Deep Prototype Learning for Robust Pattern Recognition

Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation of Chinese Academy of Sciences
liucl@nlpr.ia.ac.cn
http://www.nlpr.ia.ac.cn/liucl/
Outline

• Status of Pattern Recognition
• Robustness in Pattern Recognition
• How to Improve Robustness
• Models for Robust PR
• Learning Algorithms for Robust PR
• Deep Prototype Learning
• Future Research
Status of Pattern Recognition

- Pattern Recognition: Simulating human perception ability, to enable machine to detect and recognize objects and events in sensing data.
Evolution of PR Methods

- Core technique: pattern classification
  - Feature extraction and selection, classifier design (learning)
- Statistical PR: 1950s-
- Syntactic and structural PR: 1970s-
- Artificial neural networks: 1980s-
- SVM and kernel methods: 1990s-
- Multiple classifiers, ensemble: 1990s-
- Bayesian learning: 1990s-
- 1990s-: wide applications of PR technology
- 2000s-: semi-supervised, multi-label, probabilistic graphical models
- Recent: transfer learning, sparse representation, deep learning (deep neural networks)
Categorization of PR Methods

Two broad categories

• Statistical PR
  – Parametric (Gaussian)
  – Non-parametric (Parzen, k-NN)
  – Semi-parametric (GM)
  – Neural
  – Decision tree
  – Kernel (SVM)
  – Ensemble (Boosting)

• Structural PR
  – Syntactic parsing
  – String matching, tree
  – Graph matching
  – Hidden Markov model (HMM)
  – Markov random field (MRF)
  – Structured prediction

Hybrid Statistical-Structural: Statistical primitive/relationship

Attributed graphs, HMM and MRF/CRF are instances of hybrid models
Big Success of Deep Learning

- Speech recognition
- Face recognition
- Image classification
- Traffic sign recognition
- Character recognition

Traffic sign recognition

ImageNet Large Scale Visual Recognition Challenge
ILSVRC Object Classification Error Rates (top-5 %)

1000 categories, 1.2M training images
However

- Insufficiencies of Deep Learning
  - Black box, interpretability
  - High complexity, large redundancy
    - Brain cortex has much less layers of neurons but can recognize very quickly and accurately
  - Huge training data to guarantee generalization performance
  - Inflexible learning
    - Supervised, batch data
    - Parameter tuning tricks
  - Robustness of recognition
    - Vulnerable to outlier patterns and noise
Vulnerability to Outlier of DNN

Recognized by MNIST CNNs with high confidence

Recognized with high confidence by CNN trained on ImageNet

Deep learning is more vulnerable to adversarial examples than other kind of machine learning due to the extreme non-linearity of deep models.
Robustness in Pattern Recognition

• Robustness Received Insufficient Attention

• Major Issues of Robustness
  - Stability of performance
    • Sensitivity of output to input, usually measured by function smoothness
    • Related to model complexity, influencing generalization
  - Robustness to outlier (novelty, abnormality)
    • Related to closed world assumption
  - Rejection of ambiguity
    • Uncertainty of belonging to top-rank class

\[
\sum_{i=1}^{C} P(\omega_i | x) = 1
\]
Research Toward Stability / Generalization

- Overfitting due to high-complexity model
  - Increase sensitivity of $f(x)$
  - Deteriorate generalization

- Measuring of model complexity
  - Bias-variance, VC dimension
  - Function smoothness: measured by $\nabla^2 f(x)$

- How to improve stability/generalization
  - Use simple/smooth/parametric functions
  - Regularization: constrain the variability of parameters

T. Hastie, T. Tibshrani, J. Friedman (2001)
Ambiguity Rejection

Distance Rejection

In fact, the Tories made it worse now for the sick and needy than Labour had to make it in 1950. And as a percentage of social service expenditure, health had fallen from 28.5 to 23.1 percent.

