Learning and Vision Group, NUS, Classification task of ILSVRC 2013

Adaptive Non-parametric Rectification of Shallow and Deep Experts

Min LIN*, Qiang CHEN*, Jian DONG, Junshi HUANG, Wei XIA

Shuicheng YAN

eleyans@nus.edu.sg

National University of Singapore

NUS National University of Singapore

(* indicates equal contribution)

Task 2: Classification – NUS Solution **Overview**

ILSVRC 2013 Dataset

Finished

Unfinished due to surgery of key member, but effective

Super-coding

Shallow Experts
PASCAL VOC 2012 Solution (SVMs)

Deep Experts
Convolutional Neural Network

Bigger and Deeper

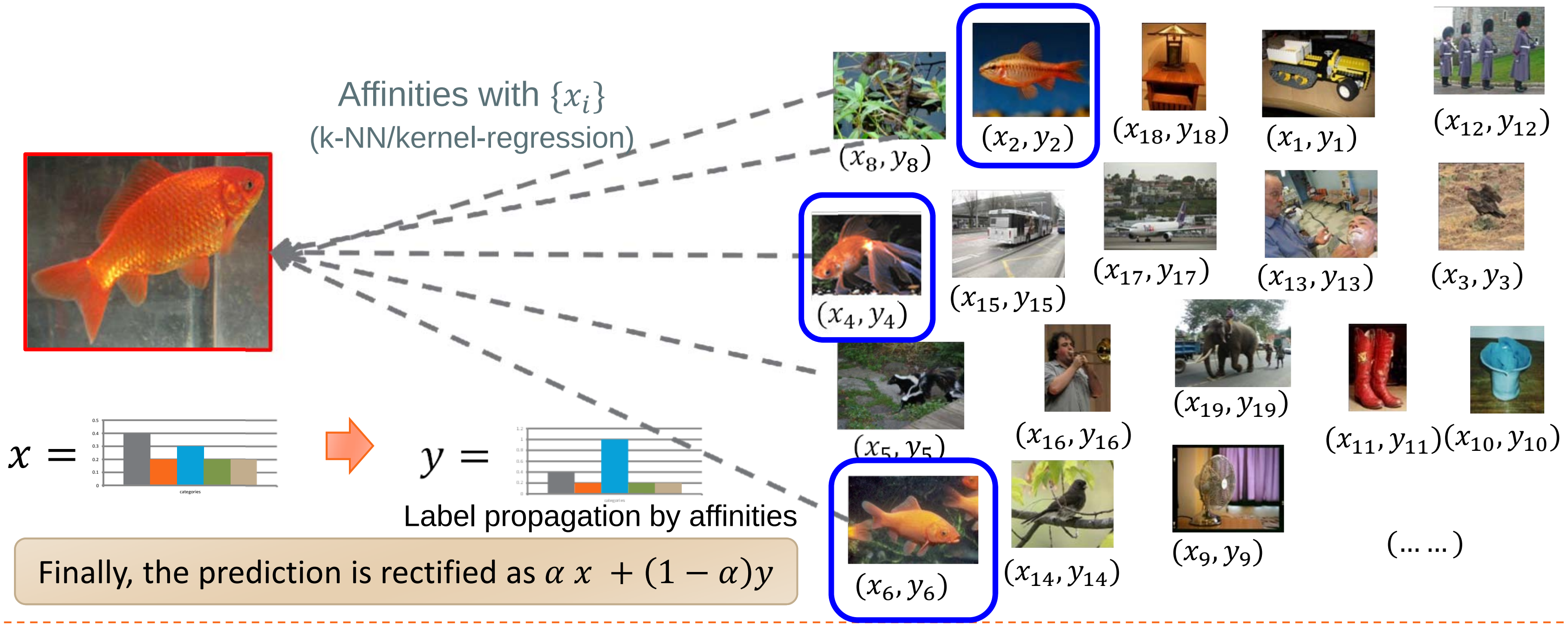
"Network in Network"
NIN: CNN with Non-linear Filters, yet No Final Fully-connected NN Layer

