SVM and ELM: Who Wins? Object Recognition with Deep Convolutional Features from ImageNet

Lei Zhang*, David Zhang
College of Communication Engineering, Chongqing University
Department of Computing, The Hong Kong Polytechnic University

Int’Conf. ELM2015, Hangzhou, China
15/12/2015
Introduction

- ELM and Deep Learning
  - Deep nets attract people’s attention in different fields, especially in computer vision
  - In general, multi-layer neural nets with more than 2 layers are called deep nets.
  - For users, one only need to specify number of layers, hidden units, type of layers and loss function (task-specific)
Introduction

- ELM and Deep Learning

A general structure of deep nets is

![Diagram of deep neural network with layers and hierarchical features]

- Feed-forward
- Back-propagation
Introduction

- **ELM and Deep Learning**

Conventional deep nets include:

- Deep belief nets (DBN)
- Stacked auto-encoder (SAE)
- Stacked denoisng auto-encoder (SDAE)
- Deep Boltzmann machine (DBM)
- Convolutional nets (CNN, LeNet)

Deep nets is a highly non-linear system which is just a “deep stack” of simpler (shallow) modules

- Hierarchical features
- Adaptive features
- Unsupervised feature learning
- End-to-End learning
Auto-encoder

- ELM and Deep Learning

Objective: \( \hat{X} \approx f(X) = s(w'x + b) \)

\[ \min ||\hat{X} - f(X)|| \]

Note: for de-noising (SDAE): noise is added on \( X \), similar to “dropout” by Hinton.
ML-ELM Auto-encoder

- Multi-layer ELM (MLELM) [1] (Kasun et al. 2013)

By stacking multiple ELMs, formulates a SAE.
In our work, we use CNN for feature learning, and ELM for classification.

CNN Training: ImageNet-1000 (Krizhevsky, NIPS’12)

The CNN feature of the test data is feed into ELM/SVM
CNN+...

- **SVM**
  \[
  \min_{w,\xi} \frac{1}{2} \|w\|^2 + C \cdot \sum_{i=1}^{N} \xi_i,
  \]
  \[s.t. \quad \xi_i \geq 0, \quad y_i \left[ w^T \phi(x_i) + b \right] \geq 1 - \xi_i\]

- **LSSVM**
  \[
  \min_{w,\xi} \frac{1}{2} \|w\|^2 + C \cdot \frac{1}{2} \sum_{i=1}^{N} \xi_i^2,
  \]
  \[s.t. \quad y_i \left[ w^T \phi(x_i) + b \right] = 1 - \xi_i, \quad i = 1, \ldots, N\]

- **ELM**
  \[
  \min_{\beta \in \mathbb{R}^{L \times C}} \frac{1}{2} \|\beta\|^2 + C \cdot \frac{1}{2} \sum_{i=1}^{N} \|\xi_i\|^2
  \]
  \[s.t. \quad h(x_i)\beta = t_i^T - \xi_i^T, \quad i = 1, \ldots, N \iff H\beta = T^T - \xi^T\]

- **KELM**

- **...**

- In this work, we target at comparing ELMs with SVMs.
CNN+

• Object Recognition

Data: Four cross-domain benchmark datasets (Amazon, DSLR, Webcam, Caltech)

Feature: 800-dim SURF features (low-level) and 4096-dim CNN features (high-level)
Amazon (Low-level feature)
Amazon (High-level feature)
Caltech (Low-level feature)
Caltech (High-level feature)
DSLR (Low-level feature)
DSLR (High-level feature)
Webcam (Low-level feature)
Webcam (High-level feature)
Experiments

• **Setting 1:** *single domain* recognition task
  - 20, 8, 8, and 8 samples per class are randomly selected for training from Amazon, DSLR, Webcam and Caltech domains, respectively.

• **Setting 2:** *cross-domain* recognition tasks (source only)
  - For example, we train a SVM/ELM on the Amazon and test on DSLR, i.e. A→D. Totally, 12 cross-domain tasks among the four domains are conducted. The training data is *source data only* (source only) without leveraging the data from target domain.

• **Setting 3:** *cross-domain* recognition tasks (both source and target).
  - Compared to setting 2, the training data includes source and a few target data (3 per class are selected).

