EmoTxMLSM Enable Machines to Understand Human Emotions and Mental States

Jian Li
Illinoise Institute of Technology
CS577
jli232@hawk.iit.edu
https://github.com/lelour/EmoTxMLSM/

November 18, 2023

Abstract

Grasping a movie's storyline requires a deep insight into the emotions and mental states of its characters. To achieve this objective, we conceptualize emotion understanding as the task of predicting a diverse and multi-label spectrum of emotions, both at the level of individual movie scenes and for each character involved. My solution, EmoTxMLSM, is a multimodal Transformer-based architecture capable of processing videos, multiple character inputs, and dialog utterances to make collective predictions, which MLSM means using MultiLabelSoftMargin loss function. By harnessing the annotations from the MovieGraphs dataset [74], I aim to forecast not only classic emotions like happiness and anger but also other nuanced mental states such as honesty and helpfulness. My experiments encompass assessments involving the most commonly occurring 10 and 25 labels, as well as a mapping that clusters 181 labels into 26 categories. Through ablation studies and comparisons with adapted state-of-the-art emotion recognition techniques, I establish the effectiveness of EmoTxMLSM. A closer examination of EmoTxMLSM's self-attention scores reveals that expressive emotions predominantly focus on character tokens, while other mental states rely more heavily on cues from video content and dialog.

1. Introduction

With the exponential growth in the availability of audiovisual content and its accompanying consumption, it is necessary to have accurate content summaries and recommendations to help audiences make the right choices. If there is a better understanding of potential emotions, a better labeling system can be applied to online content, which in turn may lead to better recommendation systems. Film story analysis is a common field of application in this field. Film story analysis requires understanding the emotions and mental states of the task. To achieve this goal, emotional understanding is defined as predicting a diverse and multi labeled emotional set at the level of movie scenes and each character [13, 31, 34, 35]. Furthermore, if we can teach machines to accurately recognize human emotions, they can better understand and interact with humans.

In the movie $The Pursuit of Happyness$, we witness the main character's tumultuous emotional journey, spanning the depths of a breakup and homelessness to the highs of securing a coveted job. These intense emotional fluctuations serve as a powerful tool to captivate the audience, allowing them to connect with the characters on a profound level of empathy. To enable machines to grasp the essence of such a film or narrative more broadly, we contend that it is vital to monitor the evolving emotional and mental states of the characters over time. In pursuit of this objective, we utilize annotations from MovieGraphs [74] and train models to observe the visual content, analyze the dialogue, and make predictions about the emotions and mental states of the characters in each scene.

As generally, in the context of a film, a scene comprises a series of shots that collectively convey a sub-story. Typically, these scenes unfold at a specific location, involve a defined set of characters, and transpire within a relatively short time frame of about 30 seconds. Consequently, scenes are notably longer in duration compared to individual dialogues or short movie clips [4,55].Our objective is to predict the emotions and mental states of all characters within the scene, while also accumulating data at the scene level. This approach is well-suited for estimating emotions and mental states over an extended time frame, naturally leading to a multi-label classification system, as characters may concurrently experience multiple emotions (e.g., curiosity and confusion) or undergo transitions in response to interactions with other characters (e.g., shifting from a state of worry to one of calmness).

In my experiments, I explore various label sets, including the Top 10 or 25 most frequently occurring emotion labels from MovieGraphs [74], as well as a mapping to the 26 labels within the Emotic space, which was introduced by the [46]. While emotions can be broadly regarded as a component of mental states, in this study, I differentiate between expressed emotions, which are visibly conveyed by characters, such as surprise, sadness, and anger, and mental states, which remain concealed and are only discernible through interactions or dialogues, such as politeness, determination, confidence, or helpfulness. I contend that classifying emotions within a comprehensive label space necessitates the consideration of multimodal context, as demonstrated by the context masking in Fig. 1. To address this, I propose EmoTxMLSM, a model that simultaneously encompasses video frames, dialogue utterances, and character appearances.

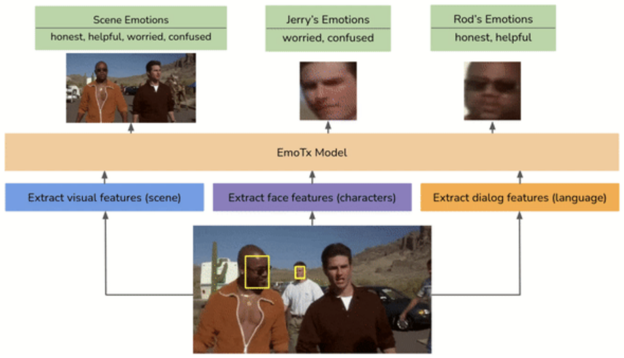
Scene Emotions
honest, helpful, worried, confused
Jerry's Emotions
worried, confused
Rod's Emotions
honest, helpful
EmoTx Model
Extract visual features (scene)
Extract face features (characters)
Extract dialog features (language)

Figure 1. Simultaneously predicting the scene and character emotions from a movie scene using EmoTxMLSM

The end-to-end NLP tasks in the past have become increasingly unified along with the increasing application of LLM. But understanding the patterns of the world is multimodal, and a large part of it is CV. Due to the perception based low dimensional characteristics in the field of CV, it is very difficult to achieve multi task unity in a single visual mode. At this point, utilizing multimodality may be one of the ideas.

I attempted to fuse on a multimodal Transformer based on pre-trained model representation and achieve a unified understanding of textual and visual content. With EmoTxMLSM, a multimodal Transformer based architecture, it can combine videos, multiple characters, and conversations for joint prediction. By leveraging annotations from the MovieGraph dataset [74], the goal is to predict typical emotions (such as happiness, anger) and other mental states (such as honesty and helpfulness). I will compare some models and perform ablation experiments on the best performing EmoTxMLSM to demonstrate the effectiveness of the model.

