Explainable Multi-class Classification Framework on Video using Interpretable Deep Learning Techniques

2410208 RENATI MULATI

1. Introduction and Background

Explainable video-based classification has become increasingly important in affective computing and other human-centric artificial intelligence applications. In these domains, models are expected not only to achieve high recognition accuracy but also to provide transparent and trustworthy decision processes. While deep learning has significantly advanced video classification through spatio-temporal architectures such as 3D convolutional networks and video transformers, these models typically operate as opaque "black boxes." Their lack of interpretability limits their applicability in real-world scenarios where understanding *why* a decision is made is as critical as the decision itself, particularly in high-stakes domains involving human behavior analysis and healthcare monitoring.

Current research in explainable artificial intelligence (XAI) typically relies on post-hoc saliency or gradient-based methods (e.g., Grad-CAM). However, extending these image-level techniques to video data introduces unique challenges. Video understanding requires handling temporal variation, motion-induced noise, and frame-level ambiguity. Existing post-hoc methods often suffer from "saliency flickering," where explanations become temporally inconsistent or semantically misaligned across consecutive frames. Furthermore, post-hoc approximations do not guarantee faithfulness to the model's actual reasoning process.

To address these limitations, this research advocates for *intrinsic interpretability*. We aim to construct a video classification framework that is transparent by design, systematically extending "glass-box" models—specifically prototype-based and dictionary-based networks—from the static image domain to dynamic video classification.

2. Proposed Methodology

This thesis proposes a unified framework that adapts *image-level interpretable deep learning models* to video-based multi-class classification. Distinct from video-native architectures that jointly model space and time in a black-box manner, our approach decomposes video understanding into two explicit stages:

(1) Frame-wise Interpretable Inference: The core of the framework is an inherently interpretable Image-Level Reasoner (ILR). We evaluate two representative paradigms within this component:

- **ProtoPNet (Prototype-based):** This model utilizes case-based reasoning. It learns a set of prototypical visual parts (e.g., specific shapes of eyes or mouths) in a latent space. During inference, the network calculates the similarity between input patches and these learned prototypes. This allows the model to generate explanations in the form of "this input looks like that prototype," effectively aligning the model's reasoning with human-understandable concepts.
- **PatchSAE (Dictionary-based):** Inspired by sparse autoencoders, this model learns a dictionary of visual primitives. It explains predictions through the activation of a sparse subset of latent atoms. Unlike ProtoPNet's semantic parts, PatchSAE captures distributed local textures and structural patterns, offering a complementary perspective on interpretability.

(2) Temporal Aggregation: To synthesize frame-level decisions into a coherent video-level prediction, we introduce a lightweight temporal aggregation mechanism. specifically, an *Exponentially Weighted Moving Average (EWMA)* is employed. Unlike complex recurrent neural networks (RNNs) that might re-introduce opacity, EWMA acts as a transparent low-pass filter. It stabilizes predictions over time, smoothing out frame-level jitter and transient noise while retaining sensitivity to the overall emotional evolution of the video sequence.

3. Experimental Setup

To systematically evaluate the proposed framework, extensive experiments were conducted on the **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, a benchmark for high-quality audio-visual emotion analysis.

A key contribution of this thesis is the formulation of the classification task under three distinct label configurations to analyze the impact of semantic granularity and visual ambiguity:

- **Original (6-class):** The standard taxonomy including Neutral, Calm, Happy, Sad, Angry, and Fearful.
- **V1 (3-class, Semantic Valence):** A clean grouping based on emotional valence (Positive vs. Negative).

- **V2 (3-class, Visual Interference):** A challenging robustness test where the *Sad* class is merged into the *Neutral* group. Since "Sad" expressions share low visual intensity (arousal) with "Neutral" faces but have negative valence, this setting creates a semantic conflict designed to test whether the model relies on high-level semantics or low-level visual cues.

Furthermore, preprocessing was treated as a first-class variable. We compared "Original frames" (with background) against "Face-only" representations. This comparison is crucial to verify that the interpretable prototypes are grounding their decisions in physiologically meaningful facial dynamics rather than spurious background correlations.

4. Results and Analysis

Experimental results demonstrate that the proposed framework successfully bridges the gap between image-level interpretability and video classification.

Quantitative Performance: While standard black-box models (e.g., VGG-16) achieved high accuracy in simpler settings, the interpretable models showed remarkable robustness in complex scenarios. In the V1 setting, ProtoPNet achieved a video-level accuracy of 64.8%. However, in the more challenging V2 (Visual Interference) setting, ProtoPNet's performance actually improved to **74.3%**. This counter-intuitive result is significant: it suggests that prototype-based reasoning aligns strongly with visual intensity (arousal) cues. By grouping low-intensity "Sad" faces with "Neutral," the model could leverage the visual similarity effectively, whereas baseline models often struggled with the semantic ambiguity. Conversely, PatchSAE generally underperformed ProtoPNet in facial analysis (57.2% in V1), indicating that semantic part-matching is more effective for faces than sparse texture encoding.

Qualitative Explainability: Visual analysis of the reasoning processes reveals a fundamental trade-off. ProtoPNet produced stable, intuitive explanations that consistently highlighted distinct facial organs (e.g., the widening of eyes or the curvature of the mouth). The prototypes acted as visual anchors, maintaining consistency across the video timeline. On the other hand, PatchSAE emphasized fine-grained local textures (e.g., skin wrinkles on the forehead). While less semantically intuitive for facial expressions, supplementary experiments on the **DIEM-A** (body skeleton) and **WLASL-10** (sign language) datasets revealed that PatchSAE outperforms ProtoPNet in domains relying on sparse, texture-less geometric data.

5. Conclusion

In conclusion, this thesis validates the feasibility of a unified, transparent pipeline for video-based classification. We demonstrate that by combining inherently interpretable frame-reasoners (ProtoPNet) with transparent temporal smoothing (EWMA), it is possible to achieve competitive performance without relying on black-box architectures.

The study establishes that prototype-based models are particularly well-suited for domains rich in semantic parts (like faces), providing explanations that align with human intuition. Meanwhile, sparse dictionary models offer greater flexibility for abstract or texture-based data. This framework provides a robust baseline for future research in trusted AI, proving that interpretability does not require a complete sacrifice of performance, even in dynamic video environments.

Keywords: Explainable AI (XAI), Facial Expression Recognition, ProtoPNet, Sparse Autoencoders, Video Classification