Practical Improvements to Automatic Visual Speech Recognition

by

Fung, Ho Long

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Computer Science and Engineering

14 November 2018, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Fung, Ho Long

14 November 2018

Practical Improvements to Automatic Visual Speech Recognition

by

Fung, Ho Long

This is to certify that I have examined the above Mphil thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Brian Kan-Wing Mak, Thesis Supervisor
Department of Computer Science and Engineering

Prof. Dit-Yan Yeung, Acting Head
Department of Computer Science and Engineering

Department of Computer Science and Engineering

14 November 2018

To

my family,

pioneers with great minds,

and those with eager to contribute to the humanity

Acknowledgments

I want to express my greatest gratitude to my family for their enormous support and encouragement to me to continue pursuing this master degree in computer science. Without their dedicated support, I would not have been able to finish the challenging yet meaningful and rewarding journey at this moment with fruitful outcomes.

I would also like to thank my supervisor Brian Mak for his kind support and advice throughout my MPhil postgraduate study. Not only did he give me important recommendations on research directions, but he also gave me crucial and useful guidance on my life and career as a computer scientist and researcher.

The work described in this thesis was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. HKUST16215816 and 16200118). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Table of Contents

List of Figures

List of Tables

Practical Improvements to Automatic Visual Speech Recognition

by Fung, Ho Long

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

Abstract

Visual speech recognition (a.k.a lip-reading) is the task of recognizing speech solely from the visual movement of the mouth. In this work, we propose multiple feasible and practical strategies, and demonstrate significant improvements to the established competitive baselines in both low-resource and resource-rich scenarios.

On one hand, one main challenge in practical automatic lip-reading is to deal with the diverse facial viewpoints in the available video data. With the recent proposal of the spatial transformer, the spatial invariance to input data in the convolutional neural network has been enhanced and it has demonstrated different levels of success in a broad spectrum of areas including face recognition, facial alignment and gesture recognition with promising results by virtue of the increased model robustness to viewpoint variations in the data. We study the effectiveness of the learned spatial transformation to our model through quantitative and qualitative analysis with visualizations and attain an absolute accuracy gain of 0.92% to our data-augmented baseline on the resource-rich Lip Reading in the Wild (LRW) continuous word recognition task with incorporation of spatial transformer.

On the other, we explore the effectiveness of convolutional neural network (CNN) and long short-term memory (LSTM) recurrent neural network in lip-reading under a low-resource scenario that has not yet been explored before. We propose an end-to-end deep learning model fusing conventional CNN and bidirectional LSTM (BLSTM) together with maxout activation units (maxout-CNN-BLSTM) and dropout, which is capable of attaining a word accuracy of 87.6% on the low-resource Ouluvs2 corpus, offering

an absolute improvement of 3.1% to the previous state-of-the-art auto-encoder-BLSTM model at that time. To emphasize, our lip-reading system does not require any separate feature extraction stage nor pre-training phase with external data resources.

Chapter 1

Introduction

In this chapter, we will cover a brief introduction of visual speech recognition, followed by the practical challenges and difficulties we encounter, and ended with a summary and outline for the remaining portions of this thesis.

1.1 Visual Speech Recognition

Visual speech recognition (a.k.a. lip-reading) is the technology of interpreting speech through mouth movement without any audio aid. Whilst this is crucial for the hearing impaired to understand speech, it is also natural for the others to employ this technique to help determine speech in situations where audio alone is ambiguous and inadequate, especially under noise-corrupted or far-field scenarios.

With the recent advancement in computational power, now it becomes feasible to accomplish this task through a wide spectrum of machine learning approaches, from latent variable models to hidden Markov models and artificial neural networks. Many of them have shown decent performance over various datasets, including the low-resource Ouluvs2 [2] and resource-rich Lip Reading in the Wild (LRW) [3] corpora.

1.2 Practical Challenges and Difficulties

Here we describe the challenges and difficulties we face in the low-resource and resource-rich circumstances respectively.

1.2.1 Low-Resource: Model Overfitting

The problem of overfitting is a well-known and common issue in the field of machine learning, particularly for the low-resource corpora, in which the situation is even worse owing to the lack of enough training resources for making a model with a considerable size and capacity to learn and work well to the expectation.

In this thesis, we are going to present a pure end-to-end deep neural network that makes use of convolutional neural network (CNN) and long short-term memory (LSTM) recurrent neural network with maxout activation units and dropout for the low-resource Ouluvs2 lip-reading task. While the maxout activation unit is free of the high zero saturation rate problem that occurs in other activation units like ordinary ReLU, its combination in tandem with dropout has a more accurate approximate model averaging in comparison with others. Unlike the previous work [4], we are able to obtain superior results on the task with this architecture without employing any additional training resources.

1.2.2 Resource-Rich: Diversity of Viewpoints

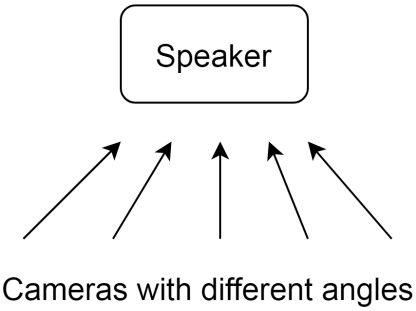

Figure 1.1: Non-frontal face relative to the cameras with angle differences.

Similar to other areas where localization of objects of interest in the three-dimensional space is required with the provision of plain image or video data only, one of the difficulties in visual speech recognition lies in overcoming the diverse viewpoint differences (i.e. non-frontal faces relative to camera as depicted in Figure 1.1) within the training samples, which ordinary convolutional neural network (CNN) does not perform well because of its limited spatial invariant robustness to the input samples. While one may perform intensive data preprocessing prior to training to normalize the lip region in the faces to some particular universal settings, not only does it remain unclear on how to decide such a universal reference that is beneficial to the later training stage, but it also poses an extra overhead to any online lip-reading production system targeted at real-time processing.

In contrast, spatial transformer network [5] serves as a soft and flexible approach by allowing the network to learn the trainable and differentiable parameters of transformation

that suit best to the data itself on its own through end-to-end back propagation. To the best of our knowledge, this is the first work that utilizes spatial transformer in the field of visual speech recognition to resolve the issue of viewpoint diversity by strengthening the neural network model with additional spatial invariance.

1.3 Thesis Summary and Outline

In this thesis, we propose various practical strategies for improving the current state-of-the-art neural network models in the area of visual speech recognition (lip-reading). The major contributions in the work can be briefly summarized as below:

- Showcase the capability and feasibility of designing an end-to-end deep neural network for the low-resource lip-reading task using CNN and BLSTM with incorporation of maxout activation units.
- Achieve a state-of-the-art accuracy of 87.6% on the low-resource Ouluvs2 10-phrase corpus without using any external data resources.
- Demonstrate the practicality of spatial transformer in the field of visual speech recognition under the neural network model with 3D-ResNet and BLSTM.
- Attain an absolute accuracy gain of 0.92% to our established baseline on the resource-rich LRW 500-word corpus.
- Conduct both qualitative and quantitative analyses on the effectiveness of spatial transformer to the capacity of the resulting model.
- Inspire the lip-reading community in future's work on the more general and challenging sequence-to-sequence continuous visual speech recognition task.

The remaining parts of the thesis are organized as follows

In Chapter 2, we give a brief review of the latest literature on visual speech recognition, maxout network, and spatial transformer respectively in disparate subsections.

In Chapter 3, we describe the settings of two corpora we have used in this thesis, the low-resource corpus Ouluvs2 and the resource-rich corpus LRW included.

In Chapter 4, we give highlighted details of the auto-encoder-BLSTM model from an existing research that works on the low-resource Ouluvs2 corpus.

In Chapter 5 and Chapter 6, we give highlighted details of the 3D-Conv-ResNet-BLSTM and the attention-based BLSTM models from two existing researches that work on the resource-rich LRW corpus.

In Chapter 7 and Chapter 8, we depict our data preprocessing steps and network architecture with various figures, and describe our detailed experiment setups with the final evaluation results on the low-resource Ouluvs2 and resource-rich LRW corpora respectively.

In Chapter 9, we carry out analysis on our model in comparison with a previous state-of-the-art model with autoencoder, and list the difficulties we encountered when training the low-resource Ouluvs2 corpus.

In Chapter 10, we conduct both qualitative and quantitative analyses on the effectiveness of the incorporation of spatial transformer to our baseline model, with illustration of the original and spatial-transformed image samples provided for contrast, and tabulation of word accuracy gain by extensive number of classes.

In Chapter 11, we propose two directions for our future works, including the addition of multi-head attention and self-attention to our model intended for visual speech recognition, with some other suggested possible directions.

In Chapter 12, we conclude by summarizing the major contributions and results we attain in this work, and hope the techniques we employed and described in this work can inspire the lip-reading community in resolving and overcoming similar obstacles that can happen in the more general sequence-to-sequence visual speech recognition task.

Chapter 2

Literature Review

In this chapter, we will review the evolution and trends in visual speech recognition, and describe the ideas of maxout network and spatial transformer respectively in separate sections below.

2.1 Visual Speech Recognition

Conventional methods on lip-reading include latent-variable models [6] and hidden Markov models (HMM) [7, 8]. The latter approaches require a prior separate stage of feature extraction that involves handcrafted features. In [7], an active contour model called "snake" was used to extract handcrafted features by curve deforming and moving towards the lips. In [8], an appearance-based model was used to extract handcrafted features with parameters on the shape and intensity of the lips in the human faces.

Along with the evolution of empirical methods in machine learning, the mainstream of the latest research interest in numerous fields has switched to artificial neural networks, and visual speech recognition is no exception. With the growing popularity of artificial neural network models, deep belief networks (DBN) such as auto-encoder and restricted Boltzmann machine (RBM) have been used as feature extractors, in conjunction with support vector machine (SVM) as the label classifier [9, 10]. Recently, there is a neural network connecting encoding layers from a separate RBM pre-trained auto-encoder to LSTM layers, followed by an end-to-end training stage [4]. This model is able to reach an accuracy of 84.5% in the low-resource lip-reading Ouluvs2 corpus, which is the state-of-the-art result without recourse to external training data at that time.

On the other hand, numerous models pre-trained with extra data resources succeed in getting better accuracies in low-resource lip-reading tasks over the aforementioned models. For example, a frame concatenated model [11] that uses a number of deep pre-trained CNNs such as GoogLeNet [12] is able to achieve a better accuracy of 85.6% on Ouluvs2. Another work [3] exploiting multi-tower 3D convolutional neural network (3D CNN) that resembles the very deep convolutional network (VGG) [13] and multiple layer perceptron

(MLP) further improves the accuracy to 93.2%, in which a separate pre-training stage using the very large LRW corpus is required. However, as far as we know, experiments of end-to-end networks combining CNN with LSTM without external data on low-resource corpora such as Ouluvs2 have not yet been reported. Moreover, leveraging the power of maxout activation units [14] in both CNN and LSTM in one single deep neural network has not been attempted before.

With the recent development and release of the large Lip Reading in the Wild (LRW) dataset [3], which is a collection of videos extracted from programs owned by BBC, lip-reading research can be performed at a much larger scale. For example, a very deep convolutional neural network exploited by VGG [15] obtained a word classification accuracy of 76.2% and a residual convolutional neural network (ResNet) [16] together with a long short-term memory (LSTM) backend [17] further improved the accuracy to 83.0%. A more recent work [18] with a primary focus on deep word embeddings even achieved 88.1% and 84.3% word accuracies with and without leveraging the extra word boundary information under the ResNet-LSTM architecture.

Through leveraging visual speech recognition in tandem with audio speech recognition, a system of audio-visual speech recognition can be built to improve the performance of audio speech recognition alone. A recent work [19] has examined the influence of the respective audio and video streams to the network predictions through class saliency maps, and shown that the saliency of video stream increases as the noise presented in the audio stream increases in an end-to-end multi-modal audio-visual speech recognition system. Other notable works on this application include [20] that fuses ResNet with bidirectional gated recurrent units (BGRU) to create a multi-modal system for LRW dataset, [21] that experiments on How-To audio visual corpus with sequence-to-sequence and connectionist temporal classification (CTC) framework, and a latest work [22] that utilizes transformer decoder in its sequence-to-sequence and CTC variants for the challenging sequence-to-sequence Lip Reading Sentences of BBC programs (LRS2-BBC) [15] and TED talks (LRS3-TED) [23] corpora.

