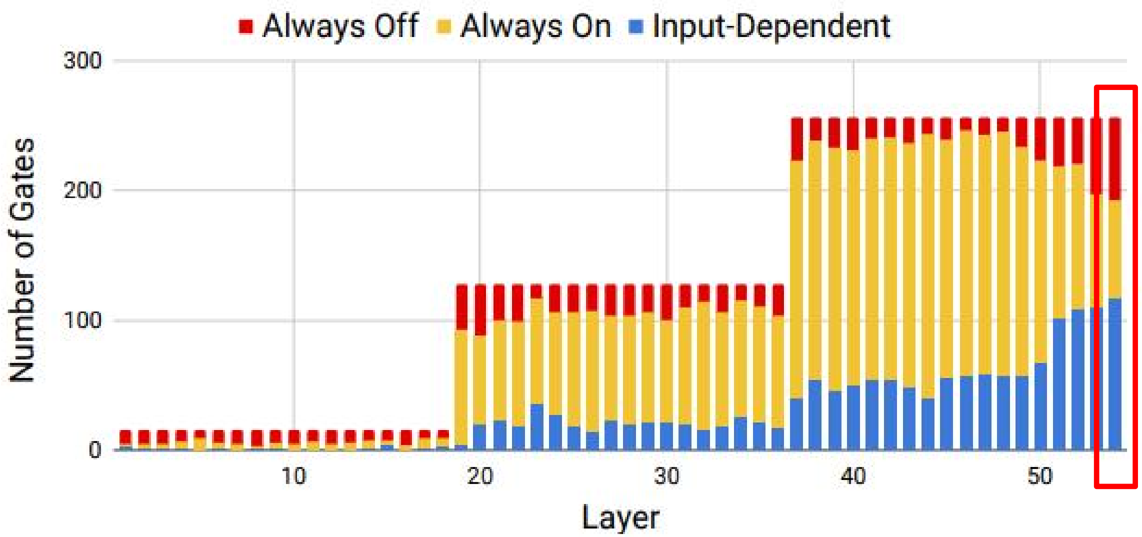
(b) Mixed training

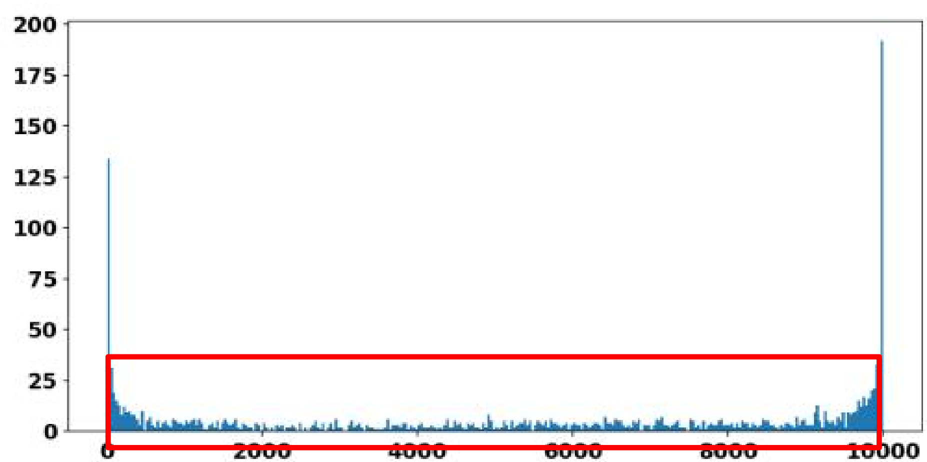
Discussion

- Scalability
- Privacy & Security
- Interpretability
- Effectiveness?

数据挖掘实验室
Data Mining Lab

■ Activation





Thank you