Learning Emotions and Mental States in Movie Scenes

Thesis submitted in partial fulfilment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

Dhruv Srivastava
2021701021
dhruv.srivastava@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2024

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Learning Emotions and Mental States in Movie Scenes" by Dhruv Srivastava, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Prof. Makarand Tapaswi

To my Friends and Family

Acknowledgments

I want to extend my sincere thanks to all those who accompanied me on my Research journey over the past two years. This period has been an amazing experience, brimming with priceless knowledge, personal development, and cherished memories.

Above all, I want to express my profound thanks and gratitude for Prof. Makarand Tapaswi. His support played a pivotal role in the realization of my ambitions, enabling me to reach heights I never thought possible. Thank you, for all your time and patience, for clearing all my trivial doubts and patiently listening to all I had to say. Not only in academics, but your guidance and personality have facilitated my personal growth.

Big shoutout to my amazing squad – Aditya, Amruth, Bhoomeendra, Dhaval, Prateek and Nayan! These awesome folks turned the academic rollercoaster into a joyride. From the *special* late night lectures in felicity ground to sleepless research submissions, we did it all together. They aren't just friends; they're my academic family. We laughed through the highs and lows, celebrated victories, and turned stress into a comedy show. Thanks for being the sunshine in my research storm, and the reason my research journey is a tale worth telling. You guys are the real heroes of this academic saga!

Heartfelt gratitude to my rock-solid family – Charchika, Mehul, Mom and Dad, you were my cheering squad, my emotional rescue team, and the reason I kept pushing forward. Your unwavering support and endless encouragement were the compass guiding me through uncharted research territories. Thank you for being my pillars of strength.

Abstract

In this thesis, we delve into the analysis of movie narratives, with a specific focus on understanding the emotions and mental states of characters within a scene. Our approach involves predicting a diverse range of emotions for individual movie scenes and each character within those scenes. To achieve this, we introduce EmoTx, a novel multimodal Transformer-based architecture that integrates video data, multiple characters, and dialogues for making comprehensive predictions.

Leveraging annotations from the MovieGraphs dataset, our model is tailored to predict both classic emotions (e.g., happiness, anger) and nuanced mental states (e.g., honesty, helpfulness). Our experiments concentrate on evaluating performance across the ten most common and twenty-five most common emotional labels, along with a mapping that clusters 181 labels into 26 categories. Through systematic ablation studies and a comparative analysis against established emotion recognition methods, we demonstrate the effectiveness of EmoTx in capturing the intricacies of emotional and mental states in movie contexts.

Additionally, our investigation into EmoTx's attention mechanisms provides valuable insights. We observe that when characters express strong emotions, EmoTx focuses on character-related elements, while for other mental states, it relies more on video and dialogue cues. This nuanced understanding enhances the interpretability and contextual relevance of EmoTx in the domain of movie story analysis. The findings presented in this thesis contribute to advancing our comprehension of character emotions and mental states in cinematic narratives.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Movies are a powerful medium for storytelling, engaging audiences through a rich collection of characters and emotions. Understanding the intricacies of characters' emotions and mental states within movie narratives is essential for unraveling the layers of storytelling and enhancing our grasp of cinematic experiences. This thesis embarks on a journey into the realm of movie story analysis, specifically focusing on the nuanced task of character emotion and mental state prediction.

The ability to decipher the emotional nuances of characters in movies extends beyond mere entertainment; it opens avenues for exploring the psychological depths of storytelling, character development, and audience engagement. This research takes a significant step forward by formulating emotion understanding as a predictive task, not only at the level of entire movie scenes but also at the individual character level. Our primary focus is on predicting a diverse and multi-label set of emotions, ranging from classic emotions like happiness and anger to more intricate mental states such as honesty and helpfulness.

To address this challenge, we introduce EmoTx [2], a novel multimodal Transformer-based architecture that harnesses the combined power of video data, multiple characters, and dialogues. This architecture facilitates joint predictions for scene and character emotions and mental states, allowing for a holistic understanding of the emotional landscape within movie scenes. Leveraging annotations from the MovieGraphs dataset [3], our model is trained to capture the spectrum of emotions and mental states, providing a comprehensive and granular analysis.

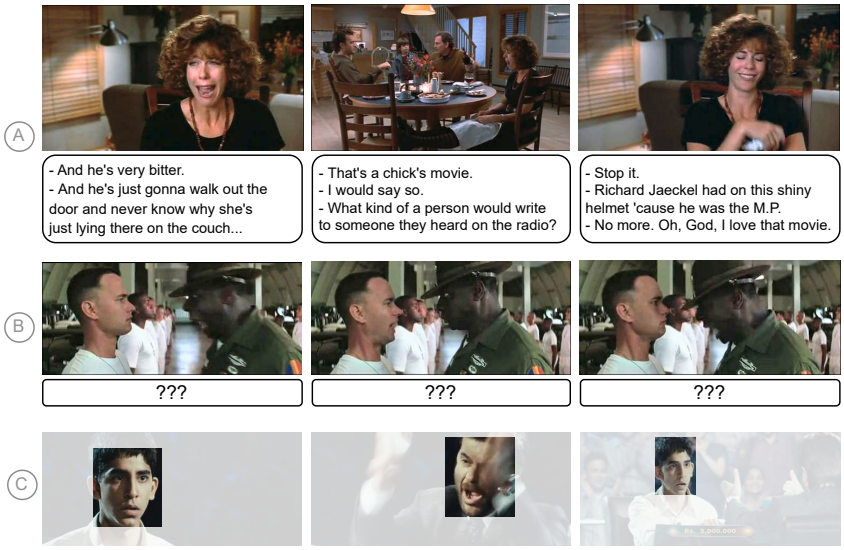

Figure 1.1: Multimodal models and multi-label emotions are necessary for understanding the story. **A**: What character emotions can we sense in this scene? Is a single label enough? **B**: Without the dialog, can we try to guess the emotions of the Sergeant and the Soldier. **C**: Is it possible to infer the emotions from the characters' facial expressions (without subtitles and visual background) only? Check the footnote below for the ground-truth emotion labels for these scenes.

1.1 Motivation

Emotions are a deeply-studied topic. From ancient Rome and Cicero's 4-way classification [4], to modern brain research [5], emotions have fascinated humanity. Psychologists use of Plutchik's wheel [6] or the proposal of universality in facial expressions by Ekman [7], structure has been provided to this field through various theories. Affective emotions are also grouped into mental (affective, behavioral, and cognitive) or bodily states [8].

A recent work on recognizing emotions with visual context, Emotic [9] identifies 26 label clusters and proposes a *multi-label* setup wherein an image may exhibit multiple emotions (*e.g. peace, engagement*). An alternative to the categorical space, valence, arousal, and dominance are also used as three continuous dimensions [9].

Predicting a rich set of emotions requires analyzing multiple contextual modalities [9, 10, 11]. Popular directions in multimodal emotion recognition are Emotion Recognition in Conversations (ERC)

Ground-truth emotions and mental states portrayed in Fig. 1.1: **A**: excited, curious, confused, annoyed, alarmed; **B**: shocked, confident (sergeant praises the soldier with agressive tone); **C**: happy, excited, amused, shocked, confident, nervous.

that classifies the emotion for every dialog utterance [12, 13, 14]; or predicting a single valence-activity score for short ∼10s movie clips [15, 16].

Classification in a rich label space of emotions requires looking at multimodal context as evident from masking context in Fig. 1.1. To this end, we propose EmoTx that jointly models video frames, dialog utterances, and character appearance.

EmoTx operates at the level of a *movie scene*: a set of shots telling a sub-story, typically at one location, among a defined cast, and in a short time span of 30-60s. Thus, scenes are considerably longer than single dialogs [12] or movie clips in [15]. EmoTx predict emotions and mental states for all characters in the scene and also by accumulating labels at the scene level. Estimation on a larger time window naturally lends itself to multi-label classification as characters may portray multiple emotions simultaneously (*e.g*. *curious* and *confused*) or have transitions due to interactions with other characters (*e.g*. *worried* to *calm*).

1.2 Movie understanding

In recent years, the field of movie understanding has undergone significant transformations, progressing beyond conventional tasks like clustering individuals and identifying characters [17, 18, 19, 20, 21, 22] to delve into the intricate analysis of storytelling. Many exciting areas have emerged, each contributing to a more comprehensive understanding of the cinematic experience.

The scope of movie understanding encompasses various dimensions, ranging from scene detection [23, 24, 25, 26, 27] and question-answering to tasks [28, 29, 30] like movie captioning [31, 32] with named entities [33], modeling interactions and/or relationships [34, 35, 36], aligning text and video storylines [37, 38, 39], and even tackling the complexities of long-form video understanding [40]. These diverse areas collectively strive to unravel the richness embedded within cinematic narratives.

The advancements in this field owe much to the availability of robust datasets that have fueled research and innovation. Datasets like Condensed Movies [41], MovieNet [42], VALUE benchmark (which extends beyond traditional movies) [43], and MovieGraphs [3] have played pivotal roles in propelling research forward. These datasets provide researchers with the necessary tools to explore and push the boundaries of what can be achieved in the realm of movie understanding.

Building on the foundations laid by MovieGraphs [3], we focus on another pillar of story understanding complementary to the above directions: identifying the emotions and mental states of each character and the overall scene in a movie.

1.3 Visual emotion recognition

The domain of visual emotion recognition has evolved significantly, initially rooted in the identification of Ekman's classic six emotions [7] predominantly through facial expressions. This foundation gained traction with influential datasets like MMI [44], CK, and CK+ [45, 46], marking a paradigm shift in our understanding of emotional cues in images.

Around a decade ago, benchmark challenges like EmotiW [47], FER [48], and AFEW [49] emerged as crucial benchmarks for in-the-wild emotion recognition. Simultaneously, deep learning approaches [50, 51] were introduced to emotion recognition, showcasing notable performance improvements. These benchmarks still focused on classic emotion label sets. Breaking away from this trend, the Emotic dataset [9] introduced the use of 26 labels for emotion understanding in images while highlighting the importance of context for emotion understanding in images.

