VALSE 2018, 大连, 2018.4.22

深度学习实践:
庖丁解牛与盲人摸象

吴建鑫
南京大学 教授 & MINIEYE 首席科学家

Learning And Mining from DatA

MINIEYE

What is my interest

in

deep learning?

Co-occurrence is **not** enough

（妈妈每天晚上）：彬彬好好睡觉哦，睡着以后会有狗狗猫猫陪彬彬玩的哦

（爸爸中午）：现在给你脱衣服，过会儿彬彬该干嘛呢？

（彬彬）：狗狗猫猫！

感谢彬彬小朋友提供此例

DL engineering- research heavy, too!

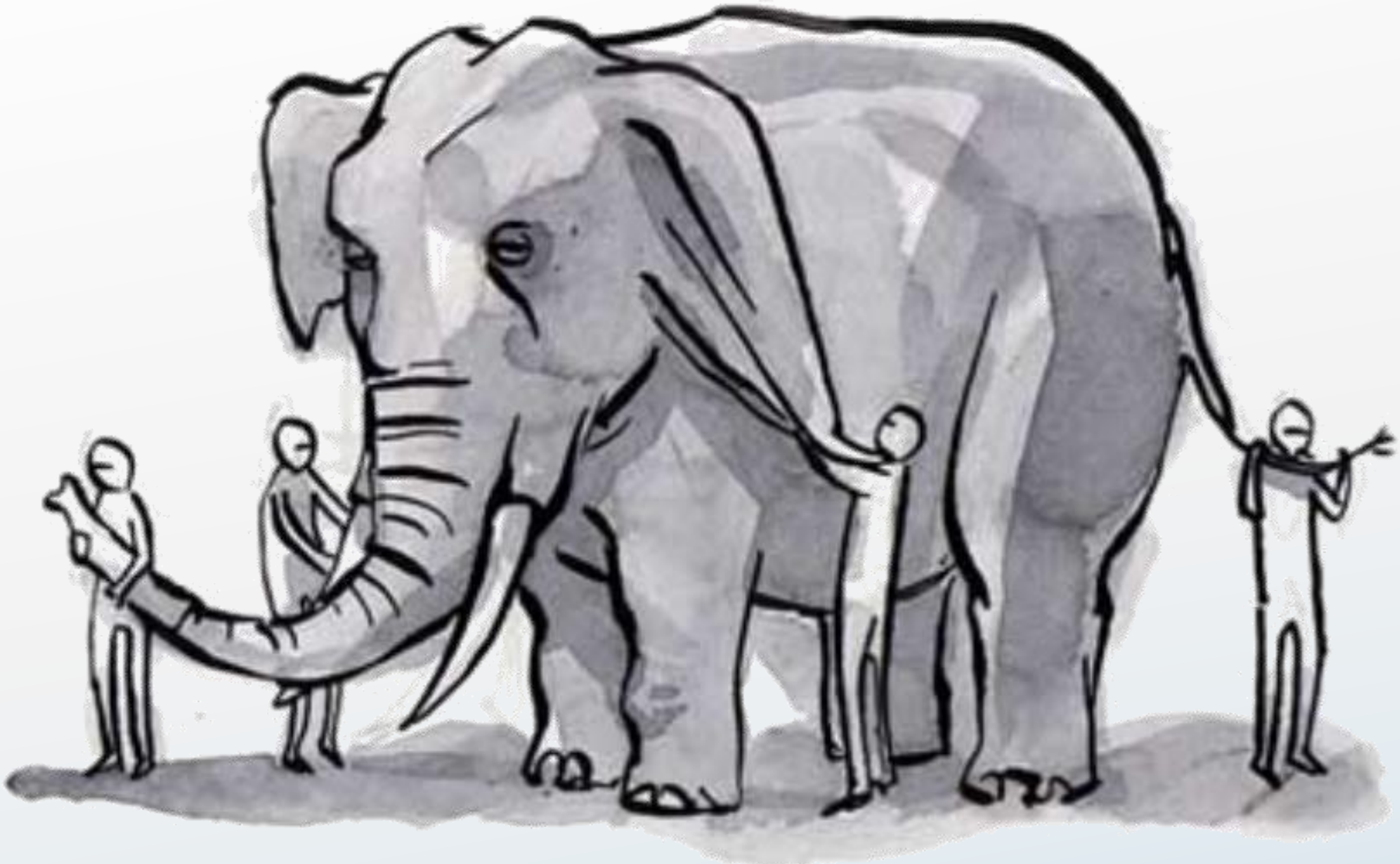
比亚迪的ADAS前装系统为MINIEYE提供

MINIEYE

Understanding & Accelerating

- What is in the representation?
- How is the representation generated?
  - Can we improve or customize it?

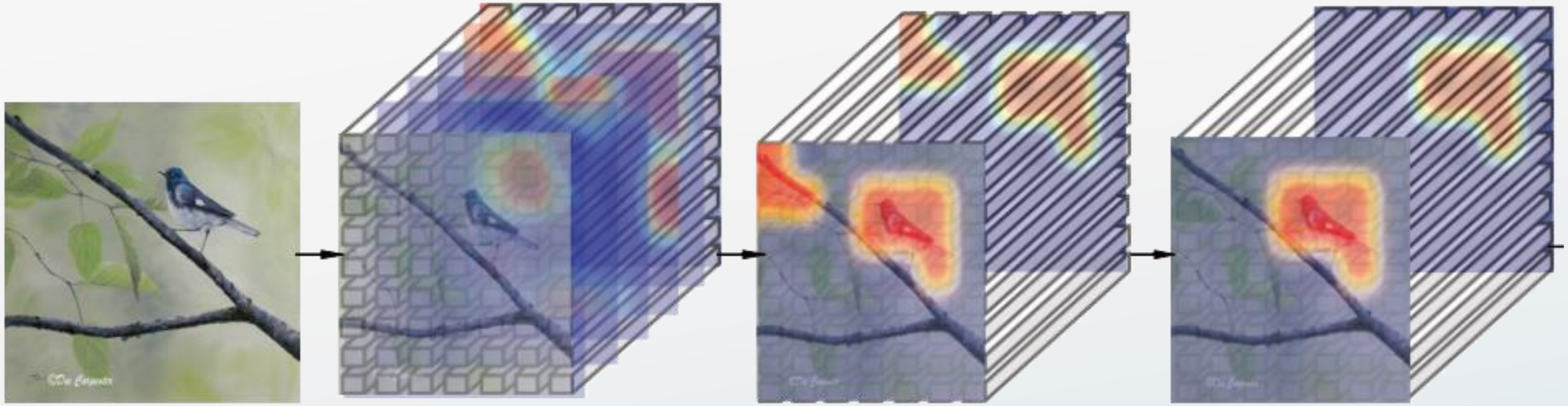
- How to accelerate the inference?
- What is the redundancy?



盲人摸象：利用已有DL模型或方法，即便我们不理解该表示，我们是否能做些什么？

SCDA: What is object?

Object vs. non-object (background)