Other new and exciting applications making use of the LRW corpus include but not limit to speech separation and talking face synthesis, in which the first work [24] proposes an audio-visual enhancement network to enhance the audio signal towards the interested speaker's speech with the additional information from video, and the second work [25]

synthesizes the talking face with an audio encoder, a pre-trained VGG encoder and a separate deblurring CNN component.

2.2 Maxout Network

Maxout unit is a simple yet elegant activation function that is believed to work better in combination with dropout owing to its more accurate approximate model averaging capability [14]. It is proposed as a plausible choice for replacing ReLU, which is criticized for its high saturation rate at zero, and as an alternative to ReLU's other improved versions such as leaky [26] and randomized ReLUs. For a neural network with the previous hidden layer of size d, the current hidden layer of size m, and a number of k feature maps, the output of the i-th node of the current hidden layer, denoted as $h_i(\mathbf{x})$, can be characterized by the following simple formula:

$$h_i(\mathbf{x}) = \max_{j \in [1,k]} \{\mathbf{x}^T \mathbf{W}_{\cdot ij} + b_{ij}\} \ , \tag{2.1}$$

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

Whilst maxout units for CNN (maxout-CNN) would be equivalent to max-pooling across channels with stride equal to k, LSTM can incorporate maxout unit by replacing the hyperbolic tangent activation in the memory gate, resulting in the following peephole-variant maxout-LSTM equations:

$$\begin{aligned}
i_t &= \sigma(\mathbf{W}_i \cdot [C_{t-1}, \mathbf{h}_{t-1}, \mathbf{x}_t] + b_i) && (2.2)\\
f_t &= \sigma(\mathbf{W}_f \cdot [C_{t-1}, \mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) && (2.3)\\
o_t &= \sigma(\mathbf{W}_o \cdot [C_{t-1}, \mathbf{h}_{t-1}, \mathbf{x}_t] + b_o) && (2.4)\\
\tilde{C}_t &= \max_{j \in [1,k]} \{\mathbf{W}_{Cj} \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_{Cj}\} && (2.5)\\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t && (2.6)\\
h_t &= o_t \cdot \tanh(C_t) && (2.7)
\end{aligned}$$

We notice that another work has also utilized maxout activation units in the LSTM alone [27] with good performance.

2.3 Spatial Transformer

Recently, spatial transformer [5] is proposed to improve the spatial invariance to input data in the neural network, which is a differentiable network that is composed of three major components. The first component is a localization network that estimates the transformation parameters based on the input characteristics derived from a convolutional neural network. In the second component, the learned parameters will be used to map every pixel in its homogeneous form to a new space called sampling grid through a transformation. Finally, a sampler with a sampling kernel serving as the interpolation smoother is required to obtain the final output pixels due to their discrete nature. For detailed description of each component, please refer to the paper [5].

Albeit it was originally proposed for digit recognition and bird classification, it has now been employed in a wide range of recognition areas including face recognition [28], facial alignment [29] and gesture recognition [30] with impressive gains. On top of the above direct applications, a broad emergence of usages in prediction and generation areas has also been observed in image compositing [31], image drawing [32] and video prediction [33].

However, as far as we know, whether the same applies to the field of visual speech recognition remains unknown and this is the first work to examine its strength in handling the viewpoint discrepancies present in the LRW corpus.

$$\begin{pmatrix} x_i^{(t)} \\ y_i^{(t)} \\ 1 \end{pmatrix} = \mathbf{A}\mathbf{x}_i^{(s)} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_i^{(s)} \\ y_i^{(s)} \\ 1 \end{pmatrix} \tag{2.8}$$

In this work, we focus solely on the affine transformation from the source ($\mathbf{x}_i^{(s)}$) to target ($\mathbf{x}_i^{(t)}$) coordinates owing to its demonstrated sufficiency in bird classification [5]. The affine transformation that supports translation, rotation, scale, skew and cropping can be simply but fully characterized by 6 trainable parameters θ_n, where $n = 1, 2, \ldots, 6$, as shown above in Equation 2.8, albeit one can always experiment on more general transformations like projective transformation that is heavily used in areas such as computer vision and computer graphics.

Chapter 3

Corpora

In this chapter, we present the low-resource and resource-rich corpora we used in our work for training and evaluation with detailed description.

3.1 Low-Resource: Ouluvs2 Dataset

We chose Ouluvs2 because it is a low-resource corpus consisting of a reasonable number of disparate subjects for training. The corpus is composed of 3 distinct parts, 5 views and 52 subjects, of which 39 are male and 13 are female. The first part of the corpus is a set of 10 different strings of digits from 0 to 9 in random order, while the second part is 10 different daily-use short phrases, and the third part is 10 random sentences adopted from the TIMIT corpus [34]. Similar to the previous work [4], we used only the frontal view of the second part of the corpus in our evaluation section, which comprises the following 10 phrases: 'Excuse me', 'Goodbye', 'Hello', 'How are you', 'Nice to meet you', 'See you', 'I am sorry', 'Thank you', 'Have a good time', and 'You are welcome'.

Amongst the 10 distinct phrases, every subject repeats each of them 3 times; therefore, a total number of 156 samples are provided for each phrase, as compared with 800-1100 samples for each word in the high-resource LRW corpus. Following the traditional data splitting scheme as suggested by the author [11], we reserved subjects 06, 08, 09, 15, 26, 30, 34, 43, 44, 49, 51 and 52 for testing, in which 10 of them are male and 2 of them are female, and the remaining for training.

3.2 Resource-Rich: LRW Dataset

Table 3.1: Statistics of the two corpora used in this work.

	Classes	Train/Validation/Test samples per class	Number of speakers
Ouluvs2	10	108/12/36	52
LRW	500	800-1000/50/50	n/a

The Lip Reading in the Wild (LRW) [3] corpus has a vocabulary of 500 words, and there are a maximum of 1000 video samples for every word. Each of the samples consists of 29 frames as extracted from the utterances spoken by a wide variety of speakers in the BBC programs.

This dataset is more challenging than some previous lip-reading corpora because the words are extracted from continuous speech in which the presence of co-articulations between each word with its preceding and following words result in great variations at the word boundaries in each video clip. In this work, we have not exploited the word boundary information, which otherwise, may help in distinguishing visually similar classes [18], for example, "other" and "others". The diversity in viewing angles as introduced by different head poses of individual speakers makes it another difficulty for the model to adjust to scenarios where faces in the samples are not, or even far from being frontal, which is a problem that we need to face when developing a reliable and robust application and system in practice.

Moreover, the video data in the training, validation and test sets in the corpus were recorded over non-overlapping periods of time. This is intended to maximize the dissimilarity in their program contexts to make the evaluation more independent to the particular contexts appeared in the training data.

Chapter 4

Highlight (I): Auto-Encoder-BLSTM Model with Low Resource

In this chapter, we highlight and provide the details of the auto-encoder-BLSTM model from an existing research [4] that works on the low-resource Ouluvs2 corpus.

4.1 Data Preprocessing

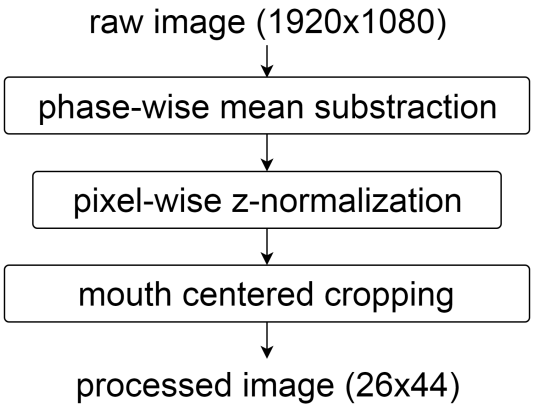

Figure 4.1: Preprocessing flow-chart.

The preprocessing phase involves three disparate steps, in which the first aims at reducing and minimizing the influence and dependence from the subject characteristics among the speakers in the corpus. This can be done by a phase-wise mean image subtraction in every individual image sequence provided, as calculated by dividing the sum of every pixel along the temporal direction by the total number of frames.

The second step involves pixel-wise normalization, which can be done by z-normalization across every pixel by global mean subtraction and standard deviation division. As mentioned in the paper [4], this is the recommended setting for training the Restricted Boltzmann Machines (RBMs) in the later stage.

The last step downsamples the cropped mouth in every image in the sequences to a universal size of 26×44 to preserve the aspect ratio to a constant before feeding to the neural network for training.

4.2 Network Architecture

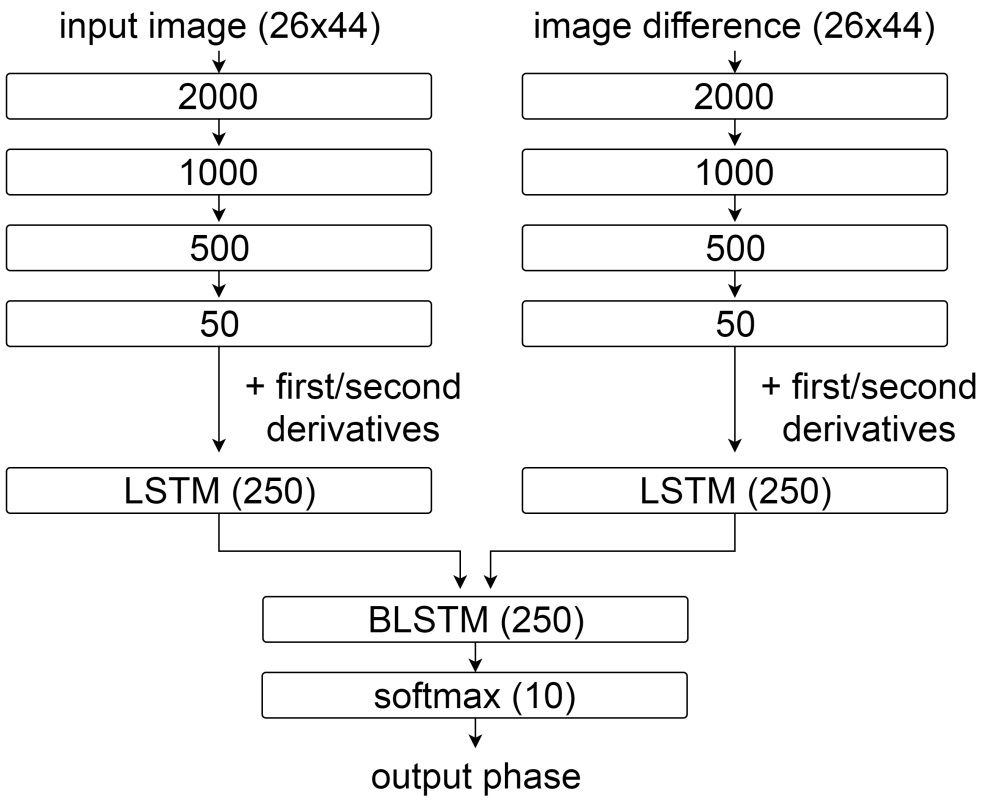

Figure 4.2: Network architecture for the auto-encoder-BLSTM model.

The whole architecture of the network can be depicted in Figure 4.2, which is composed of two separate and independent streams in the front-end component. The first stream serves as the feature extractor for the raw cropped mouth images, while the second serves as the feature extractor of the temporal information in the differences of every pair of successive and consecutive image frames.

Each of the stream is comprised of 4 encoding layers with sizes of 2000, 1000, 500 and 50 respectively, where each is followed by the sigmoid activation except for the last one. The weights of these layers are obtained by pre-training with the Restricted Boltzmann

Machines (RBMs). Afterwards, the bottleneck features are concatenated together with its computed first and second derivatives and fed to a LSTM layer with a hidden size of 250 separately for each stream.

