Deep Learning of Invariant Spatiotemporal Features from Video

- Presenter: Bo Chen
- Supervisor: Nando de Freitas
Introduction

• What:
  • Unsupervised learning
  • Invariance Features
  • Distributed representations

• How:
  • Density model of videos
  • Deep hierarchy
  • Undirected neural network

• Contributions
Overview

• Background
  • Restricted Boltzmann Machines (RBMs)
  • Convolutional RBMs
• Our Model: Spatiotemporal Deep Belief Networks (STDBNs)
• Experiments
  • Measuring Invariance
  • Action Recognition for KTH dataset
  • Denoising
  • Filling-in from “saccades”
• Conclusions and Acknowledgements
Overview

- **Background**
  - Restricted Boltzmann Machines (RBMs)
  - Convolutional RBMs
  - Our Model: Spatiotemporal Deep Belief Networks (STDBNs)
- **Experiments**
  - Measuring Invariance
  - Action Recognition for KTH dataset
  - Denoising
  - Filling-in from “saccades”
- **Conclusions and Acknowledgements**
Background: Restricted Boltzmann Machines

\[ P(h_i = 1|v) = \text{logistic} \left( (W^j)^T v \right) \]

Binary \( v \):
\[ P(v_i = 1|h) = \text{logistic} \left( (W_i)^T h \right) \]

Continuous \( v \):
\[ P(v_i = 1|h) = \mathcal{N} \left( (W_i)^T h, 1 \right) \]
Background:
Convolutional Restricted Boltzmann Machines
Background:
Convolutional Restricted Boltzmann Machines

Standard RBMs:

\[
P(h_i = 1 | v) = \text{logistic} \left( (W^j)^T v \right)
\]

\[
P(v_i = 1 | h) = \text{logistic} \left( (W_i)^T h \right)
\]

Convolutional RBMs:

\[
P(h^g, p^g | v) = \text{ProbMaxPool} (W^g \ast v^g)
\]

\[
P(v | h) = \text{logistic} \left( \sum_{g=1}^{\lvert W \rvert} W^g \ast h^g \right)
\]

Norouzi09cvpr, Lee09icml
Overview

• Background
  • Restricted Boltzmann Machines (RBMs)
  • Convolutional RBMs
  • Our Model: Spatiotemporal Deep Belief Networks (STDBNs)

• Experiments
  • Measuring Invariance
  • Action Recognition for KTH dataset
  • Denoising
  • Filling-in from “saccades”

• Conclusions and Acknowledgements
Our Model:
Spatiotemporal Deep Belief Networks
Our Model:
Spatiotemporal Deep Belief Networks

- Spatial Pooling Layer
- Temporal Pooling Layer

Wednesday, June 16, 2010
Our Model: STDBNs: Training

- DBN: Greedy Layer-wise Pre-training (hinton06neurocomp)
- Each layer: Contrastive Divergence (CD) (hinton05aistats)
- Sparsity Regularization (e.g. olshausen96nature)
- Some numbers:
  - Number of Parameters: $2K+10K+260K+30K\sim300K$ (STDBN without convolution: 3B)
  - Training time: $20h+12h+72h+20h\sim5$ days
Overview

- Background
  - Restricted Boltzmann Machines (RBMs)
  - Convolutional RBMs
- Our Model: Spatiotemporal Deep Belief Networks (STDBNs)
- Experiments
  - Measuring Invariance
  - Action Recognition for KTH dataset
  - Denoising
  - Filling-in from “saccades”
- Conclusions and Acknowledgements
STDBN as a Discriminative Feature Extractor: Measuring Invariance

Firing Rate of Unit 1

 invariant

Overly Selective

Not Selective

Degree of Transformation

Invariance scores for common transformations in natural videos, computed for layer 1 (S1) and layer 2 (S2) of a CDBN and layer 2 (T1) of STDBN (Higher is better).
STDBN as a Discriminative Feature Extractor: Unsupervised Feature Learning for Action Recognition

- KTH Dataset: 2391 videos (~1GB)
- Previous approaches: dean09Iism, laptev08cvpr, dollar05vspets, liu08cvpr, taylor10snowbird, wang09ijcai...
STDBN as a Discriminative Feature Extractor: Unsupervised Feature Learning for Action Recognition

Pipeline

Input Video → Local and Global Contrast Normalization → Input to STDBN → Unsupervised Feature Extraction using STDBN → Dimensionality standardization → STDBN Responses → Input to SVM

Wednesday, June 16, 2010
STDBN as a Discriminative Feature Extractor: Unsupervised Feature Learning for Action Recognition

• show bases
**STDBN as a Discriminative Feature Extractor: Unsupervised Feature Learning for Action Recognition**

<table>
<thead>
<tr>
<th>Layer 1</th>
<th>Layer 2</th>
<th>Layer 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Box</strong></td>
<td><strong>Clap</strong></td>
<td><strong>Wave</strong></td>
</tr>
<tr>
<td>89.4</td>
<td>4.7</td>
<td>2.7</td>
</tr>
<tr>
<td>2.8</td>
<td><strong>89.7</strong></td>
<td>6.7</td>
</tr>
<tr>
<td>0.8</td>
<td>2.0</td>
<td><strong>96.9</strong></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-layer ST-DBN</td>
<td>90.3 ± 0.83</td>
</tr>
<tr>
<td>3-layer ST-DBN</td>
<td><strong>91.13 ± 0.85</strong></td>
</tr>
<tr>
<td>2-layer ST-DBN</td>
<td>89.73 ± 0.18</td>
</tr>
<tr>
<td>1-layer ST-DBN</td>
<td>85.97 ± 0.94</td>
</tr>
<tr>
<td>Liu &amp; Shah [21]</td>
<td>94.2</td>
</tr>
<tr>
<td>Wang &amp; Li [31]</td>
<td>87.8</td>
</tr>
<tr>
<td>Dollár et al. [18]</td>
<td>81.2</td>
</tr>
</tbody>
</table>
STDBN as a Generative Model: De-noising

