Deep Machine Learning: Panel Presentation

Honglak Lee
Computer Science & Engineering Division
University of Michigan

MSR Faculty Summit
7/16/2013
Mining meaningful structures from data

- Multimedia (images, videos, speech, music, text, etc.)

- Healthcare data (medical imaging data, preoperative conditions, time series measurements, etc.)

- Multi modal sensor networks (e.g., robotics, surveillance, etc.)
Learning Representations

• Key ideas:
  – **Unsupervised Learning**: Learn **statistical structure or correlation** of the data from **unlabeled** data (and some labeled data)
  – **Deep Learning**: Learn **multiple levels** of representation of increasing complexity/abstraction.
  – The learned representations can be used as **features** in **supervised** and **semi-supervised** settings.

• I will also talk about how to go beyond supervised (or semi-supervised) problems, such as:
  – Weakly supervised learning
  – Structured output prediction
Unsupervised learning with sparsity

[NIPS 07; ICML 07; NIPS 08]

Natural Images

Learned bases: "Edges"

Test example

\[ [0, 0, \ldots, 0, 0.8, 0, \ldots, 0, 0.3, 0, \ldots, 0, 0.5, \ldots] \]

= coefficients (feature representation)

Compact & easily interpretable
Learning object representations

• Learning objects and parts in images

• Large image patches contain interesting higher-level structures.
  – E.g., object parts and full objects

• Challenge: high-dimensionality and spatial correlations
Illustration: Learning an “eye” detector

“Eye detector”

Advantage of shrinking
1. Filter size is kept small
2. Invariance

“Shrink” (max over 2x2)

“Filtering” output

filter1

filter2

filter3

filter4

Example image
Convolutional RBM (CRBM) [ICML 2009]

For “filter” $k$,

- **Max-pooling layer $P$**
- **Detection layer $H$**
- **Input data $V$**
- **“max-pooling” node (binary)**
- **Hidden nodes (binary)**
- **“Filter” weights (shared)**

$P(v, h) \propto \exp \left( \sum_{i,j,k} h^k_{i,j} (\tilde{W}^k \ast v)_{i,j} \right)$

subj. to $\sum_{(i,j) \in \text{“cell}(y)\text{”}} h^k_{i,j} \leq 1, \forall k, y.$

- **RBM (probabilistic model)**
- **Convolutional structure**
- **Probabilistic max-pooling (“mutual exclusion”)**
Convolutional deep belief networks illustration

Input image

Layer 1

Layer 2

Layer 3

Layer 1 activation (coefficients)

Layer 2 activation (coefficients)

Layer 3 activation (coefficients)

Filter visualization

Example image
Learning object-part decomposition
Applications

• Classification (ICML 2009, NIPS 2009, ICCV 2011, Comm. ACM 2011)

• Verification (CVPR 2012)

• Image alignment (NIPS 2012)

• The algorithm is applicable to other domains, such as audio (NIPS 2009)
Ongoing Work

• Investigating theoretical connections and efficient training (ICCV 2011)
• Robust feature learning with weak supervision (ICML 2013)
• Representation learning with structured outputs (CVPR 2013)
• Learning invariant representations (ICML 2009; NIPS 2009; ICML 2012)
• Multi-modal feature learning (ICML 2011)
• Life-long representation learning (AISTAST 2012)
Ongoing Work

• Investigating theoretical connections and efficient training (ICCV 2011)
• Robust feature learning with weak supervision (ICML 2013)
• Representation learning with structured outputs (CVPR 2013)
  • Learning invariant representations (ICML 2009; NIPS 2009; ICML 2012)
  • Multi-modal feature learning (ICML 2011)
• Life-long representation learning (AISTAST 2012)
Theoretical Connections and Efficient Training

• Connections between unsupervised learning methods
  – Clustering vs. distributed representation [Coates, Lee, Ng, AISTATS 2011]
  – Can we develop better learning algorithms using the links?
• Explore the connections between mixture models and RBMs.
  – We provide an efficient training method for RBMs via the connection.
  – This is the first work showing that RBMs can be trained so that they are no worse than Gaussian Mixture models (GMMs).
• State-of-the-art results on object classification tasks.
Spherical Gaussian Mixtures is equivalent to RBM with softmax constraints

\[
P(v, h) = \frac{1}{Z} \exp(-E(v, h))
\]

\[
E(v, h) = \frac{1}{2\sigma^2} \sum_i (v_i - c_i)^2 - \frac{1}{\sigma} (\sum_{i,j} v_i W_{ij} h_j + \sum_j b_j h_j)
\]

subj. to \[\sum_j h_j \leq 1\]