Moving beyond isolated facial features, the field explored novel directions, including the combination of face features and contextual information. Two-stream Convolutional Neural Networks (CNNs) [10] and methods incorporating person detections with depth maps [11] gained attention for their potential in capturing nuanced emotional states. Recognizing the importance of context, particularly in dynamic scenarios, became a pivotal focus in enhancing the robustness of emotion recognition systems.

Recent trends in emotion recognition have expanded the scope beyond discrete labels to include estimating continuous variables such as valence and arousal. Approaches have emerged to predict emotional states from faces with limited contextual information, and researchers have explored learning representations through webly supervised data to overcome biases [52] inherent in existing datasets or improving them further through a joint text-vision embedding space [53].

In contrast to the prevailing trends in visual emotion recognition, our work EmoTx [2] takes a distinctive approach by concentrating on multi-label emotions and mental states recognition in movies. Exploiting the rich multimodal context available in cinematic scenes and character interactions, our focus extends beyond single facial expressions to capture the complexity and diversity of emotional experiences in a cinematic context.

1.4 Multimodal datasets for emotion recognition

In the landscape of multimodal emotion datasets, several initiatives have aimed to capture the intricacies of human emotions across different contexts. In this section, we delve into key datasets relevant to our research focus, highlighting their distinctive features.

Acted Facial Expressions in the Wild (AFEW) [49]: AFEW is designed to predict emotions solely from facial expressions, emphasizing the spontaneous nature of emotional reactions. However, it lacks contextual information, providing a limited scope for understanding emotions within the broader context of a narrative.

Stanford Emotional Narratives Dataset [54]: This dataset captures participant-shared narratives of positive and negative events in their lives. While multimodal in nature, incorporating both textual and visual elements, the narratives differ substantially from the edited and scripted content of movies and stories, which constitutes the primary focus of our research.

Multimodal EmotionLines Dataset (MELD) [12]: MELD is an example of Emotion Recognition in Conversations (ERC), concentrating on estimating emotions for individual dialog utterances in TV episodes from the show *Friends*. Differing from MELD, our research operates at the time-scale of cohesive story units, specifically movie scenes, allowing for a more comprehensive understanding of emotions within a narrative context.

Annotated Creative Commons Emotional DatabasE (LIRIS-ACCEDE) [15]: LIRIS-ACCEDE provides emotion annotations for short movie clips, making it closely aligned with our focus on cinematic content. However, the clips in LIRIS-ACCEDE are relatively small (8-12 seconds), and annotations are obtained in the continuous valence-arousal space, offering a different perspective compared to our multi-label approach that includes both classic emotions and mental states.

MovieGraphs dataset [3] MovieGraphs dataset features 51 movies and 7637 movie scenes with detailed graph annotations. Like other annotations in the MovieGraphs dataset, emotions are also obtained as free-text leading to a huge variability and a long-tail of labels (over 500). It also includes annotations such as the situation label, or character interactions and relationships [36].

In essence, the exploration of these multimodal emotion datasets illuminates the varied approaches undertaken in understanding emotional expressions.

1.5 Multimodal emotion recognition methods

The realm of multimodal emotion recognition has witnessed the application of various methodologies, each attempting to harness the synergies between audio, visual, and textual data for a holistic understanding of emotional expressions. In this section, we provide an overview of prevalent methods, drawing inspiration from early techniques to recent advancements.

Recurrent Neural Networks (RNNs) and Graph Networks: RNNs have played a pivotal role in Emotion Recognition in Conversations (ERC) [55, 13, 56, 57], particularly when coupled with graph networks [58, 59]. This combination has proven effective in amalgamating information from audio, visual, and textual modalities. The sequential nature of RNNs facilitates capturing temporal dependencies in emotional expressions, allowing for nuanced understanding.

Transformational Leap with Transformers: Inspired by the remarkable success of Transformer architectures in various natural language processing tasks, they have found adoption in the domain of ERC [60, 61]. Transformers offer advantages in modeling long-range dependencies and capturing contextual information effectively. Recent approaches leverage these architectures to enhance the performance of multimodal emotion recognition systems.

External Knowledge Graphs and Topic Modeling: To imbue systems with commonsense knowledge, some methods incorporate external knowledge graphs [62]. Additionally, the integration of topic modeling with Transformers has shown promise in improving the accuracy of emotion recognition models [14]. These techniques go beyond raw data and tap into external knowledge sources for a more nuanced understanding of emotional expressions.

Challenges in Multi-Label Prediction: Efforts have been made to extend multimodal emotion recognition to the realm of multi-label prediction by considering a sequence-to-set approach [63]. However, this approach, often employed for multi-label scenarios, may face scalability challenges with an increasing number of labels. This scalability concern underscores the need for innovative approaches to handle a diverse and expansive set of emotional and mental state labels.

Our method- EmoTx [2]: In our research, we adopt a Transformer-based architecture for joint modeling, aligning with the latest trends in multimodal emotion recognition. However, our focus diverges from traditional ERC, as we aim to predict emotions and mental states specifically within the context of movie scenes and characters. We adapt and compare our approach against some of the methods mentioned above in our experiments, evaluating their efficacy in the unique context of movie emotion

and mental state prediction. By leveraging insights from established techniques, we aim to contribute to the evolving landscape of multimodal emotion recognition, with a particular emphasis on cinematic storytelling.

Related Approaches in Movie Understanding: Close to our work, the MovieGraphs dataset [3] has been utilized for emotion annotations, focusing on tracking changing emotions across entire movies and proposing methods for Temporal Emotion Localization. However, it's noteworthy that the former typically tracks a single emotion in each scene, while the latter introduces a distinct direction by exploring the temporal dynamics of emotions in movies.

1.6 Multi-label Emotion and Mental State Recognition

We assume that movies have been segmented automatically [23] or with a human-in-the-loop process [27, 3] into coherent *scenes* that are self-contained and describe a short part of the story.

Consider such a movie scene $\mathcal{S}$ that consists of a set of video frames $\mathcal{V}$, characters $\mathcal{C}$, and dialog utterances $\mathcal{U}$. Let us denote the set of video frames as $\mathcal{V} = \{f_t\}_{t=1}^{T}$, where T is the number of frames after sub-sampling. Multiple characters often appear in any movie scene. We model N characters in the scene as $\mathcal{C} = \{\mathcal{P}^i\}_{i=1}^{N}$, where each character $\mathcal{P}^i = \{(f_t, b_t^i)\}$ may appear in some frame f_t of the video at the spatial bounding box b_t^i. We assume that b_t^i is empty if the character $\mathcal{P}^i$ does not appear at time t. Finally, $\mathcal{U} = \{u_j\}_{j=1}^{M}$ captures the dialog utterances in the scene. For this work, we use dialogs directly from subtitles and thus assume that they are unnamed. While dialogs may be named through subtitle-transcript alignment [17], scripts are not always available or reliable for movies.

Task formulation. Given a movie scene $\mathcal{S}$ with its video, character, and dialog utterance, we wish to predict the emotions *and* mental states (referred as labels, or simply emotions) at both the scene, $\mathbf{y}^{\mathcal{V}}$, and per-character, $\mathbf{y}^{\mathcal{P}^i}$, level. We formulate this as a multi-label classification problem with K labels, *i.e.* $\mathbf{y} = \{y_k\}_{k=1}^{K}$. Each $y_k \in \{0, 1\}$ indicates the absence or presence of the k^{th} label in the scene $y_k^{\mathcal{V}}$ or portrayed by some character $y_k^{\mathcal{P}^i}$. For datasets with character-level annotations, scene-level labels are obtained through a simple logical `OR` operation, *i.e.* $\mathbf{y}^{\mathcal{V}} = \bigoplus_{i=1}^{N} \mathbf{y}^{\mathcal{P}^i}$.

Chapter 2

Extending face tracks in MovieGraphs dataset

The MovieGraphs dataset [3] stands as a valuable dataset for understanding character emotions in movies, due to its rich annotations. However, upon close examination, it became apparent that the face tracks within the dataset faced challenges due to the limitations of face detection quality. Many instances revealed a frequent occurrence of missed character detections, leading to fragmented face tracks within a clip. Moreover, multiple track IDs were often assigned to the same character within a single shot, introducing complexities in tracking consistency. Furthermore, some shots lacked any face detections altogether, yet held untapped potential in offering a broader perspective on character emotions within a given scene.

Motivated by these observations and recognizing the need for enhanced character tracking precision, this chapter embarks on the task of extending face tracks within the MovieGraphs dataset [3]. Our approach involves addressing the gaps in the ground-truth tracks within a shot, mitigating issues related to missed detections and multiple track IDs. Additionally, we aim to augment the dataset by incorporating shots with initially zero detections, unlocking valuable insights into character emotions in scenes that might have been overlooked.

The extension process unfolds in two key phases: first, we systematically recompute face detections and tracks for the movie scenes, aiming to improve the accuracy and continuity of character representation. Subsequently, a subset of the newly generated face tracks is assigned names based on their overlap with the original tracks present in the MovieGraphs dataset [3]. To further enrich our understanding, we employ hierarchical clustering techniques to group all detections within a clip, allowing us to assign names to previously unnamed tracks based on the clustering results.

In essence, this chapter seeks to remedy the limitations of the existing face tracks in the MovieGraphs dataset [3], laying the groundwork for a more robust and comprehensive analysis of character emotions

Figure 2.1: Example face detections. The original face tracks do not work for dark scenes or profile faces, while our new detections and tracks are able to find them. Scene-036 from Forrest Gump, 1994.

in cinematic narratives. By extending and refining face tracks, we aim to bridge gaps in character tracking consistency, providing a more accurate representation of the emotional journeys characters undertake throughout the unfolding scenes. Through these efforts, we strive to contribute to the nuanced exploration of emotions within the cinematic medium and offer an improved foundation for subsequent analyses in character-centric emotion recognition.

Fig. 2.1 shows an example where original tracks did not have a single detection (due to the dark scene) for a scene in the "Forrest Gump, 1994" movie, whereas our track-extension pipeline was able to correctly associate names to the unlabelled characters.