*Both the low-level feature and high-level CNN feature have been tested on Setting 2 and Setting 3!*
Experiments

- **Results on Setting 1 (CNN feature)**
  - The kernel parameters and penalty coefficients have been tuned for the best accuracy. \( C = \{1, 100, 10000\} \) and \( \sigma = \{0.0001, 0.01, 1, 100\} \). \( L = 5000 \).
  - 20 random splits of training and testing have been employed and the average is computed.

| Method | CNN_layer | Amazon | DSLR  | Webcam | Caltech | CNN_layer | Amazon | DSLR  | Webcam | Caltech |
|--------|-----------|--------|-------|--------|---------|-----------|--------|-------|--------|---------|
| NN     | f₆        | 91.0±0.3 | 97.3±0.6 | 95.0±0.4 | 75.0±0.4 | f₇        | 92.4±0.2 | 96.8±0.5 | 95.3±0.5 | 76.2±0.5 |
| SVM    | f₆        | 92.9±0.1 | 97.6±0.6 | 96.7±0.3 | 83.9±0.4 | f₇        | 93.2±0.1 | 96.9±0.5 | 96.5±0.4 | 83.2±0.5 |
| LSSVM  | f₆        | 92.9±0.2 | 97.5±0.4 | 96.4±0.4 | 84.6±0.3 | f₇        | 93.5±0.1 | 96.3±0.6 | 95.4±0.4 | 83.9±0.4 |
| ELM    | f₆        | 92.9±0.1 | **98.0±0.3** | 97.7±0.2 | 84.8±0.3 | f₇        | 93.6±0.1 | 97.2±0.4 | **97.4±0.3** | 85.0±0.3 |
| KELM   | f₆        | **93.7±0.1** | 97.8±0.3 | **97.8±0.2** | **86.0±0.3** | f₇        | **94.2±0.1** | **97.5±0.4** | 97.3±0.4 | **85.7±0.3** |

- **Increment**: Amazon: 0.8%; DSLR: 0.2%; Webcam: 1.1%; Caltech: 2.1%;
Experiments

- Results on Setting 2 (CNN feature)
  - 12 cross-domain recognition tasks are conducted, each task is run on 20 randomly splits of training and testing. (the average improvement is 4%)

| Method | CNN_layer | A→D | C→D | W→D | A→C | C→A | W→A | C→W | D→W | A→W |
|--------|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| NN     | f₆        | 71.9±0.9 | 72.0±1.7 | 92.7±0.5 | 76.8±0.3 | 56.6±0.9 | 64.4±0.4 | 75.1±0.7 | 64.0±0.6 | 78.1±0.8 | 61.5±1.1 | 95.8±0.4 | 65.1±1.0 |
|        | f₇        | 78.7±0.5 | 75.6±1.3 | 96.9±0.4 | 77.2±0.4 | 66.2±0.5 | 70.7±0.4 | 75.0±0.7 | 66.3±0.8 | 83.6±0.4 | 60.7±1.2 | 95.2±0.4 | 68.5±0.8 |
| SVM    | f₆        | 79.6±0.7 | 75.1±1.8 | 96.7±0.4 | 79.5±0.4 | 59.5±0.9 | 67.3±1.2 | 77.0±1.0 | 66.8±1.0 | 85.8±0.4 | 67.1±1.1 | 95.4±0.4 | 70.6±0.8 |
|        | f₇        | 80.6±0.8 | 76.4±1.4 | 96.7±0.4 | 79.6±0.4 | 68.1±0.6 | 74.3±0.6 | 81.8±0.5 | 73.4±0.7 | 86.5±0.5 | 67.8±1.1 | 95.3±0.5 | 71.0±0.8 |
| LSSVM  | f₆        | 77.1±0.9 | 76.8±1.2 | 96.1±0.3 | 77.5±0.6 | 61.1±0.7 | 70.6±1.0 | 80.0±0.8 | 68.2±1.1 | 86.5±0.4 | 67.8±1.2 | 96.4±0.4 | 65.5±0.8 |
|        | f₇        | 82.6±0.5 | 79.2±0.8 | 95.9±0.4 | 79.8±0.5 | 66.0±1.3 | 73.7±0.9 | 80.8±0.7 | 72.0±1.1 | 87.4±0.3 | 69.9±1.1 | 95.1±0.3 | 69.4±0.6 |
| ELM    | f₆        | 80.6±0.6 | 79.5±1.2 | 96.7±0.2 | 80.4±0.3 | 67.2±0.5 | 75.6±0.5 | 83.7±0.4 | 72.2±0.9 | 87.3±0.4 | 70.1±0.9 | 97.2±0.3 | 71.1±0.6 |
|        | f₇        | 82.3±0.5 | 81.2±0.7 | 97.0±0.4 | 81.8±0.3 | 74.0±0.3 | 79.5±0.2 | 85.8±0.3 | 76.7±0.9 | 88.3±0.2 | 72.3±0.9 | 96.8±0.3 | 72.4±0.8 |
| KELM   | f₆        | 82.3±0.5 | 80.7±0.9 | 96.5±0.3 | 82.6±0.3 | 69.5±0.4 | 77.8±0.4 | 85.3±0.4 | 73.8±1.1 | 88.0±0.4 | 72.3±1.0 | 97.6±0.2 | 72.9±0.7 |
|        | f₇        | 84.0±0.4 | 82.2±0.9 | 97.3±0.3 | 83.4±0.2 | 75.7±0.3 | 81.1±0.2 | 87.1±0.2 | 78.2±0.8 | 89.1±0.3 | 73.3±0.9 | 96.9±0.3 | 74.7±0.8 |
Experiments