Gaussian Softmax RBM

= GMM with shared covariance \[\sigma^2 I\]
Relaxing the constraints

\[
P(v, h) = \frac{1}{Z} \exp(-E(v, h))
\]

\[
E(v, h) = \frac{1}{2\sigma^2} \sum_{i} (v_i - c_i)^2 - \frac{1}{\sigma} \left( \sum_{i,j} v_i W_{ij} h_j + \sum_{j} b_j h_j \right)
\]

subject to \( \sum_{j} h_j \leq 1 \)

subject to \( \sum_{k=1}^{K} h_k \leq \alpha \),

GMM = \text{Softmax Gaussian RBM} < \text{Activation-constrained RBM}
Relaxing the constraints

\[ P(v, h) = \frac{1}{Z} \exp(-E(v, h)) \]

\[ E(v, h) = \frac{1}{2\sigma^2} \sum_i (v_i - c_i)^2 - \frac{1}{\sigma} \left( \sum_{i,j} v_i W_{ij} h_j + \sum_j b_j h_j \right) \]

subj. to \( \sum_j h_j \leq 1 \)

Gaussian Softmax RBM

subj. to \( \sum_{k=1}^{K} h_k \leq \alpha \),
activation constrained RBM

sparse RBM:
(regularize in training)
\[ \frac{1}{K} \sum_{k=1}^{K} h_k \approx \frac{\alpha}{K} \]
Experiments – Analysis

- Effect of *sparsity* to the classification performance (Caltech 101).

- The sparsity > 1/K showed the best CV accuracy.

- **Practical guarantee** that the sparse RBM lead to comparable or better classification performance than Gaussian mixtures.
Ongoing Work

• Investigating theoretical connections and efficient training (ICCV 2011)

• Robust feature learning with weak supervision (ICML 2013)

• Representation learning with structured outputs (CVPR 2013)

• Learning invariant representations (ICML 2009; NIPS 2009; ICML 2012)

• Multi-modal feature learning (ICML 2011)

• Life-long representation learning (AISTATS 2012)
Learning from scratch

• Unsupervised feature learning
  – Powerful in **discovering** features from unlabeled data.
  – However, not all patterns (or data) are equally important.
    • When data contains lots of distracting factors, learning meaningful representations can be challenging.

• Feature selection
  – Powerful in **selecting** features from labeled data.
  – However, it assumes existence of discriminative features.
    • There may not be such features at hand.

• We develop a **joint model** for feature learning and feature selection
  – allows to learn **task-relevant high-level features** using (weak) supervision.
Experiments – visualizations

- Learning from noisy handwritten digits with PGBM

Learned task-relevant hidden unit weights: mostly *pen-strokes*

Learned task-irrelevant hidden unit weights: noisy patterns

Noisy digit images (mnist-back-image)

Inferred switch variables
Experiments – visualizations

• Learning from noisy handwritten digits with PGBM

Learned task-relevant hidden unit weights: mostly *pen-strokes*

Learned task-irrelevant hidden unit weights: noisy patterns

Noisy digit images (mnist-back-image)

Inferred switch variables
We can distinguish between task-relevant and irrelevant features with point-wise gating idea while feature learning.
Experiments – weakly supervised object segmentation

• Learned set of filters (task-relevant/irrelevant)

Caltech101 - Faces

Caltech101 – car side

• (Weakly supervised) object localization

1st row: switch unit activation map,
2nd row: predicted and ground truth bounding box.
Experiments – weakly supervised object segmentation

1\textsuperscript{st} row: switch unit activation map,
2\textsuperscript{nd} row: predicted and ground truth bounding box.
Ongoing Work

• Investigating theoretical connections and efficient training (ICCV 2011)
• Robust feature learning with weak supervision (ICML 2013)
• **Representation learning with structured outputs** (CVPR 2013)
  • Learning invariant representations (ICML 2009; NIPS 2009; ICML 2012)
  • Multi-modal feature learning (ICML 2011)
  • Life-long representation learning (AISTATS 2012)
Enforcing Global and Local Consistencies for Structured Output Prediction

- Task: scene segmentation
- Problem: only enforces local consistency
- Our model can enforce both local and global consistency

(CVPR 2013)
Combining Global and Local Consistencies for Structured Output Prediction

(CVPR 2013)

\[ P(Y|X) = \frac{1}{Z} \sum_h \exp \{-E(X, Y, h; I)\} \]

\[ E(X, Y, h; I) = E_{\text{crf}}(X, Y) + E_{\text{rbm}}(Y, h) \]

\[ E_{\text{rbm}}(Y, h; I) = -\sum_{r=1}^{R^2} \sum_{l=1}^{L} \sum_{k=1}^{K} \bar{y}_{rl} W_{rlk} h_k \]

\[ -\sum_{k=1}^{K} b_k h_k - \sum_{r=1}^{R^2} \sum_{l=1}^{L} c_{rl} \bar{y}_{rl} \]

where \( \bar{y}_{rl} \triangleq \sum_{s=1}^{S(I)} p_{rs}^{(I)} y_{sl} \)
Experimental results

- Visualization of segmentation

- LR: singleton potential
- CRF: singleton + pairwise potential
- Ours: singleton + pairwise + RBM potential

(CVPR 2013)
Summary

• Generative learning of convolutional feature hierarchy
• Better training algorithms
• Learning representations with weak supervision
• Learning representations with structured outputs

Funding support:
  – NSF, ONR, Google, Toyota, Bosch Research