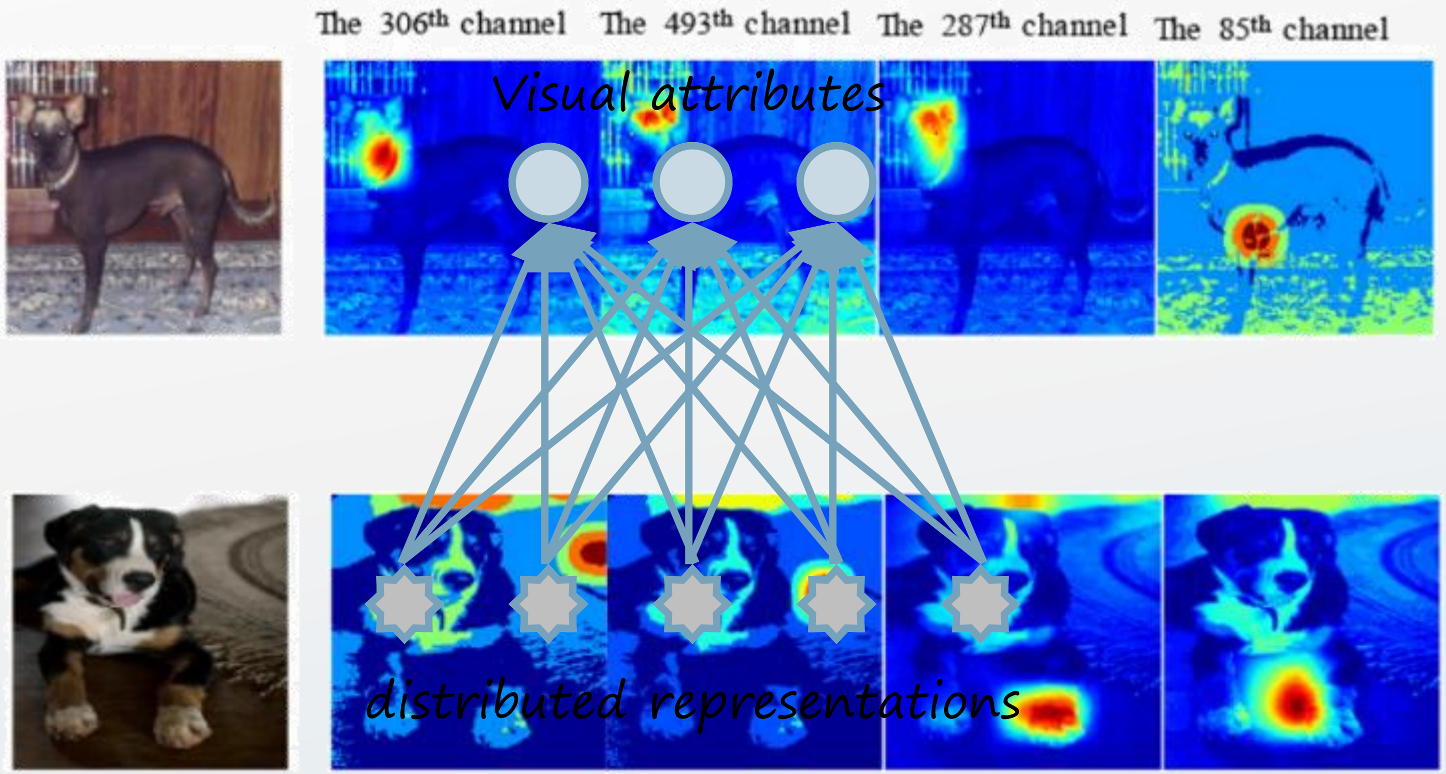


Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval

Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, Zhi-Hua Zhou

IEEE Transactions on Image Processing, 2017, 26(6): 2868-2881

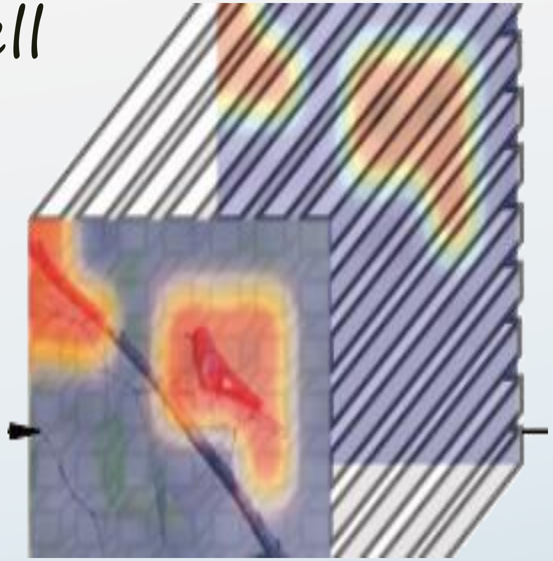
Distributed representation → Knowledge

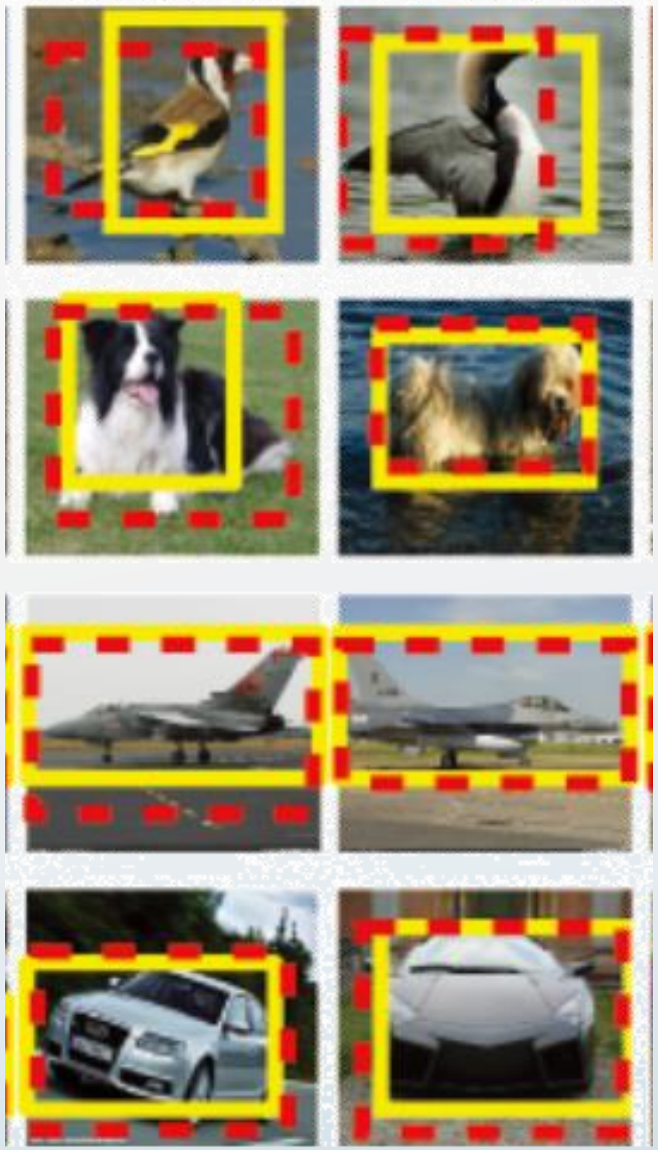


Objectness

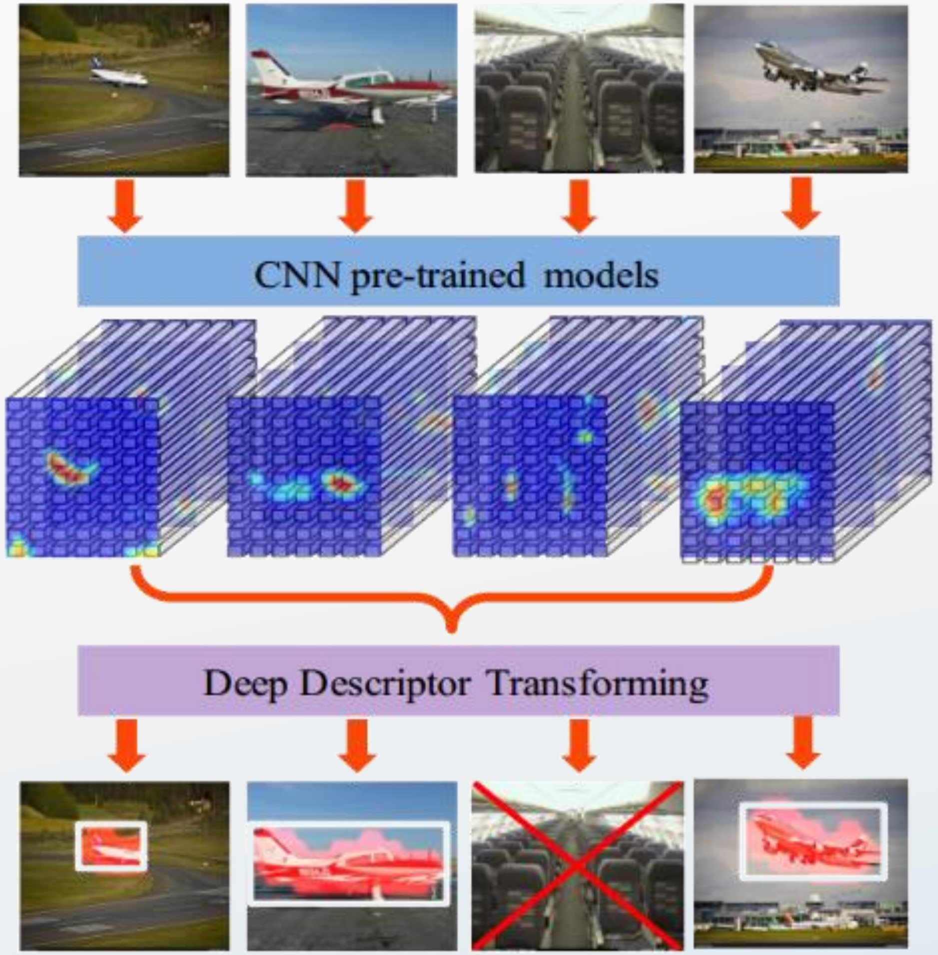
- Sum of channel responses as objectness
  - One score for every spatial position
  - Average scores over positions as threshold

- Can localize objects well





DDT: What is this object?



