CAP 6412 Advanced Computer Vision

Boqing Gong
April 14th, 2016
The source code from David & Mahdi

• LSTM demo in Keras:
  • Training text self-contained

• LSTM demo in Chainer:
  • Training videos & labels available upon request
Listening With Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines

Chao Sui, Mohammed Bennamoun, Roberto Togneri
University of West Australia

ICCV 2015

Presented by Javier Lores
Outline

- Motivation
- Problem Statement
- Main Contributions
- Methods
- Results
Motivation

• Next generation of Human-Computer Interaction will require perceptual intelligence
  – What is the environment?
  – Who is in the environment?
  – Who is speaking?
  – What is being said?
  – What is the state of the speaker?
Motivation

- A famous exchange (HAL's “premature” audio-visual speech processing capability):
  - HAL: I knew that you and David were planning to disconnect me, and I'm afraid that's something I cannot allow to happen
  - Dave: Where the hell did you get that idea, HAL?
  - HAL: Dave- although you took very thorough precautions in the pod against my hearing you, I could see your lips move.
Motivation

- Although the visual speech information content is less than audio
  - Phonemes: Distinct speech units that convey linguistic information
    - 47 in English
  - Visemes: Visually distinguishable classes of phonemes
    - 6-20 in English
- The visual channel provides important complementary information to the audio
  - Consonant confusions in audio drop by 76% when compared to audio-visual
Motivation

- Human speech production and perception is bimodal
  - We lip read in noisy environments to improve intelligibility
  - E.g. Human speech perception experiment by Summerfield (1979): Noisy word recognition at low SNR
Motivation

- Audio-visual automatic speech recognition (AV-ASR)
  - Utilizes both audio and visual signal inputs from the video of a speaker's face to obtain the transcript of the spoken utterance
  - Issues: Audio extraction, visual feature extraction, audio-visual integration
Motivation

- Audio-visual speech synthesis (AV-TTS)
  - Given text, create a talking head
  - Should be more natural and intelligence than audio-only TTS
- Audio-visual speaker recognition
  - Authenticate speaker
- Audio-visual speaker localization
  - Which person in the video is talking?
Problem Statement

• Operation on basis of traditional audio-only information
  – Lacks robustness to noise
  – Lags human performance significantly, even in ideal environments

• Joint audio+visual processing system can help bridge the usability gap
Main Contributions

• Create Audio-Visual Speech Recognition (AVSR) system
  – Uses both audio and visual signals to enrich the visual feature representation
  – Although, audio and visual are both required for training, only visual is used for testing since the missing audio modality is able to be inferred.
Architecture

- **Training**
  - Train Multimodal Deep Boltzmann Machine (DBM)
    - Visual Feature
      - Extract Local Binary Patterns-Three Orthogonal Planes (LBP-TOP)
    - Audio Feature
      - Extract Mel-Frequency Cepstral Coefficients (MFCC) from audio signal
  
- **Testing**
  - Generate missing audio features from DBM
  - Input generated audio features and visual feature into DBM
  - Concatenate DBM features with Discrete Cosine Transform (DCT)
  - Perform Linear Discriminant Analysis (LDA) to decorrelate the feature and reduce feature dimensions
  - Input feature into Hidden Markov Model (HMM) to perform classification
Figure 2. Block diagram of our proposed system. The left side of the figure shows the training phase. The visual feature is learned from both the audio and video stream using a multimodal DBM. The right side of the figure shows the testing phase, where the audio signal is not used. In the testing phase, the audio is generated by clamping the video input and sampling the audio input from the conditional distribution.
Boltzmann Machine

- Network of symmetrically stochastic binary units
- It is a type of stochastic Recurrent Neural Network (RNN) and Markov Random Field (MRF)
- Contains hidden and visible units
BM Energy and Probability

\[ \nu \in \{0,1\}^D \quad \eta \in \{0,1\}^P \quad \theta = \{W, L, J\} \]

\[ E(\nu, \eta | \theta) = -\frac{1}{2} \nu' L \nu - \frac{1}{2} \eta' J \eta - \nu' W \eta \]

\[ p(\nu, \eta | \theta) = \frac{p^*(\nu; \theta)}{Z(\theta)} = \frac{1}{Z(\theta)} \exp\left(-E(\nu, \eta; \theta)\right) \]

\[ Z(\theta) = \sum_{\nu} \sum_{\eta} \exp\left(-E(\nu, \eta; \theta)\right) \]
Boltzmann Machine