- Results on Setting 2 (Low-level feature)

From the results, the performance is very bad when low-level feature used, but ELMs are still the best.
Experiments

- Results on Setting 3 (CNN feature)
  - 12 cross-domain recognition tasks are conducted, each task is run on 20 randomly splits of training and testing. (the average improvement is 1.5%)

| Method  | CNN_layer | A→D   | C→D   | W→D   | A→C   | W→C   | D→C   | D→A   | W→A   | C→A   | C→W   | D→W   | A→W   |
|---------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| NN      | f₆       | 89.4±0.7 | 90.1±0.8 | 97.0±0.4 | 78.1±0.4 | 69.0±0.9 | 72.8±0.8 | 83.8±0.5 | 83.3±0.7 | 85.4±0.4 | 86.9±0.6 | 97.2±0.4 | 86.1±0.8 |
|         | f₇       | 93.0±0.5 | 90.9±0.9 | 98.6±0.2 | 78.9±0.4 | 73.6±0.6 | 75.6±0.4 | 86.7±0.5 | 84.0±0.5 | 87.9±0.2 | 87.8±0.9 | 96.3±0.2 | 89.1±0.6 |
| SVM     | f₆       | 94.5±0.4 | 92.9±0.8 | 99.1±0.2 | 84.0±0.3 | 81.7±0.5 | 83.0±0.3 | 90.5±0.2 | 90.1±0.2 | 90.0±0.2 | 91.5±0.6 | 97.9±0.3 | 90.4±0.8 |
|         | f₇       | 94.0±0.6 | 92.7±0.8 | 98.9±0.2 | 83.4±0.4 | 81.2±0.4 | 82.7±0.4 | 90.9±0.3 | 90.6±0.2 | 90.3±0.2 | 90.6±0.8 | 98.0±0.2 | 91.1±0.8 |
| LSSVM   | f₆       | 92.6±0.5 | 93.1±0.6 | 98.8±0.2 | 82.3±0.5 | 80.7±0.5 | 82.3±0.4 | 90.9±0.2 | 89.7±0.2 | 90.3±0.1 | 90.9±0.6 | 97.8±0.3 | 87.7±0.8 |
|         | f₇       | 91.9±0.5 | 92.4±0.8 | 98.4±0.2 | 82.9±0.4 | 81.7±0.3 | 82.6±0.5 | 90.9±0.4 | 90.0±0.2 | 90.7±0.2 | 90.4±0.5 | 97.2±0.3 | 89.5±0.7 |
| ELM     | f₆       | 94.6±0.5 | 93.7±0.6 | 99.2±0.2 | 83.4±0.3 | 81.2±0.3 | 83.5±0.3 | 91.1±0.2 | 90.3±0.2 | 90.5±0.1 | 91.6±0.7 | 98.3±0.2 | 90.5±0.6 |
|         | f₇       | 94.9±0.4 | 93.0±0.6 | 99.0±0.2 | 84.1±0.2 | 82.2±0.4 | 84.1±0.2 | 91.7±0.2 | 90.8±0.2 | 90.9±0.1 | 91.5±0.7 | 97.9±0.2 | 91.7±0.7 |
| KELM    | f₆       | 95.7±0.4 | 94.1±0.6 | 99.2±0.2 | 85.0±0.3 | 83.0±0.3 | 84.9±0.2 | 91.9±0.2 | 90.8±0.2 | 91.1±0.1 | 92.2±0.7 | 98.6±0.2 | 91.3±0.6 |
|         | f₇       | 95.5±0.4 | 93.9±0.6 | 99.1±0.1 | 85.4±0.3 | 83.4±0.3 | 85.3±0.3 | 92.1±0.2 | 91.5±0.2 | 91.5±0.1 | 91.9±0.6 | 98.2±0.3 | 92.2±0.6 |
Experiments