Finally, the outputs from the LSTM layers obtained from both streams are then concatenated and served as the input to the BLSTM layer with a hidden size of 250 to capture the temporal information for both streams in the forward and backward directions and the softmax classification layer of 10 classes is then appended to the output from the last frame of each image sequence.

4.3 Hyper-Parameters

In this section, we will cover the hyper-parameters used in different training stages, namely the RBM pre-training and the final end-to-end training.

4.3.1 RBM Training

In each of the independent streams, the first encoding layer is pre-trained by a Gaussian-Bernoulli RBM, the second and third layers are pre-trained by two respective Bernoulli-Bernoulli RBMs, and the final bottleneck layer is pre-trained by the Bernoulli-Gaussian RBM. Every of the RBMs is trained with a minibatch size of 100 for 20 epochs. The L2-regularization is used with a coefficient of 0.0002, and the learning rate is 0.1 for the RBMs without real values and 0.001 otherwise. As the detailed pre-training steps are not the primary focus of this thesis, please refer to the paper [4] for the complete and thorough descriptions.

4.3.2 End-to-End Training

The AdaDelta is chosen as the gradient optimization scheme and a minibatch size of 20 samples is used. Gradient clipping is employed in the LSTM layers alone, and early stopping with a 5-epoch delay is used to alleviate the well-known problem of overfitting.

Table 4.1: Classification accuracy of various setups on the Ouluvs2 corpus in the end-to-end auto-encoder-BLSTM model.

Setup	Accuracy (%)
Input Image Stream Only	78.0
Image Difference Stream Only	75.8
Both	84.5

4.4 Evaluation Results

As we can see in Table 4.1, the input image stream alone is sufficient to obtain a reasonable performance of 78.0% in terms of phase classification accuracy. While the image difference stream alone performs worse than the input image stream with only an accuracy of 75.8%, the combination of both streams can complement the learnt information from the features to each other and give the best accuracy of 84.5%. This proves the benefit of utilizing both streams in the Ouluvs2 phase classification task.

Chapter 5

Highlight (II): 3D-Conv-ResNet-BLSTM Model with Rich Resource

In this chapter, we highlight and provide the details of the 3D-conv-ResNet-BLSTM model from an existing research [17] that works on the resource-rich LRW corpus.

5.1 Data Preprocessing

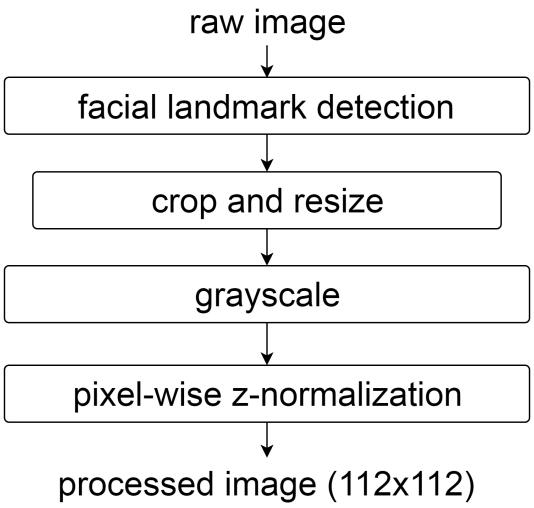

Figure 5.1: Preprocessing flow-chart.

The preprocessing stage involves two different steps, in which the first locates the mouth portion within every image by facial landmark detection. With such landmark information, the images within the same video clip are cropped and resized to a universal size of 112 $\times$ 112, where the common cropping region is calculated by the median coordinates of the landmarks in each frame.

Afterwards, each image in the video samples is converted into the grayscale format and the z-normalization is applied to every image through a global mean subtraction and standard deviation division on every pixel.

The whole preprocessing stage can be depicted in Figure 5.1.

5.2 Network architecture

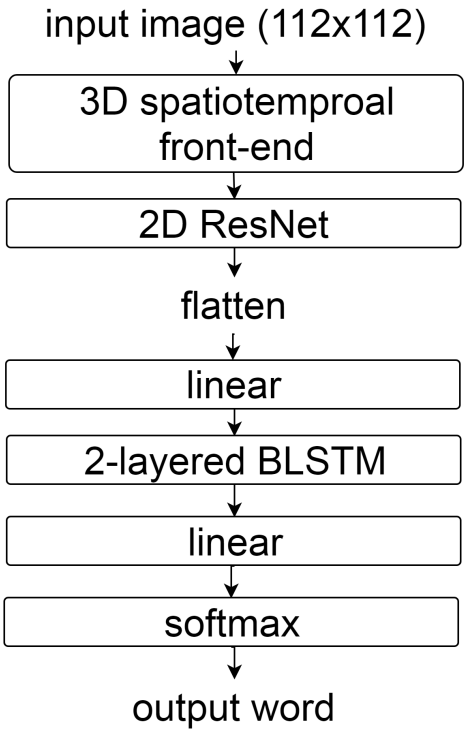

Figure 5.2: Network architecture for the 3D-conv-ResNet-BLSTM model.

The overall architecture of the network can be divided into three components, namely the spatiotemporal convolutional front-end, the residual convolutional network and the bidirectional LSTM back-end, which can be depicted in Figure 5.2.

The spatiotemporal convolutional front-end component begins with a 3-dimensional convolutional layer with 64 filters where each has a size of $5 \times 7 \times 7$ along the temporal depth, width and height respectively. This convolutional layer is followed by a batch normalization layer and the rectified linear unit (ReLU) as the activation function. Afterwards, the feature maps are passed to the spatiotemporal max-pooling layer to reduce

its sizes along the 3 dimensions. The parameter size of this component is about 16K according to the paper.

The residual convolutional network (ResNet) component resembles the 34-layered ResNet originally proposed for the use in ImageNet, where each block is comprised of two convolutional layers, batch normalization layers, ReLU activations and one residual connection from the input at the beginning to the output at the end of each block. As every input image goes through this component, the spatial size keeps decreasing and becomes 1×1 eventually, which is then flattened and fed to the next component. The parameter size of this component is about 21M according to the paper.

The final bidirectional LSTM back-end component consists of two different streams, where each of them is a stack of two LSTM layers. The first stream processes the original sequence without manipulating its temporal order. In contrast, the second stream reverses the temporal order of the original sequence. Through concatenation of the outputs gathered from both streams, the final output can then be fed to the softmax classification layer of 500 classes. With accordance to the preliminary experiments mentioned in this work, the approach of appending softmax classification layer to the output of the LSTM in every time frame is preferred. In comparison with the approach of appending the softmax classification layer only to the last frame of every image sequence, it can alleviate the issue of gradient vanishing when the error is backpropagated through time in the LSTM layers.

5.3 Training Details and Hyper-Parameters

In this section, we will cover the training details and hyper-parameters of the work.

5.3.1 Training Details

The whole training scheme can be divided into three disparate phases, in which the first replaces the bidirectional LSTM component with the temporal convolutional one and trains the resulting network in an end-to-end manner.

After the convergence of the first phase, the bidirectional LSTM component is replaced back and the temporal convolutional one would be removed. The network is then trained for 5 more epochs with the network parameters from the spatiotemporal convolutional

front-end and ResNet components remain fixed.

In the final phase, the whole network will be trained in a fully end-to-end manner with no limitations on the parameter updates until complete convergence.

The combination of all the abovementioned phases takes no more than 20 epochs until full convergence.

5.3.2 Hyper-Parameters

The gradient optimization scheme used in this work is the standard SGD with an initial learning rate of 0.005 for the first phase, 0.0005 for the second phase and 0.00005 for the final phase respectively. A momentum of 0.9 is used in all training phases and dropout is not applied throughout the network.

5.4 Evaluation Results

Table 5.1: Classification accuracy of various phases on the LRW corpus in the end-to-end 3D-conv-ResNet-BLSTM model.

	Accuracy (%)
Phase 1	74.6
Phase 2	79.6
Phase 3	83.0

As we can see in Table 5.1, the temporal convolutional back-end can still give a reasonable performance of 74.6% accuracy without the use of any BLSTM layers. With the incorporation of the BLSTM layers, the classification accuracy increases to 79.6% with 5.0% absolute gain, even the network is not trained in a fully end-to-end manner. Finally, when the network is trained fully end-to-end that all parameters are allowed to be updated, the model can reach an accuracy of 83.0%.

Chapter 6

Highlight (III): Attention-Based BLSTM Model with Rich Resource

In this chapter, we highlight and provide the details of the attention-based BLSTM model from an existing research [15] that works on the resource-rich LRW corpus.

6.1 Data Preprocessing

As the primary focus on this work is on the sequence-to-sequence visual speech recognition task of Lip Reading Sentences in the Wild (LRS) that is out of the scope of this thesis, they do not give data preprocessing details on their experiments on the LRW corpus. However, the expected size of the every input image is 120×120 by cropping around the mouth region, as mentioned in the network architecture.

6.2 Network Architecture

This attentional-based model utilizes the general encoder-decoder framework as depicted in Figure 6.1, except that the encoder is enhanced with an extra VGG network of 6 layers similar to their previous work [3].

The whole network architecture can be depicted in Figure 6.2 which is composed of three components, namely the image encoder, the audio encoder and the character decoder. As the audio encoder is intended for the audio-visual speech recognition task in their work, which is not within the scope of this thesis, we will cover only the other two components below.

The image encoder component is comprised of a 6-layered network similar to [3] and a LSTM module. The former part contains 5 convolutional layers and 1 fully connected layer in total to reduce the initial image spatial size from 120×120 to 1×1 and increase the number of kernels from 5 to 512. The initial kernel size of 5 is done by stacking every 5 consecutive grayscale images, obtained from a sliding and overlapping window of 1 frame

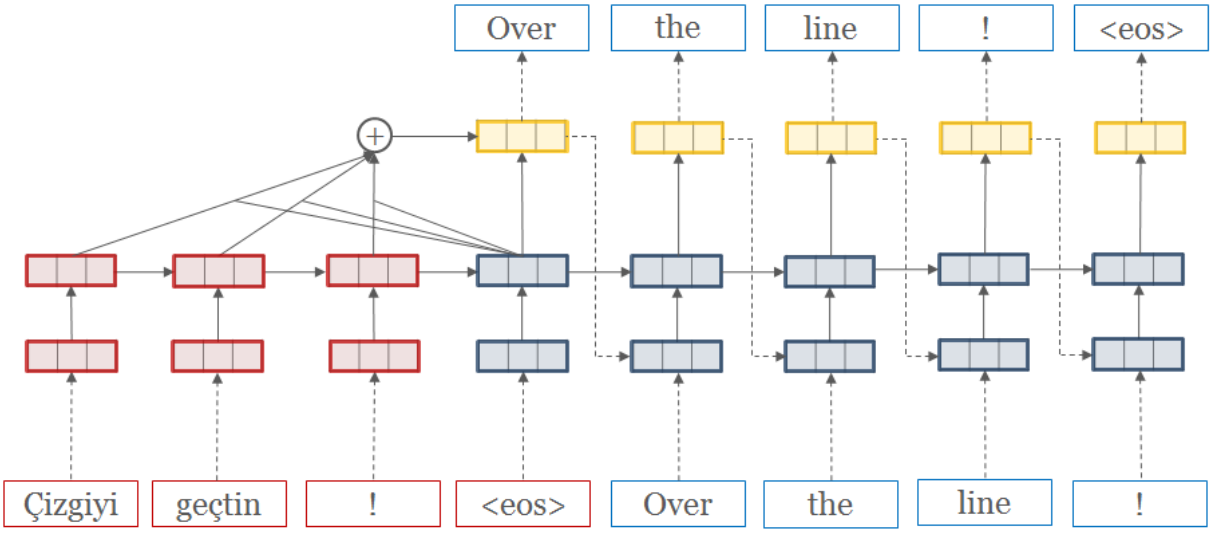

Figure 6.1: Network architecture for the general attention-based encoder-decoder framework adopted from the OpenNMT toolkit [1].

in each step. The filter size in each convolutional layer is 3×3 and a max-pooling layer is appended only after the first, second and the last convolutional layers. The flattened vector with size of 512 is then fed to the 3-layered LSTM with hidden size of 256 in each layer.