English is outlier for a digit recognizer

経営不振に陥っているソニーのパソコン事業は国内投資ファンドが買い取り、開発と製造を担がける新会社バイオをつくった。ソニーの国内向け通販サイトと同社の直営店を通じ、個人から専門家までにアピールする関係者を増やさんと意図している。

These are outlier for an English recognizer
Robustness to Outlier

- PR system involves detection and segmentation, which produces non-object proposals
- Character string recognition involves candidate character segmentation

- In character string recognition, the classification scores $f_i(x)$ of character candidates are used to evaluate the candidate paths
- Non-characters should be assigned low scores to all classes
Rejection Strategies in PR

- Rejection in Bayesian decision
  - "reject" as a new decision
    \[
    \lambda(\alpha_i | \omega_j) = \begin{cases} 
    0, & i = j \\
    \lambda_s, & i \neq j \\
    \lambda_r, & \text{reject if } \lambda_r < \lambda_s
    \end{cases}
    \]
  - Minimum risk decision
    \[
    R(\alpha_i | x) = \sum_{j=1}^{c} \lambda(\alpha_i | \omega_j) P(\omega_j | x)
    \]
    \[
    R_i(x) = \begin{cases} 
    \lambda_s [1 - P(\omega_i | x)], & i = 1, K, c \\
    \lambda_r, & \text{reject}
    \end{cases}
    \]
    \[
    \arg \min_i R_i(x) = \begin{cases} 
    \arg \max_i P(\omega_i | x), & \text{if } \max_i P(\omega_i | x) > 1 - \lambda_r / \lambda_s \\
    \text{reject}, & \text{otherwise}
    \end{cases}
    \]

Rejection Strategies in PR

• Robustness to outlier
  – Closed world assumption: \( \sum_{i=1}^{C} P(\omega_i|x) = 1 \)
  – Does \( x \) really belong to \( C \) classes? (outlier, novelty)
    • Decision based on probability density

• Two types of rejection
  – Ambiguity rejection \( \max_i P(\omega_i|x) < \tau \)
  – Distance (outlier) rejection \( p(x) < T_d, \quad p(x|\omega_i) < T_i \)

• Approximate rejection rules
  
  – **Ambiguity rejection**: assuming soft-max and top 2 rank scores dominate

\[
P(\omega_i \mid x) \approx \frac{\sum_{j=1}^{C} e^{\gamma f(x,\omega_j)}}{e^{\gamma f(x,\omega_i)}}
\]

\[
f(x, \omega_{i_1}) > f(x, \omega_{i_2}) \gg f(x, \omega_j)
\]

\[
\max_i P(\omega_i \mid x) \approx \frac{e^{\gamma f(x,\omega_{i_1})}}{e^{\gamma f(x,\omega_{i_1})} + e^{\gamma f(x,\omega_{i_2})}} = \frac{1}{1 + e^{-\gamma [f(x,\omega_{i_1}) - f(x,\omega_{i_2})]}}
\]

\[
\max_i P(\omega_i \mid x) < 1 - c_r \iff f(x,\omega_{i_1}) - f(x,\omega_{i_2}) < T
\]

 – **Distance rejection**: assuming likelihood (class output or minus distance) proportional to logarithm of density

\[
f(x, \omega_i) \propto \log p(x \mid \omega_i)
\]

\[
p(x \mid \omega_i) < T_i \iff f(x, \omega_i) < t_i
\]
Sources of Inrobustness

• Closed World Assumption
  – $C$ classes, $\sum_{i=1}^{C} P(\omega_i | x) = 1$

• Improper Model
  – $\max P(\omega_i | x)$ on outlier can be large
    • Discriminative models directly approximate posterior probabilities

• Improper Learning Algorithm
  – Trained model deviates from the pattern class distribution or structure
    • E.g., learning vector quantization (nearest prototype classifier)
How to Improve Robustness

• Open World Assumption

\[ \sum_{i=1}^{C} P(\omega_i \mid x) \leq 1 \quad \sum_{i=1}^{C+1} P(\omega_i \mid x) = 1 \]

  – Hard to model \( \omega_{C+1} \) explicitly (insufficient samples)

• Generative Models

  – Density/template-based
  – Structure model

• Learn to represent classes in addition to discrimination
Models for Robust PR

- **Density-Based Classifier**
  - Classification: \( \max_i P(\omega_i) p(x | \omega_i) \)
  - Ambiguity rejection: \( \max_i \frac{P(\omega_i) p(x | \omega_i)}{\sum_{j=1}^c P(\omega_j) p(x | \omega_j)} < \tau \)
  - Outlier (distance) rejection: \( \max_i P(\omega_i) p(x | \omega_i) < T \)

