Abstract

In recent years deep learning techniques and their applications have attracted much research attraction globally. Aiming to prompt academic collaborations amongst universities and industries, this talk is mainly served as an introduction to the various on-going research projects on deep learning of our research group based on London South Bank University, including

- Feature extraction and labelling large data sets,
- Speaker-independent lip reading,
- Dimensionality reduction techniques, and
- Optimal CNN topology design.

Real-world case studies and the relevant systems demos will be provided.

Speaker-independent lip reading: This project is about the use of deep learning to automate the visual speech recognition or lip reading of a person speaking using purely visual lip movements without any audio input. There are many obstacles to lip reading which. These include the speaker dependency of neural network based lip reading where the performance of a lip reading system is dependent on the speaker whom it was tested on and would vary when implemented on different speakers; that lack of available training data that covers a wide variety of vocabulary and contexts required to train lip reading systems that are suitable for natural spoken language and the inability of existing models to distinguish between homopheme words or words that produce identical lip movements when uttered.

A neural network model is proposed for the specific classification of phonemes and visemes which are the most fundamental units of speech with the model itself being a stacked configuration of convolutional neural networks and recurrent neural networks. A phoneme corresponds to a spoken character or sound such that each one has an associated acoustical signal, whereas a viseme is a distinct lip movement or visual units of sound that is produced for every spoken character of which there are around a dozen.

Work that has been carried out to addresses such challenges which include the use of contour mapping, which is an edge detection pre-processing procedure for extracting an outline of someone's lips to use as the feature input; a three-dimensional convolutional neural network for classifying visemes; and a stacked recurrent neural network with word vectors and embeddings for deciphering homopheme words as well as review of which feature representations are the most ideal for deep learning based lip reading.

Dimensionality reduction techniques: A supervised version of *t*-SNE algorithm has been proposed which can be applied in any high dimensional datasets for visualisation and/or as a feature extraction for classification problems in a (much) lower dimensional space. The super performance of this new algorithm can be demonstrated by applying it to three different high dimensional datasets: MNIST, Chest x-ray, and SEER Breast Cancer. The embedded data generated by the algorithm in a 2-dimensional space has shown a better visualization and a significant improvement in terms of classification accuracy in comparison to the original t-SNE.

Optimal CNN topology design: Existing Convolutional Neural Network (CNN) models come with different number of layers. This made the CNN application less portable to different fields and data sizes. An ongoing study in LSBU is focussing to address this issue. The study aims to investigate how much each layer can contribute to the overall feature learning by using different data separability measures in order to quantify the layers learning capacity.