2.1 Face and person detection and tracking

In the pursuit of refining character representation within the MovieGraphs dataset [3], we initiate a comprehensive process of face and person detection across every movie scene. Leveraging advanced deep learning methodologies, we employ the Multi-Task Cascaded Convolutional Neural Networks (MTCNN) [1] for face detection and Cascade-RCNN pretrained on cast annotations from MovieNet [42] for person detection.

Figure 2.2 shows the false positive bounding boxes predicted by MTCNN [1] detector when given the entire frame. MTCNN [1] is sensitive to threshold and using using higher threshold does not guarantee perfect face detection.

Figure 2.2: False positive face detections by MTCNN [1] given the entire frame at once with threshold set to 0.95. The detector still predicts incorrect face bounding boxes.

Figure 2.3: Confined face detections within the person bounding boxes. This allows us to have a correspondence between the person and face detections and ensemble of both the models considerably reduces the false positives.

To mitigate this issue, we undertake a two-step process to seamlessly integrate person detections. First, we utilize the Cascade-RCNN to compute person bounding boxes within each movie scene. Subsequently, we extract face detections confined within these person bounding boxes, establishing a crucial mapping between face and person detections. Figure 2.3 shows the face detections confined within the person bounding boxes. This considerably resolves the issue of incorrect face detections. In the event of multiple faces coexisting within a single person bounding box, our methodology prioritizes accuracy by selecting the face with the highest detection probability. This ensures that the ensuing face detections maintain fidelity to the most salient facial representation within the confined space.

The resulting face and person bounding boxes are subjected to tracking mechanisms for continuous identity mapping across frames. To achieve this, we implement the Kalman-filter based Simple Online and Realtime Tracking (SORT) algorithm [64], enabling real-time tracking of these bounding boxes. Importantly, our approach establishes a direct mapping between face and person tracks, sharing the same track ID for seamless coordination between the two modalities. Figure 2.4 shows the consistent track id for multiple characters in different frames within the same shot of a movie scene. The tracking ids however gets updated at the shot boundaries. We move on to the next step to establish the character mapping across shots within the movie scene.

Figure 2.4: Character detections with consistent track IDs across frames for both the actors. Our detection and tracking pipeline results in a consistent and reliable character tracks within a shot of a movie scene.

2.2 Assigning character names to the new character tracks

In the process of assigning names to new character tracks, we begin with the existing tracks in the MovieGraphs dataset, which exclusively pertain to faces. These original tracks provide a foundation for understanding the characters emotions and mental states in the scenes. Now, when we identify new face tracks, we want to connect them to the existing dataset to leverage the character names already associated with those faces.

To achieve this, we compare each new face detection with the original tracks by assessing the degree of overlap between them. This overlap is quantified using a metric called Intersection over Union (IoU), a measure that helps us determine how much the new detection aligns with the original track. We've set a threshold of 0.7 for IoU, meaning that we consider a match only when the overlap is substantial, ensuring a reliable connection.

Once we identify these matches, the next step involves associating the names from the original track with the corresponding new track. This association is based on the idea that if a face in the new detection aligns significantly with a face in the original track, they likely represent the same character.

Now, given that there might be multiple names associated with faces in the original track, we employ a democratic approach—a majority vote. This means that if a new track is linked to several names from the original track, we choose the name that the majority of these associations agree upon. In simpler terms, it's like asking a group of people for their opinion and going with the name that most people think fits the best.

In summary, this method allows us to seamlessly transfer character names from the original MovieGraphs dataset to the new character tracks we've identified, enhancing the richness of our character analysis in the broader context of the movie scenes.

2.3 Face clustering and naming other tracks.

In our effort to extract dense face tracks from MovieGraphs dataset [3], we encountered a challenge: not all characters could be given names using the tricks discussed till now because some detections were missed in the dataset itself. To fix this, we introduced a clustering method to find names for the new characters we discovered.

Identity Features Extraction: Our process commences with the extraction of robust identity features using the individual face detections. To achieve this, we leverage an InceptionResNetV1 model [65] pretrained on the VGGFace2 [66] dataset. These features serve as distinctive representations for each face, capturing nuanced characteristics crucial for identity clustering.

Clustering with C1C [21] Algorithm: For clustering, we employ the C1C [21] algorithm, which incorporates track information to establish must and cannot links between the extracted face features. This integration of track information ensures that the clustering process is sensitive to the temporal continuity of faces, enhancing the accuracy of identity assignments.

Silhouette Score and Representative Partition: The C1C [21] algorithm generates multiple partitions, each containing varying numbers of clusters. To determine the most representative partition, we calculate the Silhouette score for each partition. The partition with the highest Silhouette score is selected, signifying optimal cluster cohesion and separation.

Probability Assignment to Clusters: Now, based on the character names assigned through the earlier method, each cluster is endowed with a probability distribution corresponding to the distinct names found within the cluster. In cases where a cluster lacks any named detection, an equal probability is distributed across all names present in the scene.

Name Probability Thresholding: The cluster name-probabilities corresponding to the detections of unnamed tracks are then extracted. To consolidate these probabilities, we calculate the average of these soft scores, reflecting the likelihood of each name for the newly discovered tracks. This process yields name probabilities for the extended tracks.

In the final step, a threshold of 0.7 is applied to these name probabilities to select the definitive name for the new tracks. This ensures that only names with a high confidence level are assigned to the extended character tracks.

With this method, we're making sure that even the characters we missed at first get names, helping us better understand and analyze emotions in the extended MovieGraphs [3] face tracks.

Chapter 3

EmoTx: Our Approach

In this chapter, we introduce EmoTx, a sophisticated method designed to jointly predict multi-label emotions and mental states for both movie scenes and individual characters within the scene. EmoTx utilizes a Transformer-based architecture to achieve accurate and comprehensive emotion recognition.

The process begins with video pre-processing and feature extraction pipeline. This initial phase is crucial for distilling relevant representations from the complex visual and auditory information embedded in movie scenes. By carefully extracting key features, EmoTx sets the stage for a more nuanced understanding of the emotional dynamics at play.

The core of EmoTx lies in its Transformer encoder, that facilitates the seamless integration of information across various modalities. This encoder enables the model to capture intricate relationships and dependencies within the feature representations essential to capture the emotions. The Transformer architecture is particularly adept at handling sequential and contextual information, making it well-suited for the complex nature of emotional expression in movie scenes.

Building on these integrated representations, EmoTx incorporates a classification module, drawing inspiration from prior advancements in multi-label classification with Transformers [67]. This module serves as the final computational layer, responsible for making predictions about emotions associated with each scene and character. By leveraging the Transformer's capacity for contextual understanding, EmoTx excels in discerning the intricate nuances of emotions, providing a robust framework for comprehensive emotion recognition in the realm of cinematic storytelling.

An overview of the approach is presented in Fig. 3.1.

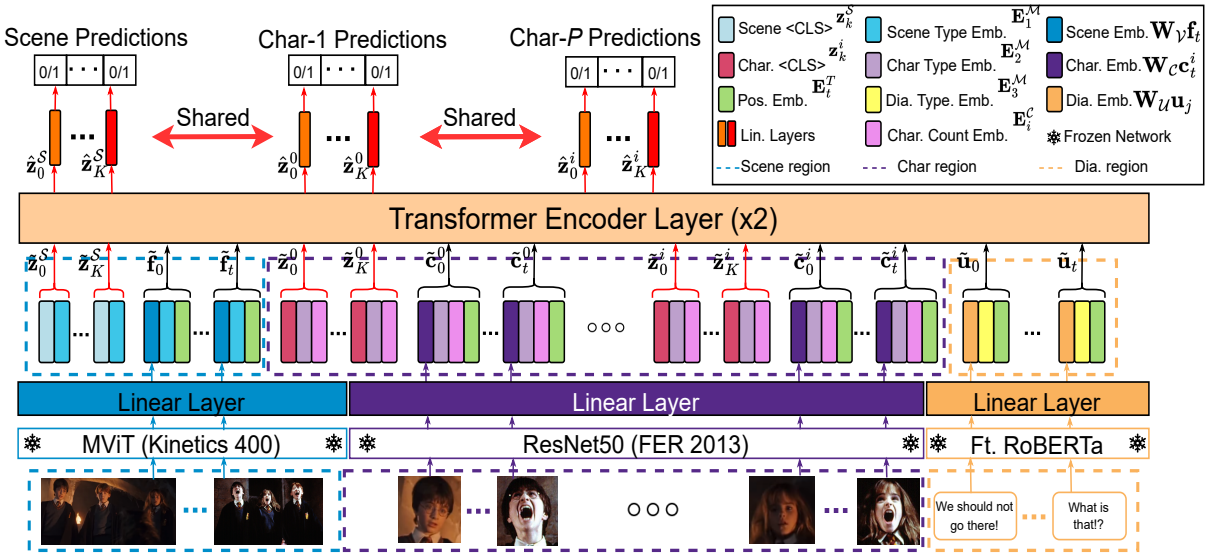

Figure 3.1: An overview of EmoTx. **A**: Video features (in blue region), character face features (in purple region), and utterance features (in orange region) are obtained using frozen backbones and projected with linear layers into a joint embedding space. **B**: Here appropriate embeddings are added to the tokens to distinguish between modalities, character count, and to provide a sense of time. We also create per-emotion classifier tokens associated with the scene or a specific character. **C**: Two Transformer encoder layers perform self-attention across the sequence of input tokens. **D**: Finally, we tap the classifier tokens to produce output probability scores for each emotion through a linear classifier shared across the scene and characters.

3.1 Preparing multimodal representations

Recognizing complex emotions and mental states (*e.g. nervous, determined*) requires going beyond facial expressions to understand the larger context of the story. To facilitate this, we encode multimodal information through multiple lenses:

(i) the video is encoded to capture where and what event is happening.

(ii) the character faces are encoded to represent their expressions; and

(iii) we encode the dialog utterances as information complementary to the visual domain.

