Brain-like Unsupervised Feature Learning for Convolutional Neural Networks

Takashi Shinozaki

CiNet, National Institute of Information and Communications Technology (NICT)
Graduate School of Information Science and Technology, Osaka University

Summary

- Integrates competitive learning to CNNs
  - Unsupervised pre-training by competitive learning
  - Supervised fine-tuning by error based learning without BP signals
- Validate the method with MNIST, CIFAR-10 and ImageNet
  - State-of-the-art performance as a biologically-motivated NN
- Could apply for various types of data
  - Video data, 1D-temporal sequence data, medical data, etc.

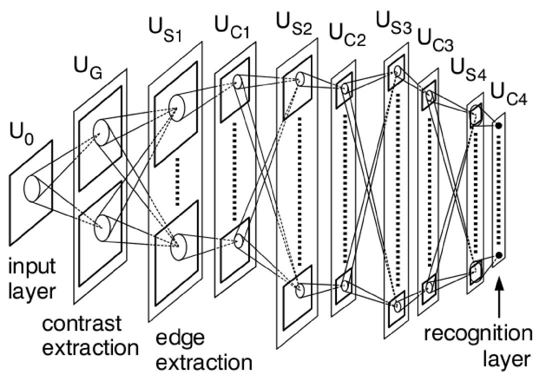
1. Introduction

1.1 Deep Neural Networks

- Key technologies of the recent revolutionary advance of AI
  - Convolutional Neural Network (CNN) [LeCun+1989]
  - Long Short Term Memory (LSTM) [Hochreiter & Schmidhuber 1997]
- Both of them use Back Propagation (BP) learning [Rumelhart+1986]
  - Tremendously good for fine-tuning
  - Extracts features for the discrimination → Limits the generalization

1.2 Competitive Learning

- Traditional unsupervised learning method for NNs
- Used for a couple of classical neural networks
  - Self Organizing Map (SOM) [Kohonen 1982]
  - Neocognitron [Fukushima 1980]
- Extracts bases of input data like as ICA
- Good for pre-training, but not good for fine-tuning



Neocognitron [Fukushima 1980]

Combine competitive learning with error based learning!!

2. Competitive Learning in Convolutional Layers

- Employ the simplest competitive learning
  - Just use winner-takes-all (WTA) algorithm
  - Do not use any spatial information over the filter space
- Basically behave alike conventional convolutional layers
- Weight gradient is calculated with the feedforward propagation

$$\Delta \boldsymbol{w}_{l,i} = \begin{cases} \rho z_{l-1}, & \text{if } i = \mathrm{argmax}_k(u_{l,k} + b_k) \\ 0, & \text{otherwise,} \end{cases} \qquad (1)$$

- $\boldsymbol{u}_l$: output vector of l-th layer described as $\boldsymbol{u}_l = \boldsymbol{W}_l z_{l-1}$
- $\boldsymbol{W}_l$: connection matrix, z_{l-1}: output vector of the previous layer
- ρ: learning coefficient of competitive learning, 0.01