Deep Descriptor Transforming for Image Co-Localization

Xiu-Shen Wei, Chen-Lin Zhang*, Yao Li, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, Zhi-Hua Zhou*

International Joint Conference on Artificial Intelligence (IJCAI 2017), pp. 3048-3054

Representation for the common object

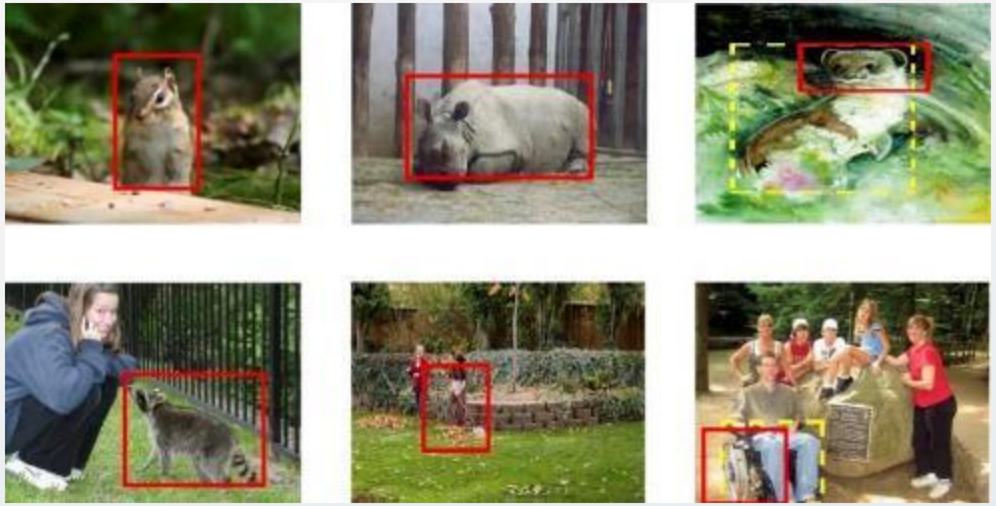
- *Given n images containing the same object*
  - What shall be the representation for this object?

- *The answer is simple*
  - One image → $h \times w \times d$ visual descriptor
  - n images → a large collection of visual descriptors

  - Find its principal component !
    - *Named as DDT --- "deep descriptor transforming"*
    - *Threshold is 0 !!*

DDT results

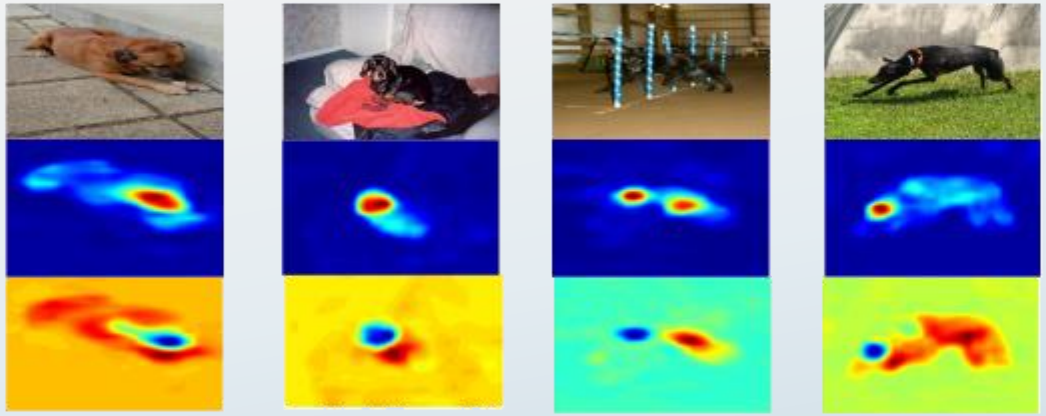
- Localizing 6 objects that are not in ILSVRC
  - Yes DDT generalized well!



- Numerical results

Methods	*Chipmunk*	*Rhino*	*Stoat*	*Racoon*	*Rake*	*Wheelchair*	**Mean**
[Cho *et al.*, 2015]	26.6	81.8	44.2	30.1	8.3	35.3	37.7
SCDA	32.3	71.6	52.9	34.0	7.6	28.3	37.8
[Li *et al.*, 2016]	44.9	81.8	67.3	41.8	14.5	39.3	48.3
Our DDT	**70.3**	**93.2**	**80.8**	**71.8**	**30.3**	**68.2**	**69.1**

- What about the 2^{nd} eigenvector?

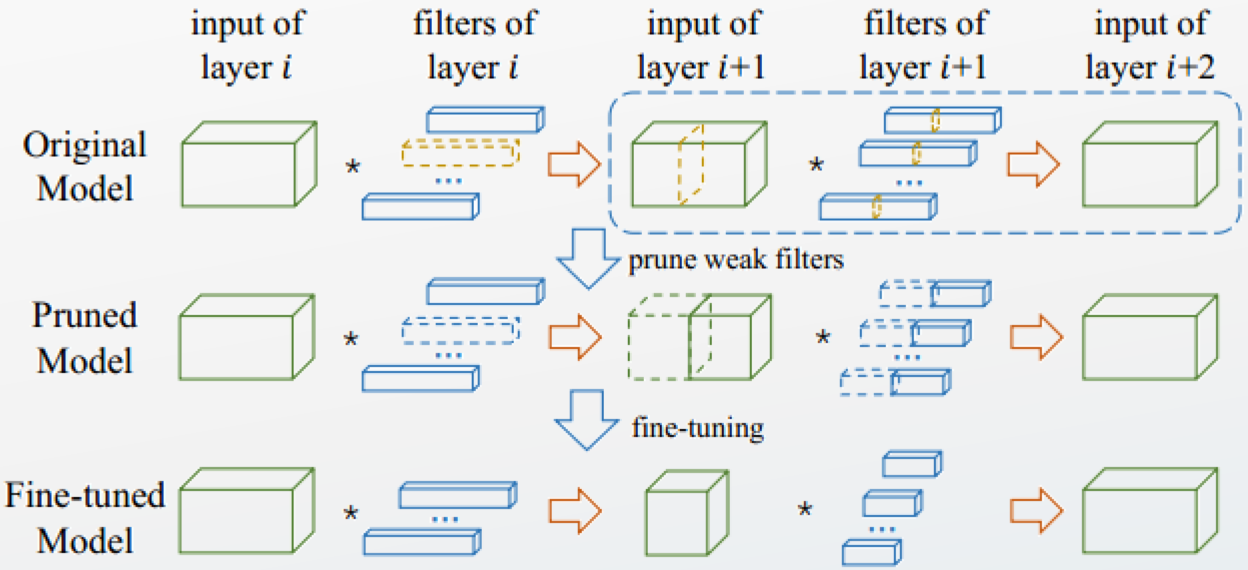


Where else are they useful?

- SCDA
  - Fine-grained image retrieval
  - Classic general purpose image instance retrieval
  - Grouping images for different attributes!

- DDT
  - Remember DDT is robust to exclude noise?
  - Improve webly supervised images
    - *Details: arXiv 1707.06397*

ThiNet: fast inference on hardware



ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression

Jian-Hao Luo, Jianxin Wu, Weiyao Lin

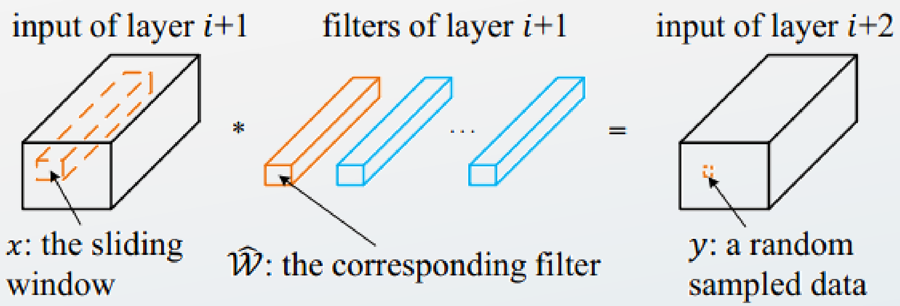
International Conference on Computer Vision (ICCV 2017), pp. 5058-5066

What to prune?