A pretrained encoder $\phi_{\mathcal{V}}$ extracts relevant visual information from a single or multiple frames as $\mathbf{f}_t = \phi_{\mathcal{V}}(\{f_t\})$. Similarly, a pretrained language model $\phi_{\mathcal{U}}$ extracts dialog utterance representations as $\mathbf{u}_j = \phi_{\mathcal{U}}(u_j)$. Characters are more involved as we need to first localize them in the appropriate frames.

Given a valid bounding box b_t^i for person $\mathcal{P}^i$, we extract character features using a backbone pretrained for emotion recognition as $\mathbf{c}_t^i = \phi_{\mathcal{C}}(f_t, b_t^i)$.

Linear projection. Since the extracted features have different embeddings dimensions, we first bring all modalities to the same dimension with linear layers. Specifically, we project visual representation $\mathbf{f}_t \in \mathbb{R}^{D_{\mathcal{V}}}$ using $\mathbf{W}_{\mathcal{V}} \in \mathbb{R}^{D \times D_{\mathcal{V}}}$, utterance representation $\mathbf{u}_j \in \mathbb{R}^{D_{\mathcal{U}}}$ using $\mathbf{W}_{\mathcal{U}} \in \mathbb{R}^{D \times D_{\mathcal{U}}}$, and character representation $\mathbf{c}_t^i \in \mathbb{R}^{D_{\mathcal{C}}}$ using $\mathbf{W}_{\mathcal{C}} \in \mathbb{R}^{D \times D_{\mathcal{C}}}$.

3.2 Additional embeddings

Token representations in a Transformer often combine the core information (*e.g.* visual representation) with meta information such as the timestamp through position embeddings (*e.g.* [68]). This section lists all the embeddings that are used in EmoTx.

Modality Embeddings: We learn three embedding vectors $\mathbf{E}^{\mathcal{M}} \in \mathbb{R}^{D \times 3}$ to capture the three modalities corresponding to (1) video, (2) characters, and (3) dialog utterances. We also assist the model in identifying tokens coming from characters by including a special character count embedding, $\mathbf{E}^{C} \in \mathbb{R}^{D \times N}$. Note that the modality and character embeddings do not encode any specific meaning or imposed order (*e.g.* higher to lower appearance time, names in alphabetical order) - we expect the model to use this only to distinguish one modality/character from the other.

Time embeddings: The number of tokens depend on the chosen frame-rate. To inform the model about relative temporal order across modalities, we adopt a discrete time binning strategy that translates real time (in seconds) to an index. Thus, video frame/segment and character box representations fed to the Transformer are associated with their relevant time bins. For an utterance u_j, binning is done based on its middle timestamp t_j. We denote the time embeddings as $\mathbf{E}^{T} \in \mathbb{R}^{D \times \lceil T^*/\tau \rceil}$, where T^* is the maximum scene duration and τ is the bin step. For convenience, $\mathbf{E}_t^{T}$ selects the embedding using a discretized index $\lceil t/\tau \rceil$.

Classifier tokens: Similar to the classic CLS tokens in Transformer models [69, 70] we use learnable classifier tokens to predict the emotions. Furthermore, inspired by Query2Label [67], we use K classifier tokens rather than tapping a single token to generate all outputs (see Fig. 3.1D). This allows capturing label co-occurrence within the Transformer layers improving performance. It also enables analysis of per-emotion attention scores providing insights into the model's workings. In particular, we

use K classifier tokens for scene-level predictions (denoted $\mathbf{z}_k^{\mathcal{S}}$) and $N \times K$ tokens for character-level predictions (denoted $\mathbf{z}_k^i$ for character $\mathcal{P}^i$, one for each character-emotion pair).

Token representations: Combining the features with relevant embeddings provides rich information to EmoTx. The token representations for each input group are as follows:

$$\text{scene cls. tokens: } \tilde{\mathbf{z}}_k^{\mathcal{S}} = \mathbf{z}_k^{\mathcal{S}} + \mathbf{E}_1^{\mathcal{M}}, \tag{3.1}$$

$$\text{char. cls. tokens: } \tilde{\mathbf{z}}_k^i = \mathbf{z}_k^i + \mathbf{E}_2^{\mathcal{M}} + \mathbf{E}_i^C, \tag{3.2}$$

$$\text{video: } \tilde{\mathbf{f}}_t = \mathbf{W}_{\mathcal{V}}\mathbf{f}_t + \mathbf{E}_1^{\mathcal{M}} + \mathbf{E}_t^T, \tag{3.3}$$

$$\text{character box: } \tilde{\mathbf{c}}_t^i = \mathbf{W}_{\mathcal{C}}\mathbf{c}_t^i + \mathbf{E}_2^{\mathcal{M}} + \mathbf{E}_i^C + \mathbf{E}_t^T, \tag{3.4}$$

$$\text{and utterance: } \tilde{\mathbf{u}}_j = \mathbf{W}_{\mathcal{U}}\mathbf{u}_j + \mathbf{E}_3^{\mathcal{M}} + \mathbf{E}_{t_j}^T\,. \tag{3.5}$$

Fig. 3.1B illustrates this addition of embedding vectors. We also perform LayerNorm [71] before feeding the tokens to the Transformer encoder layers, not shown for brevity.

3.3 Transformer Self-attention.

In the core architecture of EmoTx, the Transformer self-attention mechanism allows the fusion of information across visual, facial and language modalities for nuanced emotion recognition.

The process unfolds through the concatenation of all relevant embeddings to the feature vectors, representing different aspects of the input data, as they traverse through $H{=}2$ layers of the Transformer encoder [72]. These layers are instrumental in facilitating self-attention, a mechanism that allows the model to weigh and prioritize different parts of the input sequence based on their contextual relevance. The self-attention mechanism is particularly powerful in capturing intricate relationships and dependencies across various modalities, enabling a holistic understanding of the complex information present in movie scenes.

Within the context of emotion prediction, our focus narrows down to specific outputs generated by the Transformer encoder. These outputs correspond exclusively to the classification tokens, which are strategically chosen to encapsulate the most salient information for the task at hand. By selectively tapping into these outputs, EmoTx hones in on the essential elements relevant to predicting emotions, optimizing the model's efficiency and effectiveness in this specific domain.

$$[\hat{\mathbf{z}}_k^{\mathcal{S}}, \hat{\mathbf{z}}_k^i] = \mathsf{TransformerEncoder}\left(\tilde{\mathbf{z}}_k^{\mathcal{S}}, \tilde{\mathbf{f}}_t, \tilde{\mathbf{z}}_k^i, \tilde{\mathbf{c}}_t^i, \tilde{\mathbf{u}}_j\right)\,. \tag{3.6}$$

We jointly encode all tokens spanning $\{k\}_1^K$, $\{i\}_1^N$, $\{t\}_1^T$, and $\{j\}_1^M$.

Emotion labeling. The contextualized representations for the scene $\hat{\mathbf{z}}_k^{\mathcal{S}}$ and characters $\hat{\mathbf{z}}_k^i$ are sent to a shared linear layer $\mathbf{W}^E \in \mathbb{R}^{K \times D}$ for classification. Finally, the probability estimates through a sigmoid activation $\sigma(\cdot)$ are:

$$\hat{y}_k^{\mathcal{S}} = \sigma(\mathbf{W}_k^E \hat{\mathbf{z}}_k^{\mathcal{S}}) \;\text{ and }\; \hat{y}_k^i = \sigma(\mathbf{W}_k^E \hat{\mathbf{z}}_k^i),\; \forall k, i\,. \tag{3.7}$$

3.4 Training and Inference

Training. EmoTx undergoes end-to-end training with the *BinaryCrossEntropy* (BCE) loss. This comprehensive approach considers the entire model architecture for optimization. The primary objective is to equip EmoTx with the capability to seamlessly understand and predict emotions within the diverse context of movie scenes.

The BCE loss function serves as the guiding force during EmoTx's training. This loss function is well-suited for binary classification tasks, aligning with our goal of predicting emotions in a binary fashion. It quantifies the difference between predicted and ground truth labels, steering the model towards more accurate emotion predictions.

EmoTx encounters the challenge of class imbalance, where certain emotional labels may be overrepresented or underrepresented in the training data. To counteract this imbalance, we introduce weights (ω_k) for positive labels. These weights are determined based on the inverse of proportions, strategically assigning higher weights to underrepresented emotional classes. This thoughtful adjustment ensures that EmoTx pays due attention to all emotional categories, preventing bias towards frequently occurring labels and enhancing its sensitivity to the diversity of emotions present in movie scenes.

The scene and character prediction losses are combined as

$$\mathcal{L} = \sum_{k=1}^{K} \mathsf{BCE}(\omega_k, y_k^{\mathcal{V}}, \hat{y}_k^{\mathcal{S}}) + \sum_{i=1}^{N} \sum_{k=1}^{K} \mathsf{BCE}(\omega_k, y_k^{\mathcal{P}^i}, \hat{y}_k^i)\,. \tag{3.8}$$

Inference. At test time, we follow the procedure outlined in Sec. 3.1, 3.2, 3.3 and generate emotion label estimates for the entire scene and each character as indicated in Eq. 3.7.

Variations. As we will see empirically, our model is very versatile and well suited for adding/removing modalities or additional representations by adjusting the width of the Transformer (number of tokens). It can be easily modified to act as a unimodal architecture that applies only to video or dialog utterances by disregarding other modalities.

Chapter 4

Experiments and Discussion

We present our experimental setup in Sec. 4.1 before diving into the implementation details in Sec. 4.2. A series of ablation studies motivate the design choices of our model (Sec. 4.3) while we compare against the adapted versions of various SoTA models for emotion recognition in Sec. 4.3.4. Finally, we present some qualitative analysis and discuss how our model switches from facial expressions to video or dialog context depending on the label in Sec. 4.4.

4.1 Dataset and Setup

We use the MovieGraphs dataset [3] that features 51 movies and 7637 movie scenes with detailed graph annotations. We focus on the list of characters and their emotions and mental states, which naturally affords a multi-label setup. Other annotations such as the situation label, or character interactions and relationships [36] are ignored as they cannot be assumed to be available for a new movie.