2. Related Work

2.1. Movie Understanding

In recent years, the understanding of movies has evolved beyond simple character recognition [6, 7, 19, 29, 47, 67], with a focus on tasks such as scene identification [11, 56, 57, 59, 68], question answering [35, 70, 79], and aligning text with video storylines [69, 78, 86]. This shift has been driven by datasets like Condensed Movie [3], MovieNet [27], the VALUE benchmark [37], and MovieGraphs [74]. Now, in addition to these advancements, there is a growing interest in identifying and analyzing the emotions, mental states, and overall mood of characters within movie scenes, enhancing our comprehension of storytelling.

2.2. Visual Emotion Recognition

Traditionally, emotion recognition relied on Ekman's six classic emotions [18], but in recent years, the field has evolved with in-the-wild benchmarks like EmotiW [16], FER [24], and AFEW [15], accompanied by the integration of deep learning techniques. The Emotic dataset [31] introduced a novel approach by incorporating 26 emotion labels for image-based emotion comprehension, emphasizing context through techniques like two-stream CNNs [34] and depth map-based person detection [45]. In contrast to existing trends, this research focuses on recognizing multi-label emotions and mental states in the context of movies, leveraging multimodal context at both scene and character levels [75].

2.3. Multimodal Datasets for Emotion Recognition

Recent developments in emotion recognition have seen the use of multimodal datasets. While some datasets focus on facial expressions (e.g., Acted Facial Expressions in the Wild [15]) or conversations (e.g., MELD [55]), the research discussed here centers on emotion recognition within the context of edited movies and narratives, specifically movie scenes. In contrast to other datasets, this work not only predicts emotions but also character-level mental states, highlighting the importance of considering video and dialogue context in achieving these labels.

2.4. Multimodal Emotion Recognition Methods

Recent advancements in Emotion Recognition in Conversations (ERC [28, 42, 64, 76]) have seen the utilization of RNNs and Transformer architectures [12, 63] for integrating audio, visual, and textual data, often enhanced by external knowledge graphs [22] and topic modeling. Multi-label prediction in ERC has been explored, but scalability challenges arise with a growing number of labels. In this context, the work discussed here employs a Transformer [73] for comprehensive modeling and focuses on predicting emotions and mental states in movie scenes and characters, distinguishing it from traditional ERC approaches. The MovieGraphs [74] project, although related, primarily tracks a single emotion within each scene or explores temporal emotion localization, providing alternative directions for research in the field.

2.5. EmoTxMLSM

EmoTx [88] is proposed on Learning Emotions and Mental States in Movie Scenes task. But the emotion labels are imbalanced actually. My EmoTxMLSM proposed better performance with MultiLabelSoftMargin loss function based on EmoTx.

3. Method

EmoTxMLSM utilizes the self-attention mechanism inspired by Transformers to forecast emotions and mental states. I first define the task (Sec. 3.1) and then describe my proposed approach (Sec. 3.2), before ending this section with details regarding training and inference (Sec. 3.3).

3.1. Problem Statement

We presume that movies have been segmented, either automatically [56] or with a human-in-the-loop process [68, 74] , into coherent scenes, each representing a self-contained segment of the story. The primary focus of this study lies in characterizing emotions within movie scenes, which are typically of substantial duration, ranging from 30 to 60 seconds and involving multiple shot changes.

Consider a movie scene denoted as $\mathcal{S}$, comprising a collection of video frames $\mathcal{V}$, characters $\mathcal{C}$, and dialog utterances $\mathcal{U}$. We represent the set of video frames as $\mathcal{V} = \{f_t\}_{t=1}^{T}$, where T signifies the number of frames after sub-sampling. Movie scenes often feature multiple characters, and we model N characters within the scene as $\mathcal{C} = \{\mathcal{P}^i\}_{i=1}^{N}$. Each character $\mathcal{P}^i$ is defined as $\mathcal{P}^i = \{(f_t, b_t^i)\}$, denoting their potential appearance in some frame f_t of the video within the spatial bounding box b_t^i . If a character $\mathcal{P}^i$ is absent at a given time t, we assume that b_t^i is empty. Additionally, $\mathcal{U} = \{u_j\}_{j=1}^{M}$ represents the dialog utterances in the scene. In this study, we utilize dialogues directly from subtitles [19] and assume that they are not associated with character names. While it is possible to name dialogues through subtitle-transcript alignment , the availability and reliability of scripts for movies can be inconsistent.

Task formulation. In the context of a movie scene, denoted as $\mathcal{S}$, comprising its video, characters, and dialog utterances, our objective is to predict both the emotions and mental states, collectively referred to as ”emotions,” at two levels: for the entire scene $\mathbf{y}^{\mathcal{V}}$ and for individual characters $\mathbf{y}^{\mathcal{P}^i}$. We frame this as a multi-label classification task involving K labels, where each $y_k \in \{0, 1\}$ signifies the absence or presence of the k^{th} label in either the scene $y_k^{\mathcal{V}}$ or expressed by a particular character $y_k^{\mathcal{P}^i}$. In datasets that include character-level annotations, scene-level labels are derived through a straightforward logical **OR** operation, i.e., $\mathbf{y}^{\mathcal{V}} = \bigoplus_{i=1}^{N} \mathbf{y}^{\mathcal{P}^i}$.

3.2. EmoTxMLSM : My Approach

Introducing EmoTxMLSM, my Transformer-driven approach designed to detect emotions both at the movie scene and individual character levels. My process begins with initial video pre-processing and feature extraction, which extracts pertinent representations. Subsequently, a Transformer encoder amalgamates information from various modalities, followed by the implementation of a classification module influenced by previous research in multi-label classification with Transformers. A visual overview of our approach is outlined in Fig. 2.