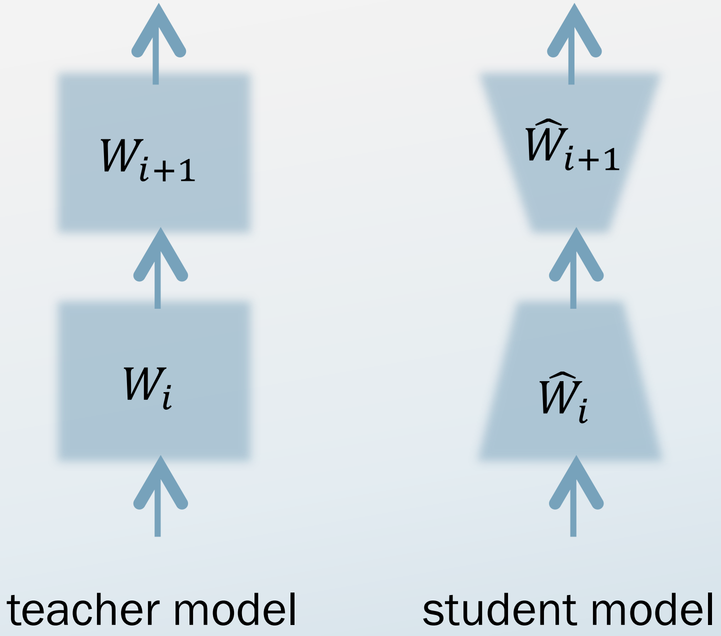
- Connections
  - $\begin{pmatrix} 0 & 0.18 & 0 \\ 0.23 & 0 & 0 \\ 0 & 0 & -0.07 \end{pmatrix}$
- Flops vs. running time?
  - 80% sparsity
  - Dedicated sparse convolution software
  - Speed versus dense kernel?
    - *Sparse is 3x slower*
- Filter
  - Can treat as group sparse
  - Best implementation in
    - CPU
    - GPU
    - FPGA
    - ASIC
    - …

How to prune?

- A technical view
  - Key: use the next layer's activation to guide which filter should be removed in the current one



- An alternative explanation
  - Intuitive but slow

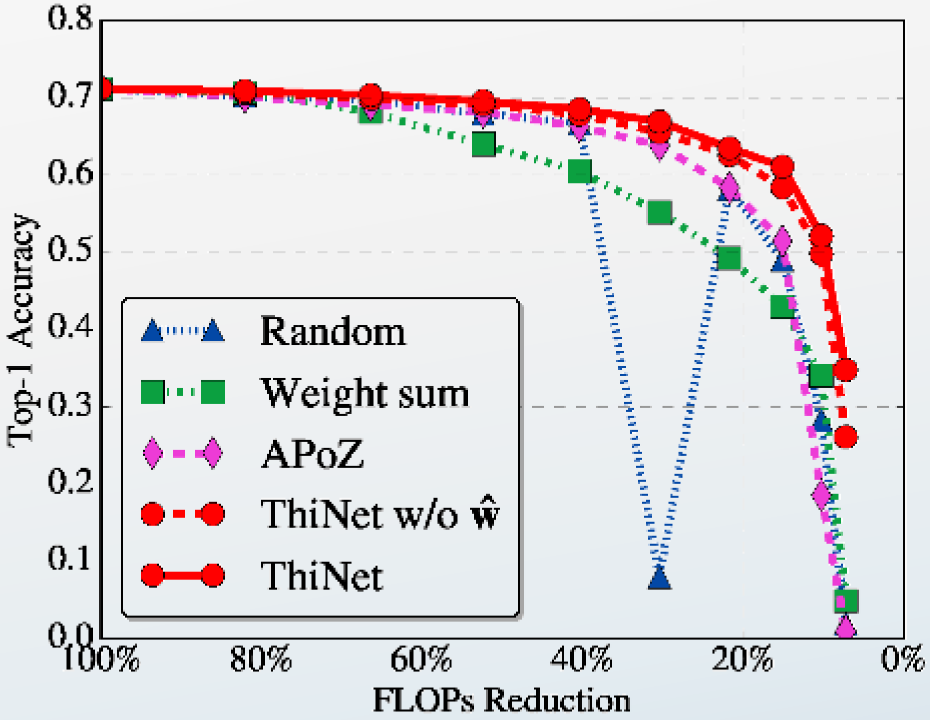


How to evaluate?

- Speed
  - On actual hardware
- Accuracy
  - Two most competitive baselines
    - Random pruning!
    - Train from scratch!

    &

    - Generalization ability!



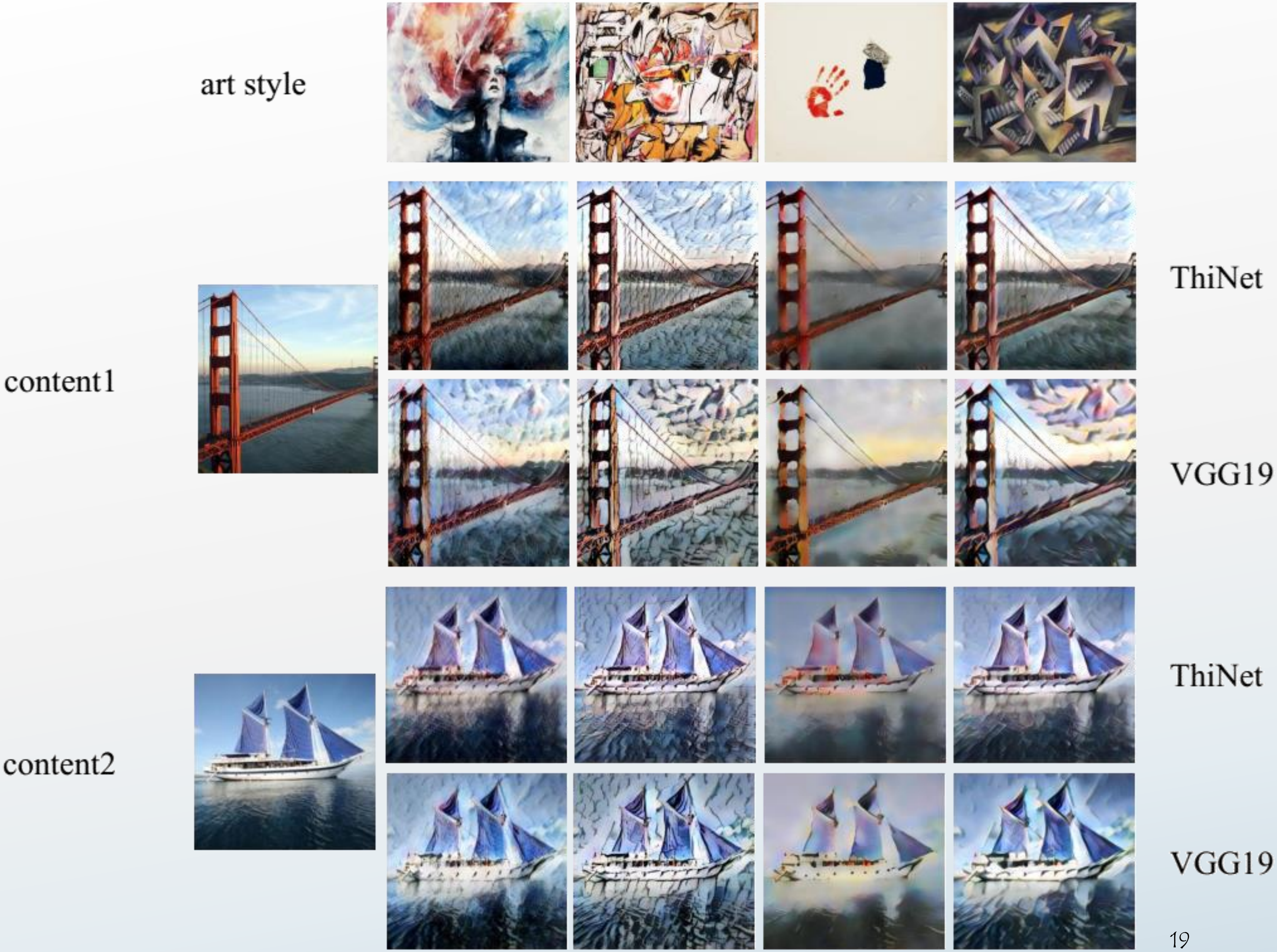
ThiNet applications

Ongoing work (improved upon ICCV paper)