Label sets. Like other annotations in the MovieGraphs dataset, emotions are also obtained as free-text leading to a huge variability and a long-tail of labels (over 500). We focus our experiments on three types of label sets: (i) *Top-10* considers the most frequently occurring 10 emotions; (ii) *Top-25* considers frequently occurring 25 labels; and (iii) *Emotic*, a mapping from 181 MovieGraphs emotions to 26 Emotic labels provided by [16]. In Fig. 4.2 we show the number of movie scenes that contain top-10 and 25 emotions.

Statistics. We first present row max-normalized co-occurrence matrices for the scene and characters (Fig. 4.1). It is interesting to note how a movie scene has high co-occurrence scores for emotions such as *worried* and *calm* (perhaps owing to multiple characters), while *worried* is most associated with *confused* for a single character. Another high scoring example for a single character is *curious* and

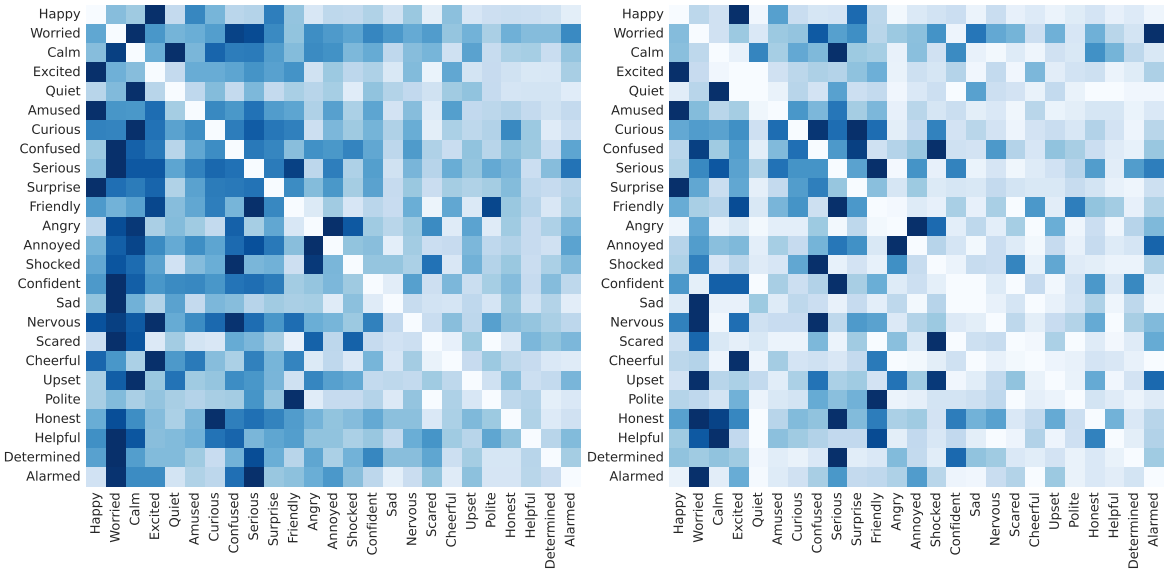

Figure 4.1: Row normalized label co-occurrence matrices for the top-25 emotions in a *movie scene* (left) or for a *character* (right).

surprise, while a movie scene has *curious* with *calm* and *surprise* with *happy*. In Fig. 4.3, we show the number of movie scenes that contain a specified number of emotions. Most scenes have 4 emotions.

Evaluation metric. We use the original splits from MovieGraphs. As we have K binary classification problems, we adopt mean Average Precision (mAP) to measure model performance (similar to Atomic Visual Actions [73]). Note that AP also depends on the label frequency.

4.2 Implementation Details

Feature representations play a major role on the performance of any model. We describe different backbones used to extract features for video frames, characters, and dialog.

Video features $\mathbf{f}_t$: The visual context is important for understanding emotions [9, 10, 11]. We extract spatial features using ResNet152 [74] trained on ImageNet [75], ResNet50 [74] trained on Place365 [76], and spatio-temporal features, MViT [77] trained on Kinetics400 [78].

Dialog features $\mathbf{u}_j$: Each utterance is passed through a RoBERTa-Base encoder [69] to obtain an utterance-level embedding. We also extract features from a RoBERTa model fine-tuned for the task of multi-label emotion classification (based on dialog only).

Character features $\mathbf{c}_t^i$: are represented based on face or person detections. We perform face detection with MTCNN [1] and person detection with Cascade RCNN [79] trained on MovieNet [42]. Tracks are

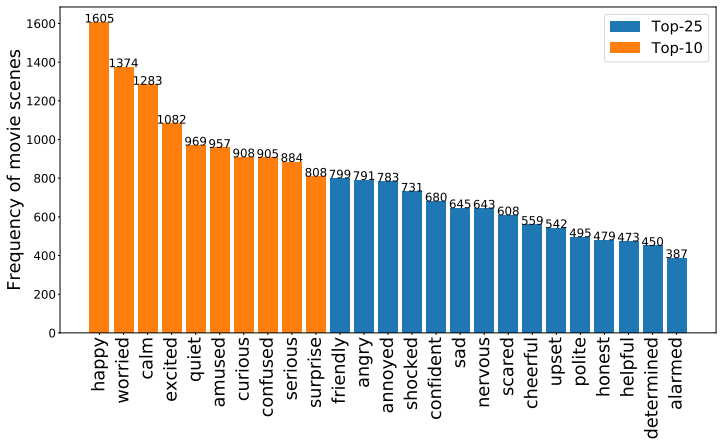

Figure 4.2: Number of movie scenes containing top-10 and 25 emotions. Note, the top-25 label set includes the top-10 emotions.

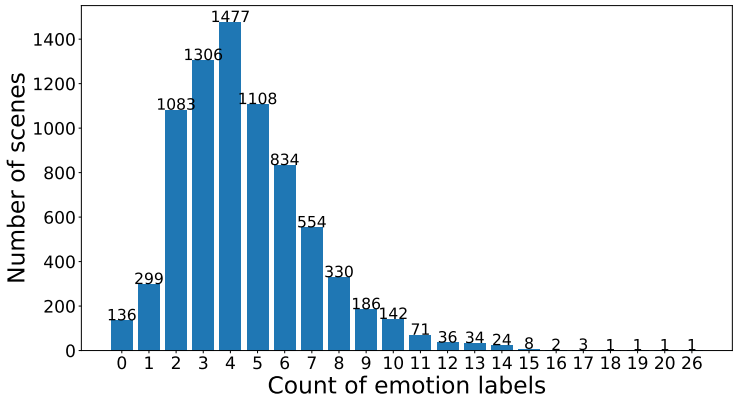

Figure 4.3: Bar chart showing the number of movie scenes associated with a specific count of annotated emotions.

obtained using SORT [64], a simple Kalman filter based algorithm, and clusters using C1C [21]. Details of the character processing pipeline are presented in Chapter 2. ResNet50 [80] trained on SFEW [81] and pretrained on FER13 [48] and VGGFace [82], VGGm [80] trained on FER13 and pretrained on VGGFace, and InceptionResnetV1 [65] trained on VGGFace2 [66] are used to extract face representations.

Frame sampling strategy: We sample up to T=300 tokens at 3 fps (100s) for the video modality. This covers $\sim$99% of all movie scenes. Our time embedding bins are also at 3 per second, *i.e.* τ=1/3s. During inference, a fixed set of frames are chosen, while during training, frames are randomly sampled from 3 fps intervals which acts as data augmentation. Character tokens are treated in a similar fashion, however are subject to the character appearing in the video.

Architecture details: We experiment with the number of encoder layers, $H \in \{1, 2, 4, 8\}$, but find H=2 to work best (perhaps due to the limited size of the dataset). Both the layers have same configuration - 8 attention heads with hidden dimension of 512. The maximum number of characters is N=4 as it covers up to 91% of the scenes. Tokens are padded to create batches and to accommodate shorter video clips. Appropriate masking prevents self-attention on padded tokens. Put together, EmoTx encoder looks at K scene classification tokens, T video tokens, $N \cdot (K + T)$ character tokens, and T utterance tokens. For K=25, N=4 (Top-25 label set), this is up to 1925 padded tokens.

Training details: EmoTx is implemented in PyTorch [83], a versatile deep learning framework. TheThe training process is conducted on a single NVIDIA GeForce RTX-2080 Ti GPU, optimizing computational efficiency. We set a maximum training duration of 50 epochs, each comprising a batch size of 8 samples. The hyperparameters are thoughtfully tuned to attain optimal performance on the validation set. For optimization, we employ the Adam optimizer [84], a robust choice for fine-tuning model parameters. The initial learning rate is set at 5×10^{-5}, establishing an effective starting point for the training process. To dynamically adjust the learning rate during training, we incorporate the `ReduceLROnPlateau` learning rate scheduler, a mechanism that reduces the learning rate by a factor of 10 when the model's performance plateaus. Throughout the training process, the model continually refines its understanding of emotional dynamics in movie scenes and character interactions. The effectiveness of the training is evaluated based on the geometric mean of scene and character mean Average Precision (mAP). The best checkpoint, representing the model's peak performance, is determined by maximizing this geometric mean. This approach ensures that EmoTx not only excels in capturing emotional nuances within individual scenes but also adeptly handles the complexities of character-level emotion prediction, contributing to its overall effectiveness in understanding emotions in movie scenes.

4.3 Ablation Studies

We perform ablations across three main dimensions: architectures, modalities, and feature backbones. When not mentioned, we adopt the defaults: (i) MViT trained on Kinetics400 dataset to represent video; (ii) ResNet50 trained on SFEW, FER, and VGGFace for character representations; (iii) fine-tuned RoBERTa for dialog utterance representations; and (iv) EmoTx with appropriate masking to pick modalities or change the number of classifier tokens.