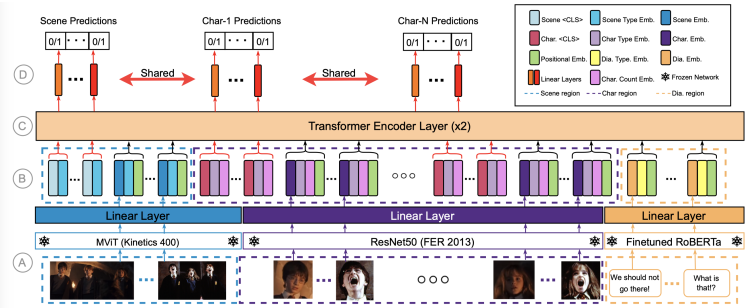

Figure 2. An ovrview of EmoTxMLSM. I detail the comprehensive approach in Sec. 3 while offering a concise overview here. A: Video features (depicted in the blue region), character facial features (in the purple region), and utterance features (found within the orange region) are acquired using fixed backbones and then mapped into a unified embedding space via linear layers. B: In this phase, pertinent embeddings are added to the tokens to differentiate between modalities, character distinctions, and to introduce a temporal context. Additionally, specific tokens for each emotion classifier associated with the scene or a particular character are generated. C: Two layers of Transformer encoders conduct self-attention across the sequence of input tokens. D: Ultimately, we utilize the classifier tokens to generate output probability scores for each emotion employing a shared linear classifier spanning across both scenes and characters.

Preparing multimodal representations. Recognizing intricate emotions and mental states, such as nervousness or determination, necessitates a deeper understanding of the broader narrative context beyond just facial expressions. To enable this, we employ a multi-faceted approach to encode information: (i) Video content is encoded to capture the spatial and temporal context of events. (ii) Characters are detected, tracked, clustered, and represented based on their facial and full-body appearance. (iii) Dialog utterances are encoded to provide complementary information to the visual domain.

I utilize a pretrained encoder, denoted as $\phi_{\mathcal{V}}$, to extract relevant visual information from single or multiple frames, expressed as $\mathbf{f}_t = \phi_{\mathcal{V}}(\{f_t\})$. Similarly, a pretrained language model $\phi_{\mathcal{U}}$ is used to extract representations of dialog utterances, defined as $\mathbf{u}_j = \phi_{\mathcal{U}}(u_j)$. The handling of characters involves a more comprehensive process, beginning with their localization within the appropriate frames. When a valid bounding box, b_t^i, is available for a character $\mathcal{P}^i$, character features are extracted using a backbone pretrained for emotion recognition, leading to $\mathbf{c}_t^i = \phi_{\mathcal{C}}(f_t, b_t^i)$.

Linear projection. In Transformers, token representations typically merge fundamental information, like visual representation, with additional meta-information, such as the timestamp, using position embeddings (e.g., [65]). My approach harmonizes all modalities to a uniform dimension by employing linear layers. To be specific, we use linear projections to transform visual representation $\mathbf{f}_t \in \mathbb{R}^{D_{\mathcal{V}}}$ with $\mathbf{W}_{\mathcal{V}} \in \mathbb{R}^{D \times D_{\mathcal{V}}}$, utterance representation $\mathbf{u}_j \in \mathbb{R}^{D_{\mathcal{U}}}$ with $\mathbf{W}_{\mathcal{U}} \in \mathbb{R}^{D \times D_{\mathcal{U}}}$, and character representation $\mathbf{c}_t^i \in \mathbb{R}^{D_{\mathcal{C}}}$ with $\mathbf{W}_{\mathcal{C}} \in \mathbb{R}^{D \times D_{\mathcal{C}}}$. For the sake of brevity, we omit mentioning the linear layer biases.

Modality embeddings. To encompass the three distinct modalities—video, characters, and dialog utterances—I train three embedding vectors, $\mathbf{E}^{\mathcal{M}} \in \mathbb{R}^{D \times 3}$. In addition, we aid the model in distinguishing tokens originating from characters by incorporating a dedicated character count embedding, denoted as $\mathbf{E}^{\mathcal{C}} \in \mathbb{R}^{D \times N}$. It's essential to note that these modality and character embeddings do not carry specific meaning or enforced order, such as higher or lower appearance times or alphabetical naming. The model is anticipated to utilize these embeddings solely for the purpose of differentiation between modalities and characters.

Time embeddings. The number of tokens is contingent on the chosen frame rate. To provide the model with information regarding the relative temporal sequence across different modalities, we employ a discrete time binning approach that maps real-time values (in seconds) to an index. Consequently, representations of video frames or segments and character bounding boxes fed into the Transformer are associated with their corresponding time bins. In the case of an utterance u_j, binning is determined based on its middle timestamp, denoted as t_j. The embeddings of these time bins are represented as $\mathbf{E}^T \in \mathbb{R}^{D \times \lceil T^*/\tau \rceil}$, where T^* denotes the maximum scene duration, and τ is the bin step. For convenience, $\mathbf{E}_t^T$ selects the appropriate embedding using a discretized index $\lceil t/\tau \rceil$.

Classifier tokens. In a manner akin to the classic **CLS** tokens within Transformer models [17,87] , we employ learnable classifier tokens for the purpose of emotion prediction. Building on the concept introduced by Query2Label [40] , we opt for K classifier tokens rather than a single token to generate all the outputs, as depicted in Figure 2D. This approach not only enables the model to capture label co-occurrence within the Transformer layers, thereby enhancing performance, but also facilitates the analysis of per-emotion attention scores, offering valuable insights into the model's functioning. Specifically, we employ K classifier tokens for scene-level predictions, denoted as $\mathbf{z}_k^{\mathcal{S}}$, and $N \times K$ tokens for character-level predictions, denoted as $\mathbf{z}_k^i$ for each character-emotion pair pertaining to character $\mathcal{P}^i$.