The character decoder component is comprised of an attention mechanism and a LSTM module. In each decoder step, the context vector obtained from the attention mechanism together with the output from LSTM are concatenated and fed to the softmax classification layer. In addition, the context vector and one-hot input character vector from this step are concatenated and served as input to the LSTM in the next decoder step. The LSTM module in the decoder has a hidden size of 512 in the three stacked layers.

6.3 Training Strategies and Hyper-Parameters

In this section, we will cover the training strategies and hyper-parameters of the work.

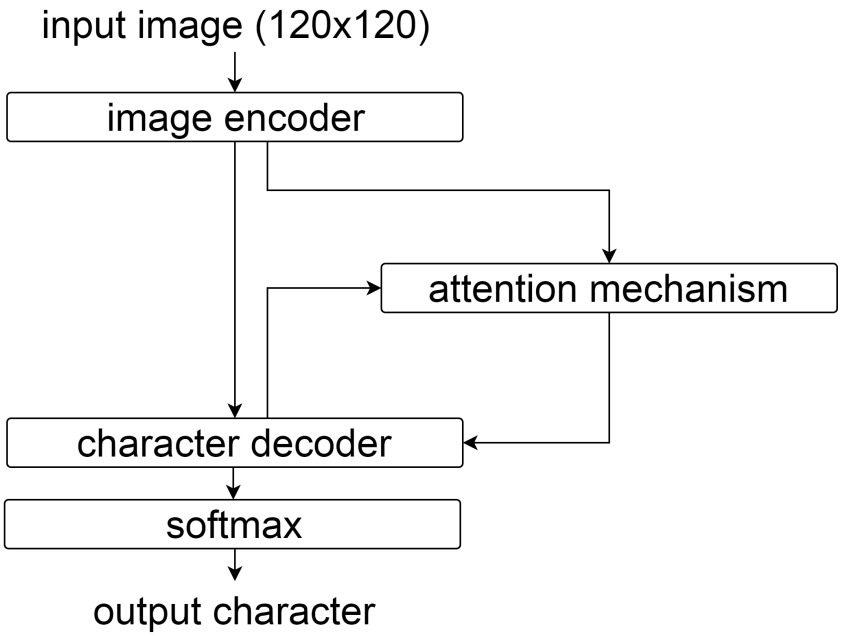

Figure 6.2: Network architecture for the attention-based BLSTM model.

6.3.1 Training Strategies

There are two main strategies that help the training of the challenging sequence-to-sequence visual speech recognition task, namely curriculum learning and scheduled sampling.

The idea of curriculum learning is to increase the input sequence length to the model gradually. The training is first started with samples of single words only and continued with short sentences extracted from the complete ones in the dataset, which can help increase the convergence speed of the network by several times.

Another training strategy is scheduled sampling which feeds the ground-truth character to the decoder only by a certain probability. During the validation and test stages, the ground-truth character is not available and only the inferenced character from the previous decoder step can be fed to the next one. To mimic the this condition, the ground-truth character is fed to the decoder step only with a certain probability and randomly sampled character is fed otherwise. This techniques is applied only on the full sentence training with a probability no greater than 0.25.

6.3.2 Hyper-Parameters

The standard stochastic gradient descent (SGD) optimization scheme is used in this work with an initial learning rate of 0.1, which is decreased by 10% when the training error does not improve for every 2000 iterations. A minibatch size of 64 is used with dropout and label smoothing. The training is stopped when the full sentence training does not improve the validation error for 5000 iterations and the whole model involves training for around 500K iterations on the LRS corpus, after which the whole network is fine-tuned for 1 more epoch on the LRW dataset.

6.4 Evaluation Results

The primary interest of this work is on the sequence-to-sequence lip-reading task which is out of the scope of this thesis. Here, we only give the word accuracy result on the LRW corpus which is 76.2%. Notice that this work involves the use a huge amount of external data resources in the LRS corpus to pre-train the model before fine-tuning on the LRW corpus.

Chapter 7

Experiment (I): Low-Resource Ouluvs2 Corpus

We present our experiments in three parts, namely data preprocessing, network architecture and hyper-parameters, and evaluation results.

7.1 Data Preprocessing

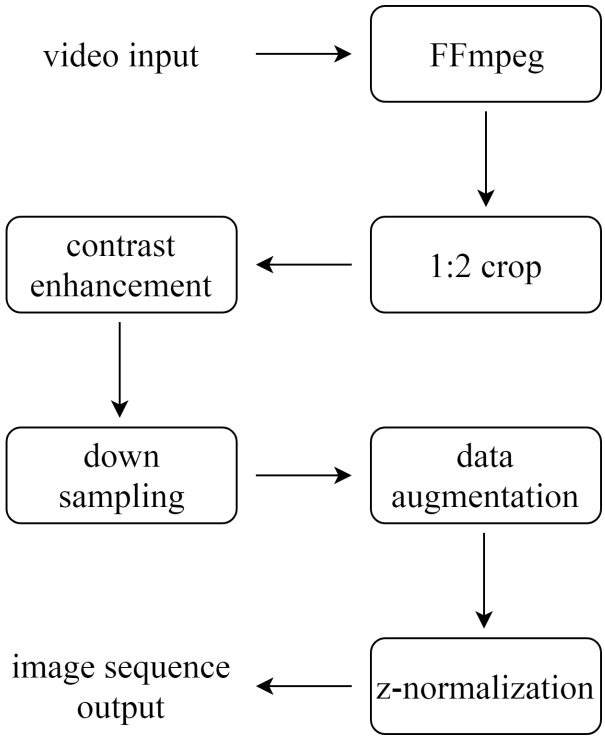

Figure 7.1: Preprocessing flow-chart

Each video clip was converted to a sequence of lossless grayscale images of variable length using the FFmpeg [35] software. Then we performed a 1:2 crop around the mouth region found by dlib [36] and contrast enhancement in each image. Afterwards, each image

was downsampled to 16×32 using bicubic interpolation smoother. Data augmentation was carried out by shifting the cropping area to 8 different directions (top-left, top, top-right, right, bottom-right, bottom, bottom-left, and left) by 10 pixels, followed by a further augmentation through horizontal flipping, to create a total of 16 different copies from the original image. Finally, each image was z-normalized across each pixel through a global mean subtraction and variance normalization. The whole preprocessing procedure is depicted in Figure 7.1.

7.2 Network Architecture and Hyper-Parameters

Our end-to-end deep neural network is comprised of two parts. The first part contains 8 layers of convolutional layers as the visual front-end and the second part contains one layer of bidirectional LSTM (BLSTM) as the sequence learning back-end.

Each of the convolutional layers is a spatial-temporal convolution (3D convolution) with no zero-padding or stride, followed by an activation function, either a maxout or ReLU unit, without any pooling layer.

For the BLSTM layer, either the common bidirectional peephole LSTM using the hyperbolic tangent activation, or its maxout version described in the aforementioned section was used. Finally, outputs of the forward and backward LSTM of the last frame of each input sequence were concatenated together into a vector, which serves as the input to the softmax classification layer of 10 targets.

In order to demonstrate the effectiveness of the maxout activation units in the deep neural network, we carried out the experiment under four different setups, namely ReLU-CNN with tanh-BLSTM, ReLU-CNN with maxout-BLSTM, maxout-CNN with tanh-BLSTM, and maxout-CNN with maxout-BLSTM (maxout-CNN-BLSTM) respectively. Note that the input at each time step is a stack of 8 consecutive images obtained from a sliding window along the image sequence, i.e. a tensor of 16×32×8. Figure 7.2 gives the network architecture of the maxout-CNN-BLSTM as an example.

To alleviate the problem of overfitting, we employed a number of regularization methods including batch normalization, dropout and L2-regularization. Whilst batch normalization layer was inserted between various layers in CNN, a dropout rate of 0.5 was applied to the whole network starting from the 4th epoch, and an L2-regularization with weight

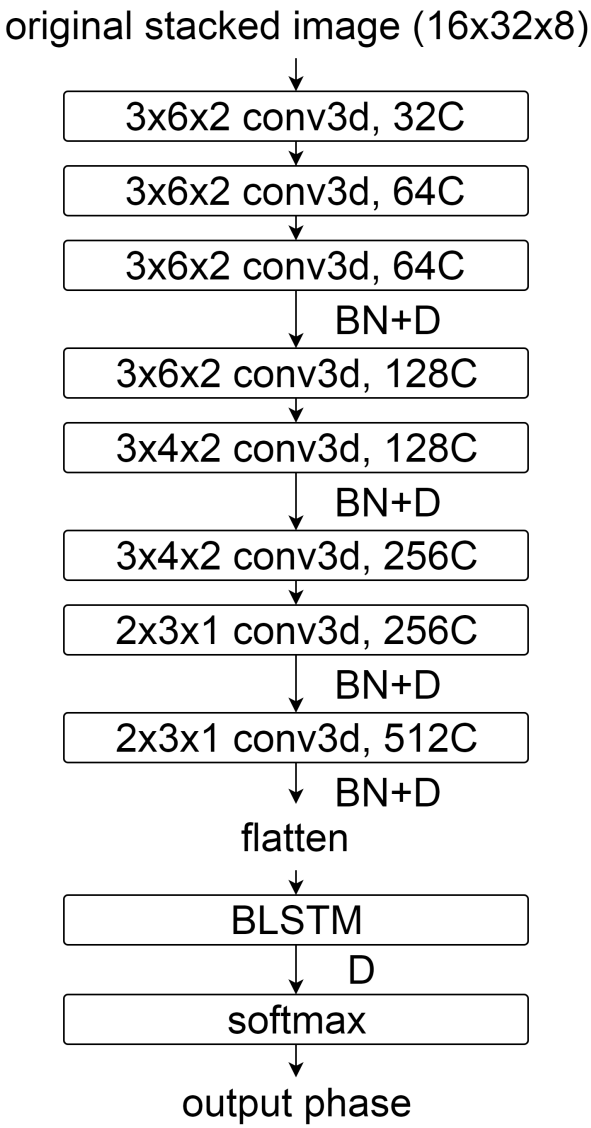

Figure 7.2: Network architecture of the maxout-CNN-BLSTM model. C: Channel; BN: Batch Normalization; D: Dropout.

0.00155 was applied to all trainable parameters to penalize highly positive and negative values. Along with batch normalization, a momentum of 0.6 was used in the first 10 epochs followed by 0.9 in the remaining epochs to speed up convergence in the training stage. Initial learning rate was set to 0.01 and was reduced by roughly half after every 2 epochs. A mini-batch size of 256 images, not image sequences, was used, and a total number of 15 epochs were run in every setup.

7.3 Evaluation Results

We implemented and evaluated our models using CNTK [37], which takes great advantage of the parallel computations in GPUs to improve training speed. To improve performance reliability, each of the above experimental setups was repeated 12 times. During each run, the training set of 40 subjects was randomly partitioned into two non-overlapping groups of 4 and 36 subjects respectively. The small and large partitions were used as the validation and training data respectively. The reported result of each setup is the average of testing accuracies under 12 respective runs, where each was evaluated on the epoch with the lowest validation error.

Table 7.1: Classification accuracy of various models.

Method ($k = 4$ for maxout)	Accuracy (%)
Auto-encoder with tanh-BLSTM [4]	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

It can be seen from Table 7.1 that our proposed maxout-CNN-BLSTM model is the best among the tested models and is able to obtain a state-of-the-art accuracy of 87.6% in the low-resource Ouluvs2 task without resorting to any other external data resources. This also confirms the superior performance of maxout unit over the conventional ReLU and tanh in deep neural network, probably because it is free of the high zero saturation rate problem that occurs in ReLU, and has more accurate approximate model averaging with dropout.

Table 7.2: 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

From Table 7.2, it can be seen that CNN with maxout units increases the training

time to more than 3 times to that with ReLU. This confirms the use of maxout units involves a k-time increase in the network parameter size, which in turn leads to many more computations. On the other hand, the difference in training time between BLSTM with hyperbolic tangent activations and that with maxout units is minor. Nonetheless, maxout units are still beneficial due to the abovementioned accuracy gain.

