Postgraduate Seminar Series

Venue: Graduate Seminar Room
Date & Time: August 09, 2025 at 3:00 PM

Speaker Information

Sadia Afrin Purba (Std No. 0422052056) is a part time M.Sc. student in the department of CSE, BUET. She completed her undergraduate studies from Ahsanullah University of Science and Technology (AUST) in 2019. Her research interest lies in the field(s) of Quantum Computing, Machine Learning, and Deep Learning. She is currently doing her postgraduate thesis under the supervision of Dr. Md. Monirul Islam. She will be speaking about her ongoing research in this talk.

Multimodal Sentiment Analysis and Emotion Recognition in Conversation Using Contrastive Learning

The explosive growth of online video and social media platforms has fundamentally transformed the nature of human communication, resulting in an abundance of multimodal data—spanning text, audio, and visual streams. Traditional sentiment analysis methods, focused primarily on textual content, have proven insufficient for understanding the nuanced emotional dynamics present in contemporary conversational media. To address these challenges, we propose a novel framework for multimodal sentiment analysis and emotion recognition: the Contrastive Multimodal Variational Autoencoder (CM-VAE). Our approach systematically fuses audio, visual, and textual features, leveraging state-of-the-art representation learning methods including Contrastive Language-Image Pre-training (CLIP) for visual and textual data and Wav2Vec (a framework for self-supervised learning of speech representations) for audio. We integrate these modalities using a unified fusion architecture, and employ a variational autoencoder (VAE) to learn compact, probabilistic latent representations. Crucially, we enhance this framework with a contrastive learning mechanism based on InfoNCE loss, which explicitly encourages the separation of latent representations according to their semantic content. This enables the model to robustly disentangle sentiment and emotion cues, even in the presence of noise or ambiguous input. We validate our CM-VAE framework through extensive experiments on five challenging benchmark datasets, covering both English and multilingual conversational scenarios. Our results demonstrate that the proposed method consistently outperforms traditional early fusion and existing state-of-the-art baselines, particularly when handling cross-modal, asynchronous, or noisy data. The contrastive learning component proves critical for achieving discriminative and structured latent spaces, leading to significant improvements in both sentiment analysis and emotion recognition tasks. Our findings highlight the value of jointly optimizing generative and discriminative objectives within a multimodal setting and establish CM-VAE as a robust, generalizable solution for nuanced affective computing in complex, real-world conversational environments.