International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
Volume 47, Issue 01, Quarter 01 (January to March 2018)
An Indexed and Referred Journal with Impact Factor: 2.80
ISSN (Online): 2319-6564
www.ijesonline.com

Deep Multimodal Learning for Audio Visual Speech Recognition

Archana Panda[1], Yogo Maya Mahapatra[2] and Jagannath Ray[3]

[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar
[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Publishing Date: 3rd March, 2018

Abstract

In this paper, we present methods in deep multimodal learning for fusing speech and visual modalities for Audio-Visual Automatic Speech Recognition (AV-ASR). First, we study an approach where uni-modal deep networks are trained separately and their final hidden layers fused to obtain a joint feature space in which another deep network is built. While the audio network alone achieves a phone error rate (PER) of 41% under clean condition on the IBM large vocabulary audio-visual studio dataset, this fusion model achieves a PER of 35:83% demonstrating the tremendous value of the visual channel in phone classification even in audio with high signal to noise ratio. Second, we present a new deep network architecture that uses a bilinear soft max layer to account for class specific correlations between modalities. We show that combining the posteriors from the bilinear networks with those from the fused model mentioned above results in a further significant phone error rate reduction, yielding a final PER of 34:03%.

Keywords: *Audio-Visual Automatic Speech Recognition (AV-ASR), Multimodal Learning, Deep Neural Networks.*

Introduction

Human speech perception is not only about hearing but also about seeing: our brain integrates the waveforms representing the speech information as well as the lips poses and motions, often called visemes, which carry important visual information about what is being said. This has been demonstrated by the so called McGurk effect [1], which shows that a voicing of ba and a mouthing of ga is perceived as being da. In the presence of noise and multiple speakers (cocktail party effect), humans rely on lip reading in order to enhance speech recognition [2]. The visual information is also important in a clean speech scenario as it helps in disambiguating voices with similar acoustics [3]. In Audio-Visual Automatic Speech Recognition (AV- ASR), both audio recordings and videos of the person talking are available at training time. It is challenging to build models that integrate both visual and audio information. This work was done while Youssef M roueh was an intern in the Speech and Algorithms Group at IBM T.J Watson Research Center enhance the recognition performance of the overall system. While most previous works in AV-ASR focused on enhancing the performance in the noisy case [4, 5], where the visual information can be crucial, we focus in this paper on showing that the visual information is indeed helpful even in the clean speech scenario.

Multimodal learning consists of fusing and relating in- formation coming from different sources, hence AV-ASR is an important multimodal problem. Finding correlations be- tween different modalities, and modeling their interactions, has been addressed in various learning frameworks and has been applied to AV-ASR [6, 7, 8, 9, 10]. Deep Neural Networks (DNN) have shown impressive performance in both audio and visual classification tasks, which is why we restrict ourselves to the deep multimodal learning framework.

In this paper, we propose methods in deep learning to fuse modalities, and validate them on the IBM AV-ASR Large Vocabulary

Studio Dataset. First we consider the training of two networks on the audio and the visual modality separately. Then, considering the last layer of each network as a better feature space, and concatenating them, we train a classifier on that joint representation, and obtain gains in Phone Error Rates (PER), with respect to an audio-only trained network. We then propose a new bilinear network that accounts for correlations between modalities and allows for joint training of the two networks, we show that a committee of such bilinear networks, fused at the level of posteriors, achieves a better PER in a clean speech scenario.

The paper is organized as follows: we present the IBM AV-ASR large vocabulary studio dataset, our feature extraction pipeline for the audio and the visual channels. Next, we present results for the fusion of net-works separately trained on each modality. We introduce the bilinear DNN that allows for a joint training and captures correlations between the two modalities, and derive its back-propagation algorithm. Finally we present posterior combination of bimodal and bilinear bimodal DNNs.

Feature Extraction

For the audio channel we extract 24 MFCC coefficients at 100 frames per second. Nine consecutive frames of MFCC coefficients are stacked and projected to 40 dimensions using an LDA matrix. Input to the audio neural network is formed by concatenating 4 LDA frames to the central frame of interest, resulting in an audio feature vector of dimension 360.

Jones algorithm: We then do a mouth carving by an open CV mouth detection model. Both these utilize the ENCARA2 model as described in. In order to get an invariant representation to small distortions and scales we then extract level 1 and level 2 scattering coefficients on the 64 64 mouth region of interest and then reduce their dimension to 60 using LDA (Linear discriminant Analysis). In order to match the audio frame rate we replicate video frames according to audio and video time stamps. We also add 4 context frames to the central frame of interest, and obtain finally a visual feature vector of dimension 540.

Context-dependent Phoneme Targets

Each audio + video frame is labeled with one of 1328 targets that represent context dependent phonemes. 42 phones in phonetic context of 2 are clustered using decision trees down to 1328 classes. We measure classification error rate at the level of these 1328 classes, this is referred to as phone error rate (PER).

Conclusion

In this paper we have studied deep multimodal learning. We can think of those terms as messages passed between networks through the bilinear term. In that way, one network influences the weights of the other one. For the rest of the updates, it follows standard back-propagation in both networks; we give it here for com acoustic conditions using visual channel in addition to speech results in significantly improved classification performance. A bilinear bimodal DNN is introduced which leverages correlation be-tween the audio and visual modalities, and leads to further error rate reduction.

References

[1] H. McGurk and J. MacDonald, Hearing lips and seeing voices, Nature, vol. 264, pp. 746–748, 1976.
[2] S. Cox, R. Harvey, Y. Lan, and J. Newman, The challenge of multispeaker lip-reading, in International Conference on Auditory-Visual Speech Processing, 2008.
[3] Q. Summerfield, Lipreading and audio-visual speech perception in Trans. R. Soc., London, 1992.
[4] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, Audio-visual automatic speech recognition: An overview., in Issues in Visual and Audio-Visual Speech Processing. MIT Press, 2004.
[5] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee and Andrew Y. Ng, Multimodal deep learning,‖ in

International Conference on Machine Learning (ICML), Bellevue, USA, June 2011.

[6] Mihai Gurban and et al., Information theoretic feature extraction for audio-visual speech recognition, IEEE Transactions on signal processing, 2009.

[7] Patrick Lucey and Sridha Sridharan, Patch-based representation of visual speech, in Proceedings of the HC-SNet Workshop on Use of Vision in Human- computer Interaction-Volume 56, Darlinghurst, Australia, Australia, 2006, VisHCI '06, pp. 79–85, Australian Computer Society, Inc.

[8] Uwe Meier, Wolfgang Hrst, and Paul Duchnowski, Adaptive bimodal sensor fusion for automatic speech reading, 1996.

[9] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos, ―Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition,‖ in IEEE 9th Workshop on Multi- media Signal Processing, MMSP 2007, Chania, Crete, Greece, October 1-3, 2007, 2007, pp. 264–267.

[10] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos, ―Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition,‖ IEEE Transactions on Audio, Speech & Language Processing, vol. 17, no. 3, pp. 423–435, 2009.