Adaptive Non-parametric Rectification

NUS National University of Singapore

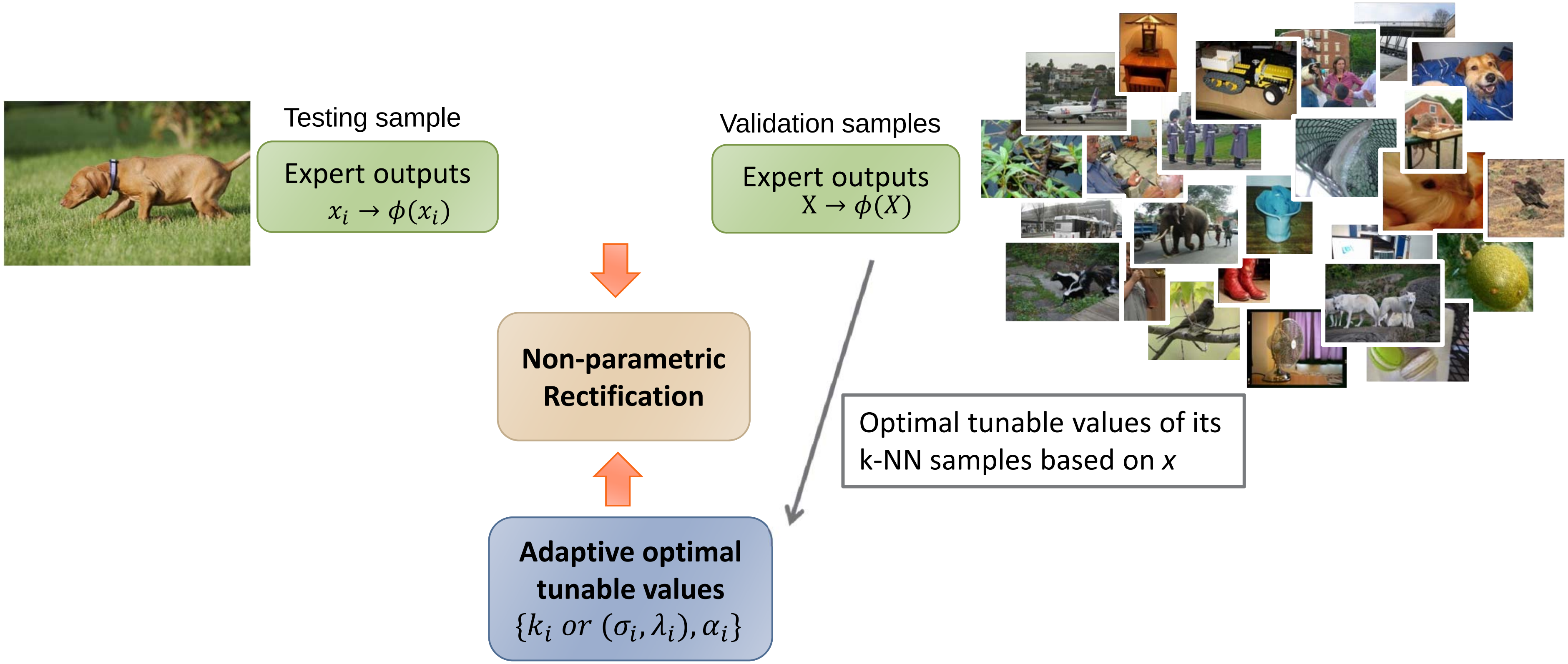
Non-parametric **Rectification**

- Motivation
  - Each validation-set image has a **pair** of **outputs-from-experts** (x_i) and **ground-truth label** (y_i), possibly inconsistent
  - For a testing image, rectify the experts based on priors from validation-set pairs (**experts errors are often repeated**)



Finally, the prediction is rectified as $\alpha\, x + (1-\alpha) y$

Adaptive Non-parametric Rectification



- Determine the optimal tuneable values for each test sample
  - For each test sample, refer to the **k-NN** in the validation set
  - Optimal tuneable values for validation samples are obtained through cross-validation

NUS National University of Singapore

Shallow Experts

Shallow Experts
PASCAL VOC 2012 Solution (SVMs)

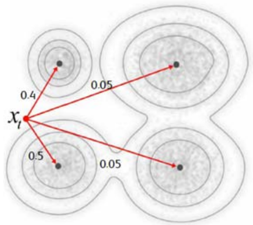
Handcrafted Features → Coding + Pooling → SVMs Learning → Prediction