• Problems
  – Exact inference is intractable
    • Exponential in number of hidden units
  – Approximate inference is slow
    • E.g. Gibbs Sampling
Restricted Boltzmann Machines

- Restrict the architecture of a BM
  - Nodes are arranged in layers
  - No connections between nodes in same layer
  - Creates a bipartite graph

- New Energy Function:

\[ E(v, h | \theta) = -v'Wh \]
BM vs RBM

Figure 1: **Left**: A general Boltzmann machine. The top layer represents a vector of stochastic binary “hidden” features and the bottom layer represents a vector of stochastic binary “visible” variables. **Right**: A restricted Boltzmann machine with no hidden-to-hidden and no visible-to-visible connections.
DBMs and DBNs

• Stacked RBMs

"Deep Boltzmann Machines", by Ruslan Salakhutdinov
DBM Energy and Probability

\[ \theta = \{ W^{(1)}, W^{(2)}, \ldots, W^{(n-1)} \} \]

\[ E(v, h|\theta) = -v' W^{(1)} h^{(1)} - \sum_{i=2}^{n} h^{(i-1)}' W^{(i-1)} h^{(i)} \]

\[ P(v|\theta) = \sum_{h^{(1)}, \ldots, h^{(n)}} P(v, h^{(1)}, \ldots, h^{(n)}|\theta) \]

\[ = \frac{1}{Z(\theta)} \sum_{h^{(1)}, \ldots, h^{(n)}} \exp(-E(v, h^{(1)}, \ldots, h^{(n)}|\theta)) \]
DBM Pretraining

- Greedy layer-wise pretraining of DBM
  - Using Contrastive Divergence (CD)
- Low-level RBM lacks top-down input
  - Input must be doubled
- Top-level RBM lacks bottom-up input
  - Hidden units must be doubled
- Intermediate layers
  - RBM weights are doubled
"Listening With Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines", by Chao Sui et al.

Figure 2: Left: Pretraining a DBM with three hidden layers consists of learning a stack of RBM’s that are then composed to create a Deep Boltzmann Machine. Right: Resulting Deep Boltzmann Machine, where the parameters \(\{R^1, R^2, R^3\}\) define the recognition model.
DBM Fine-Tuning

• Exact maximum likelihood learning is intractable

• Approximate learning $Q(h_i|v_i)$
  - Approximate the posterior distribution
    • Using mean-field inference $Q(h_i|v_i)$
  - Then use the marginals of to augment the input vector
"Listening With Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines", by Chao Sui et al.
Generating Missing Modality

- DBM is an undirected generative model
  - Audio signals can be inferred from visual signals
- Given the observed visual features
  - Clamp the visual feature at input
  - Gibbs sample the hidden units from the conditional distribution
Inferring Audio Diagram

Figure 3. The generation of the missing audio signals can be divided into two steps: 1. Infer the audio signal from the given visual features. 2. Generate a joint representation using both the reconstructed audio and the given visual features.
Conditional Distribution Eq.

\[ P(h_j^k = 1| h^{k-1}, h^{k+1}) = \sigma \left( \sum_i W_{ij} h_i^{k-1} + \sum_m W_{jm} h_m^{k+1} \right) \]

\[ P(h_m^n = 1| h^{n-1}) = \sigma \left( \sum_j W_{jm} h_j^{n-1} \right) \]

\[ P(v_i = 1| h^1) = \sigma \left( \sum_i W_{ij} h_j^1 \right) \]
Inferring Audio Algorithm

\begin{algorithm}
\begin{algorithmic}
\STATE \textbf{Algorithm 1} Process of generating the missing audio feature
\STATE Clamp the observed visual feature $v_v$ at the input.
\FOR {Each hidden layer $k$ in visual stream}
\STATE Gibbs Sample the hidden layer state in a bottom-up manner, and estimate $P(h_i^k = 1|h_{k-1}, h_{k+1}^1)$ using Eq. 7
\ENDFOR
\STATE Gibbs sample the joint layer state, and estimate $P(h_{m}^n = 1|h_{n-1}^1)$ using Eq. 8
\FOR {Each hidden layer $k$ in audio stream}
\STATE Gibbs Sample the hidden layer state in a top-down manner, and estimate $P(h_j^k = 1|h_{k-1}, h_{k+1}^1)$ using Eq. 9
\ENDFOR
\STATE Infer the missing audio feature using Eq. 9.
\STATE Gibbs Sample the joint representation in a bottom-up manner by feeding both the reconstructed audio and observed visual features into the network.
\end{algorithmic}
\end{algorithm}
Visual Features