Show video of de-noising

Clean Video  Noisy Video  Spatially Denoised Video  Spatiotemporally Denoised Video
NMSE = 1  NMSE=0.175  NMSE=0.15
STDBN as a Generative Model: Filling-in from “Saccades”
STDBN as a Generative Model: Filling-in from “Saccades”
STDBN as a Generative Model: Filling-in from “Saccades”

Missing

Prediction

Truth (Fully Observed)

Hidden Layer for Missing

Hidden Layer for Truth

Wednesday, June 16, 2010
Overview

• Background
  • Restricted Boltzmann Machines (RBMs)
  • Convolutional RBMs
• Our Model: Spatiotemporal Deep Belief Networks (STDBNs)
• Experiments
  • Measuring Invariance
  • Action Recognition for KTH dataset
  • Denoising
  • Filling-in from “saccades”
• Conclusions and Acknowledgements
Discussion

- Limitations
  - computation
  - trade off invariance vs reconstruction
  - alternating vs joint space-time convolution
- Extensions
  - leveraging feature detection to reduce training data size
  - optimize gaze planning
Conclusions

- STDBN has:
  - feature invariance (invariance score)
  - hierarchical distributed representations
  - good discriminative performance (action recognition)
  - generative properties (de-noising, reconstruction, prediction, etc)
Publications


Benjamin Marlin, Kevin Swersky, Bo Chen and Nando de Freitas. Inductive Principles for Restricted Boltzmann Machine Learning. *International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*

Haiyang Wang, Jiangchuan Liu, Bo Chen and Ke Xu. Understand the Tracker Selection for BitTorrent Traffic Locality. *The IEEE International Conference on Peer-to-Peer Computing 2010*


Acknowledgements

- Jo-Anne Ting
- Benjamin Marlin
- Kevin Swersky
- Nando de Freitas
- All of you!
Probabilistic Max

\[ I^g = \hat{W}^g \ast v \]

\[
P(h_{i,j}^g = 1|v) = \frac{\exp(I_{i,j}^g)}{1 + \sum_{r,s \in B_\alpha} \exp(I_{r,s}^g)}
\]

\[
P(p_{\alpha}^g = 1|v) = \frac{\sum_{r,s \in B} \exp(I_{r,s}^g)}{1 + \sum_{r,s \in B} \exp(I_{r,s}^g)}
\]
Invariance Measure

\[ L(i) = \frac{1}{|Z|} \sum_{z \in Z} \frac{1}{|T(z)|} \sum_{x \in T(z)} f_i(x) \]  
Local Firing Rate

\[ G(i) = E[f_i(x)] \]  
Global Firing Rate

\[ S(i) = \frac{L(i)}{G(i)} \]  
Invariance Measure (Goodfellow09nips)
\[ \partial W = v \ast \tilde{h} - \tilde{v} \ast \tilde{h} \]
Figure 8, Denoising results, uav Test frame; ubv Test frame corrupted with noise; ucv Reconstruction using 4zlayer STzD]N; udv Reconstruction with 5zlayer STzD]Nfi

comes at a cost when inferring missing parts of framesy it is crucial for good discriminative perforz offame. Future research must address this fundamental tradezoff. The results in the figure, though apparently simple, are quite remarkable. They represent an important step toward the design of attentional mechanisms for gaze planning. While gazing at the subject's head, the model is able to infer where the legs are. This coarse resolution gist may be used to guide the placement of high-resolution detectors.

Figure 9, Top video shows an observed sequence of gazes; uifiefiy frames 5z7vfi. Bottom video shows reconstructions within the gaze windows and predictions outside them.

5 Conclusions

In this paper, we introduced a hierarchical distributed probabilistic model for learning invariant features from spatio-temporal data. Using 'RMs as a building block, our model, the Space-Time Deep Belief Network, pools over space and time. It fulfills all the four desirable properties of a feature extractor that we reviewed in the introduction. In addition to possessing feature invariance for selectivity and robustness to input transformations and a hierarchical, distributed representation, our model is generative and shows good discriminative performance on a simple human action recognition task. Testing on larger video databases is an obvious immediate avenue for further research. Interestingly, the max-pooling operation that allows feature invariance to be captured hierarchically from spatio-temporal data has an adverse effect for predicting missing parts of a video sequence. To address this issue, future work will examine how to minimize the information loss associated with max-pooling when performing inference. We conjecture that combinations of models with and without pooling might be required. Additionally, precautions should be taken to ensure representations are not made too compact with too many layers in the architecture. Model selection is an open challenge in this line of research. Finally, we plan to build on the gaze prediction results. Our intention is to use planning to optimize the gaze locations so as to solve various recognition and verification tasks efficiently.

References


[5] Ifi Goodfellow, Qfi Ley [fi Saxey and [fiY fi Ng, Measuring invariances in deep networks. NIPS. 522:.
Filling-in