Token representations. The amalgamation of features with the corresponding embeddings furnishes EmoTxMLSM with a wealth of information. The token representations for each input group can be outlined as follows:

$$\text{scene cls. tokens} : \tilde{\mathbf{z}}_k^{\mathcal{S}} = \mathbf{z}_k^{\mathcal{S}} + \mathbf{E}_1^{\mathcal{M}} \tag{1}$$

$$\text{char. cls. tokens} : \tilde{\mathbf{z}}_k^{i} = \mathbf{z}_k^{i} + \mathbf{E}_2^{\mathcal{M}} + \mathbf{E}_i^{\mathcal{C}} \tag{2}$$

$$\text{video} : \tilde{\mathbf{f}}_t = \mathbf{W}_{\mathcal{V}}\mathbf{f}_t + \mathbf{E}_1^{\mathcal{M}} + \mathbf{E}_t^{T} \tag{3}$$

$$\text{character box} : \tilde{\mathbf{c}}_t^i = \mathbf{W}_{\mathcal{C}}\mathbf{c}_t^i + \mathbf{E}_2^{\mathcal{M}} + \mathbf{E}_i^{C} + \mathbf{E}_t^{T} \tag{4}$$

$$\text{and utterance} : \tilde{\mathbf{u}}_j = \mathbf{W}_{\mathcal{U}}\mathbf{u}_j + \mathbf{E}_3^{\mathcal{M}} + \mathbf{E}_{t_j}^{T} \tag{5}$$

Additionally, as demonstrated in Fig. 2B, it's worth noting that we apply LayerNorm [2] prior to the input tokens being fed into the Transformer encoder layers, although this step is omitted here for the sake of brevity.

Transformer Self-attention. We concatenate all tokens and route them through two Transformer encoder layers $H = \mathit{2}$ that facilitate self-attention across all modalities, as described in [73]. In the context of emotion prediction, we exclusively utilize the outputs associated with the classification tokens as

$$\left[\hat{\mathbf{z}}_k^{\mathcal{S}}, \hat{\mathbf{z}}_k^{i}\right] = \text{TransformerEncoder}\left(\tilde{\mathbf{z}}_k^{\mathcal{S}}, \tilde{\mathbf{f}}_t, \tilde{\mathbf{z}}_k^{i}, \tilde{\mathbf{c}}_t^{i}, \tilde{\mathbf{u}}_j\right) \tag{6}$$

This collective encoding encompasses all tokens across sets $\{k\}_1^K$, $\{i\}_1^N$, $\{t\}_1^T$, and $\{j\}_1^M$.

Emotion labeling. The contextualized representations for the scene, denoted as $\hat{\mathbf{z}}_k^{\mathcal{S}}$, and the characters, denoted as $\hat{\mathbf{z}}_k^{i}$, are both directed to a common linear layer represented by $\mathbf{W}^E \in \mathbb{R}^{K \times D}$ for classification. Ultimately, the probability estimates are derived through a sigmoid activation function, denoted as $\sigma\left(\cdot\right)$:

$$\hat{y}_k^{\mathcal{S}} = \sigma\left(\mathbf{W}_k^E \hat{\mathbf{z}}_k^{\mathcal{S}}\right) \text{ and } \hat{y}_k^{i} = \sigma\left(\mathbf{W}_k^E \hat{\mathbf{z}}_k^{i}\right), \forall k, i \tag{7}$$

3.3. Training and Inference

Training. EmoTxMLSM is trained in an end-to-end manner using the MultiLabelSoftMargin loss to address class imbalance. The losses for scene and character predictions are then aggregated as follows:

$$\mathcal{L} = -\frac{1}{K}\Sigma_{k=1}^{K} y_k \log\left(\frac{1}{1 + exp\left(-\hat{y}_k\right)}\right) + (1 - y_k)\log\left(\frac{exp\left(-\hat{y}_k\right)}{1 + exp\left(-\hat{y}_k\right)}\right) \tag{8}$$

Inference. During the testing phase, we adhere to the procedure detailed in Section 3.2 and produce emotion label predictions for both the entire scene and each character, as outlined in Equation 7.

Variations. As we will observe through empirical analysis, our model demonstrates remarkable versatility and is amenable to the incorporation or removal of modalities or additional representations simply by adjusting the Transformer's width, i.e., the number of tokens. It can be readily adapted to function as a unimodal architecture, focusing exclusively on video or dialog utterances while disregarding other modalities.

4. Experiments and Discussion

My experimental configuration is introduced in Section 4.1, followed by a more detailed exploration of implementation particulars in Section 4.2. I then proceed with a series of ablation studies in Section 4.3 to elucidate the rationale behind our model's design choices. Finally in Section 4.4, I provide a comparative analysis with adapted versions of several state-of-the-art models for emotion recognition.

4.1. Dataset and Setup

We employ the MovieGraphs dataset [74] , which comprises 51 movies and encompasses 7,637 movie scenes, each enriched with comprehensive graph annotations. My specific emphasis is on character-related data, including their emotions and mental states, which naturally aligns with a multilabel classification approach. I disregard other annotations, such as situation labels and character interactions and relationships [32] , as these cannot be presumed to be accessible for new movies.

Label sets. Similar to various annotations within the MovieGraphs dataset, emotions are obtained in free-text format, resulting in substantial variability and a long-tail distribution of labels, totaling over 500. Our experiments concentrate on three distinct types of label sets: (i) Top10, comprising the ten most frequently occurring emotions; (ii) Top-25, encompassing the frequently occurring 25 labels; and (iii) Emotic, which involves mapping 181 MovieGraphs emotions to 26 Emotic labels, as provided by [46].