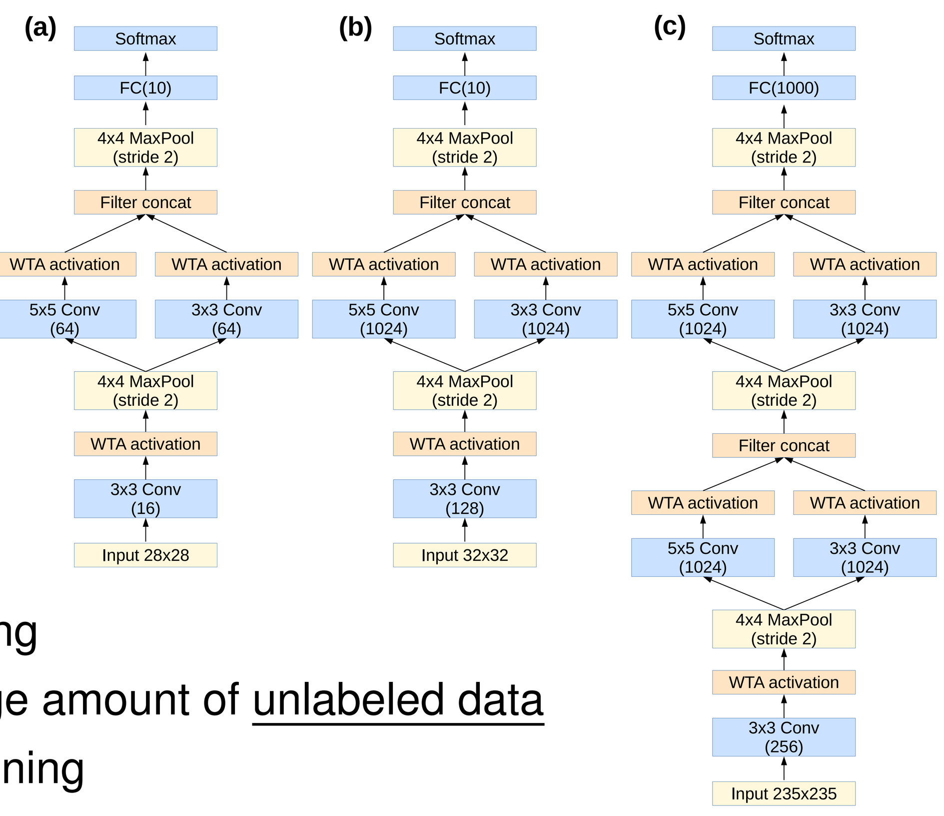
$$b_i = C(1/N - p_i). \qquad (2)$$

- b_i: conscience factor [Desieno 1988]
  - Prevents for some filters to dominate over the layer
  - C: conscience coefficient, 5.0, N: the number of filters at the layer
  - p_i: the probability of winning for the i-th unit in the mini-batch
- Update weights with the gradient by conventional method: e.g. SGD, Adam
  - The weight vector is normalized by L2-norm at every update
- WTA activation function
  - ReLU must required well controlled threshold learned by BP

3. Experiments

3.1 Network Structure

- LeNet5 based networks
  - Two or three *conv* layers
    - w/inception-like structure
  - WTA activation function
    - Instead of ReLU
  - Max-pooling
  - One *fc* layer w/softmax



- Pre-training with competitive learning
  - Unsupervised learning with a huge amount of unlabeled data
- Evaluation in fine-tuning by BP learning
  - Supervised learning with labeled data
  - Just for the last FC layer → **No BP signal is required!!**