Table 7.3: 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

To further investigate the maxout activation units, we have conducted experiments on the effect of the number of feature maps, denoted as k, in the maxout-CNN-BLSTM architecture. As shown in Table 7.3 we can observe that the time of computation increases with the number of feature maps, and $k = 4$ offers a slightly better accuracy in comparison with others. It also confirms that even with only two feature maps, it is already sufficient to approximate arbitrarily sophisticated and non-linear functions, having a similar effect to other activation functions such as ReLU and tanh.

Chapter 8

Experiment (II): Resource-Rich LRW Corpus

We implement and validate our idea using the elegant and popular PyTorch [38] toolkit, and present our experiment in three disparate parts, namely data preprocessing, network architecture and results, in the subsections below.

8.1 Data Preprocessing

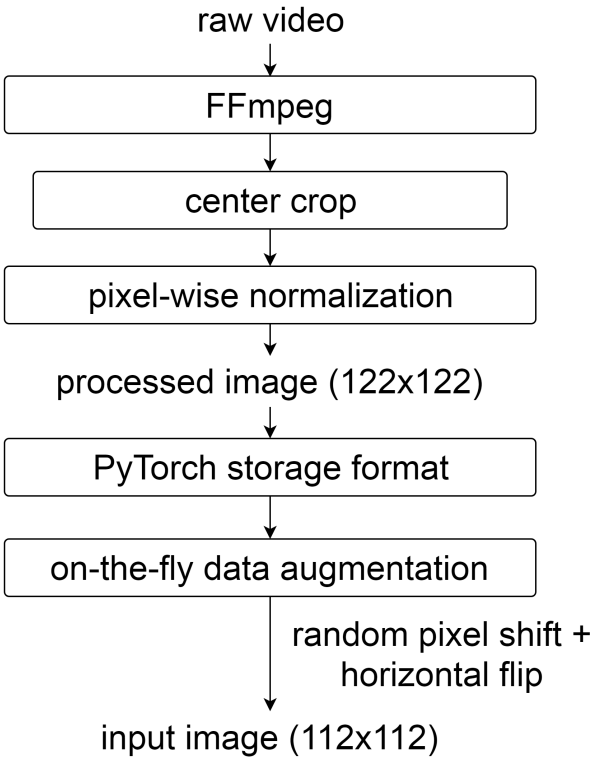

Figure 8.1: Preprocessing flow-chart.

We first utilize FFmpeg [35] to convert each video clip to its corresponding lossless image sequence of 29 frames. Afterwards, each grayscale image is universally center-cropped

to a size of 122×122, and a global pixel normalization is applied to those images with a mean and standard deviation equal to 0.4161 and 0.1688 respectively, as recommended in an earlier work [17]. Eventually, the processed samples are stored in a designated format that facilitates efficient reloading in PyTorch during the training stage. Note that the 122×122 center cropping is intended for the later on-the-fly batch-based data augmentation with a further random cropping of 112×112 that performs translations within a maximum window of 10 pixels along both the width and height dimensions, in tandem with a random horizontal flipping with a probability of 0.5 to alleviate the well-known problem of overfitting when every data chunk is reloaded at each new epoch. The complete preprocessing procedure is depicted in Figure 8.1.

8.2 Network Architecture

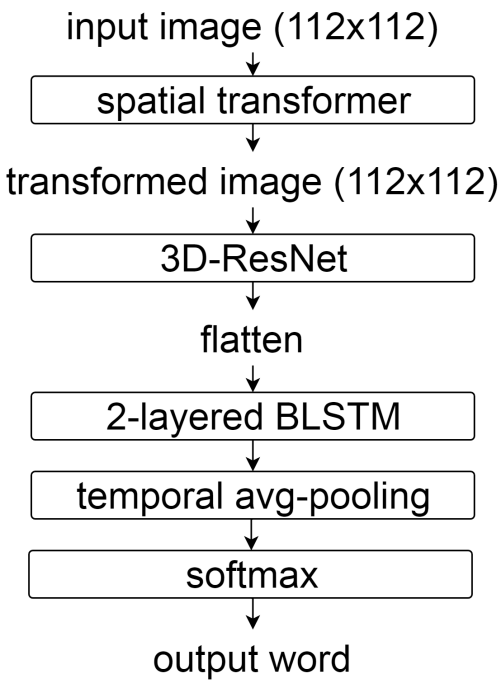

Figure 8.2: Network architecture of the overall ST-3D-ResNet-BLSTM model.

Our network consists of four major components, which constitute a spatial transformer (ST), a 3-dimensional 18-layered residual convolutional neural network (3D-ResNet), a bidirectional long-short term memory (BLSTM) back-end and a temporal average pooling

layer connected to the final softmax classification layer, as illustrated in Figure 8.2.

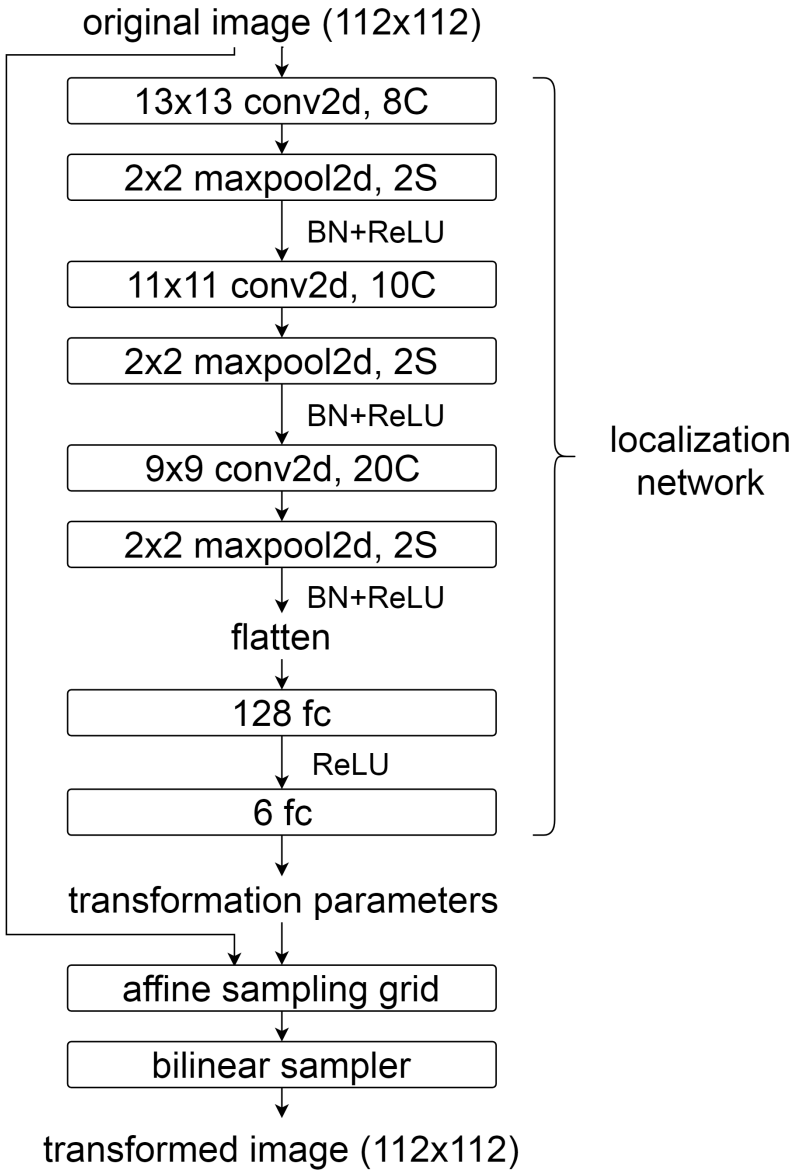

Figure 8.3: Spatial transformer with a localization network, affine sampling grid and bilinear sampler used in this work. (BN: batch normalization; C: channel; S: stride)

The spatial transformer resembles the original one as proposed in [5] except we make use of three convolutional layers in the localization network with kernel sizes decreasing from 13, 11 to 9, and number of feature maps increasing from 8, 10 to 20. Every subsequent max-pooling layer with a stride equal to 2 is followed by a batch normalization layer and a rectified linear activation unit (ReLU). A bilinear smoother with zero padding is used

in the sampler on the affine sampling grid to obtain the transformed images, which serve as inputs to the next component. The detailed structure of this component is depicted in Figure 8.3.

On top of the suggestion from [17] to replace the first 2D convolutional layer by a 3D convolutional layer with a filter size of 5×7×7 in the ordinary 2D-ResNet, we substitute every other 2D convolutional layer with a 3D convolutional layer with a filter size of 3×3×3, and with a stride being applied only along the width and height dimensions. The original 2D batch normalization layers are modified accordingly to their 3D version as well. We also change the kernel size of the last average pooling layer to 1×3×3 to obtain a 2×2 output along the width and height per feature map. Note that we have also removed the last fully connected layer as found in the usual ResNet, with no dropout being applied at this stage.

We then flatten the 512×2×2 output from the ResNet to a 2048-dimensional vector per image and pass it as input to the subsequent BLSTM layers. Our 2-layered BLSTM has a hidden size of 512 on each direction with a dropout rate of 0.2 being used. Finally, the output from the BLSTM layers goes through a temporal average pooling layer that collapses it along the temporal axis prior to the softmax classification layer of 500 classes.

8.3 Evaluation Results

The whole training scheme involves two phases, in which the first takes 15 epochs to train the baseline model from scratch with an initial learning rate of 0.001 that decays with the Adam optimizer until fully convergence. In the latter phase, training continues from the previous trained baseline model for 9 more epochs by resetting the learning rate to its initial, with the addition of data augmentation and spatial transformer to the original network. The baseline word accuracy and the word accuracy gains attained by the improving setups can be found in Table 8.1.

As shown in Table 8.1, the batch-based on-the-fly data augmentation described in the section above gives a 0.31% absolute word accuracy gain from the baseline setup that is better and stronger than [15]. Through the integration of the spatial transformer, we obtain another 0.92% absolute gain as compared with the setup using data augmentation

Table 8.1: Test word accuracy and model parameter size in various setups on the LRW corpus. (DA: data augmentation; ST: spatial transformer)

	Word accuracy (%)	**Parameter size**
Baseline	78.35	50.9M
[+DA]	78.66 (+ 0.31)	50.9M
[+DA +ST]	**79.58 (+ 0.92)**	51.0M

alone, which gives an accumulative absolute word accuracy gain of 1.23% from the baseline. Therefore, with only a mere increase in model parameter size (of 0.1M), the spatial transformer module is capable of demonstrating a convincing word accuracy gain that is additive but not complementary to the concurrent data augmentation applied.

Chapter 9

Analysis (I): Low-Resource Ouluvs2 Corpus

In this section, we will make some key comparisons to the auto-encoder-BLSTM model, and explain the difficulties in coming up with our final maxout-CNN-BLSTM architecture that can successfully outperform that pervious work.

9.1 Comparisons to Auto-Encoder-BLSTM

We propose using a convolutional neural network (CNN) as a replacement of the auto-encoder employed in the previous work [4] chiefly because of its capability of capturing spatial correlations present in the image sequences. We have a strong belief that a CNN, of which each convolutional layer is designed and intended to work as a filter to capture local correlations along the spatial dimensions, will not work worse than encoding layers in an auto-encoder in its abilities in extracting discriminative features for the final classification later, and this has been proved and confirmed with the aforementioned evaluation results.

Using a CNN as the front-end component also allows us to extend the 2D convolution (using 2D filters) to 3D convolution (using 3D filters) by taking into account the additional temporal dimension present in the video data, so as to capture the temporal correlations among successive images in the sequence on top of the spatial correlations in each individual image. There have been a wide range of works showing and demonstrating that 3D convolution can provide a substantial gain across various applications and tasks in different fields. The authors in [4] had trained two individual autoencoder respectively on the individual images and differences between successive images, which should also model the temporal correlations among successive images to some extent that contributes to a significant performance improvement in their work as well. Nevertheless, we found the use of 3D convolution in our work is more effective as confirmed in the superior results obtained by our model.