- ThiNet models
  - Tiny: 2.66MB disk space
  - Small: 4.67MB
- Detection

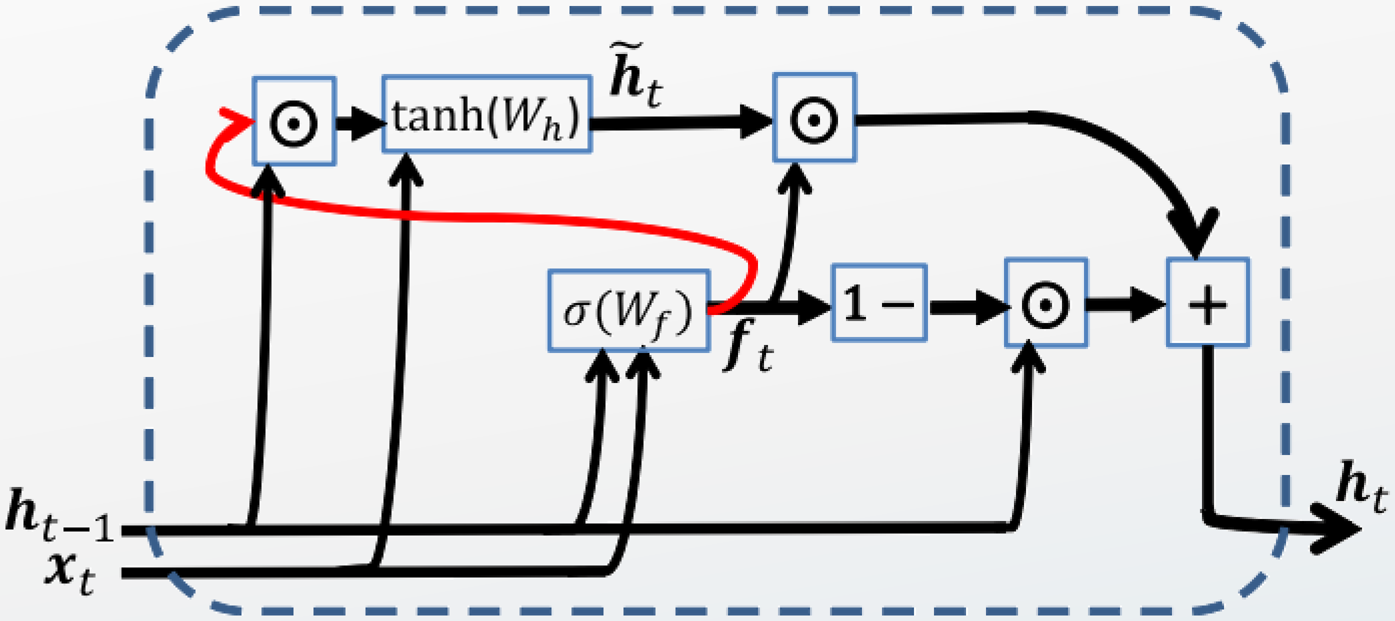
Model	Size	FPS	mAP
AlexNet	21.3MB	93	51.7
SqueezeNet	17.1MB	68	59.1
ThiNet-Tiny	13.5MB	69	55.0
SqueezeNet-DSD	17.1MB	68	43.9
ThiNet-Small	16.1MB	45	66.4
SSD300 [44]	105.2MB	22	77.2
Fast-YOLO [46]	180MB	89	52.7
Tiny-YOLO [47]	63.5MB	66	57.1

- More
  - Tested on CPU/GPU/ARM/FPGA
  - More applications
    - Image classification
    - Image retrieval
    - Semantic segmentation
    - Style transfer


art style
content1
content2
ThiNet
VGG19
ThiNet
VGG19

庖丁解牛：改进DL模型或方法，我们能解决什么问题？或者，获得什么理解？

MGU— RNN Gates: to have or not to have



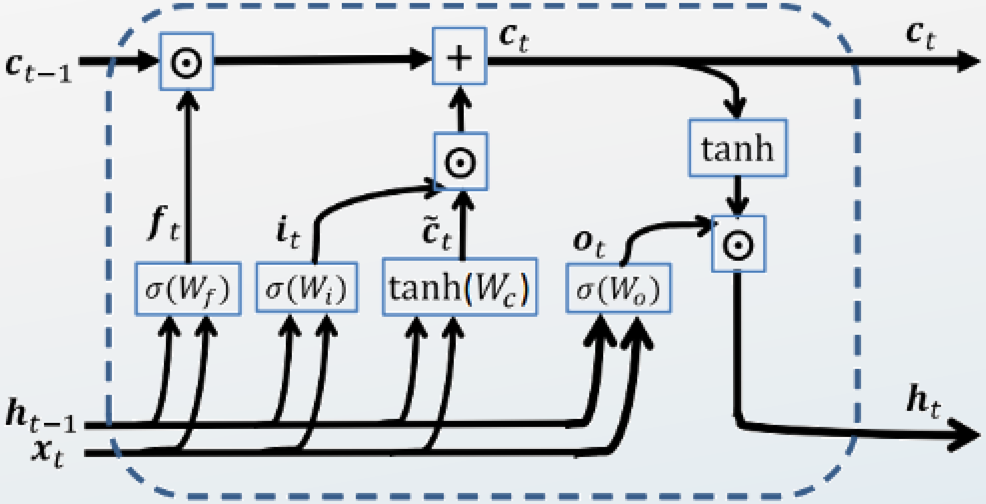
Minimal Gated Unit for Recurrent Neural Networks

Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, Zhi-Hua ZHou

International Journal of Automation and Computing, 13(3), 2016: pp. 226-234

LSTM: 3G

- Long short-term memory
  - Avoids gradient vanishing / exploding



- 3G – 3 gates

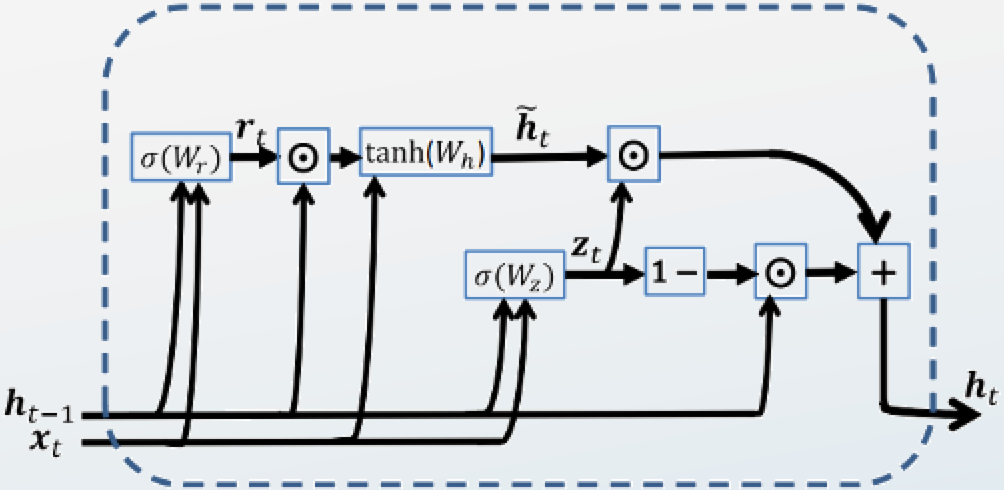
$$\boxed{\begin{aligned} \boldsymbol{f}_t &= \sigma\left(W_f\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right), \\ \boldsymbol{i}_t &= \sigma\left(W_i\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right), \\ \boldsymbol{o}_t &= \sigma\left(W_o\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_o\right), \end{aligned}}$$

$$\begin{aligned} \tilde{\boldsymbol{c}}_t &= \tanh\left(W_c\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_c\right), \\ \boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \\ \boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t). \end{aligned}$$

Long ShortTerm Memory
Sepp Hochreiter and Jurgen Schmidhuber
Neural Computation, 9(8):1735-1780,1997

GRU: 2G

- *Gated recurrent unit*
  - Also widely used
  - Similar accuracy with LSTM
  - Faster speed



- *2G – 2 gates*

$$\boxed{\begin{aligned} \boldsymbol{z}_t &= \sigma\left(W_z\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_z\right), \\ \boldsymbol{r}_t &= \sigma\left(W_r\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_r\right), \end{aligned}}$$

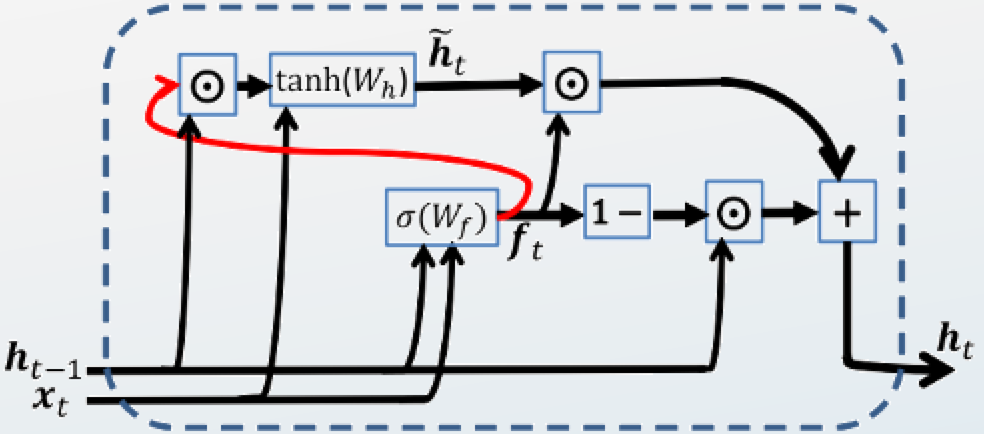
$$\tilde{\boldsymbol{h}}_t = \tanh\left(W_h\left[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_h\right),$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t.$$

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
Kyunghyun Cho, Bart van Merienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio
Empirical Methods in Natural Language Processing (EMNLP), pages 1724-1735, 2014

Minimal gated unit: 1G

- Easier for analysis
- Good performance
  - Faster speed & smaller unit
  - Comparable accuracy