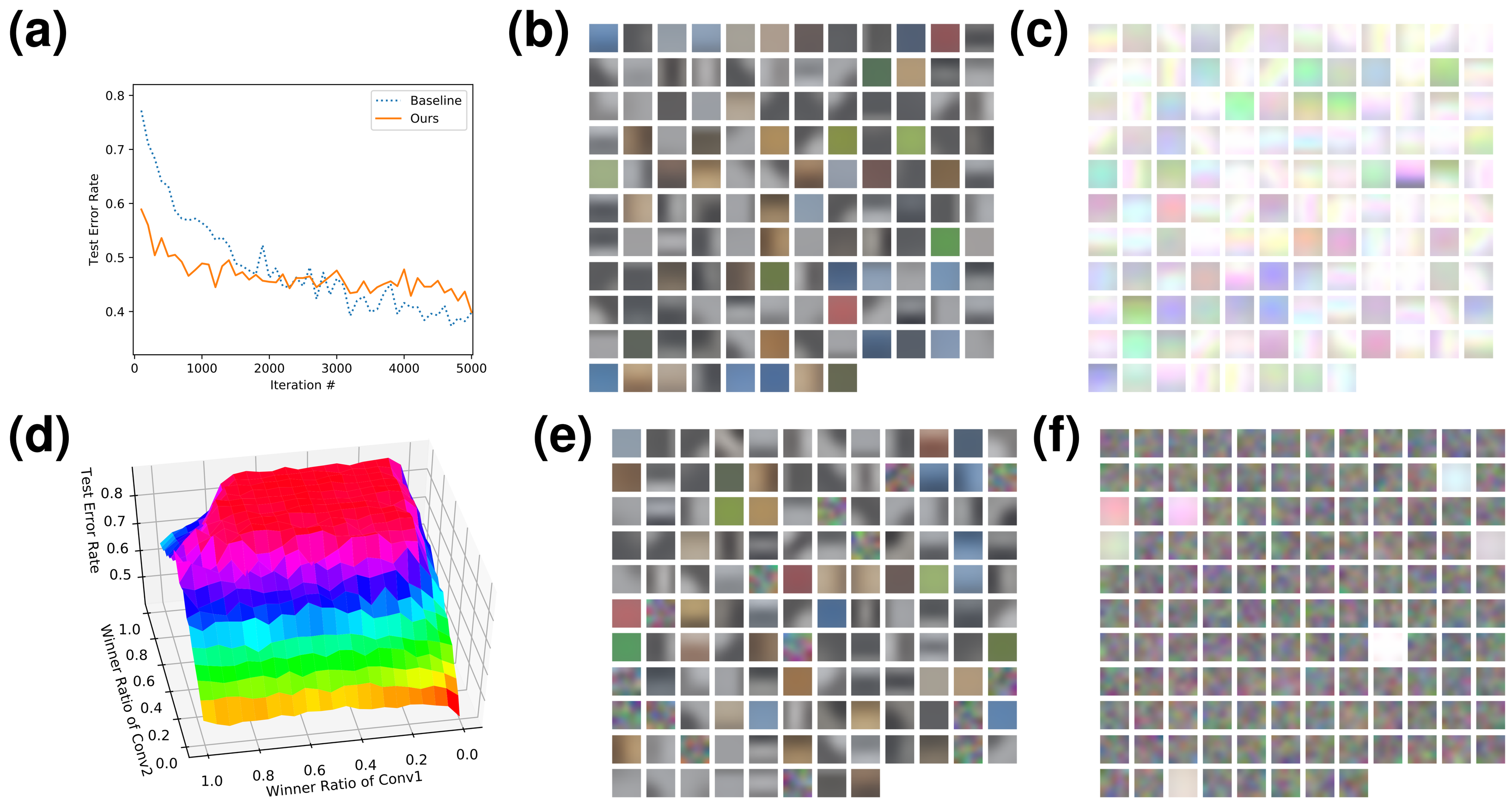
3.2 Results

- Calculated by NVIDIA Tesla P100 with *Chainer v4* [Tokui+2012]
  - The code is available at `https://github.com/t-shinozaki/convcp`
- Validate the proposed method with three datasets
  - MNIST dataset [LeCun 1998]
  - CIFAR-10 dataset [Krizhevsky+2009]
  - ImageNet dataset [Russakovsky+2015]
- State-of-the-art performance as a biologically-motivated NN

Method	MNIST	CIFAR-10	ImageNet Top-1	Top-5
Baseline [Bartunov+2018]	**0.90**	**37.74**	**63.93**	**40.17**
Ours	1.79	39.31	**87.72**	**73.84**
DTP, Parallel, LC [Bartunov+2018]	**1.52**	39.47	98.34	94.56
SDTP, Parallel, LC [Bartunov+2018]	1.98	46.63	99.28	97.15
FA, LC [Bartunov+2018]	1.85	**37.44**	93.08	82.54

Table 1: *Test errors of image discrimination tasks.*

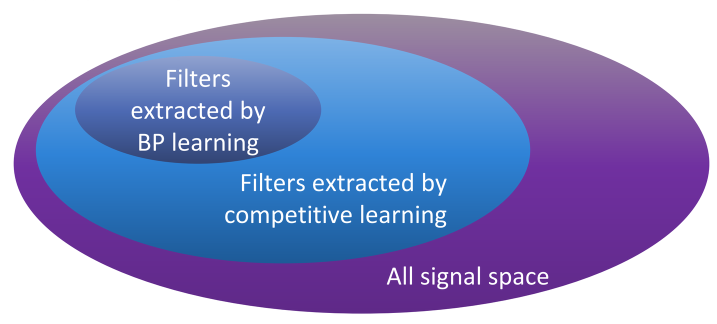
3.3 CIFAR-10



- (a) Transitions of test error rates during fine-tuning
  - The competitive learning accelerated the initial speed of fine-tuning
    - Though the competitive layers do not perform BP learning at all
- (b,c) Obtained learning representations for conv1 & conv2
  - Competitive learning robustly extracted bases of images
  - WTA activation function enables hierarchical competitive learning
- (d) Comparison over several winner ratios
  - Winners Share All (WSA) activation function
    - Only the upper units in the value order has output (controlled by winner ratio)
  - 0.0 is the best, meaning WTA condition
- (e,f) Obtained learning representations without the conscience factor
  - Stronger filters dominated and obstructed weaker filters
  - Some filters couldn't get clear spatial patterns (especially in conv2)

4. Discussion

- Integrates competitive learning into conventional CNNs
  - Apply for unsupervised pre-training
- Powerful representation learning
  - Acquires task-independent filters
    - Much more generalized filters
  - Requires much more filters
    - Most filters are **redundant** for a specific task
    - Maybe tolerance for adversarial attacks?
  - Fundamentally, equivalent to BP learning ($\Delta \boldsymbol{w}_{l,i} = \delta_{l,i} z_{l-1}$)?
  - Utilize a huge amount of unlabeled data
    - Application for many kind of time-series signals
- More effective when the number of filters was sufficiently large
- Enables seamless switching between unsupervised and supervised learning
  - Applicable for semi-supervised learning?



[Athalye+2018]

- Future works
  - Apply for Various types of data: Video data, 1D-temporal sequence data, etc.
  - Utilize SOM like organization over the filter space
    - The interpolation may enrich expressions of the filters
    - Must require an explosive amount of memory and calculation power
- Pruning method for a specific task

International Symposium on Artificial Intelligence and Brain Science, Oct 10-12, 2020, ~~Tokyo, JAPAN~~ Online. This work was supported by JST ERATO Grant Number JPMJER1801, Japan.