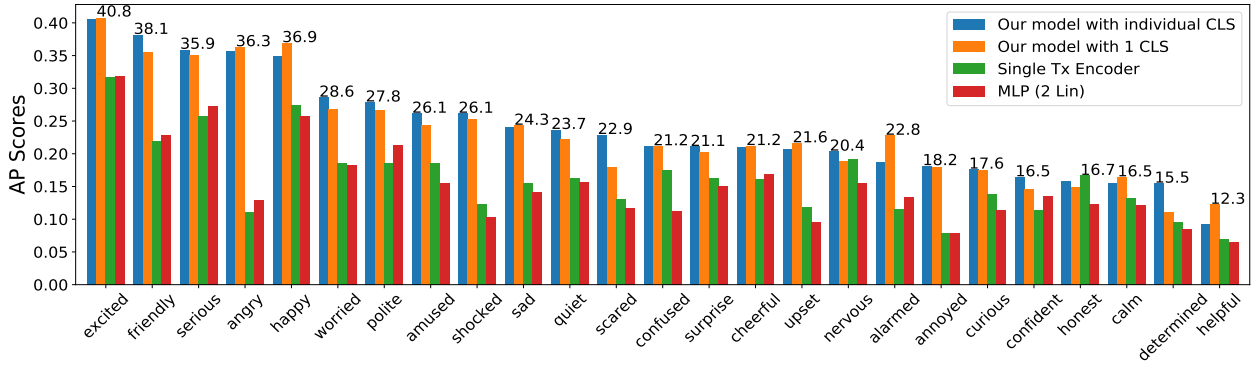

Figure 4.4: Comparing scene-level per class AP of EmoTx against baselines (Table 4.1) shows consistent improvements. We also see that our model with K classifier tokens outperforms the 1 CLS token on most classes. AP of the best model is indicated above the bar. Interestingly, the order in which emotions are presented is not the same as the frequency of occurrence (see 4.2).

4.3.1 Architecture ablations

We compare our architecture against simpler variants in Table 4.1. The first row sets the expectation by providing scores for a *random* baseline that samples label probabilities from a uniform random distribution between $[0, 1]$ with 100 trials. Next, we evaluate *MLP (2 Lin)*, a simple MLP with two linear layers with inputs as max pooled scene or character features. An alternative to max pooling is self-attention. The *Single Tx encoder* performs self-attention over features (as tokens) and a classifier token to which a multi-label classifier is attached. Both these approaches are significantly better than random, especially for individual character level predictions which are naturally more challenging than scene-level predictions. Finally, we compare multimodal EmoTx that uses 1 classifier token to predict all labels (EmoTx: 1 CLS) against K classifier tokens (last row). Both models achieve significant improvements, *e.g.* in absolute points, +8.5% for Top-10 scene labels and +2.3% for the much harder Top-25 character level labels. We believe the improvements reflect EmoTx's ability to encode multiple modalities in a meaningful way. Additionally, the variant with K classifier tokens (last row) shows small but consistent +0.5% improvements over 1 classifier token on Top-25 emotions.

Fig. 4.4 shows the scene-level AP scores for the Top-25 labels. Our model outperforms the MLP and Single Tx encoder on 24 of 25 labels and outperforms the single classifier token variant on 15 of 25 labels. EmoTx is good at recognizing expressive emotions such as *excited, serious, happy* and even mental states such as *friendly, polite, worried*. However, other mental states such as *determined* or *helpful* are challenging.

Method	Top-10		Top-25	
	Scene	Char	Scene	Char
Random	16.87±0.23	12.49±0.15	9.73±0.101	5.84±0.05
MLP (2 Lin)	23.94±0.03	20.39±0.01	15.26±0.02	10.57±0.02
Single Tx encoder	25.66±0.02	20.95±0.09	16.14±0.03	11.08±0.18
EmoTx: 1 CLS	*34.11*±0.34	*23.81*±0.24	*23.34*±0.11	12.86±0.11
EmoTx (Ours)	**34.22**±0.18	**24.35**±0.23	**23.86**±0.10	**13.36**±0.11

Table 4.1: Architecture ablation. Emotions are predicted at both movie scene and individual character (Char) levels. We see that our multimodal model significantly outperforms simpler baselines. Best numbers in bold, close second in italics.

4.3.2 Modality ablations

We evaluate the impact of each modality (video, characters, and utterances) on scene- and character-level emotion prediction in Table 4.2. We observe that the character modality (row 4, R4) outperforms any of the video or dialog modalities (R1-R3). Similarly, dialog features (R3) are better than video features (R1, R2), common in movie understanding tasks [3, 28].

Interistingly, we also observe that having additional modality does not always help and the choice of feature backbone is important to get the desired results. Scene features V_r, extracted from ResNet50 pretrained on Places365 dataset, which are more representative of the environment where the movie scenes may be happening, are consistently worse than action features V_m, which are extracted from a MViT_v1 model pretrained on Kinetics400 dataset and is expected to be more representative of the actions happening within the movie scenes. Comparing R1, R2 or R5, R6 or R8, R9 reflect that V_m assists model and works well with character and scene modalities for emotion recognition whereas V_r makes it difficult for model to desipher emotions.

Finally, our observations reveal that the utilization of all modalities (R9) yields superior performance compared to other combinations. This outcome strongly suggests that the task of emotion recognition is inherently multimodal.

	V_r	V_m	D	C	Top 10 (mAP)		Top 25 (mAP)	
					Scene	Char	Scene	Char
1	✓	-	-	-	22.81±0.02	15.90±0.19	14.85±0.02	7.98±0.05
2	-	✓	-	-	25.73±0.02	17.88±0.12	16.11±0.05	8.96±0.12
3	-	-	✓	-	27.28±0.01	20.25±0.14	20.20±0.08	11.09±0.12
4	-	-	-	✓	31.38±0.40	21.22±0.50	20.32±0.05	11.23±0.14
5	✓	-	✓	-	27.19±0.07	19.45±0.10	19.72±0.03	10.67±0.08
6	-	✓	✓	-	28.93±0.02	21.41±0.15	21.29±0.05	12.03±0.23
7	-	-	✓	✓	33.59±0.10	23.54±0.16	23.40±0.09	13.01±0.08
8	✓	-	✓	✓	33.60±0.02	22.89±0.02	22.76±0.02	12.21±0.02
9	-	✓	✓	✓	**34.22** ±0.18	**24.35**±0.23	**23.86**±0.10	**13.36**±0.11

Table 4.2: Modality ablation. V_r: ResNet50 (Places365), V_m: MViT (Kinetics400), D: Dialog, and C: Character.

4.3.3 Backbone ablations

We compare several backbones for the task of emotion recognition.

(i) MViT_V1 model [77] pre-trained on Kinetics400 [78] dataset, ResNet50 [74] pre-trained on Places365 dataset [76] and ResNet152 [74] pre-trained on ImageNet dataset [75] for video features.

(ii) ResNet50 [80] pre-trained on FER [48], SFEW [81] and VGGFace [82] datasets, VGG-m [80] pre-trained on FER13 [48] dataset and InceptionResNetV1 [65] pre-trained on VGGFace2 [66] dataset for character features; and

(iii) A pre-trained and finetuned version on RoBERTa for dialogue features.

The effectiveness of the fine-tuned RoBERTa model is evident by comparing pairs of rows R6, R12 and R5, R15 and R3, R13 of Table 4.3, where we see a consistent improvement of 1-3%. Character representations with ResNet50-FER show improvement over VGGm-FER as seen from R11, R18 or R15, R17. Finally, comparing R17, R18 shows the benefits provided by action features as compared to places.

In conclusion, ResNet50 trained on FER appears to be a good representation for characters, and the MViT trained on Kinetics400 provides better results for both the label sets, while ResNet50 trained on Places365 is a close second.

	Video			Character			Dialog		Metrics (mAP)			
	MViT	R50	R152	R50	VGG-M	IRv1	RB	RB	Top-10		Top-25	
	K400	P365	INet	FER	FER	VGG-F	FT	PT	Scene	Char	Scene	Char
1	-	✓	-	-	-	✓	-	✓	25.07±0.12	15.48±0.15	16.41±0.24	8.31±0.17
2	-	-	✓	-	-	✓	-	✓	25.85±0.24	15.63±0.21	16.45±0.09	8.31±0.09
3	-	-	✓	-	✓	-	-	✓	29.20±0.22	19.88±0.27	18.93±0.38	10.16±0.17
4	✓	-	-	-	-	✓	-	✓	29.27±0.08	18.07±0.22	18.35±0.09	0.09±0.08
5	-	✓	-	-	✓	-	-	✓	29.30±0.21	19.73±0.17	19.05±0.19	10.31±0.00
6	✓	-	-	-	✓	-	-	✓	29.34±0.08	20.50±0.04	19.07±0.19	10.34±0.17
7	-	✓	-	-	-	✓	✓	-	29.34±0.17	19.49±0.03	20.73±0.08	10.75±0.02
8	-	-	✓	-	-	✓	✓	-	29.47±0.14	19.29±0.10	20.74±0.11	10.79±0.07
9	-	✓	-	✓	-	-	-	✓	29.69±0.38	20.25±0.14	20.16±0.29	11.06±0.12
10	-	-	✓	✓	-	-	-	✓	30.19±0.38	20.27±0.26	19.83±0.07	11.06±0.16
11	✓	-	-	✓	-	-	-	✓	31.39±0.34	21.18±0.18	20.88±0.28	11.46±0.08
12	✓	-	-	-	✓	-	✓	-	31.50±0.36	21.60±0.09	21.49±0.30	11.64±0.20
13	-	-	✓	-	✓	-	✓	-	31.96±0.20	21.81±0.37	21.28±0.25	11.58±0.26
14	✓	-	-	-	-	✓	✓	-	32.23±0.07	21.45±0.07	22.10±0.11	11.63±0.06
15	-	✓	-	-	✓	-	✓	-	32.42±0.26	22.32±0.27	21.45±0.17	11.62±0.05
16	-	-	✓	✓	-	-	✓	-	33.44±0.33	22.89±0.24	22.75±0.18	12.52±0.12
17	-	✓	-	✓	-	-	✓	-	33.46±0.21	22.98±0.16	22.69±0.22	12.48±0.20
18	✓	-	-	✓	-	-	✓	-	**34.22**±0.18	**24.35**±0.23	**23.86**±0.10	**13.36**±0.11

Table 4.3: Extended feature ablations. The different feature backbones are (MViT, K400): MViT pretrained on Kinetics400, (R50, P365): ResNet50 on Places365, (R152, INet): ResNet152 on ImageNet, (R50, FER): ResNet50 on Facial Expression Recognition (FER), (VGG-M, FER): VGG-M on FER, (IRv1, VGG-F): InceptionResNet-v1 trained on VGG-Face dataset, (RB, FT): pretrained RoBERTa finetuned for emotion recognition and (RB, PT): pretrained RoBERTa. Best numbers are in bold.