Statistics. I commence by showcasing row max-normalized co-occurrence matrices for both scenes and characters, as depicted in Figure 3. It is noteworthy to observe that in a movie scene, there are elevated co-occurrence scores for emotions like "worried" and "calm," which might be attributed to the presence of multiple characters. Additionally, for an individual character, "worried" is most closely associated with "confused." Another notable instance is the high co-occurrence of "curious" with "surprise" for a single character, whereas in a movie scene, we observe "curious" paired with "calm" and "surprise" alongside "happy." In Figure 4, we present data on the frequency of movie scenes containing a specific number of emotions, with most scenes incorporating four emotions.

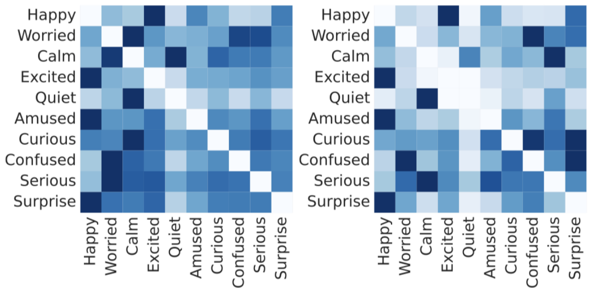

Figure 3. Normalized matrices displaying the co-occurrence of top-10 emotions within a movie scene (on the left) or for a character (on the right).

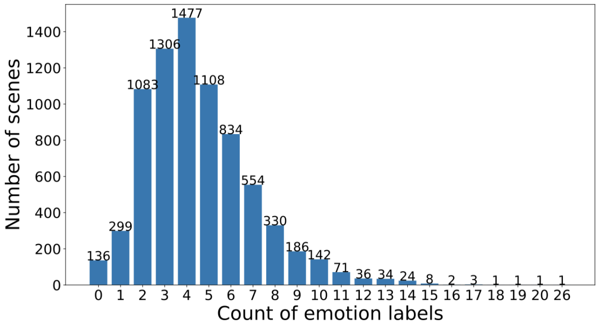

Figure 4. A bar chart depicting the frequency of movie scenes linked to various counts of annotated emotions.

Evaluation metric. I adhere to the original splits provided by MovieGraphs. Given the K binary classification tasks at hand, we employ the mean Average Precision (mAP) as our performance metric, akin to the approach used in Atomic Visual Actions [25]. It's worth noting that the Average Precision (AP) metric is influenced by label frequency.

4.2. Implementation Details

Feature representations. The choice of backbones plays a pivotal role in determining the performance of any model. In this section, we delineate the various backbones utilized for feature extraction from video frames, characters, and dialog.

Video Features $\mathbf{f}_t$: Visual context holds significant importance in the comprehension of emotions [31, 34, 45]. To capture spatial features, we employ ResNet152 [26], pretrained on ImageNet [61], and ResNet50 [26], pretrained on Place365 [84]. For spatiotemporal features, we utilize MViT [20], which has been trained on Kinetics400 [10].

Dialogue Features $\mathbf{u}_j$: Every utterance is subjected to encoding through a RoBERTa-Base model [87], resulting in an utterance-level embedding. In addition, we extract features from a RoBERTa model fine-tuned specifically for the multi-label emotion classification task, primarily based on dialog content.

Character Features $\mathbf{c}_t^i$: Character representations are established based on either facial or person detections. Facial detection is accomplished with MTCNN [83], while person detection relies on Cascade RCNN [8] trained on MovieNet [27] data. To obtain character tracks, we apply the SORT [5] algorithm, which is a straightforward Kalman filter-based approach, and perform clustering using C1C [29]. Face representations are extracted using ResNet50 [1], pretrained on SFEW [14] and further pretrained on FER13 [24] and VGGFace [52], as well as VGGm [1], which is trained on FER13 and pretrained on VGGFace. Additionally, InceptionResnetV1 [66], trained on VGGFace2 [9], is employed to extract face representations.

Frame sampling strategy. For the video modality, we sample a maximum of $T = 300$ tokens at a rate of 3 frames per second (equivalent to 100 seconds), encompassing approximately 99% of all movie scenes. The time embedding bins are set at a rate of 3 per second, corresponding to $\tau = 1/3$ second. During inference, we select a fixed set of frames, while during training, frames are randomly sampled at 3 frames per second intervals, serving as a form of data augmentation. Character tokens are treated in a comparable manner, with the additional condition that they are contingent upon the character's presence in the video.

Architecture details. We conducted experiments varying the number of encoder layers, H, in the set $\{1, 2, 4, 8\}$, and determined that $H = 2$ yields the best results, possibly due to the dataset's limited size. Both layers share the same configuration, featuring 8 attention heads with a hidden dimension of 512. The maximum number of characters, N, is set at 4, covering up to 91% of the scenes. Tokens are padded to create batches and accommodate shorter video clips. Effective masking is applied to prevent self-attention on the padded tokens. In summary, the EmoTxMLSM encoder processes K scene classification tokens, T video tokens, $N \cdot (K + T)$ character tokens, and T utterance tokens. For the specific case of $K = 25$ and $N = 4$ (Top-25 label set), this amounts to a maximum of 1925 padded tokens.

Training details. My model is developed using PyTorch [53] and trained on a single NVIDIA V100 GPU for a maximum of 20 epochs, utilizing a batch size of 8. Hyperparameters are fine-tuned to optimize performance on the validation set. We employ the Adam optimizer [30] with an initial learning rate of 1×10^{-5}, which is then dynamically reduced by a factor of 10 through the learning rate scheduler ReduceLROnPlateau. The best checkpoint is determined based on the maximum geometric mean of scene and character mAP.

4.3. Ablation Studies

My ablation experiments encompass three primary dimensions: architectural variations, modalities, and feature backbones. In cases where specific configurations are not specified, we adhere to the default settings, which include (i) employing MViT trained on the Kinetics400 dataset to represent video, (ii) utilizing ResNet50 trained on SFEW, FER, and VGGFace for character representations, (iii) employing fine-tuned RoBERTa for dialog utterance representations, (iv) utilizing EmoTx with appropriate masking to select modalities or modify the number of classifier tokens, and (v) utilizing EmoTxMLSM with appropriate masking to select modalities or modify the number of classifier tokens.