- Results on Setting 3 (Low-level feature)

| Method    | A→D   | C→D   | W→D   | A→C   | W→C   | D→C   | D→A   | W→A   | C→A   | C→W   | D→W   | A→W   |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| NN        | 46.2±0.9 | 47.0±0.8 | 67.7±1.0 | 31.8±0.4 | 27.7±0.4 | 29.8±0.4 | 37.3±0.5 | 36.8±0.5 | 37.5±0.5 | 49.5±1.3 | 68.3±0.6 | 52.1±1.1 |
| SVM       | 50.9±0.9 | 47.9±1.3 | 72.4±0.8 | 42.1±0.5 | 33.7±0.5 | 35.5±0.3 | 42.3±0.4 | 41.6±0.5 | 42.8±0.8 | 47.4±1.2 | 75.5±0.6 | 52.7±0.8 |
| LSSVM     | 47.6±1.2 | 53.5±0.9 | 71.5±0.8 | 40.1±0.4 | 32.6±0.4 | 35.1±0.4 | 41.0±0.4 | 40.0±0.5 | 47.3±0.6 | 53.9±1.1 | 75.0±0.7 | 51.6±0.9 |
| ELM       | 44.9±0.9 | 49.3±1.0 | 70.2±0.7 | 37.6±0.4 | 31.2±0.5 | 32.9±0.4 | 40.0±0.5 | 39.4±0.4 | 42.5±0.5 | 51.9±0.8 | 74.7±0.5 | 49.2±1.0 |
| KELM      | 51.3±0.9 | 54.9±1.0 | 71.5±0.8 | 42.3±0.4 | 33.2±0.3 | 35.5±0.4 | 42.4±0.4 | 41.3±0.4 | 47.1±0.5 | 56.8±1.0 | 78.3±0.5 | 55.7±0.9 |
Codes and Resources Released

Lei Zhang (张磊)

Chongqing University, 博士, 博士后, 特聘研究员, 百人计划学者, 香江学者, 博士生导师

Research Interests: Machine Learning, Machine Olfaction, Electronic Nose, Artificial Intelligent, Signal Processing

http://www.escience.cn/people/lei/index.html

More details about data and codes can be found here

[news] Our paper titled "L. Zhang and D. Zhang, Visual Understanding via Multi-Feature Shared Learning with Global Consistency" is accepted for publication in IEEE Transactions on Multimedia, 2015. [paper and code coming soon]

[news] Our paper titled "X. Yin, L. Zhang, F.C. Tian and D. Zhang, Temperature Modulated Gas Sensing E-nose System for Low-cost and Fast Detection" is accepted for publication in IEEE Sensors Journal, 2015. [paper, data and code are available]

[news] Our paper titled "X. Peng, L. Zhang", F. Tian, and D. Zhang, A novel sensor feature extraction based on kernel entropy component analysis for discrimination of indoor air contaminants" is published in Sensors and Actuators A, vol.234, pp.143-149, 2015. The code is available. [paper][code]

[news] Our paper titled "L. Zhang, X.W. Peng, F. Tian, and D. Zhang, Time series estimation of gas sensor baseline drift using ARMA and Kalman based models" is accepted for publication in Sensor Review, 2015.

[news] Our paper titled "L. Zhang, D. Zhang and F.C. Tian, SVM and ELM: Who Wins? Object Recognition with Deep Convolutional Features from ImageNet" is accepted for presentation in The 6th Int Conf ELM, 2015. The code is available. [code]

http://www.leizhang.tk
Thank you all! 😊