- Only 1 gate
  - which is minimal
  - it is necessary to be gated

$$\boxed{\boldsymbol{f}_t = \sigma\left(W_f\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right)\ ,}$$

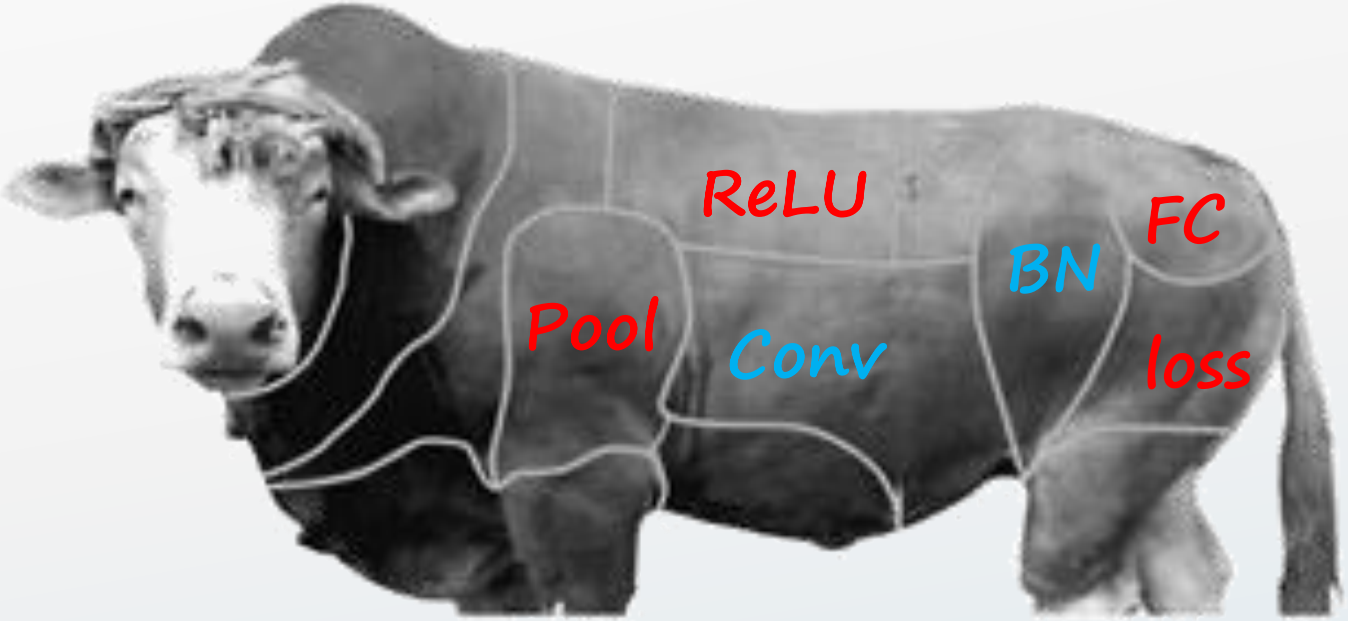
$$\tilde{\boldsymbol{h}}_t = \tanh\left(W_h\left[\boldsymbol{f}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_h\right)\ ,$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{f}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{f}_t \odot \tilde{\boldsymbol{h}}_t\ .$$

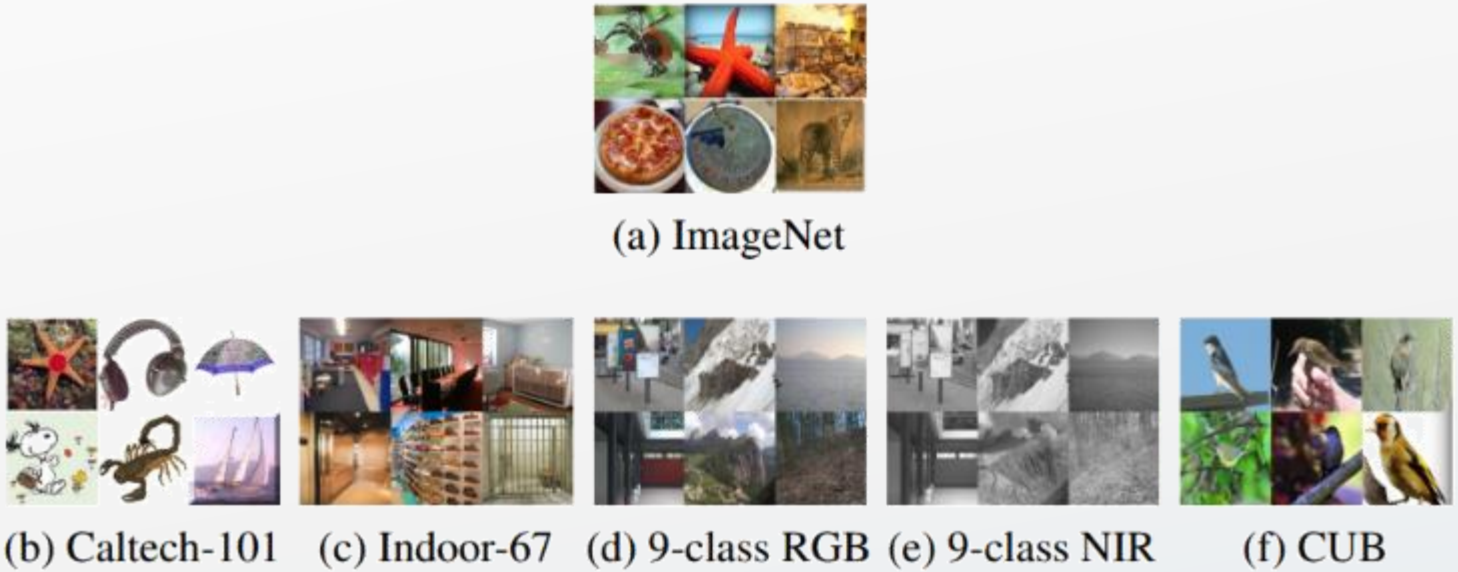
"Improving speech recognition by revising gated recurrent units", InterSpeech 2017

"Parameter Compression of Recurrent Neural Networks and Degradation of Short-term Memory", IJCNN 2017

The CNN “cattle”



FC: this obscured layer is a firewall



In Defense of Fully Connected Layers in Visual Representation Transfer

Chen-Lin Zhang, Jian-Hao Luo, Xiu-Shen Wei, Jianxin Wu

Pacific-Rim Conference on Multimedia (PCM 2017)

Better generalization via FC

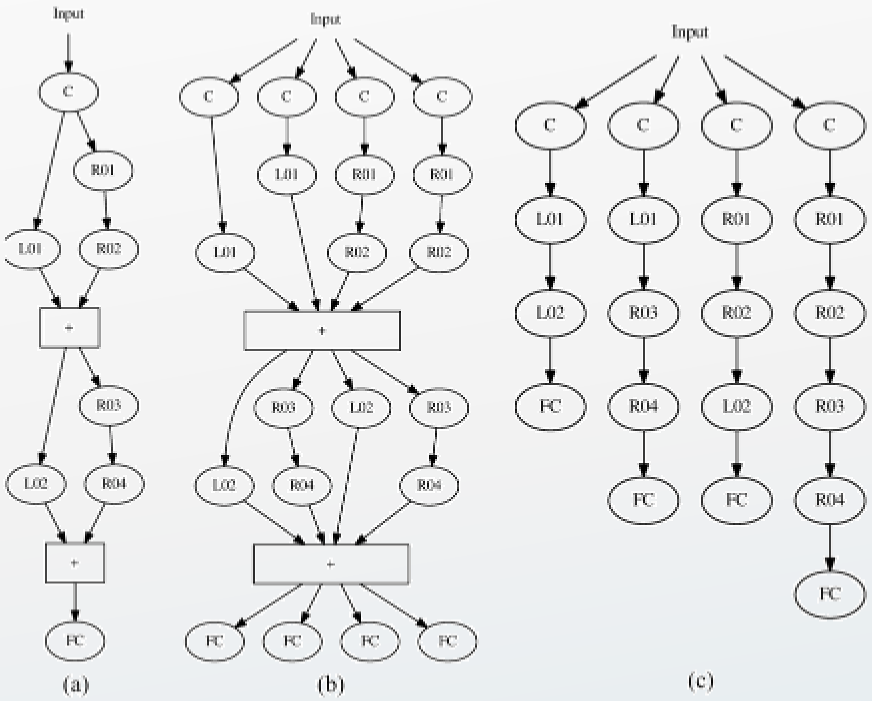
Recognition

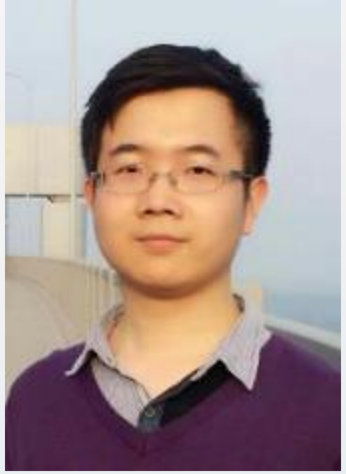
	FC	Caltech-101	indoor-67	RGB scene	NIR scene	CUB
VGG-w.-FC	✓	87.24%	**66.27%**	**80.20%**	**76.40%**	**73.24%**
VGG-w/o-FC	✗	**88.17%**	64.97%	78.80%	75.56%	71.90%
VGG-w.-FC-fix	✓	88.64%	**66.56%**	**81.60%**	**79.12%**	**68.42%**
VGG-w/o-FC-fix	✗	**89.40%**	64.86%	77.76%	76.52%	67.90%
ResNet-w.-FC	✓	90.89%	**74.75%**	**90.20%**	**87.87%**	**81.81%**
ResNet-w/o-FC	✗	**91.03%**	74.44%	89.90%	86.86%	81.50%