Architecture ablations. In Table 1, we present a comparative analysis of our architecture against simpler variants. I evaluate the performance of MLP ($\mathit{2\ Lin}$), which is a straightforward Multi-Layer Perceptron with two linear layers using max-pooled scene or character features as inputs. As an alternative to max pooling, we employ self-attention. The Single Tx encoder employs self-attention over features, treating them as tokens, and connects to a classifier token with a multi-label classifier. Both of these approaches exhibit substantial improvements over the random baseline, especially for individual character-level predictions, which tend to be inherently more challenging than scene-level predictions.

Furthermore, I compare the multimodal EmoTx [88] model that utilizes a single classifier token to predict all labels (EmoTx: 1 CLS) against the variant with K classifier tokens. Both models demonstrate significant improvements, with absolute gains of $+7.5\%$ for Top-10 scene labels and $+2.3\%$ for the more demanding Top-25 character-level labels. These improvements underscore EmoTx's proficiency in encoding multiple modalities effectively. Notably, the variant with K classifier tokens (last row) consistently exhibits small but noteworthy $+0.7\%$ enhancements over the 1 classifier token setup for Top-25 emotions.

For the imbalanced labels, I improved EmoTx [88] with Multi Label Soft Margin loss function which called EmoTxMLSM. And it performs $+2\%$ for Top-10 scene labels and $+0.5\%$ for the Top-25 character-level labels compare to EmoTx.

Figure 5 showcases the scene-level AP scores for the Top-25 labels. My model outperforms the MLP and Single Tx encoder for 24 out of 25 labels and surpasses the single classifier token variant for 15 out of 25 labels. EmoTxMLSM excels in recognizing expressive emotions such as ”excited,” ”serious,” ”happy,” and even mental states like ”friendly,” ”polite,” and ”worried.” However, it encounters challenges in identifying other mental states like ”determined” or ”helpful.”

Backbone ablations. We conduct a comparison of various

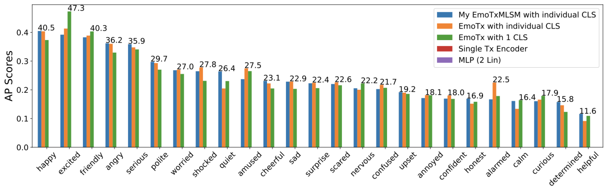

Figure 5. When I contrast EmoTxMLSM's scene-level per-class AP with the baselines (Table 1), it reveals a consistent enhancement. Additionally, my model, featuring K classifier tokens, demonstrates superior performance across most classes in comparison to the EmoTx. The AP of the top-performing model is marked atop the bar. Intriguingly, the sequence of emotions presented doesn't align with their frequency of occurrence.

Method	Top-10		Top-25	
	Scene	Char	Scene	Char
MLP(2 Lin)	25.60	20.87	16.74	10.74
Single Tx Encoder	25.57	21.24	16.46	11.17
EmoTx: 1 CLS	33.12	23.24	23.22	13.11
EmoTx	*33.37*	*23.38*	*23.97*	*13.34*
EmoTxMLSM(mine)	**34.01**	**23.86**	**24.02**	**13.35**

Table 1. displays the results of our architecture ablation, featuring emotion predictions at both the movie scene and individual character (Char) levels. It is evident that mine multimodal model excels significantly, surpassing simpler baseline approaches. The best-performing results are highlighted in bold, while close contenders are marked in italics.

backbones for the emotion recognition task. The efficacy of the fine-tuned RoBERTa model is evident when comparing pairs of rows R2, R5, R3, R7, and R4, R8 in Table 2, revealing a consistent improvement ranging from 1% to 3%. In addition, character representations utilizing ResNet50-FER demonstrate improvements over VGGm-FER, as observed in R5, R8, and R6, R7. Lastly, the benefits of action features over place features become evident when comparing R8. The patterns across all trends remain consistent: fine-tuning RoBERTa consistently contributes positively. It's evident that ResNet50 trained on FER serves as a strong representation for characters, while MViT trained on Kinetics400 yields superior results for both label sets. Additionally, ResNet50 trained on Places365 closely follows as a strong contender.

	Video			Character			Dialogue		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	×	✓	×	×	×	✓	×	✓	24.96	15.33	16.71	8.04
2	×	×	✓	×	×	✓	×	✓	25.67	15.38	16.67	8.11
3	×	×	✓	×	×	✓	×	✓	29.42	20.24	18.66	10.47
4	✓	×	×	×	×	✓	×	✓	29.37	18.49	18.47	9.44
5	×	✓	×	×	✓	×	×	✓	29.57	20.04	19.06	10.47
6	✓	×	×	×	✓	×	×	✓	30.46	21.37	19.26	10.32
7	×	✓	×	×	×	✓	✓	×	29.97	19.51	21.12	10.56
8	×	×	✓	×	×	✓	✓	×	30.04	19.80	21.01	10.84
9	×	✓	×	✓	×	×	×	✓	29.40	19.69	19.99	10.93
10	×	×	✓	✓	×	×	×	✓	30.63	20.80	19.87	11.11
11	✓	×	×	✓	×	×	×	✓	32.21	21.82	21.27	11.83
12	✓	×	×	×	✓	×	✓	×	32.67	22.36	21.37	11.65
13	×	×	✓	×	✓	×	✓	×	31.43	21.91	22.15	11.84
14	✓	×	×	×	×	✓	✓	×	32.13	21.39	22.51	11.81
15	×	✓	×	×	✓	×	✓	×	32.00	22.52	21.12	11.35
16	×	×	✓	✓	×	×	✓	×	32.30	22.92	22.18	12.07
17	×	✓	×	✓	×	×	✓	×	33.77	23.02	22.38	12.17
18	✓	×	×	✓	×	×	✓	×	33.80	23.53	24.03	13.56

Table 2. Feature ablations are conducted using various backbones and are denoted as follows: (MViT, K400) denoting MViT pre-trained on Kinetics400, (R50, P365) representing ResNet50 trained on Places365, (R152, INet) indicating ResNet152 pre-trained on ImageNet, (R50, FER) signifying ResNet50 specialized in Facial Expression Recognition (FER), (VGGM, FER) referring to VGG-M specifically trained for FER, (IRv1, VGG-F) representing InceptionResNet-v1 trained using the VGG-Face dataset, (RB, FT) representing pre-trained RoBERTa fine-tuned for emotion recognition, and (RB, PT) indicating pre-trained RoBERTa.

