Restricted Boltzmann Machines and their extensions

P(v, h) = exp(−E(v, h))/Z

• Binary RBM
E(v, h) = −v⊤W h − b⊤v − a⊤h

• Gaussian RBM
E(v, h) = ∑_{i,j} (v_i - h_i)^2 - ∑_{i,j} W_{i,j} h_j - a⊤h

• Replicated Softmax Model
E(v, h) = −∑_{k,l} v_k W_{k,l} h_j - ∑_{k,l} v_k b_l - M ∑_{j} h_j a_j

Multimodal Deep Belief Net

• Multimodal “pathways” combined with a top-level RBM
  - First layer RBMs are modality-specific - Gaussian for images, Replicated Softmax for text
  - Each successive layer learns higher-level features, abstracts away modality-specific correlations
  - Top-level RBM jointly models high-level image and text features
  - Easier to discover cross-modal relations since both sets of features are now binary and sparse.
  - In contrast, input representations were widely different, which makes it difficult for shallow models to find cross-modal relations.

Learning - Greedy layer-wise training with Persistent Contrastive Divergence

Generative Tasks - Sample conditional models using MCMC methods
  - Retrieve images using P(v_m|v_t)
  - Annotate images using P(v_t|v_m)

Discriminative Tasks - Use DBN to initialize feedforward network
  - Multimodal Inputs - use both pathways
  - Unimodal Inputs - infer unknown pathway with Gibbs Sampling (DBN-GenText)

Data pre-processing - Images - extract SIFT, Gist, MPEG-7 descriptors, 2000 most frequent tags

Goal: Use unlabeled multimodal data to
  - Learn joint “modality-free” representation
  - Infer missing modalities given some observed ones

Method: Build a joint density model using a DBN
  - Use states of top level hidden units as joint representation
  - Sample from conditional density model to fill in missing data

Data: Multimedia Information Retrieval Flickr dataset
  - 1M images with noisy (sometimes missing) user-assigned tags
  - 25K annotated with 38 topics e.g. sky, tree, animals, baby, water (Used only for classification experiments).

Joint Distribution

The Multimodal DBN implies the following joint distribution

P(v_m, v_t) = ∑_{h^{(2)}_m} P(h^{(2)}_m|k^{(2)}_m) P(k^{(2)}_m|k^{(3)}_m) × ∑_{h^{(2)}_t} P(v_t|h^{(2)}_t) P(h^{(2)}_t|h^{(1)}_t) × ∑_{h^{(1)}_m} P(v_m|h^{(1)}_m) P(h^{(1)}_m|v^{(1)}_m)

Multimodal Deep Belief Net

Sampling from conditional model P(v_t|v_m)

Model Description

Multimodal DBN ≡ Unimodal “pathways” combined with a top-level RBM

Joint Distribution

Classification Results on MIR-Flickr dataset

Task: Predict whether input belongs to a user-annotated topic. Results are averaged over all topics.

Multimodal Inputs

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
<th>Prec@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.124</td>
<td>0.124</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>0.482</td>
<td>0.751</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>0.475</td>
<td>0.754</td>
</tr>
<tr>
<td>DBN-Labeled-Data</td>
<td>0.503</td>
<td>0.741</td>
</tr>
<tr>
<td>Deep Autoencoder</td>
<td>0.547</td>
<td>0.794</td>
</tr>
<tr>
<td>DBN</td>
<td>0.563</td>
<td>0.785</td>
</tr>
</tbody>
</table>

Unimodal Inputs

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
<th>Prec@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image-SVM</td>
<td>0.375</td>
<td>-</td>
</tr>
<tr>
<td>Image-DBN</td>
<td>0.413</td>
<td>0.718</td>
</tr>
<tr>
<td>Text-DBN</td>
<td>0.471</td>
<td>0.723</td>
</tr>
<tr>
<td>DBN-ZeroText</td>
<td>0.484</td>
<td>0.730</td>
</tr>
<tr>
<td>DBN-GenText</td>
<td>0.492</td>
<td>0.762</td>
</tr>
</tbody>
</table>

Learning Representations for Multimodal Data with Deep Belief Nets

Nitish Srivastava, Ruslan Salakhutdinov
Department of Computer Science, University of Toronto

Introduction

• Real world data is often multimodal - Captioned images, video, sensory perception
• Strong associations exist across modalities but hard to discover in terms of low-level features
• Goal: Use unlabeled multimodal data to
  - Learn joint “modality-free” representation
  - Infer missing modalities given some observed ones

Method: Build a joint density model using a DBN
  - Use states of top level hidden units as joint representation
  - Sample from conditional density model to fill in missing data

Data: Multimedia Information Retrieval Flickr dataset
  - 1M images with noisy (sometimes missing) user-assigned tags
  - 25K annotated with 38 topics e.g. sky, tree, animals, baby, water (Used only for classification experiments).

Restricted Boltzmann Machines and their extensions

P(v, h) = exp(−E(v, h))/Z

• Binary RBM
E(v, h) = −v⊤W h − b⊤v − a⊤h

• Gaussian RBM
E(v, h) = ∑_{i,j} (v_i - h_i)^2 - ∑_{i,j} W_{i,j} h_j - a⊤h

• Replicated Softmax Model
E(v, h) = −∑_{k,l} v_k W_{k,l} h_j - ∑_{k,l} v_k b_l - M ∑_{j} h_j a_j

Multimodal Deep Belief Net

• Multimodal “pathways” combined with a top-level RBM
  - First layer RBMs are modality-specific - Gaussian for images, Replicated Softmax for text
  - Each successive layer learns higher-level features, abstracts away modality-specific correlations
  - Top-level RBM jointly models high-level image and text features
  - Easier to discover cross-modal relations since both sets of features are now binary and sparse.
  - In contrast, input representations were widely different, which makes it difficult for shallow models to find cross-modal relations.

Learning - Greedy layer-wise training with Persistent Contrastive Divergence

Generative Tasks - Sample conditional models using MCMC methods
  - Retrieve images using P(v_m|v_t)
  - Annotate images using P(v_t|v_m)

Discriminative Tasks - Use DBN to initialize feedforward network
  - Multimodal Inputs - use both pathways
  - Unimodal Inputs - infer unknown pathway with Gibbs Sampling (DBN-GenText)

Data pre-processing - Images - extract SIFT, Gist, MPEG-7 descriptors, 2000 most frequent tags