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

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**.

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:

$$h_i(\mathbf{x}) = \max_{j\in[1,k]}\{\mathbf{x}^T\mathbf{W}_{\cdot ij} + b_{ij}\}\ ,$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{W} \in \mathbb{R}^{d\times m\times k}$ and $\mathbf{b} \in \mathbb{R}^{m\times k}$

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×)
- z-normalization across each pixel

Input formation:

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

RESULTS

Table 1: Classification accuracy of various models.

Method ($k = 4$ for maxout)	Accuracy (%)
Auto-encoder with tanh-BLSTM	84.5
ReLU-CNN with tanh-BLSTM	84.6
ReLU-CNN with maxout-BLSTM	84.4
maxout-CNN with tanh-BLSTM	85.6
maxout-CNN-BLSTM	87.6

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

Table 2: Training time (hr) of various models (each run).

Method ($k = 4$ for maxout)	Time (hr)
ReLU-CNN with tanh-BLSTM	2.4
ReLU-CNN with maxout-BLSTM	2.5
maxout-CNN with tanh-BLSTM	7.8
maxout-CNN-BLSTM	7.8

There is a significant training time increase in CNN with maxout due to parameter size increase, while that in BLSTM is relatively minor.

Table 3: Effect of various number of maxout feature maps, k.

maxout-CNN-BLSTM	Accuracy (%)	Time (hr)
$k = 2$	85.6	4.2
$k = 3$	86.1	6.2
$k = 4$	87.6	7.8
$k = 5$	86.3	10.0

Training time increases along with k, and $k = 4$ gives a slightly better accuracy among the others.

NETWORK ARCHITECTURE

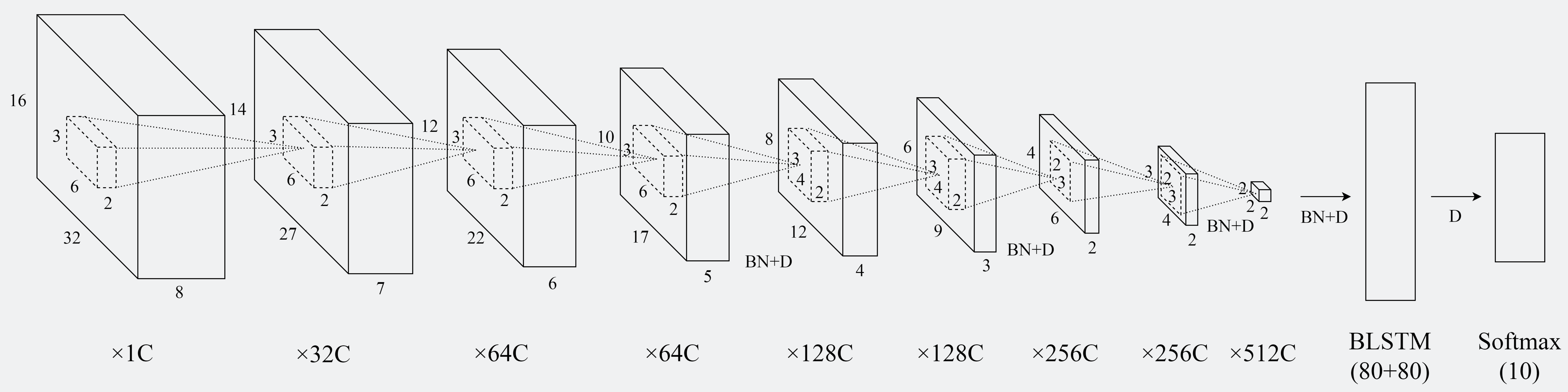

Figure 1: Network architecture of the maxout-CNN-BLSTM model. C: Channel; BN: Batch Normalization; D: Dropout.

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)

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