Modality ablations. Table 3 delves into the evaluation of the individual impact of each modality—video, characters, and utterances—on emotion prediction at both the scene and character levels. Notably, the character modality (row 4, R4) surpasses the performance of any of the video or dialog modalities (R1-R3). Additionally, dialog features (R3) exhibit superior performance compared to video features (R1, R2), which aligns with common trends in movie-related tasks [70, 74]. The selection of visual features holds significance, with scene features V_r consistently underperforming action features V_m, as evidenced in comparisons R1, R2, R5, R6, and R8, R9. Ultimately, the utilization of all modalities (R9) outperforms other modality combinations, underscoring that emotion recognition is inherently a multimodal task.

4.4. Analyzing Self-Attention Scores

EmoTxMLSM offers a user-friendly method to comprehend the modalities used in making predictions. The self-attention scores matrix, termed α, allows us to scrutinize specific rows and columns. By segregating the K classifier tokens, we can discern attention-score-based evidence for each predicted emotion, observed in the $\alpha_{Z_k^S}$ row within the matrix.

In Fig. 6, EmoTxMLSM demonstrates its predictive

	V_r	V_m	D	C	Top-10(mAP)		Top-25(mAP)	
					Scene	Char	Scene	Char
1	✓	×	×	×	21.76	14.67	14.60	7.82
2	×	✓	×	×	25.20	17.30	16.73	9.36
3	×	×	✓	×	26.89	19.99	19.43	11.02
4	×	×	×	✓	30.56	20.78	19.53	11.10
5	✓	×	✓	×	27.20	19.05	19.44	10.61
6	×	✓	✓	×	28.01	20.44	21.29	12.08
7	×	×	✓	✓	33.04	23.77	23.22	13.48
8	✓	×	✓	✓	33.15	22.86	22.49	12.29
9	×	✓	✓	✓	33.22	23.16	23.95	13.30

Table 3. Modality ablation is performed with the following notations: V_r for ResNet50 (Places365), V_m for MViT (Kinetics400), D for Dialog, and C for Character.

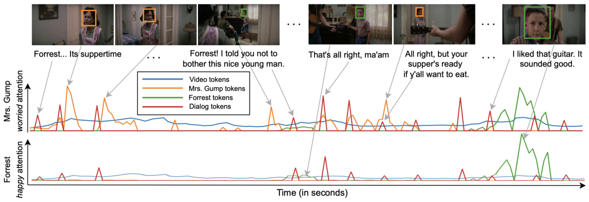

Figure 6. A sequence from the film Forrest Gump displaying the multimodal self-attention scores for two predictions: Mrs. Gump expressing concern and Forrest feeling happy. The token associated with worry focuses on Mrs. Gump's character at the scene's onset, whereas the token linked to Forrest's happiness concentrates on him towards the scene's conclusion. The video frames exhibit relatively consistent attention scores, while the dialog contributes emotional cues, such as "told you not to bother" or "it sounded good."

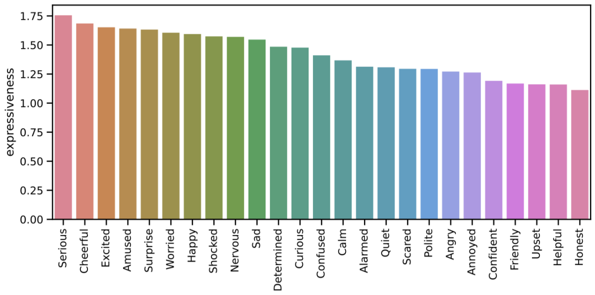

Figure 7. The ordered expressiveness scores for the Top-25 emotions reveal a distinction: higher scores denote expressive emotions, signaling the model's focus on character representations, whereas lower scores point to more attention given to the video and dialogue context, particularly for mental states.

capacity, indicating Forrest's happiness and Mrs. Gump's concern in a movie scene. The model distinctly focuses on relevant moments and modalities crucial for accurate predictions.

Distinguishing between expressive emotions and mental states, we posit that the self-attention module may emphasize character tokens for expressive emotions, while considering overall video frames and dialogues for more abstract mental states. We propose an expressiveness score represented as:

$$e_k = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{Z_k^S, C_t^i}}{\sum_{t=1}^{T} \alpha_{Z_k^S, f_t} + \sum_{j=1}^{M} \alpha_{Z_k^S, u_j}} \qquad (9)$$

Where $\alpha_{z_k^s, c_t^i}$ signifies the self-attention score between the scene classifier token for emotion $k(z_k^s)$ and character $\mathcal{P}^i$'s appearance in the video frame as b_t^i; $\alpha_{z_k^s, c_t^i}$ pertains to the video f_t, and $\alpha_{z_k^s, u_j}$ relates to dialog utterance u_j. Higher scores denote expressive emotions, where the model concentrates on character features, while lower scores indicate mental states, analyzing video and dialogue contexts.