Layer 1 | Layer 2

- Two-layer feature representation
  - Layer 1: Traditional handcrafted features
    - We exact dense-SIFT, HOG and color moment features within patches
  - Layer 2: Coding + Pooling
    - Derivative coding: Fisher-Vector
    - Parametric coding: **Super-Coding**

Shallow Experts: GMM-based **Super-Coding**

- Two basic strategies to obtain the patch based GMM coding [1]
  - **Derivative**: Fisher-Vector (w.r.t. μ_i and σ_i, high-order), Super-Vector (w.r.t. μ_i only)



FV formulas:

$$\mathcal{G}^X_{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right)$$

$$\mathcal{G}^X_{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right]$$

Image from [F Perronnin, 2012]

  - **Parametric**: use adapted model parameters, e.g. Mean-Vector (1st order)
- High-order parametric coding
  - The **Super-Coding:** $C_a = [\frac{\mu_1^a}{\sqrt{\sigma_1}}; \cdots ; \frac{\mu_K^a}{\sqrt{\sigma_K}}; \frac{\sigma_1^a}{\sigma_1}; \cdots ; \frac{\sigma_K^a}{\sigma_K}]$
  - The inner product of the codings is an approximate of the KL-divergence

- Advantages
  - Comparable and **complementary** performance with Fisher-Vector
  - It is very efficient to compute Super-Coding along with Fisher-Vector

[1] Derivative and Parametric Kernels for Speaker Verification, C. Longworth and M. Gales, INTERSPEECH, 2007

Shallow Experts: **Early-stop SVMs**

Shallow Experts
PASCAL VOC 2012 Solution (SVMs)

Handcrafted Features → Coding + Pooling → SVMs Learning → Prediction

Layer 1   Layer 2

- Two-layer feature representation
  - Layer 1: Traditional handcrafted features
    - We use dense-SIFT, HOG and color moment
  - Layer 2: Coding + Pooling
    - Derivative coding: Fisher-Vector
    - Parametric coding: Super-Coding
- Classifier learning
  - Dual coordinate descent SVM [2]
  - Model averaging for early stopped SVMs

[2] A Dual Coordinate Descent Method for Large-scale Linear SVM, Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, S. Sundararajan, ICML 2008

Shallow Experts: **Performance**

- Results on validation set
  - 1024-component GMM
  - Average early-stopped SVMs
    - For each round, 1) randomly select 1/10 of the negative samples, and 2) stop the SVMs at around 30 epochs **[balance efficiency and performance]**
    - Train 3 rounds, and average

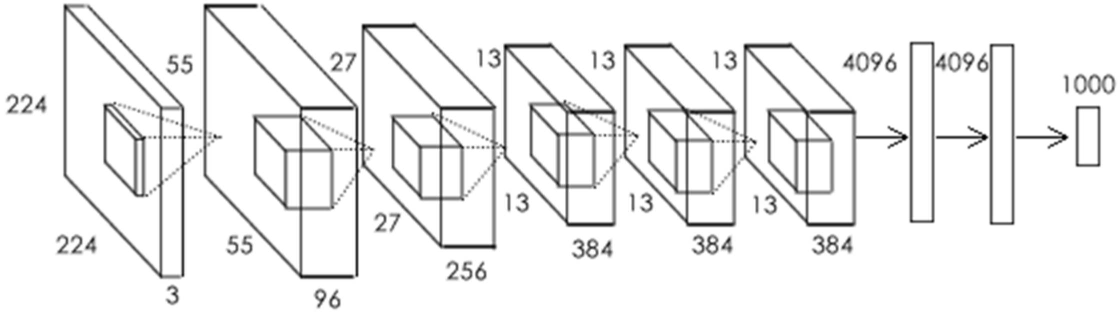
	Fisher-Vector (FV)	Super-Coding (SC)	FV+SC	3 FV+SC
Top 1	47.93%	47.67%	45.3%	43.27%
Top 5	25.93%	25.54%	24.0%	22.5%