One may question on the use of 3D convolution in the front-end component given that

the back-end LSTM can also learn the temporal dependency among the images in the input image sequence. Our results show that it is advantageous and beneficial to utilize both of them together instead. We believe that the 3D convolutions performed in the front-end component can probably capture the short-term temporal correlations among successive images, and the resulting feature maps can thus provide the back-end LSTM with a more global view of the image sequence to help it capture both the long-term and short-term correlations from the image sequence eventually.

9.2 Difficulties in Training with CNN-BLSTM

It is interesting that the previous work [4] has mentioned but failed to reach a better accuracy when they use a CNN as the visual feature extractor as in our work, in comparison with the autoencoder adopted in their work, probably due to multiple reasons in its network architecture design and training strategies.

First and foremost, we used a number of convolutional layers to reduce each stacked image input to a small enough size before feeding it to the BLSTM, which in our case is $2\times2\times2$ along width, height and temporal depth of the image sequences. We found that it would lead to a worse performance if we choose not to reduce it to such a small size.

Second, the maxout activation works better in comparison with both the conventional ReLU activation in CNN and tanh activation in BLSTM. As demonstrated by the maxout-CNN-BLSTM architecture, the maxout activation provides a considerable absolute gain of 3.0% in accuracy when compared to its counterparts, which confirms its free of zero saturation problem and more accurate approximate model averaging with dropout.

Third, techniques of preventing overfitting are important across the whole network. Among the aforementioned three methods, L2-regularization has the most direct impact in addressing this problem, which can prevent the network weights from having too positive or negative values as the network is being trained.

Finally, data augmentation is important for training such a deep network for a low-resource corpus. Though a deep neural network is well-known for its ability in learning high-level and abstract features, that happens only when a sufficient number of training samples is provided and data augmentation is amongst one of the most useful and important strategy to handle this given that no extra data is available.

Chapter 10

Analysis (II): Resource-Rich LRW Corpus

In this section, we provide both qualitative and quantitative analysis on the influence of the spatial transformer module to our model, as compared to the data augmented setup.

10.1 Qualitative: Transformation Samples for Test

(a) Original samples

(b) Spatial-transformed samples

Figure 10.1: Qualitative comparison between original (up) and spatial-transformed image samples (bottom), displayed with reverse normalization. Notice that padding has been applied outside the transformation boxes near the edges of every transformed image sample.

Figure 10.1 shows some contrastive pairs of original and spatial-transformed images produced by our proposed model. It can be noticed that the objects of interest (lips) in the spatial-transformed images tend to be moved closer to the horizontal center, and rotated closer to the horizontal balance. Although data augmentation has already increased the data translation invariance during the training stage, it turns out this seemingly small but significant degree of transformation learned by the spatial transformer is complementary and still contributes to a persuasive gain in our overall word accuracy result.

As directly compared with the spatial-transformed samples in the digit recognition task experimented in [5], one can notice that the degree of transformation is not as huge as those digit samples. However, this is probably because the data collection stage in this corpus has already discarded an enormous number of highly non-frontal samples in advance. To emphasize, this seemingly slight degree of transformation is already helpful and beneficial for the model learning during the training stage with only a limited increase in model parameters.

10.2 Quantitative: Word Accuracy Gain by Class

Table 10.1: Words with the largest accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
MILITARY	82	92	10
PAYING	72	82	10
THOSE	54	64	10
THREE	74	84	10
UNDER	42	52	10
WARNING	88	98	10
WHERE	60	70	10
ANYTHING	72	84	12
WRONG	82	94	12
CRISIS	74	88	14

According to our test statistics, we can observe that while fewer than one third of the classes perform worse with spatial transformer in terms of word accuracy, around half of the classes give better performance instead. We have provided those with the largest class accuracy gains in the list in Table 10.1 due to space limitation. This indicates the

incorporation of spatial transformer has an overall positive effect on the model, and it is not confined only to particular word classes, as proved and confirmed by the final overall accuracy result as mentioned in the above section. Please see the Appendix A for the complete list of word accuracy gain of the 500 classes for details.

Chapter 11

Future Works

In this chapter, we present some possible future research directions for visual speech recognition and other relevant applications.

11.1 Multi-Head Attention Mechanism

One future direction of our research is to examine the effectiveness of the multi-head attention idea in the field of visual speech recognition. Since the concept of multi-head attention was proposed [39], it has been widely adopted in disparate areas under a broad spectrum of architectures. With minor and minimal modification to the original Transformer, it has been proved practical in audio speech recognition [40], video captioning [41], and even multi-task learning [42] in image captioning, object recognition, speech recognition and machine translation. Apart from the abovementioned non-recurrent approaches, works resorting to multi-head attentional recurrent model or its variants have also been observed in speech recognition [43, 44] and voice search [45]. In view of all of the above success, we have borrowed this concept to the area of machine translation and successfully demonstrated its applicability and practicability in one of our works [46] with convincing results of 0.40 gain in BLEU and 0.32% reduction in TER to the baseline model on the WMT'16 English-to-German translation task. However, it remains unknown whether this concept can be applied in the field of visual speech recognition.

To facilitate understanding on the concept, we describe the brief idea of multi-head attention in following subsections:

11.1.1 Major Variants in General Attention Mechanism

The majority of attention mechanisms can be categorized into two groups, namely Bahdanau's additive [47] and Luong's multiplicative [48] styles. Here we choose the latter, which is the default and recommended setting in many popular toolkits such as OpenNMT [1] and TensorFlow [49]:

$$score(\mathbf{h}_t, \hat{\mathbf{h}}_s) = \mathbf{h}_t^\top \mathbf{W}_{sc} \hat{\mathbf{h}}_s \tag{11.1}$$

where $\mathbf{W}_{sc} \in \mathbb{R}^{d_h \times d_h}$, and $\mathbf{h}_t, \hat{\mathbf{h}}_s \in \mathbb{R}^{d_h}$ are the t^{th} and s^{th} hidden state of decoder and encoder with hidden size of d_h, respectively.

11.1.2 Single-Head Attention

The usual single-head attention can be characterized by the following simple formulas:

$$\alpha_{ts} = \frac{\exp[score(\mathbf{h}_t, \hat{\mathbf{h}}_s)]}{\sum_{s'=1}^{|S|} \exp[score(\mathbf{h}_t, \hat{\mathbf{h}}_{s'})]} \tag{11.2}$$

$$t = \sum_{s=1}^{|S|} \alpha_{ts} \hat{\mathbf{h}}_s \tag{11.3}$$

$$\tilde{\mathbf{a}}_t = \tanh(\mathbf{W}_{sm}[t; \mathbf{h}_t]) \tag{11.4}$$

$$\bar{\mathbf{a}}_t = \tanh(\mathbf{W}_{fd}[t; \mathbf{h}_t]) \tag{11.5}$$

where $\alpha_{ts} \in \mathbb{R}$ is the attentional weight of the t^{th} decoder hidden state to the s^{th} encoder hidden state; $|S| \in \mathbb{R}$ is the source sentence length; $t \in \mathbb{R}^{d_h}$ is the context vector of the decoder; $\tilde{\mathbf{a}}_t \in \mathbb{R}^{d_{sm}}$ and $\bar{\mathbf{a}}_t \in \mathbb{R}^{d_{fd}}$ are the input to the softmax layer and input to the next decoder step from the t^{th} decoder hidden state, respectively.

Note that $\mathbf{W}_{sm} \in \mathbb{R}^{d_{sm} \times 2d_h}$ and $\mathbf{W}_{fd} \in \mathbb{R}^{d_{fd} \times 2d_h}$ are two independent linear projections, which are different from the original ones as implemented in OpenNMT and TensorFlow.

11.1.3 Multi-Head Attention

Through replicating the above single-head operations, we can create multiple instances of attention with different parameters, which can potentially learn disparate representations and information between the same hidden states of encoder and decoder. Following the same idea of concatenation in Transformer, we have:

$$\tilde{\mathbf{a}}'_t = concat([\tilde{\mathbf{a}}^1_t, ..., \tilde{\mathbf{a}}^h_t]) \tag{11.6}$$

$$\bar{\mathbf{a}}'_t = concat([\bar{\mathbf{a}}^1_t, ..., \bar{\mathbf{a}}^h_t]) \tag{11.7}$$

where $\tilde{\mathbf{a}}^i_t \in \mathbb{R}^{d'_{sm}}$, $\bar{\mathbf{a}}^j_t \in \mathbb{R}^{d'_{fd}}$, $d'_{sm} = d_{sm}/h$, $d'_{fd} = d_{fd}/h$, and h is the number of heads.

Note that dividing by h has the effect of conserving the dimension of $\tilde{\mathbf{a}}'_t$ and $\bar{\mathbf{a}}'_t$.

11.2 Self-Attention Mechanism

The idea of the self-attention mechanism is to obtain an attentional matrix on the basis of operations on the hidden states itself with trainable and differentiable parameters. This is originally designed as a more interpretable and meaningful sentence embedding in a form 2-D matrix to replace its ordinary weaker vector representation. This novel concept has been successfully applied in multiple areas, author profiling, sentiment analysis and textual entailment included. For author profiling, the model has to predict the age and gender of writers from the content of tweets they write; while the sentiment analysis aims at predicting the rating based on the content in reviews, and the textual entailment targets at labeling the sentence pairs in labels with either contradiction or neutral. With incorporation of this idea in to the model, promising results have been attained consistently across these multiple areas. Afterwards, this self-attention idea becomes a natural extension to other models in a wide spectrum of areas and the performance in tasks such as constituency parsing [50], semantic role labeling [51, 52], question answering [53] and even graph learning [54] has been enhanced and improved. Nevertheless, whether this new technique can be applied to our proposed architecture in continuous word visual speech recognition on LRW corpus remains unknown, and we will carry out experiments on substitution of the last temporal average pooling layer with the self-attentive layer just before the softmax classification layer.

In the following, we describe how the self-attention mechanism can be formulated.

11.2.1 Attention Matrix

$$A = softmax(W_2 \tanh(W_1 H^T)) \tag{11.8}$$

$$M = AH \tag{11.9}$$

The above formula shows how the attention matrix can be computed, with d_a controls the hidden dimension of attention, r represents the number of heads and u is the number of hidden units in each direction of LSTM. While $H \in \mathbb{R}^{n \times 2u}$ is the hidden states in the LSTM, $W_1 \in \mathbb{R}^{d_a \times 2u}$ and $W_2 \in \mathbb{R}^{r \times d_a}$ are the trainable and learnable parameters to decide the attentional behavior. Afterwards, the resulting sentence embeddings can be computed as $M \in \mathbb{R}^{r \times 2u}$.

11.2.2 Penalty Term

$$P = \left\| AA^T - I \right\|_F^2 \tag{11.10}$$

In addition to the computation of attention matrix alone, a penalty term denoted as P has to be applied through a matrix norm of the difference between the multiplication of A and A^T and the identity matrix. This penalty has an effect of favoring each head to focus on only a few number of words, together with penalizing high similarities between different pairs of heads. The penalty term is directly added to the training loss and has shown to be advantageous with a coefficient of 1.0.

11.3 Other Directions

On top of the aforementioned two proposed directions, in the future, we are going to explore the opportunities of applying the maxout units in other more challenging and difficult lip-reading tasks such as the visual speech recognition of sentences using the Lip Reading Sentences [15] and GRID corpora [55], through utilizing other advanced end-to-end networks and architectures.

Moreover, we are also looking for the possibilities of replacing and upgrading the current spatial transformer to its 3D version with additional transformation along the temporal axis, and other feasible and viable approaches that have potential in further

pushing the performance of the current state-of-the-art visual speech recognition system forward on top of building a much more competitive baseline.

