**OVERVIEW**

Lip-reading is the task of recognizing speech solely from the visual movement of the mouth.

Our neural network consists of the following parts:
- convolutional neural network (CNN)
- bidirectional long short-term memory (BLSTM)
- maxout activation unit

The whole network is trained end-to-end under low-resource scenario with no pretraining or extra data.

**DATA PREPROCESSING**

Ouluvs2 corpus:
- 52 subjects (40 train, 12 test)
- 10 phrases (3 samples/subject)
- 156 samples/phrase

Major preprocessing steps (for each video clip):
- loseless grayscale image sequence conversion
- 1:2 crop around mouth region
- contrast enhancement
- down sampling (16×32)
- data augmentation (16×32)
- z-normalization across each pixel

Input formation:
- slide a window along each image sequence
- concatenate every 8 consecutive frames

**MAXOUT UNIT**

Maxout unit has the following advantages over ReLU:
- more accurate approximate model averaging
- less affected by high saturate rate at zero

It can be characterized by this simple formula:
\[ h_i(x) = \max_{j \in [1,k]} \{ x^T W_{ij} + b_{ij} \}, \]
where \( x \in \mathbb{R}^d, W \in \mathbb{R}^{d \times m \times k} \) and \( b \in \mathbb{R}^{m \times k} \)

**RESULTS**

<table>
<thead>
<tr>
<th>Method (( k = 4 ) for maxout)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-encoder with tanh-BLSTM</td>
<td>84.5</td>
</tr>
<tr>
<td>ReLU-CNN with tanh-BLSTM</td>
<td>84.6</td>
</tr>
<tr>
<td>ReLU-CNN with maxout-BLSTM</td>
<td>84.4</td>
</tr>
<tr>
<td>maxout-CNN with tanh-BLSTM</td>
<td>85.6</td>
</tr>
<tr>
<td>maxout-CNN-BLSTM</td>
<td>87.6</td>
</tr>
</tbody>
</table>

Maxout provides a significant gain in accuracy in comparison with tanh and ReLU activations.

**DISCUSSION**

Comparisons to auto-encoder-BLSTM:
- pure end-to-end: no separate pretraining stage
- 3D convolutional: capturing local spatial and temporal correlations

Techniques in training with CNN-BLSTM:
- reduce feeding size from CNN to LSTM (2×2×2)
- use a better activation (maxout)
- help prevent overfitting (batch normalization, dropout, and L2-regularization)
- create more amount of data (data augmentation)

**CONCLUSION**

In this work, we have successfully demonstrated:
- feasibility of designing an end-to-end network with CNN and LSTM
- superiority of incorporating maxout activation
- a state-of-the-art accuracy of 87.6% on the low-resource Ouluvs2 10-phrase corpus.

**FUTURE WORK**

In the future, we are going to:
- explore more challenging lip-reading tasks
- utilize other end-to-end architectures

[NOTES]

**END-TO-END LOW-RESOURCE LIP-READING WITH MAXOUT CNN AND LSTM**

Ivan Fung and Brian Mak

Department of Computer Science and Engineering - The Hong Kong University of Science and Technology

10-Phrase Ouluvs2 Corpus:
- 52 subjects (40 train, 12 test)
- 10 phrases (3 samples/subject)
- 156 samples/phrase

Major preprocessing steps (for each video clip):
- loseless grayscale image sequence conversion
- 1:2 crop around mouth region
- contrast enhancement
- down sampling (16×32)
- data augmentation (16×32)
- z-normalization across each pixel

Input formation:
- slide a window along each image sequence
- concatenate every 8 consecutive frames

**NETWORK ARCHITECTURE**

Figure 1: Network architecture of the maxout-CNN-BLSTM model. C: Channel; BN: Batch Normalization; D: Dropout.

**DISCUSSION**

Comparisons to auto-encoder-BLSTM:
- pure end-to-end: no separate pretraining stage
- 3D convolutional: capturing local spatial and temporal correlations

Techniques in training with CNN-BLSTM:
- reduce feeding size from CNN to LSTM (2×2×2)
- use a better activation (maxout)
- help prevent overfitting (batch normalization, dropout, and L2-regularization)
- create more amount of data (data augmentation)

**RESULTS**

<table>
<thead>
<tr>
<th>Method (( k = 4 ) for maxout)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-encoder with tanh-BLSTM</td>
<td>84.5</td>
</tr>
<tr>
<td>ReLU-CNN with tanh-BLSTM</td>
<td>84.6</td>
</tr>
<tr>
<td>ReLU-CNN with maxout-BLSTM</td>
<td>84.4</td>
</tr>
<tr>
<td>maxout-CNN with tanh-BLSTM</td>
<td>85.6</td>
</tr>
<tr>
<td>maxout-CNN-BLSTM</td>
<td>87.6</td>
</tr>
</tbody>
</table>

Maxout provides a significant gain in accuracy in comparison with tanh and ReLU activations.

**CONCLUSION**

In this work, we have successfully demonstrated:
- feasibility of designing an end-to-end network with CNN and LSTM
- superiority of incorporating maxout activation
- a state-of-the-art accuracy of 87.6% on the low-resource Ouluvs2 10-phrase corpus.

**FUTURE WORK**

In the future, we are going to:
- explore more challenging lip-reading tasks
- utilize other end-to-end architectures