Retrieval (SCDA)

Models	FC	Avg. pooling		Max pooling		Avg.+Max pooling	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
VGG-w.-FC 7×7	✓	**56.42%**	**63.14%**	**58.35%**	**64.18%**	**59.72%**	**65.79%**
VGG-w/o-FC 7×7	✗	22.26%	29.33%	24.44%	31.51%	26.20%	33.31%
VGG-w.-FC 14×14	✓	**55.33%**	**62.04%**	**58.03%**	**63.93%**	**59.08%**	**65.45%**
VGG-w/o-FC 14×14	✗	22.51%	30.06%	24.21%	31.48%	26.61%	33.91%

Pooling: the more info. the higher acc





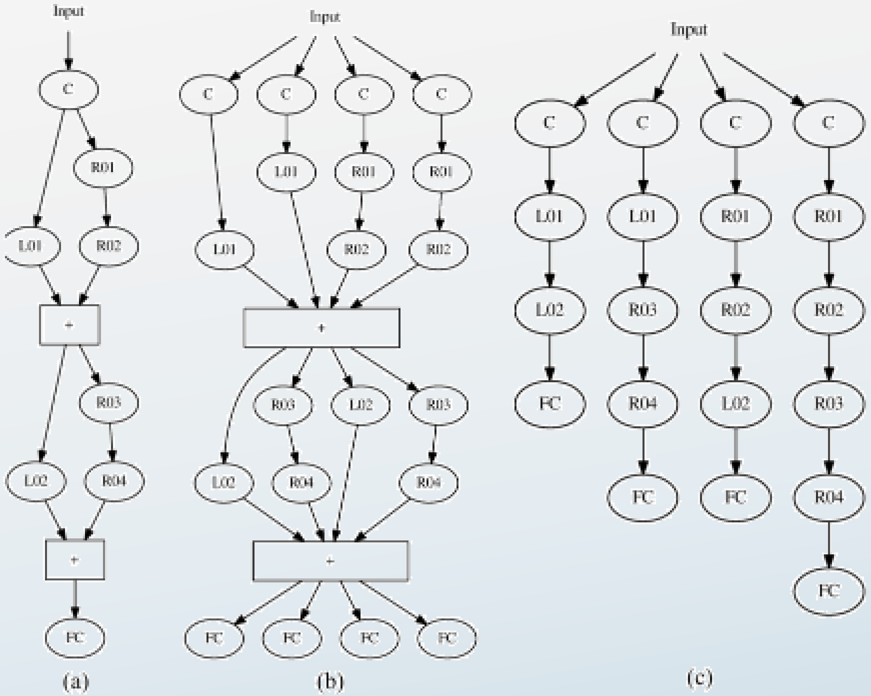
集成最大汇合：最大汇合时只有最大值有用吗?
张皓，吴建鑫
中国科学技术大学学报, 2017, 47(10): 799-807

pooling = probabilistic ensemble

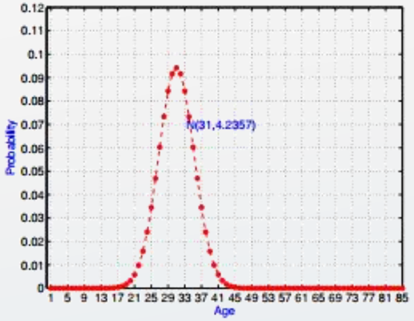
- Max pooling
  - $\begin{bmatrix} 0.0 & 0.7 \\ 0.3 & 0.0 \end{bmatrix} \rightarrow [0.7]$
- Ensemble max-pooling
  - $\begin{bmatrix} 0.0 & 0.7 \\ 0.3 & 0.0 \end{bmatrix} \rightarrow$

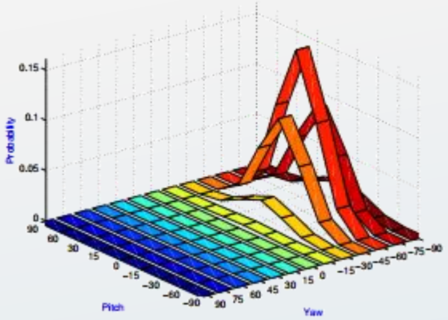
    $(1-p) \times 0.3 + p \times 0.7$

- Implicit ensemble
  - Like dropout



DLDL: treasure beneath uncertainty





Deep Label Distribution Learning with Label Ambiguity

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, Xin Geng

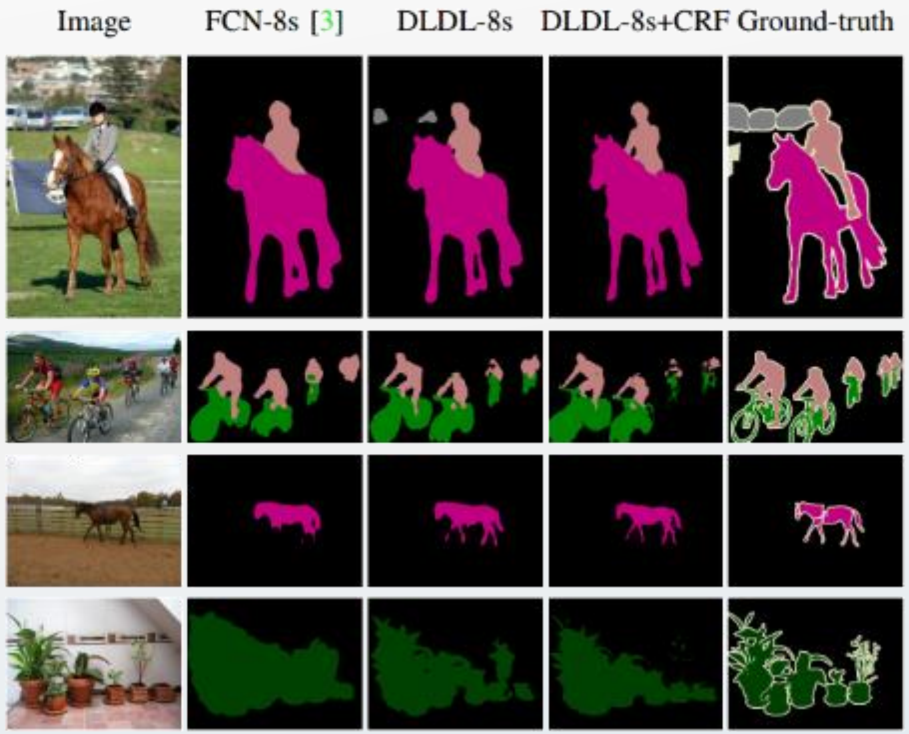
IEEE Transactions on Image Processing, 26(6), 2017: 2825-2838

Distributions: generate & compare

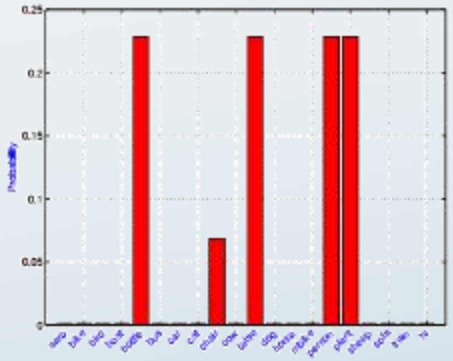
- Groundtruth distribution
  - From dataset
  - From prior knowledge
    - Age = 35
    - Normal $\mu = 35, \sigma = 3$
- Whenever there is uncertainty is labels
  - Multi-label recognition
  - Semantic segmentation
- Compare distributions (aka, loss)
  - KL divergence between them
    - Softmax: activation → predicted distribution
- Backpropagation rule
  - Simple as you can imagine
  - $\frac{\partial T}{\partial \boldsymbol{\theta}} = (\hat{\boldsymbol{y}} - \boldsymbol{y}) \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}$

