MIDI 2025 Conference

Explaining Complex Emotion Recognition in Social Robots: Integrating Explainable AI Methods into Region-Based Convolutional Networks

Abstract

The paper investigates the integration of explainable artificial intelligence (XAI) methods into region-based convolutional neural networks (R-CNNs) for complex emotion recognition in social robots.

While deep learning models such as Faster R-CNN and Mask R-CNN achieve high accuracy in detecting affective cues, their opacity limits transparency and user trust—critical aspects for socially interactive robots. To address this, four state-of-the-art XAI techniques—Grad-CAM++, Integrated Gradients, Layer-wise Relevance Propagation (LRP), and DeepSHAP—were applied to visualize and interpret emotional inference processes within three R-CNN architectures. Experiments conducted on the EMOTIC dataset demonstrated that Grad-CAM++ provided the most stable and interpretable visual explanations, while Integrated Gradients captured contextual dependencies essential for recognizing composite emotions such as sympathy and peace. A prototype robot-level interpretation module translated visual attributions into communicative outputs, enabling explanations such as gaze shifts and verbal justifications.

The findings indicate that combining XAI with region-based detectors enhances both the technical interpretability and social transparency of emotion-aware robots, forming a foundation for trust-aware and ethically responsible human–robot interaction.