- **Template (prototype, distance)-Based Classifier**
  - Sparse representation classifier is a special case
  - Classification: \( \min_i d(x, \omega_i) = \min_j \| x - m_{ij} \| \)
  - Ambiguity rejection: \( d(x, \omega_{i_2}) - d(x, \omega_{i_1}) < T_1 \)
  - Outlier rejection: \( \min_i d(x, \omega_i) < T_2 \)

- **Take Advantage of Deep Learning**
  - Generative neural networks (RBM, DBN, autoencoder)
  - Generative model in the learned feature space
• Relationship between density & template-based
  – Gaussian density

\[ P(\mathbf{x} | \omega_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right] \]

\[ 2 \log P(\omega_i)P(\mathbf{x} | \omega_i) = 2 \log P(\omega_i) - (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) - d \log 2\pi - \log |\Sigma_i| \]
\[ = C_i - (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) - \log |\Sigma_i| \quad \text{Quadratic distance} \]

– Mixture of Gaussians

\[ p(\mathbf{x} | \omega_i) = \sum_{j=1}^{K} \pi_j N(\mathbf{x} | \mu_{ij}, \Sigma_{ij}) \]

• Assume identity covariance matrices and \( \| \mathbf{x} - \mu_{in} \|^2 \ll \| \mathbf{x} - \mu_{ir_2} \|^2 \)

\[ \log p(\mathbf{x} | \omega_i) \propto - \min_j \| \mathbf{x} - \mu_{ij} \|^2 \]
Learning Algorithms for Robust PR

• Hybrid Discriminative Generative
  – Empirical risk minimization
    \[ \min_{\theta} R_{emp} = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n, \theta)) \]
  – Maximum likelihood (ML), for generative models only
    \[ \max_{\theta_i} LL_i(\theta_i) = \log p(X_i | \theta_i) = \sum_{n=1}^{N_i} \log p(x_n | \theta_i) \propto -\sum_{n=1}^{N_i} d(x_n, \omega_i) \]
  – ML regularization (a.k.a. I-smoothing in speech recognition)
    \[ \min_{\theta} R_{emp} = \frac{1}{N} \sum_{n=1}^{N} [L(y_n, f(x_n, \theta)) + \lambda d(x_n, \omega_{y_n})] \]
  – E.g. prototype classifier

Prototype classifier
Convolutional Prototype Learning

- Convolutional NN: Learning discriminative feature space
- Prototype classifier: Distance based, robust to outlier

Convolutional Prototype Learning

- **Learning Objective**
  - Prototypes representing class distribution
  - Feature space: desired to be compact within each class
  - Classification Loss (CL): MCE, MCL (margin-based classification loss, GMCL, DCE (distance-based cross-entropy))

Classification loss does not lead to compactness in feature space
Minimum Classification Error (MCE) Loss

- Misclassification error for a sample \((x, y)\):
  \[
  \mu_y(x) = -g_y(x) + \left[ \frac{1}{C-1} \sum_{j \neq y} g_j(x)^\eta \right]^{1/\eta}
  \]

- When \(\eta \to \infty\), we have:
  \[
  \mu_y(x) = -g_y(x) + g_r(x)
  \]
  \[
  = \lVert f(x) - m_{yi} \rVert^2 - \lVert f(x) - m_{rj} \rVert^2
  \]

- MCE is defined as:
  \[
  l((x, y); \theta) = \frac{1}{1 + e^{-\xi \mu_y(x)}}
  \]

\(r\) is the most competitive class, \(m_{yi}\) and \(m_{rj}\) denote the closest prototype from the correct class and incorrect class respectively.
Margin based Classification Loss (MCL)

For a sample \((x, y)\), when it is classified correctly:
\[
d(f(x), m_{yi}) < d(f(x), m_{rj})
\]
otherwise:
\[
d(f(x), m_{yi}) > d(f(x), m_{rj})
\]
classification loss is defined as:
\[
l((x,y); \theta, M) = [d(f(x), m_{yi}) - d(f(x), m_{rj})]_+
\]
Add a margin, then MCL is defined as:
\[
l((x,y); \theta, M) = [d(f(x), m_{yi}) - d(f(x), m_{rj}) + m]_+
\]
To better select \(m\), we define GMCL as:
\[
l((x,y); \theta, M) = \left[\frac{d(f(x), m_{yi}) - d(f(x), m_{rj})}{d(f(x), m_{yi}) + d(f(x), m_{rj})} + m\right]_+
\]
\[0 < m < 1\]
● Distance based Cross Entropy (DCE)