In addition, we are also considering the development and innovation of other useful and exciting applications that leverage the visual information of lips on the face. There have been latest researches that utilize the visual lip information in audio-visual speech recognition, speech separation or even talking face synthesis, but they are still very limited because this field and community is still actively growing and emerging. To obtain desirable results in these applications, it would require many advanced techniques gathered from various fields and we hope we will be able to find and innovate new strategies to handle the difficulties in those tasks. On top of this, we are also looking for opportunities in working on novel and creative applications that have not been done before, and can be benefited from the visual lip information used in the visual speech recognition.

Chapter 12

Conclusion

In short, we have successfully showcased the capability and feasibility of designing an end-to-end deep neural network for the low-resource lip-reading task using CNN and BLSTM with incorporation of maxout activation units with dropout. We are able to achieve a state-of-the-art accuracy of 87.6% on the low-resource Ouluvs2 10-phrase task without using any external data resources. While the maxout units do not have the high zero saturation rate problem which occurs in many other activation units, its combination with dropout has an effect of a more accurate approximate model averaging during training. With the above advantage in conjunction with a carefully designed end-to-end neural network with various techniques applied, the well-known problem of overfitting due to inadequacy in training data for a model with considerable size and capacity can be alleviated.

On the other hand, we have also demonstrated the practicality of spatial transformer in the field of visual speech recognition under the neural network model with 3D-ResNet and BLSTM, and conducted both qualitative and quantitative analysis on its effectiveness to the capacity of the resulting model in details. We can observe from the qualitative samples that the spatial-transformed images have a tendency to move closer to the horizontal center, and rotated closer towards the horizontal balance as trained by the network, which has a great potential in benefiting the latter and remaining parts of the network during training. At the same time, we can notice the word accuracy gain in the spatial transformer setup is not restricted to individual but nearly half of the classes, and the overall effect is positive as proved in the overall absolute accuracy gain of 0.92% to the data-augmented setup alone during our evaluation with test samples. The aforementioned analysis confirms the effectiveness of spatial transformer in dealing with the issue of diverse facial viewpoints in the video data in the field of visual speech recognition. We hope that the use of spatial transformer in this work can help and inspire the lip-reading community in future's work on the more general and challenging sequence-to-sequence continuous visual speech recognition task using the Lip Reading Sentences (LRS) corpus [15], which involves an even larger degree of diversity in facial viewpoints. We are also

expecting and looking forward to more uses of spatial transformer in other applications and fields relevant to visual speech recognition such as audio-visual speech recognition, speech separation and talking face synthesis as mentioned in the above sections.

In the future, we will continue our works in visual speech recognition through exploring other effective techniques and examine other interesting and exciting applications that can be built by leveraging the latest resource-rich lip-reading corpora. This includes but not limited to the multi-head attention and self-attention mechanism as described in the previous section.

References

[1] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proc. ACL*, 2017. [Online]. Available: https://doi.org/10.18653/v1/P17-4012.

[2] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis," in *Proceedings of the International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 1. IEEE, 2015, pp. 1–5.

[3] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proceedings of the Asian Conference on Computer Vision*, 2016.

[4] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017, pp. 2592–2596.

[5] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[6] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen, "A compact representation of visual speech data using latent variables," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, 2014.

[7] G. I. Chiou and J. N. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192–1195, 1997.

[8] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. L. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the International Conference on Machine Learning*, 2011, pp. 689–696.

[10] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann

machines," in *Advances in Neural Information Processing Systems*, 2012, pp. 2222–2230.

[11] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based CNN for visual speech recognition," in *Proceedings of the Asian Conference on Computer Vision*, 2016, pp. 277–289.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[13] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the British Machine Vision Conference*, 2014.

[14] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proceedings of the International Conference on Machine Learning*, vol. 28, 2013.

[15] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2017, pp. 3444–3453.

[16] K. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[18] ——, "Deep word embeddings for visual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 4974–4978.

[19] M. Wand, J. Schmidhuber, and N. T. Vu, "Investigations on end-to-end audiovisual fusion," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3041–3045.

[20] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," *arXiv preprint arXiv:1802.06424*, 2018.

[21] S. Palaskar, R. Sanabria, and F. Metze, "End-to-end multimodal speech recognition," *arXiv preprint arXiv:1804.09713*, 2018.

[22] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *arXiv preprint arXiv:1809.02108*, 2018.

[23] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[24] ——, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[25] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *British Machine Vision Conference*, 2017.

[26] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning*, vol. 30, 2013.

[27] X. G. Li and X. H. Wu, "Improving long short-term memory networks using maxout units for large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 4600–4604.

[28] W. Wu, M. Kan, X. Liu, Y. Yang, S. Shan, and X. Chen, "Recursive spatial transformer (ReST) for alignment-free face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3772–3780.

[29] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides, "Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses," in *The IEEE International Conference on Computer Vision*, 2017, pp. 4000–4009.

[30] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng, "Egocentric gesture recognition using recurrent 3D convolutional neural networks with spatiotemporal transformer modules," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3763–3771.

[31] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "ST-GAN: Spatial transformer generative adversarial networks for image compositing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9455–9464.

[32] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Advances in Neural Information Processing Systems*, 2016, pp. 217–225.

[33] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in neural information processing systems*, 2016, pp. 64–72.

[34] W. M. Fisher, "The DARPA speech recognition research database: specifications and status," in *Proceedings of the DARPA Workshop Speech Recognition*, 1986, pp. 93–99.

[35] FFmpeg team, "FFmpeg," https://ffmpeg.org.

[36] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[37] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 2135–2135.

[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[40] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[41] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," *arXiv preprint arXiv:1804.00819*, 2018.

[42] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit, "One model to learn them all," *arXiv preprint arXiv:1706.05137*, 2017.

[43] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[44] T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Multi-head decoder for end-to-end speech recognition," *arXiv preprint arXiv:1804.08050*, 2018.

[45] T. N. Sainath, C.-C. Chiu, R. Prabhavalkar, A. Kannan, Y. Wu, P. Nguyen, and Z. Chen, "Improving the performance of online neural transducer models," *arXiv preprint arXiv:1712.01807*, 2017.

[46] I. Fung and B. Mak, "Multi-head attention for end-to-end neural machine translation," in *2018 International Symposium on Chinese Spoken Language Processing*, 2018.

[47] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.

[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[50] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," *arXiv preprint arXiv:1805.01052*, 2018.

[51] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum, "Linguistically-informed self-attention for semantic role labeling," *arXiv preprint arXiv:1804.08199*, 2018.

[52] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *AAAI Conference on Artificial Intelligence*, 2018.

[53] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "QANet: Combining local convolution with global self-attention for reading comprehension," in *International Conference on Learning Representations*, 2018.

[54] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[55] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

List of Publications

Ivan Fung and Brian Mak, "End-to-end low-resource lip-reading with maxout CNN and LSTM," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 2511–2515.

Ivan Fung and Brian Mak, "Multi-Head Attention for End-to-End Neural Machine Translation," in *2018 International Symposium on Chinese Spoken Language Processing*.

Ivan Fung and Brian Mak, "Improving deep visual speech recognition with spatial transformer," **Submitted to** *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019.

Appendix A

Word Accuracy Gain by Class on LRW

Table A.1: Part I - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
CRISIS	74	88	14
ANYTHING	72	84	12
WRONG	82	94	12
ACTION	54	64	10
COUNTRY	78	88	10
EDITOR	80	90	10
GROWING	80	90	10
HAVING	68	78	10
MILITARY	82	92	10
PAYING	72	82	10
THOSE	54	64	10
THREE	74	84	10
UNDER	42	52	10
WARNING	88	98	10
WHERE	60	70	10
AGREE	72	80	8
BENEFITS	74	82	8
BETTER	58	66	8
CAPITAL	74	82	8
COMPANIES	74	82	8
DEGREES	86	94	8
ENOUGH	64	72	8
HEARD	52	60	8
HISTORY	76	84	8
INSIDE	76	84	8
LITTLE	56	64	8
MILLION	62	70	8
MONEY	76	84	8
NOTHING	60	68	8
QUITE	76	84	8
SPEND	46	54	8
STARTED	58	66	8

Table A.2: Part II - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
THING	38	46	8
THROUGH	66	74	8
TONIGHT	60	68	8
WAITING	82	90	8
WORDS	52	60	8
ABOUT	56	62	6
ABSOLUTELY	92	98	6
ACCESS	80	86	6
ALLEGATIONS	92	98	6
ALREADY	74	80	6
ANSWER	62	68	6
BELIEVE	88	94	6
BRITAIN	48	54	6
CHANCE	62	68	6
CONCERNS	84	90	6
COULD	50	56	6
COURSE	72	78	6
DECISION	84	90	6
FACING	86	92	6
GUILTY	74	80	6
JUDGE	68	74	6
LEVELS	90	96	6
MESSAGE	80	86	6
NATIONAL	80	86	6
POLICE	70	76	6
POLICY	86	92	6
PRICE	56	62	6
RIGHTS	62	68	6
SERIOUS	72	78	6
SHOULD	56	62	6
SOUTHERN	82	88	6
TERMS	78	84	6
TRYING	64	70	6
USING	70	76	6
WEATHER	78	84	6
WINDS	86	92	6
YEARS	52	58	6
ABUSE	82	86	4

Table A.3: Part III - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	[+DA] (%)	[+DA +ST] (%)	**Gain (%)**
ACCORDING	88	92	4
ACTUALLY	56	60	4
AGAINST	58	62	4
BRING	52	56	4
BUILDING	78	82	4
CENTRAL	80	84	4
CHANGE	74	78	4
CLAIMS	70	74	4
CONFERENCE	78	82	4
CONTROL	82	86	4
DAVID	80	84	4
ELECTION	70	74	4
EUROPE	82	86	4
EVENTS	76	80	4
EVERYONE	88	92	4
EXPECTED	78	82	4
FORCE	84	88	4
FRANCE	66	70	4
FURTHER	78	82	4
FUTURE	88	92	4
GROUP	84	88	4
HEAVY	74	78	4
IMPACT	84	88	4
JAMES	84	88	4
KILLED	82	86	4
LARGE	86	90	4
MATTER	50	54	4
MEDICAL	84	88	4
MEETING	68	72	4
MIGRANTS	90	94	4
MINUTES	70	74	4
MISSING	80	84	4
NEVER	68	72	4
OTHER	58	62	4
PERSON	52	56	4
PHONE	66	70	4
POLITICAL	84	88	4
POSITION	92	96	4

Table A.4: Part IV - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
POWER	70	74	4
QUESTION	82	86	4
REALLY	62	66	4
RECENT	72	76	4
RECORD	88	92	4
SCHOOL	70	74	4
SENIOR	74	78	4
SERIES	74	78	4
SERVICE	86	90	4
SHORT	76	80	4
SIDES	76	80	4
SPENDING	72	76	4
STAND	52	56	4
STILL	68	72	4
SYRIA	88	92	4
TAKEN	58	62	4
TAKING	56	60	4
TALKING	74	78	4
THEIR	40	44	4
THESE	48	52	4
THREAT	74	78	4
TRUST	70	74	4
WITHIN	78	82	4
WORLD	60	64	4
WOULD	72	76	4
ACCUSED	94	96	2
AFFAIRS	80	82	2
AFFECTED	74	76	2
AFRICA	92	94	2
AFTER	80	82	2
AGREEMENT	94	96	2
ALMOST	86	88	2
AMERICAN	78	80	2
AMONG	80	82	2
ASKING	82	84	2
BECAUSE	50	52	2
BECOME	90	92	2
BEING	56	58	2

Table A.5: Part V - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
BENEFIT	82	84	2
BETWEEN	92	94	2
BIGGEST	72	74	2
BRITISH	78	80	2
BROUGHT	72	74	2
BUDGET	88	90	2
CAMPAIGN	96	98	2
CERTAINLY	62	64	2
CHALLENGE	80	82	2
CHANGES	92	94	2
CLEAR	72	74	2
CLOSE	74	76	2
COMMUNITY	84	86	2
CONSERVATIVE	84	86	2
COUNTRIES	74	76	2
CUSTOMERS	92	94	2
DEFICIT	82	84	2
DETAILS	94	96	2
DIFFICULT	92	94	2
EASTERN	84	86	2
EVERYTHING	84	86	2
FIGURES	82	84	2
FINAL	84	86	2
FIRST	64	66	2
FOOTBALL	86	88	2
FORCES	84	86	2
FOREIGN	82	84	2
FORWARD	86	88	2
FRIDAY	86	88	2
FRONT	80	82	2
GAMES	70	72	2
GIVEN	70	72	2
GREAT	58	60	2
GREECE	80	82	2
HAPPENING	82	84	2
HOMES	90	92	2
HOSPITAL	92	94	2
HOURS	70	72	2