More DLDL applications

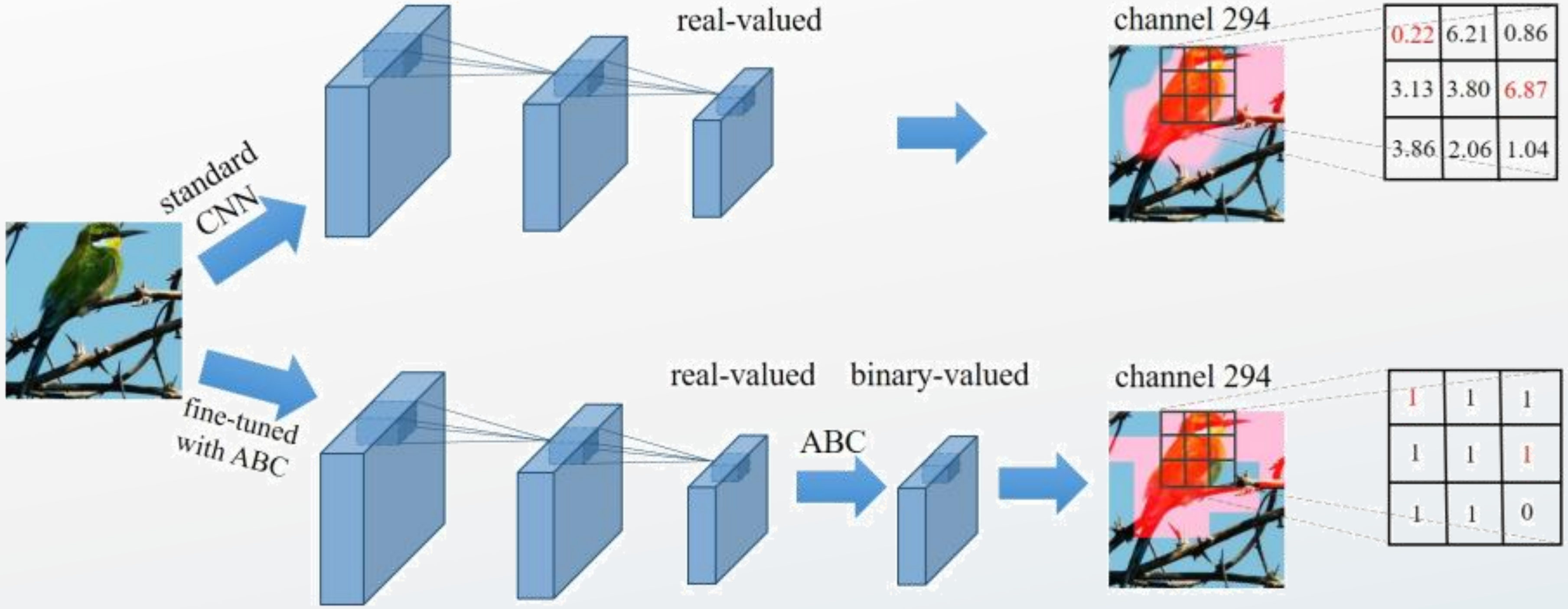
- Semantic segmentation



- Multi-label recognition



ABC: magnitude not important, sign is!



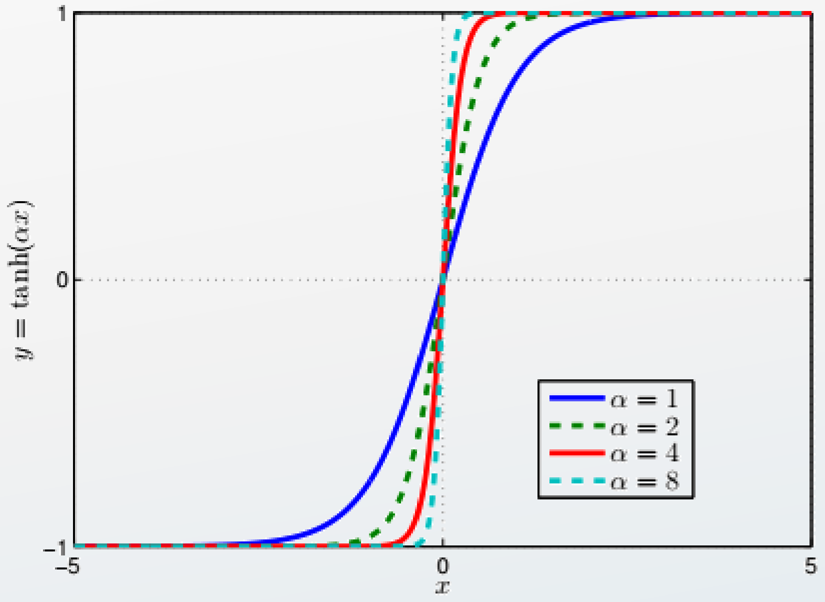
Learning Effective Binary Visual Representations with Deep Networks

Jianxin Wu, Jian-Hao Luo

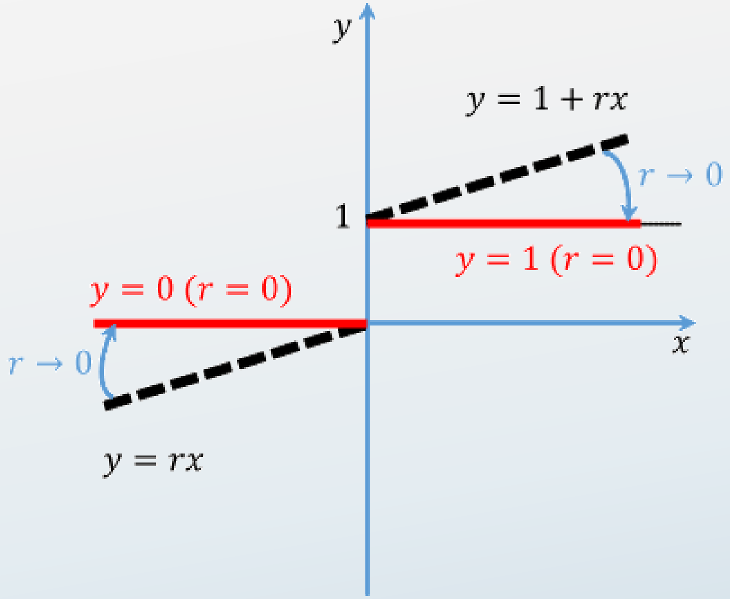
arXiv:1803.03004

Learning binary representation

- $\tanh(\alpha x)$
  - Increase α continually



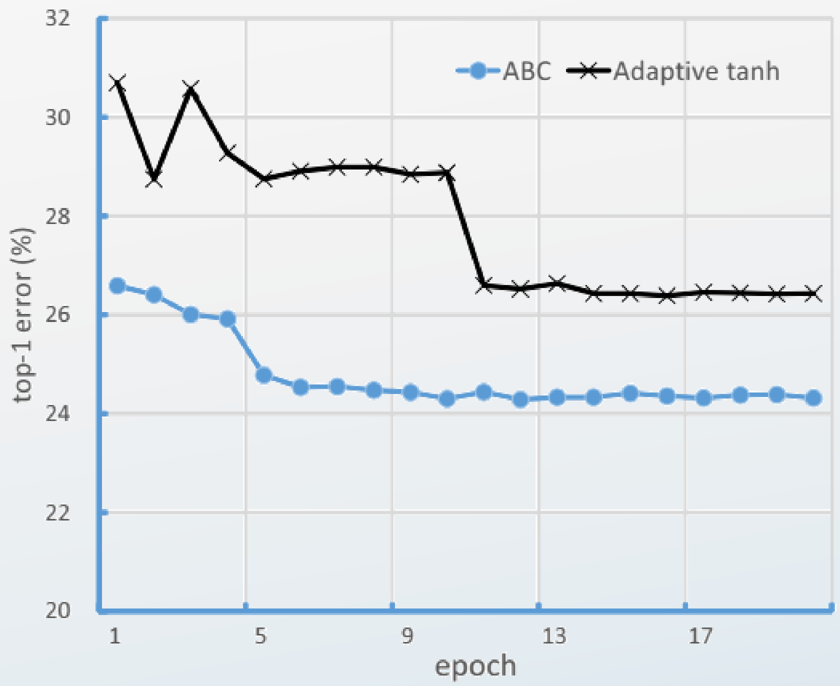
- Approximately binary clamping
  - Decrease r gradually



Deep Binary Reconstruction for Cross-modal Hashing, Li et al., ACM MM 2017
HashNet: Deep Learning to Hash by Continuation, Cao et al., ICCV 2017

ABC properties & applications

- *True binary representations*
- *Converges quickly*
  - Fine tuning ILSCRC almost converges in 5 epochs
- *Comparable accuracy*
- *Binary representations generalizes better!*
  - 1% higher mAP in Fast R-CNN object detection



Thank you!