✓ Prototype probability based on distance:

\[ p(x \in m_{ij}|x) \propto -\|f(x) - m_{ij}\|^2 \]

\[ p(x \in m_{ij}|x) = \frac{e^{-\gamma d(f(x), m_{ij})}}{\sum_{k=1}^{C} \sum_{l=1}^{K} e^{-\gamma d(f(x), m_{kl})}} \]

✓ Class probability is defined as:

\[ p(y|x) = \sum_{j=1}^{K} p(x \in m_{yj}|x) \]

✓ The DCE is defined as:

\[ l((x, y); \theta, M) = -\log p(y|x) \]

Like conditional log-likelihood (CLL)

✓ Minimize \( l \iff \) pull \( f(x) \) closer to its prototypes
Regularization: Prototype Loss (PL)

- To pull the feature closer to its prototype, decreasing the within-class distance:
  \[ pl((x, y); \theta, M) = \| f(x) - m_{yj} \|_2^2 \]

- PL is actually a maximum likelihood (ML) regularization:
  \[ \log p(f|y) = \log \mathcal{N}(f|m_{yj}, kI) \propto \| f - m_{yj} \|_2^2 \]

- Based on the Gaussian assumption, can be viewed as a generative model;

Hybrid Discriminative-Generative Learning

GCPL: \( CL + \lambda \cdot PL \)
CPL: Learned Feature Space

- Prototype Loss (PL) drives the samples of each class to be compact in feature space
- Discriminative classification loss makes different classes separate

- Compactness of each class benefits robustness to outlier (which has larger distance than within-class samples)
 çoğu inakali ve çok inakali gibi ortak bir örnekli daha yüksek doğruluk CNN ile soft-max veri

<table>
<thead>
<tr>
<th>method</th>
<th>test accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>soft-max</td>
<td>99.13</td>
</tr>
<tr>
<td>CPL (DCE)</td>
<td>99.28</td>
</tr>
<tr>
<td>GCPL(DCE+PL)</td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0.0001$</td>
<td>99.45</td>
</tr>
<tr>
<td>$\lambda = 0.001$</td>
<td>99.33</td>
</tr>
<tr>
<td>$\lambda = 0.1$</td>
<td>99.29</td>
</tr>
<tr>
<td>$\lambda = 0.1$</td>
<td>99.30</td>
</tr>
</tbody>
</table>

Table 1. Test accuracy of different methods on MNIST

<table>
<thead>
<tr>
<th>CNN structure</th>
<th>soft-max</th>
<th>CPL</th>
<th>GCPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>model C [34]</td>
<td>90.26 [34]</td>
<td>90.70</td>
<td>90.80</td>
</tr>
<tr>
<td>model C with BN</td>
<td>91.37</td>
<td>91.59</td>
<td>91.90</td>
</tr>
<tr>
<td>ResNet 20</td>
<td>91.32</td>
<td>91.46</td>
<td>91.63</td>
</tr>
<tr>
<td>ResNet 32</td>
<td>92.50</td>
<td>92.60</td>
<td>92.63</td>
</tr>
</tbody>
</table>

Table 2. The accuracy of different CNN structures and different models on CIFAR-10

<table>
<thead>
<tr>
<th>loss function</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>soft-max</td>
<td>97.55 [39]</td>
</tr>
<tr>
<td>MCE</td>
<td>97.35</td>
</tr>
<tr>
<td>MCL</td>
<td>97.61</td>
</tr>
<tr>
<td>GMCL</td>
<td>97.36</td>
</tr>
<tr>
<td>DCE</td>
<td>97.58</td>
</tr>
</tbody>
</table>

Table 3. The accuracy of GCPL on OLHWDB dataset
✓ Superior rejection to outlier samples

AR: Accept rate of in-class samples (MNIST digits)
RR: Reject rate of outlier samples (CIFAR-10)

