Multimodal Data Fusion Architecture

A thesis submitted to Faculty of Computers and Artificial Intelligence, Fayoum University in fulfillment of the requirements for the Ph.D degree in computer science

Submitted By:

Hadeer Mostafa Sayed Tolba
Assistant Lecturer, Computer Science Department,
Faculty of Computers and Artificial Intelligence, Fayoum
University

Supervised By:

Prof. Dr. Shereen Aly Taie
Professor in Computer Science Department, Vice Dean of the Faculty of Computers and Artificial Intelligence for Postgraduate Studies and Research, Fayoum University

Prof.Dr. Hesham Eldeeb Professor in Electronics Research Institute
President of New Cairo Technological University

Faculty of Computers and Artificial Intelligence Fayoum University
Egypt
2023

ABSTRACT

- In the big data era, we work with various datasets from different sources and fields that describe a particular event. These datasets comprise numerous modalities, each having distinct representations, distributions, scales, and densities. Machine learning algorithms need to have the ability to interpret and analyze multiple modes of signals together. The process of combining heterogeneous datasets with varying modalities to produce more practical and comprehensive information is known as multimodal data fusion.
- Audiovisual speech recognition (AVSR) is a technology that endeavors to identify spoken words and phrases by analyzing both audio and visual features. AVSR systems incorporate the speaker's voice's audio output along with visual cues, such as lip movements, facial expressions, and body language, to enhance speech recognition accuracy. By fusing these two modalities, AVSR can frequently attain higher accuracy levels than conventional speech recognition systems that rely solely on audio input.
- At the beginning, this thesis presents a comprehensive analysis of techniques used to merge diverse data types, also known as multimodal data fusion. It evaluates and discusses various strategies, highlighting their advantages and disadvantages. Additionally, this thesis explores the AVSR task as a specific case study, focusing on the most recent fusion models for audio and visual data. Finally, it offers a detailed assessment of recent research related to multimodal data fusion in the context of the AVSR task.
- In this thesis, three fusion models are proposed: BiVAE (Bimodal Variational Autoencoder), BiVAE_SoA (Bimodal Variational Autoencoder based on soft attention), and BiVAE_SeA (Bimodal Variational Autoencoder based on self-attention). These models aim to combine audio and visual data for effective fusion in the context of audiovisual tasks. The BiVAE model leverages the capabilities of the Variational Autoencoder (VAE) to smoothly learn latent representations and generate new data. This enables it to create a unified representation that combines both audio and visual features. Additionally, the model considers scenarios where certain modalities may be absent during supervised training and testing phases.
- The BiVAE_SoA and BiVAE_SeA models are enhanced versions of the BiVAE model that incorporate attention mechanisms. These attention mechanisms, namely soft attention and self-attention, are employed to develop models capable

of prioritizing signals with more informative content during the recognition task, rather than relying equally on all complementary signals. By leveraging different attention mechanisms, the BiVAE_SoA and BiVAE_SeA models assign a higher weight to the relevant signal than the other signal. The assignment of these weights is dynamic and based on the signal-to-noise ratio (SNR) of each observation. This allows the models to adaptively determine the importance of each signal and focus more on the one that contains valuable information.

- The experiment aimed to evaluate how well the AVSR tasks using BiVAE, BiVAE SoA, and BiVAE SeA models performed compared to the Audio Speech Recognition (ASR) tasks especially, when dealing with corrupted audio signals. The experiment used two techniques for extracting audio features, Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC), and tested the performance of three different classifiers: Long-short Term Memory (LSTM), Artificial Neural Network (ANN), and Support Vector Machine (SVM). Additionally, the experiment evaluated the performance of the proposed fusion models against other state-of-the-art models whether both modalities are available or only a single modality is available during the supervised training and test.
- The results showed that the BiVAE_SeA is the best-proposed fusion model compared to other proposed models. In a more comprehensive view, the AVSR task based on the proposed fusion models BiVAE, BiVAE_SoA, and BiVAE SeA outperformed the ASR task by an average accuracy difference up to 40.18%, 40.77%, and 41.07%, respectively. Furthermore, even though the MFCC technique performed better than GFCC in ASR, their performances were quite similar in AVSR, particularly when looking at the results for ANN and SVM classifiers. This suggests that the proposed fusion models can generalize their performance across different audio feature extraction techniques. Moreover, in both ASR and AVSR tasks, the SVM classifier performed better than LSTM and ANN classifiers in terms of speech recognition accuracy.
- The proposed fusion models outperformed the state-of-the-art models by an average accuracy difference of up to 4.84% and 16.26% for clean and noisy audio, respectively. Furthermore, in the case of missing modality, the proposed fusion models outperformed the state-of-the-art models by an average accuracy difference of up to 3.46% and 17.54% for audio-only availability and video-only availability, respectively.