Method	Top 10		Top 25		Emotic	
	Val	Test	Val	Test	Val	Test
Random	16.87±0.23	13.84±0.20	9.73±0.10	7.57±0.08	11.47±0.11	11.36±0.09
CAER [10]	18.35±0.10	15.38±0.13	11.84±0.07	9.49±0.08	13.91±0.06	12.68±0.02
ENet [53]	19.14±0.10	16.14±0.05	11.22±0.06	9.08±0.08	13.55±0.06	12.64±0.03
AANet [85]	21.55±0.18	17.55±0.16	12.55±0.15	10.20±0.13	14.71±0.19	13.37±0.20
M2Fnet [60]	24.55±0.39	19.10±0.06	16.02±0.14	13.05±0.31	18.27±0.16	16.76±0.20
EmoTx	**34.22**±0.18	**29.35**±0.18	**23.86**±0.10	**19.47**±0.10	**23.67**±0.03	**21.40**±0.03

Table 4.4: Comparison against SoTA for scene-level predictions. *AANet* denotes AttendAffectNet, while *ENet* refers to EmotionNet.

Method	Top 10		Top 25		Emotic	
	Val	Test	Val	Test	Val	Test
Random	12.49±0.15	11.37±0.14	5.84±0.05	5.36±0.05	6.40±0.05	6.32±0.05
AANet [85]	17.43±0.28	16.04±0.19	8.64±0.19	7.20±0.15	8.53±0.17	7.75±0.11
M2Fnet [60]	20.82±0.28	19.01±0.45	10.67±0.38	9.71±0.34	11.30±0.35	9.92±0.02
EmoTx (Ours)	**24.35**±0.23	**22.31**±0.11	**13.36**±0.11	**11.71**±0.05	**12.29**±0.08	**11.76**±0.10

Table 4.5: Comparison against SoTA for character-level predictions. *AANet* denotes AttendAffectNet.

4.3.4 SoTA Comparison

We compare our model against published works EmotionNet [53], CAER [10], AttendAffectNet [85], and M2Fnet [60] by adapting them for our tasks.

EmotionNet [53] employs a joint embedding training approach that aligns learned text embeddings, obtained through the word2vec model [86], with image embeddings extracted from a ResNet50 backbone. For our adaptation, we utilize the same backbones. Since we use video as input, the frame features are max-pooled to generate a consolidated representation. The embedding loss is applied, with emotion labels serving as keywords for joint embedding training. The ResNet50 is then fine-tuned for multilabel emotion recognition, where individual frame features undergo max-pooling before reaching the logits layer.

CAER (Context Aware Emotion Recognition) [10] is a deep Convolutional Network which consists of two stream encoding networks to separately extract the facial and context features which are fused using an adaptive fusion network. Detections from our extended face tracks are used as inputs for the face encoding stream and the full video frame with masked faces was used as input to context encoding stream. Since CAER is designed to extract emotions from images we adapt it to videos by applying max-pooling over the fused features from both the streams to generate a single representation for a video. This adapted model is trained to predict multiple scene-level emotions.

M2FNet [60] is a transformer based model originally developed for Emotion Recognition in Conversations (ERC) and features a fusion-attention mechanism to modulate the attention given to each utterance considering the audio and visual features. As this model is designed for utterance emotion recognition we apply a max-pooling operation over the final outputs of fusion attention module to generate a feature representation for all the utterances in a video. Since this model provides two strategies to consider visual features: one with the video frame and another that combines multiple faces in a frame, we use them to predict either scene- or character-level emotions separately.

AttendAffectNet [85] proposes two multi-modal self-attention based approaches for predicting emotions from movie clips. We adapted the proposed Feature AttendAffectNet model in our work. It leverages the transformer encoder block where every input token represents a different modality. These modality feature vectors are generated by average pooling over respective features. Following the proposed mechanism, a classification head was attached at the end of the model for predicting multi-label emotions. We adopt the same backbone representations, MViT [77] pre-trained on Kinetics400 [78] and ResNet50 pretrained on FER13 [48], for their work to extract scene and face features respectively.

Table 4.4 shows scene-level performance while the character-level performance is presented in Table 4.5. First, we note that the test set seems to be harder than val as also indicated by the random baseline, leading to a performance drop from val to test across all approaches. EmoTx outperforms all previous baselines by a healthy margin. For scene level, we see +4.6% improvement on Emotic labels, +7.8% on Top-25, and +9.7% on Top-10. Character-level predictions are more challenging, but we see consistent improvements of +1.5-3% across all label sets. Matching expectation, we see that simpler models such as EmotionNet or CAER perform worse than Transformer-based approaches of M2Fnet and AttendAffectNet. Note that EmotionNet and CAER are challenging to adapt for character-level predictions and are not presented, but we expect M2Fnet or AttendAffectNet to outperform them.

Figure 4.5: A scene from the movie *Forrest Gump* showing the multimodal self-attention scores for the two predictions: *Mrs. Gump* is *worried* and *Forrest* is *happy*. We observe that the *worried* classifier token attends to *Mrs. Gump*'s character tokens when she appears at the start of the scene, while *Forrest*'s *happy* classifier token attends to *Forrest* towards the end of the scene. The video frames have relatively similar attention scores while dialog helps with emotional utterances such as *told you not to bother* or *it sounded good*.

4.4 Analyzing Self-attention Scores

EmoTx provides an intuitive way to understand which modalities are used to make predictions. We refer to the self-attention scores matrix as α, and analyze specific rows and columns. Separating the K classifier tokens allows us to find attention-score based evidence for each predicted emotion by looking at a row $\alpha_{\mathbf{z}_k^{\mathcal{S}}}$ in the matrix.

Fig. 4.5 shows an example movie scene where EmoTx predicts that *Forrest* is *happy* and *Mrs. Gump* is *worried*. We see that the model pays attention to the appropriate moments and modalities to make the right predictions.

Expressive emotions *vs*. Mental states. We hypothesize that the self-attention module may focus on character tokens for expressive emotions, while looking at the overall video frames and dialog for the more abstract mental states. We propose an *expressiveness* score as

$$e_k = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{\mathbf{z}_k^{\mathcal{S}},\mathbf{c}_t^i}}{\sum_{t=1}^{T}\alpha_{\mathbf{z}_k^{\mathcal{S}},\mathbf{f}_t} + \sum_{j=1}^{M}\alpha_{\mathbf{z}_k^{\mathcal{S}},\mathbf{u}_j}}, \tag{4.1}$$

where $\alpha_{\mathbf{z}_k^{\mathcal{S}},\mathbf{c}_t^i}$ is the self-attention score between the scene classifier token for emotion k ($\mathbf{z}_k^{\mathcal{S}}$) and character $\mathcal{P}^i$'s appearance in the video frame as b_t^i; $\alpha_{\mathbf{z}_k^{\mathcal{S}},\mathbf{f}_t}$ is for the video f_t and $\alpha_{\mathbf{z}_k^{\mathcal{S}},\mathbf{u}_j}$ is for dialog utterance u_j. Higher scores indicate expressive emotions as the model focuses on the character features, while

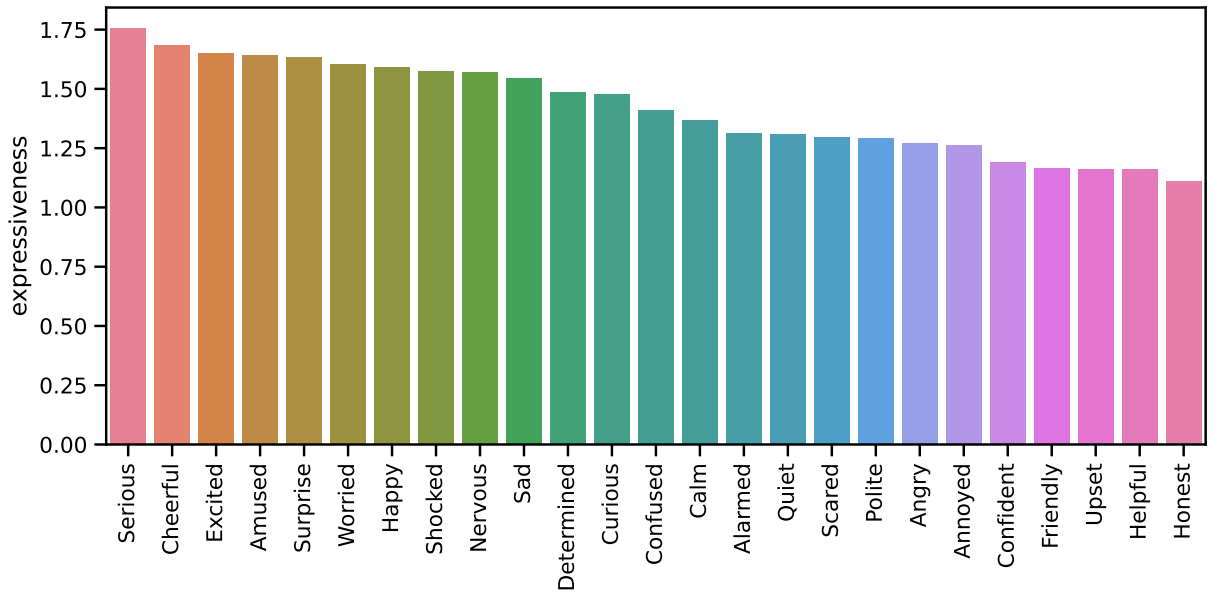

Figure 4.6: Sorted expressiveness scores for Top-25 emotions. Expressive emotions have higher scores indicating that the model attends to character representations, while mental states have lower scores suggesting more attention to video and dialog context.

lower scores identify mental states that analyze the video and dialog context. Fig. 4.6 shows the averaged expressiveness score for the Top-25 emotions when the emotion is present in the scene (*i.e.* y_k=1). We observe that mental states such as *honest, helpful, friendly, confident* appear towards the latter half of this plot while most expressive emotions such as *cheerful, excited, serious, surprise* appear in the first half. Note that the expressiveness scores in our work are for faces and applicable to our particular dataset.