Comparable & complementary

NUS National University of Singapore

Deep Experts

Deep Experts
Convolutional Neural Network



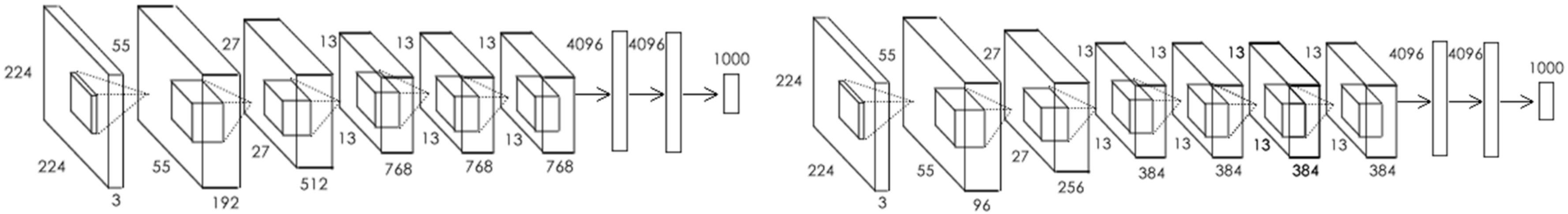
- Follow Krizhevsky et al. [3]
  - Achieved top-1 performance 1% better than reported by Krizhevsky
  - No network splitting for two GPUs, instead NVIDIA TITAN GPU card 6GB memory
  - Our network does not have PCA noise for data expansion, which is reported by Krizhevsky to improve the performance by 1%

	Krizhevsky's	Ours
Top 1	40.7%	39.7%
Top 5	18.2%	17.8%

[3] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012.

Deep Experts: **Extensions**

- Two extensions



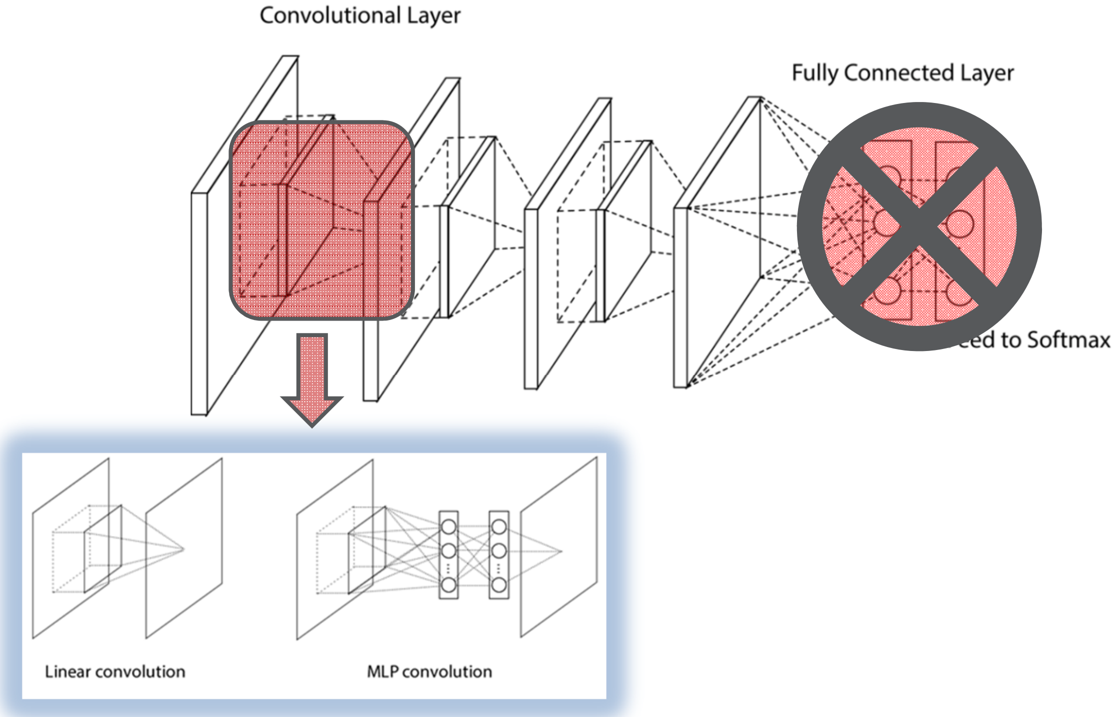
  - Bigger (**left**): Big network with doubled convolutional filters/kernels
  - Deeper (**right**): CNN with 6 convolutional layers
- Performance comparison on validation set

	CNN5 (8days)	**BigNet (30days)**	**CNN6 (12days)**	**5 CNN6**	**5 CNN6 +BigNet**
Top 1	39.7%	37.67%	38.32%	36.27%	35.96%
Top 5	17.8%	15.96%	16.52%	15.21%	14.95%