Table A.6: Part VI - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
HOUSING	78	80	2
IMPORTANT	72	74	2
INDUSTRY	72	74	2
INTEREST	80	82	2
INVOLVED	84	86	2
ISSUE	78	80	2
JUSTICE	80	82	2
LEGAL	76	78	2
LIKELY	68	70	2
LIVES	82	84	2
LONDON	58	60	2
MARKET	76	78	2
MEANS	72	74	2
MEDIA	70	72	2
MEMBERS	94	96	2
MINISTER	84	86	2
MOMENT	94	96	2
MORNING	90	92	2
NUMBERS	84	86	2
OBAMA	94	96	2
OFFICE	82	84	2
OPERATION	94	96	2
OTHERS	78	80	2
OUTSIDE	76	78	2
PARENTS	86	88	2
PARTY	86	88	2
PERHAPS	96	98	2
PERIOD	74	76	2
POLITICS	86	88	2
PRICES	82	84	2
PRIVATE	96	98	2
PROBLEM	78	80	2
PROCESS	88	90	2
PUBLIC	90	92	2
SERVICES	78	80	2
SEVEN	82	84	2
SIGNIFICANT	92	94	2
SINCE	68	70	2

Table A.7: Part VII - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	[+DA] (%)	[+DA +ST] (%)	**Gain (%)**
SINGLE	80	82	2
SITUATION	88	90	2
SMALL	92	94	2
SOCIAL	82	84	2
SOMETHING	74	76	2
SPEAKING	76	78	2
STAFF	84	86	2
STATES	54	56	2
SUPPORT	84	86	2
SYRIAN	72	74	2
TALKS	84	86	2
TOGETHER	80	82	2
UNION	88	90	2
VIOLENCE	86	88	2
WALES	86	88	2
WANTED	74	76	2
WANTS	72	74	2
WATCHING	88	90	2
WEAPONS	98	100	2
WELFARE	96	98	2
WESTERN	92	94	2
WHICH	74	76	2
WOMEN	98	100	2
WORST	72	74	2
ACROSS	76	76	0
AGAIN	72	72	0
AHEAD	82	82	0
ALLOW	70	70	0
ALLOWED	76	76	0
AMERICA	80	80	0
AMOUNT	76	76	0
ANNOUNCED	76	76	0
ATTACK	84	84	0
AUTHORITIES	86	86	0
BANKS	76	76	0
BLACK	80	80	0
BORDER	64	64	0
CAMERON	92	92	0

Table A.8: Part VIII - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
CANNOT	80	80	0
CHARGE	76	76	0
CHIEF	94	94	0
COMING	88	88	0
CONFLICT	90	90	0
DEBATE	74	74	0
DECIDED	82	82	0
DESPITE	88	88	0
DURING	72	72	0
ECONOMIC	88	88	0
EUROPEAN	92	92	0
EVENING	84	84	0
EVERY	74	74	0
EVIDENCE	92	92	0
EXAMPLE	88	88	0
EXTRA	80	80	0
FAMILIES	92	92	0
FIGHTING	82	82	0
FOCUS	82	82	0
FORMER	94	94	0
GERMANY	98	98	0
GLOBAL	82	82	0
GOING	54	54	0
GOVERNMENT	88	88	0
HUMAN	88	88	0
IMMIGRATION	92	92	0
INDEPENDENT	70	70	0
INFLATION	88	88	0
INFORMATION	96	96	0
INQUIRY	96	96	0
INVESTMENT	96	96	0
IRELAND	94	94	0
LABOUR	86	86	0
LEADER	74	74	0
LEAST	64	64	0
LEAVE	84	84	0
MAJOR	86	86	0
MAKES	58	58	0

Table A.9: Part IX - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
MANCHESTER	84	84	0
MOVING	96	96	0
NIGHT	64	64	0
NORTHERN	84	84	0
NUMBER	72	72	0
OFFICIALS	94	94	0
OFTEN	72	72	0
OPPOSITION	96	96	0
PARLIAMENT	96	96	0
PATIENTS	90	90	0
PEOPLE	90	90	0
POLITICIANS	90	90	0
POSSIBLE	96	96	0
POTENTIAL	94	94	0
PRIME	98	98	0
PRISON	72	72	0
PROBLEMS	88	88	0
PROVIDE	98	98	0
RATES	68	68	0
REPORTS	94	94	0
RESPONSE	94	94	0
RETURN	86	86	0
RIGHT	66	66	0
RUSSIAN	68	68	0
SAYING	48	48	0
SCHOOLS	80	80	0
SECOND	66	66	0
SECTOR	60	60	0
SIMPLY	92	92	0
SOCIETY	82	82	0
SPEECH	86	86	0
SPENT	56	56	0
STATE	62	62	0
STRONG	88	88	0
SUNSHINE	96	96	0
SYSTEM	90	90	0
THEMSELVES	100	100	0
THERE	40	40	0

Table A.10: Part X - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	**[+DA] (%)**	**[+DA +ST] (%)**	**Gain (%)**
THOUGHT	54	54	0
THOUSANDS	96	96	0
TOWARDS	82	82	0
UNTIL	50	50	0
VICTIMS	96	96	0
VOTERS	84	84	0
WATER	84	84	0
WEEKEND	94	94	0
WEEKS	76	76	0
WESTMINSTER	100	100	0
YOUNG	72	72	0
AFTERNOON	98	96	-2
ANOTHER	72	70	-2
AREAS	82	80	-2
AROUND	50	48	-2
ARRESTED	86	84	-2
BILLION	82	80	-2
BUSINESS	72	70	-2
BUSINESSES	88	86	-2
CALLED	70	68	-2
CASES	80	78	-2
CHARGES	78	76	-2
CHILD	64	62	-2
CHILDREN	88	86	-2
CHINA	84	82	-2
COMES	94	92	-2
COUNCIL	80	78	-2
COUPLE	86	84	-2
CRIME	90	88	-2
CURRENT	70	68	-2
DIFFERENCE	84	82	-2
DIFFERENT	70	68	-2
DOING	64	62	-2
ECONOMY	88	86	-2
EDUCATION	94	92	-2
EMERGENCY	88	86	-2
ENGLAND	86	84	-2
EXPECT	82	80	-2

Table A.11: Part XI - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	[+DA] (%)	[+DA +ST] (%)	**Gain (%)**
FAMILY	90	88	-2
FINANCIAL	86	84	-2
FOLLOWING	98	96	-2
FRENCH	96	94	-2
GEORGE	76	74	-2
GETTING	56	54	-2
GROWTH	84	82	-2
HOUSE	64	62	-2
HUNDREDS	96	94	-2
ISLAMIC	98	96	-2
LEADERSHIP	94	92	-2
LOCAL	68	66	-2
LONGER	76	74	-2
LOOKING	82	80	-2
MASSIVE	90	88	-2
MEMBER	94	92	-2
MINISTERS	92	90	-2
MONTHS	74	72	-2
NORTH	82	80	-2
PARTIES	88	86	-2
PLACES	78	76	-2
PRESIDENT	84	82	-2
PRESS	66	64	-2
PRETTY	82	80	-2
PROTECT	88	86	-2
RATHER	74	72	-2
REFERENDUM	98	96	-2
RESULT	80	78	-2
RUSSIA	84	82	-2
SCOTLAND	98	96	-2
SCOTTISH	84	82	-2
SECRETARY	90	88	-2
SPECIAL	90	88	-2
STATEMENT	94	92	-2
STORY	84	82	-2
STREET	80	78	-2
TEMPERATURES	100	98	-2
THINGS	58	56	-2

Table A.12: Part XII - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	[+DA] (%)	[+DA +ST] (%)	**Gain (%)**
THINK	40	38	-2
THIRD	82	80	-2
TIMES	84	82	-2
TOMORROW	96	94	-2
TRADE	70	68	-2
WELCOME	94	92	-2
WHILE	74	72	-2
WHOLE	82	80	-2
WITHOUT	92	90	-2
WORKERS	84	82	-2
YESTERDAY	84	82	-2
ALWAYS	80	76	-4
ATTACKS	78	74	-4
BEFORE	98	94	-4
BEHIND	82	78	-4
BUILD	82	78	-4
CANCER	72	68	-4
CLOUD	90	86	-4
CONTINUE	70	66	-4
COURT	62	58	-4
DESCRIBED	98	94	-4
ENERGY	84	80	-4
EVERYBODY	92	88	-4
EXACTLY	74	70	-4
FIGHT	72	68	-4
FOUND	94	90	-4
GENERAL	84	80	-4
HAPPEN	72	68	-4
HEALTH	88	84	-4
HEART	84	80	-4
INCREASE	84	80	-4
ITSELF	82	78	-4
LATEST	80	76	-4
LIVING	62	58	-4
MAJORITY	92	88	-4
MEASURES	90	86	-4
MILLIONS	84	80	-4
MONTH	88	84	-4

Table A.13: Part XIII - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	[+DA] (%)	[+DA +ST] (%)	**Gain (%)**
MURDER	74	70	-4
NEEDS	48	44	-4
OFFICERS	92	88	-4
ORDER	72	68	-4
PARTS	96	92	-4
PERSONAL	74	70	-4
PLANS	86	82	-4
POINT	78	74	-4
POWERS	76	72	-4
PRESSURE	88	84	-4
PROBABLY	88	84	-4
QUESTIONS	96	92	-4
REMEMBER	92	88	-4
REPORT	76	72	-4
RUNNING	72	68	-4
SECURITY	90	86	-4
SEEMS	74	70	-4
STAGE	70	66	-4
UNDERSTAND	74	70	-4
WORKING	78	74	-4
DEATH	68	62	-6
EARLY	82	76	-6
GIVING	62	56	-6
GROUND	80	74	-6
HAPPENED	74	68	-6
HIGHER	74	68	-6
ISSUES	86	80	-6
LATER	70	64	-6
MAKING	80	74	-6
MAYBE	92	86	-6
PLACE	72	66	-6
SEVERAL	76	70	-6
SOMEONE	92	86	-6
SOUTH	92	86	-6
START	78	72	-6
SUNDAY	80	74	-6
TODAY	84	78	-6
TRIAL	86	80	-6

Table A.14: Part XIV - Complete list of word classification accuracy gains from the [+DA] to [+DA +ST] setup.

Word	[+**DA**] (%)	[+**DA** +**ST**] (%)	**Gain** (%)
UNITED	72	66	-6
WHETHER	88	82	-6
ASKED	56	48	-8
COMPANY	78	70	-8
KNOWN	76	68	-8
LEADERS	76	68	-8
LEVEL	70	62	-8
MIDDLE	64	56	-8
RULES	80	72	-8
MIGHT	64	54	-10
SENSE	76	66	-10
REASON	82	70	-12

Biographical Note

Fung, Ho Long (Ivan Fung) received his bachelor degree with double majors in computer science and mathematics (pure mathematics track) in the Hong Kong University of Science and Technology (HKUST) in the late 2016. Afterwards, he decided to pursue his master degree in computer science in the same university owing to his interest towards the growing and emerging field of machine learning under the area of artificial intelligence. In the wake of receiving his master degree, he has a rigorous plan of working in a research-based startup with a primary focus on automatic speech recognition (ASR) and natural language processing (NLP) for about half a year before commencing his PhD study to continue his role as a research student in HKUST in the late 2019.

His major research interests include visual speech recognition (lip-reading) and machine translation. He is looking forward to developing and improving various novel applications that can be benefited from the idea of lip-reading, on the basis of different innovative and state-of-the-art techniques originally employed in the fields of automatic speech recognition, computer vision and natural language processing.