Chapter 5

Conclusion

In this thesis, we presented a novel task for multi-label emotion and mental state recognition at the level of a movie scene and for each character.

Our work, EmoTx [2], a Transformer encoder based model, introduces a unique dimension by aiming to predict character-level mental states in addition to emotions and obtained significant improvements over previous works adapted for this task.

Operating within the temporal framework of movie scenes, we demonstrate the influence of video and dialog context in enhancing the accuracy of predictions for these nuanced labels. This aspect differentiates our research from previous works and aligns with our goal of capturing the rich interplay of emotions and mental states within the context of cinematic storytelling.

Our learned model was shown to have interpretable attention scores across modalities – they focused on the video or dialog context for mental states while looking at characters for expressive emotions. In the future, EmoTx may benefit from audio features or by considering the larger context of the movies instead of treating every scene independently.

The evolution of movie understanding has brought forth a diverse array of tasks and challenges, with each facet contributing to a more holistic comprehension of the cinematic medium. Our exploration, anchored in emotional and mental state identification, represents a valuable contribution to the multifaceted journey of understanding movies in all their narrative richness.

Bibliography

[1] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, pp. 1499–1503, 2016. 9, 10, 19

[2] D. Srivastava, A. K. Singh, and M. Tapaswi, "How you feelin'? Learning Emotions and Mental States in Movie Scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 4, 6, 30

[3] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler, "MovieGraphs: Towards Understanding Human-Centric Situations from Videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 5, 7, 8, 9, 12, 18, 23

[4] A. M. Schmitter, "17th and 18th Century Theories of Emotions," in *The Stanford Encyclopedia of Philosophy*, 2021. 2

[5] J. E. LeDoux, "Evolution of Human Emotions," *Progress in Brain Research*, vol. 195, pp. 431–442, 2013. 2

[6] R. Plutchik, "A General Pscychoevolutionary Theory of Emotion," *Theories of Emotion*, pp. 3–33, 1980. 2

[7] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, pp. 124–9, 1971. 2, 4

[8] G. L. Clore, A. Ortony, and M. A. Foss, "The Psychological Foundations of the Affective Lexicon," *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 751–766, 1987. 2

[9] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4, 19

[10] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware Emotion Recognition Networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4, 19, 26, 27

[11] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 19

[12] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in *Association of Computational Linguistics (ACL)*, 2019. 3, 5

[13] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019. 3, 6

[14] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, "Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection," in *International Joint Conference on Natural Language Processing (IJCNLP)*, 2021. 3, 6

[15] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, pp. 43–55, 2015. 3, 5

[16] T. Mittal, P. Mathur, A. Bera, and D. Manocha, "Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 18

[17] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is ... Buffy" – Automatic Naming of Characters in TV Video," in *British Machine Vision Conference (BMVC)*, 2006. 3, 7

[18] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, ""Knock! Knock! Who is it?" Probabilistic Person Identification in TV series," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[19] A. Brown, V. Kalogeiton, and A. Zisserman, "Face, Body, Voice: Video Person-Clustering with Multiple Modalities," in *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 3

[20] A. Brown, E. Coto, and A. Zisserman, "Automated Video Labelling: Identifying Faces by Corroborative Evidence," in *Multimedia Information Processing and Retrieval (MIPR)*, 2021. 3

[21] Kalogeiton, Vicky, and Zisserman, Andrew, "Constrained video face clustering using 1nn relations," in *British Machine Vision Conference (BMVC)*, 2020. 3, 12, 20

[22] A. Nagrani and A. Zisserman, "From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script," in *British Machine Vision Conference (BMVC)*, 2017. 3

[23] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin, "A Local-to-Global Approach to Multi-modal Movie Scene Segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 7

[24] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, and R. Hamid, "Self-Supervised Learning for Scene Boundary Detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[25] D. Rotman, D. Porat, and G. Ashour, "Optimal Sequential Grouping for Robust Video Scene Detection using Multiple Modalities," *International Journal of Semantic Computing*, vol. 11, no. 2, pp. 192–208, 2017. 3

[26] Z. Rasheed and M. Shah, "Scene Detection in Hollywood Movies and TV Shows," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 3

[27] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "StoryGraphs: Visualizing Character Interactions as a Timeline," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3, 7

[28] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: Understanding Stories in Movies through Question-Answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 23

[29] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: Localized, Compositional Video Question Answering," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[30] Y. Yu, J. Kim, and G. Kim, "A Joint Sequence Fusion Model for Video Question Answering and Retrieval," in *European Conference on Computer Vision (ECCV)*, 2018. 3

[31] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie Description," *IJCV*, vol. 123, p. 94–120, 2017. 3

[32] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[33] J. S. Park, T. Darrell, and A. Rohrbach, "Identity-Aware Multi-Sentence Video Description," in *European Conference on Computer Vision (ECCV)*, 2020. 3

[34] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[35] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "LAEO-Net: revisiting people Looking At Each Other in videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[36] A. Kukleva, M. Tapaswi, and I. Laptev, "Learning Interactions and Relationships between Movie Characters," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 5, 18

[37] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Book2Movie: Aligning Video scenes with Book chapters," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[38] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books," in *International Conference on Computer Vision (ICCV)*, 2015. 3

[39] Y. Xiong, Q. Huang, L. Guo, H. Zhou, B. Zhou, and D. Lin, "A Graph-based Framework to Bridge Movies and Synopses," in *International Conference on Computer Vision (ICCV)*, 2019. 3

[40] C.-Y. Wu and P. Krähenbühl, "Towards Long-Form Video Understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[41] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed Movies: Story Based Retrieval with Contextual Embeddings," in *Asian Conference on Computer Vision (ACCV)*, 2020. 3

[42] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "MovieNet: A Holistic Dataset for Movie Understanding," in *European Conference on Computer Vision (ECCV)*, 2020. 3, 9, 19

[43] L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, *et al.*, "VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation," in *Advances in Neural Information Processing Systems (NeurIPS): Track on Datasets and Benchmarks*, 2021. 3

[44] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *International Conference on Multimedia and Expo (ICME)*, 2005. 4

[45] Y.-I. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 2, pp. 97–115, 2001. 4

[46] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94–101. 4

[47] Dhall, Abhinav and Goecke, Roland and Joshi, Jyoti and Wagner, Michael and Gedeon, Tom, "Emotion recognition in the wild challenge 2013," in *International Conference on Multimodal Interaction (ICMI)*, 2013. 4

[48] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing (ICONIPS)*, 2013. 4, 20, 24, 27

[49] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies," *IEEE Multimedia*, vol. 19, pp. 34–41, 2012. 4, 5

[50] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis," in *Asian Conference on Computer Vision (ACCV)*, 2014. 4

[51] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1805–1812. 4

[52] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias," in *European Conference on Computer Vision (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., 2018. 4

[53] Z. Wei, J. Zhang, Z. Lin, J.-Y. Lee, N. Balasubramanian, M. Hoai, and D. Samaras, "Learning Visual Emotion Representations From Web Data," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 26

[54] D. Ong, Z. Wu, T. Zhi-Xuan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset," *IEEE Transactions on Affective Computing*, 2019. 5

[55] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling," in *Interspeech*, 2010. 6

[56] W. Jiao, M. Lyu, and I. King, "Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020. 6

[57] S. Sivaprasad, T. Joshi, R. Agrawal, and N. Pedanekar, "Multimodal Continuous Prediction of Emotions in Movies using Long Short-Term Memory Networks," in *International Conference on Multimedia Retrieval (ICMR)*, 2018. 6

[58] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation," in *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 6

[59] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 6

[60] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 6, 26, 27

[61] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021. 6

[62] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion Identification in Conversations," in *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 6

[63] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal Multi-label Emotion Detection with Modality and Label Dependence," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 6

[64] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," in *International Conference on Image Processing (ICIP)*, 2016. 10, 20

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 12, 20, 24

[66] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 12, 20, 24

[67] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2Label: A Simple Transformer Way to Multi-Label Classification," *arXiv:2107.10834*, 2021. 13, 15

[68] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A Joint Model for Video and Language Representation Learning," in *International Conference on Computer Vision (ICCV)*, 2019. 15

[69] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A Robustly Optimized BERT Pre-training Approach with Post-training," in *Chinese National Conference on Computational Linguistics*, 2021. 15, 19

[70] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021. 15

[71] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv: 1607.06450*, 2016. 16

[72] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 16

[73] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 19

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 19, 24

[75] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, pp. 211–252, 2015. 19, 24

[76] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 40, no. 6, pp. 1452–1464, 2017. 19, 24

[77] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale Vision Transformers," in *International Conference on Computer Vision (ICCV)*, 2021. 19, 24, 27

[78] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 19, 24, 27

[79] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 19

[80] S. Albanie and A. Vedaldi, "Learning Grimaces by Watching TV," in *British Machine Vision Conference (BMVC)*, 2016. 20, 24

[81] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *International Conference on Computer Vision Workshops (ICCVW)*, 2011. 20, 24

[82] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *British Machine Vision Conference (BMVC)*, 2015. 20, 24

[83] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 21

[84] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015. 21

[85] H. T. P. Thao, B. Balamurali, D. Herremans, and G. Roig, "AttendAffectNet: Self-Attention based Networks for Predicting Affective Responses from Movies," in *International Conference on Pattern Recognition (ICPR)*, 2021. 26, 27

[86] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *International Conference on Neural Information Processing Systems (ICONIPS)*, 2013. 26