NUS National University of Singapore

Deep Experts: "Network in Network" (NIN)

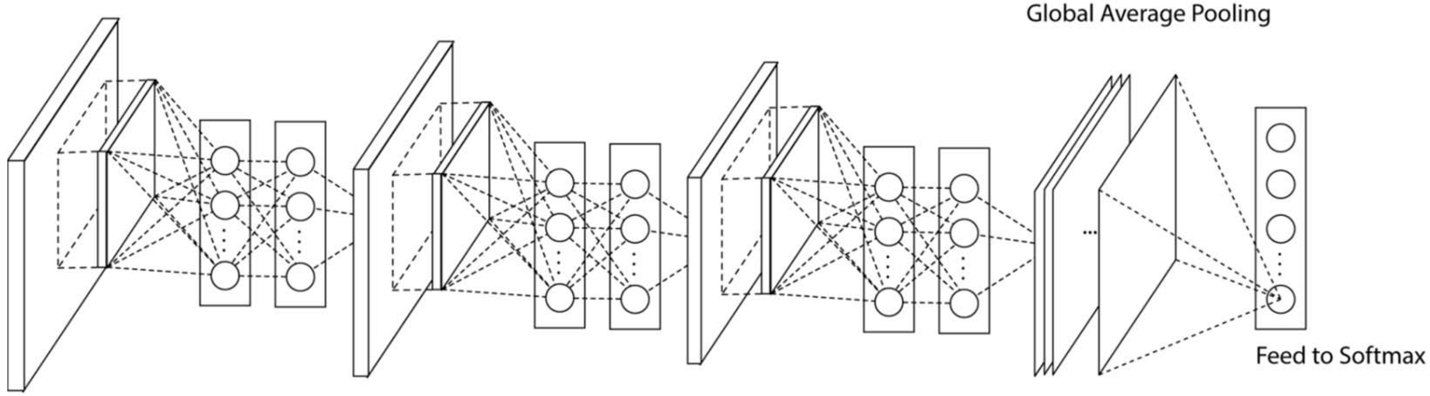
- **NIN: CNN with non-linear filters, yet without final fully-connected NN layer**



CNN

Deep Experts: "Network in Network" (NIN)

- **NIN: CNN with non-linear filters, yet without final fully-connected NN layer**



NIN

- Intuitively less overfitting globally, and more discriminative locally

(**not finally used** in our submission due to the surgery of our main team member, but very effective)

	Cifar-10	Cifar-100
Previous Best performance (Maxout) [4]	11.68%	38.57%
Our method	10.41%	36.30%

With less parameter #

More details at: http://arxiv.org/abs/1312.4400

[4] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C. Courville, Yoshua Bengio: Maxout Networks. ICML (3) 2013: 1319-1327

NUS National University of Singapore

NUS Submissions

- Results on test set

Submission	Method	Top 5 error rate
tf	traditional framework based on PASCAL VOC12 winning solution with extension of high-order parametric coding	22.39% (26.17%)
cnn	weighted sum of outputs from one large CNN and five CNNs with 6-convolutional layers	15.02% (16.42%)
weigtht tune	weighted sum of all outputs from CNNs and refined PASCAL VOC12 winning solution	13.98% (↓1.04%)
anpr	adaptive non-parametric rectification of all outputs from CNNs and refined PASCAL VOC12 winning solution	13.30% (↓0.68%)
anpr retrain	adaptive non-parametric rectification of all outputs from CNNs and refined PASCAL VOC12 winning solution, with further CNN retraining on the validation set	12.95% (↓0.35%)

Clarifai 11.74% (↓1.21%)

NUS National University of Singapore

Conclusions & Further Work

- Conclusions
  - **Complementarity** of shallow and deep experts
  - **Super-coding:** effective**, complementary** with Fisher-Vector
  - Deep learning: **deeper & bigger, better**
- Further work
  - Consider **more** validation data for adaptive non-parametric rectification

    (training data are overfit, yet only 50k validation data; training: less is more)
  - **Network in Network (NIN): CNN with non-linear filters, yet without final fully-connected NN layer** on ILSVRC data; paper draft is accessible at http://arxiv.org/abs/1312.4400

Shuicheng YAN

eleyans@nus.edu.sg

National University of Singapore