<table>
<thead>
<tr>
<th></th>
<th>softmax</th>
<th>GCPL Prob</th>
<th>GCPL Dist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR</td>
<td>RR</td>
<td>AR</td>
</tr>
<tr>
<td></td>
<td>100.0</td>
<td>0.000</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>99.98</td>
<td>0.200</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>99.72</td>
<td>8.110</td>
<td>99.80</td>
</tr>
<tr>
<td></td>
<td>99.14</td>
<td>25.17</td>
<td>99.39</td>
</tr>
<tr>
<td></td>
<td>98.52</td>
<td>40.60</td>
<td>99.30</td>
</tr>
<tr>
<td></td>
<td>97.61</td>
<td>57.54</td>
<td>99.21</td>
</tr>
<tr>
<td></td>
<td>83.95</td>
<td>71.66</td>
<td>98.96</td>
</tr>
<tr>
<td></td>
<td>76.67</td>
<td>85.97</td>
<td>98.73</td>
</tr>
<tr>
<td></td>
<td>75.49</td>
<td>98.02</td>
<td>98.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>99.20</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>97.39</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>98.07</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>98.43</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>98.57</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>98.73</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>99.09</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>99.20</strong></td>
</tr>
</tbody>
</table>

Table 4. The tradeoff between acceptance rate AR (%) and rejection rate RR (%) for different methods.
Better generalization on less training samples

Generalized accuracies with variable number of training samples

<table>
<thead>
<tr>
<th>sample size (%)</th>
<th>soft-max</th>
<th>GCPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99.13 ± 0.10</td>
<td>99.33 ± 0.10</td>
</tr>
<tr>
<td>70</td>
<td>98.37 ± 0.10</td>
<td>99.29 ± 0.10</td>
</tr>
<tr>
<td>50</td>
<td>98.07 ± 0.39</td>
<td>99.12 ± 0.10</td>
</tr>
<tr>
<td>30</td>
<td>92.68 ± 4.52</td>
<td>98.89 ± 0.10</td>
</tr>
<tr>
<td>10</td>
<td>86.12 ± 6.00</td>
<td>97.80 ± 0.22</td>
</tr>
<tr>
<td>5</td>
<td>73.95 ± 6.10</td>
<td>96.44 ± 0.40</td>
</tr>
<tr>
<td>3</td>
<td>50.79 ± 17.44</td>
<td>94.90 ± 0.58</td>
</tr>
</tbody>
</table>

Table 1. Test accuracy (%) under different percentages of training samples. It is shown that GCPL is much more robust for small sample size.
✓ Cluster of new class in feature space, facilitating incremental class classification

Extending 10-class classifier to 11-class by adding a new prototype

<table>
<thead>
<tr>
<th>new class ID (CIFAR)</th>
<th>test accuracy (11-class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.23</td>
</tr>
<tr>
<td>1</td>
<td>99.23</td>
</tr>
<tr>
<td>2</td>
<td>99.21</td>
</tr>
<tr>
<td>3</td>
<td>99.24</td>
</tr>
<tr>
<td>4</td>
<td>99.23</td>
</tr>
<tr>
<td>5</td>
<td>99.23</td>
</tr>
<tr>
<td>6</td>
<td>99.24</td>
</tr>
<tr>
<td>7</td>
<td>99.22</td>
</tr>
<tr>
<td>8</td>
<td>99.20</td>
</tr>
<tr>
<td>9</td>
<td>99.23</td>
</tr>
<tr>
<td>without new class</td>
<td>99.27</td>
</tr>
</tbody>
</table>

Table 2. Accuracy (%) of GCPL on test samples from both known and new classes
Future Research

• Better generalization on very small sample
• Better discovery of new class
• Explore more deep generative models
  – Deep generative neural networks (autoencoder, VAE)
  – Density/structural model on feature space learned by DL
• Application of discriminative generative models to learning
  – Incremental learning
  – Weakly supervised (joint detection and recognition)
  – Semi-supervised learning, with outlier
  – Transfer learning, domain adaptation
    • Domain shift: model mismatch measured by generative model
References


• Codes: https://github.com/YangHM/Convolutional-Prototype-Learning

• Contact: liucl@nlpr.ia.ac.cn
Thank You for Your Attention!