- **LBP-TOP**
  - Allows capture of spatial and temporal information
    - i.e. appearance of mouth and lip movement
  - Similar to Volume Local Binary Patterns (VLBP)
  - Take 3 orthogonal planes
    - Computationally cheaper
- Given a mouth ROI frame
  - 59-bin histogram generated for each plane using LBP-TOP
  - Concatenate 3 histograms for 177-dimensional feature vector
  - Subdivided mouth region into 2x5 subregions
  - Input units is 1770
"Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions", by Guoying Zhao et al.
"Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions", by Guoying Zhao et al.
Audio Features

- MFCC
  - Take the Fourier transform of a signal
  - Map the powers of the spectrum onto the mel scale
  - Take the logs of the powers as each mel frequency
  - Take the DCT of the list of mel log powers
  - The MFCCs are the amplitudes of the resulting spectrum
- 13 MFCCs are extracted with the zero-th coefficient appended
- Then, each 13-dimensional MFCCs were stacked across 11 consecutive frames which results in a total of 143 coefficients for each frame
Data Corpus

- **AusTalk**
  - Australian wide research project
  - Large-scale audio-visual corpus spoken Australian English
  - Only the digit sequence data subset is used
    - 12 four-digit strings are provided for people to read

- **AVLetters**
  - British English corpus
  - 10 speakers saying the letters A to Z three times each
Different Audio Features

Table 2. Connected digit recognition performance with multimodal inputs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Audio Input</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Boltzmann Machine</td>
<td>Zero padding</td>
<td>34.2%</td>
</tr>
<tr>
<td></td>
<td>Clean audio</td>
<td>79.8%</td>
</tr>
<tr>
<td></td>
<td>Inferred audio</td>
<td>59.9%</td>
</tr>
</tbody>
</table>
Learned Feature Variants Comparison

Table 3. Performance comparison between the DBM learned feature with its variants proposed in this paper.

<table>
<thead>
<tr>
<th>Visual Feature</th>
<th>Reduction</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBM learned feature</td>
<td>None</td>
<td>59.9%</td>
</tr>
<tr>
<td>DBM learned feature + LDA</td>
<td>LDA</td>
<td>64.4%</td>
</tr>
<tr>
<td>DBM learned feature + DCT</td>
<td>LDA</td>
<td>69.1%</td>
</tr>
</tbody>
</table>
Comparison to VSR Methods

Table 4. Performance comparison between our proposed method with other feature learning and extraction techniques.

<table>
<thead>
<tr>
<th>Feature Representation</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT + LDA</td>
<td>54.7%</td>
</tr>
<tr>
<td>DCT + MMI</td>
<td>52.3%</td>
</tr>
<tr>
<td>DCT + mRMR</td>
<td>52.2%</td>
</tr>
<tr>
<td>DCT + CMI</td>
<td>51.1%</td>
</tr>
<tr>
<td>LBP-TOP [29] + MMI [19]</td>
<td>52.5%</td>
</tr>
<tr>
<td>LBP-TOP [29] + mRMR [16]</td>
<td>53.1%</td>
</tr>
<tr>
<td>Deep Bottleneck Feature [23]</td>
<td>57.3%</td>
</tr>
<tr>
<td>Augmented Deep Bottleneck Feature [23]</td>
<td>67.8%</td>
</tr>
<tr>
<td><strong>Our proposed method</strong></td>
<td><strong>69.1%</strong></td>
</tr>
</tbody>
</table>
Future Directions

• Try different visual features
  – Convolutional Neural Network (CNN)
• What about inferring the visual features from the audio?
  – Can this be used to improve the audio recognition?