In Fig. 7, the averaged expressiveness score for the Top-25 emotions when the emotion is present in the scene (i.e., $y_k = 1$) is depicted. It's observed that mental states like honest, helpful, friendly, and confident appear in the latter half of the plot. In contrast, most expressive emotions such as cheerful, excited, serious, and surprise appear in the first half. It's important to note that the expressiveness scores in our study are specific to facial expressions and are applicable within our dataset context.

5. Conclusion

I introduced an innovative endeavor focusing on recognizing multiple emotions and mental states within a movie scene for individual characters. EmoTxMLSM, a Transformer encoder-based model, was devised to simultaneously integrate all modalities, yielding substantial enhancements compared to prior methodologies tailored for this purpose. My model demonstrated intelligible attention scores spanning various modalities, emphasizing video or dialogue context for mental states and character traits for expressive emotions. Moving forward, augmenting EmoTxMLSM with audio features or contextualizing within the broader movie landscape rather than treating each scene in isolation could be advantageous.

References

[1] S. Albanie and A. Vedaldi. Learning Grimaces by Watching TV. In British Machine Vision Conference (BMVC), 2016. 6

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. arXiv: 1607.06450, 2016. 4

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed Movies: Story Based Retrieval with Contextual Embeddings. In Asian Conference on Computer Vision (ACCV), 2020. 2

[4] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. LIRIS-ACCEDE: A video database for affective content analysis. IEEE Transactions on Affective Computing, pages 43–55, 2015. 1

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In International Conference on Image Processing (ICIP), 2016. 6

[6] Andrew Brown, Ernesto Coto, and Andrew Zisserman. Automated Video Labelling: Identifying Faces by Corroborative Evidence. In Multimedia Information Processing and Retrieval (MIPR), 2021. 2

[7] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, Body, Voice: Video Person-Clustering with Multiple Modalities. In International Conference on Computer Vision Workshops (ICCVW), 2021. 2

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into High Quality Object Detection. In Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 6

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In International Conference on Automatic Face and Gesture Recognition (FG), 2018. 6

[10] Jo˜ao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 5

[11] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Self-Supervised Learning for Scene Boundary Detection. In Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 2

[12] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022. 2

[13] Gerald L. Clore, Andrew Ortony, and Mark A. Foss. The Psychological Foundations of the Affective Lexicon. Journal of Personality and Social Psychology, 53(4):751–766, 1987. 1

[14] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In International Conference on Computer Vision Workshops (ICCVW), 2011. 6

[15] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. IEEE Multimedia, 19:34–41, 2012. 2

[16] 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. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations (ICLR), 2021. 4

[18] Paul Ekman and W V Friesen. Constants across cultures in the face and emotion. Journal of personality and social psychology, pages 124–9, 1971. 2

[19] Mark Everingham, Josef Sivic, and Andrew Zisserman. “Hello! My name is ... Buffy” – Automatic Naming of Characters in TV Video. In British Machine Vision Conference (BMVC), 2006. 2, 3

[20] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In International Conference on Computer Vision (ICCV), 2021. 5

[21] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning. In Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[22] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In Findings of Empirical

Methods in Natural Language Processing (EMNLP), 2020. 2

[23] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander 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.

[24] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing (ICONIPS), 2013. 2, 6

[25] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. In Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 5

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 5

[27] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A Holistic Dataset for Movie Understanding. In European Conference on Computer Vision (ECCV), 2020. 2, 6

[28] Wenxiang Jiao, Michael Lyu, and Irwin King. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In Association for the Advancement of Artificial Intelligence (AAAI), 2020. 2

[29] Kalogeiton, Vicky, and Zisserman, Andrew. Constrained video face clustering using 1nn relations. In British Machine Vision Conference (BMVC), 2020. 2, 6

[30] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, International Conference on Learning Representations (ICLR), 2015. 6

[31] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 2, 5

[32] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. Learning Interactions and Relationships between Movie Characters. In Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 5

[33] Joseph E. LeDoux. Evolution of Human Emotions. Progress in Brain Research, 195:431–442, 2013.

[34] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware Emotion Recognition Networks. In Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 1, 2, 5

[35] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, Compositional Video Question Answering. In Empirical Methods in Natural Language Processing (EMNLP), 2018. 1, 2

[36] Juncheng Li, Junlin Xie, Linchao Zhu, Long Qian, Siliang Tang, Wenqiao Zhang, Haochen Shi, Shengyu Zhang, Longhui Wei, Qi Tian, and Yueting Zhuang. Dilated Context Integrated Network with Cross-Modal Consensus for Temporal Emotion Localization in Videos. In ACM Multimedia (MM), 2022.

[37] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang 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. 2

[38] Mengyi Liu, Shaoxin Li, S. Shan, Ruiping Wang, and Xilin Chen. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In Asian Conference on Computer Vision (ACCV), 2014.

[39] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial Expression Recognition via a Boosted Deep Belief Network. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 1805–1812, 2014.

[40] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2Label: A Simple Transformer Way to Multi-Label Classification. arXiv:2107.10834, 2021. 4

[41] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara 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), pages 94–101, 2010.

[42] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In Association for the Advancement of Artificial Intelligence (AAAI), 2019. 2

[43] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo MedinaSuarez, and Andrew Zisserman. LAEO-Net: revisiting people Looking At Each Other in videos. In Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In International Conference on Neural Information Processing Systems (ICONIPS), 2013.

[45] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle. In Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2, 5

[46] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality. In Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 2, 5

[47] Arsha Nagrani and Andrew Zisserman. From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script. In British Machine Vision Conference (BMVC), 2017. 2

[48] Desmond Ong, Zhengxuan Wu, Tan Zhi-Xuan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. IEEE Transactions on Affective Computing, 2019.

[49] Rameswar Panda, Jianming Zhang, Haoxiang Li, JoonYoung Lee, Xin Lu, and Amit K. Roy-Chowdhury. Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, European Conference on Computer Vision (ECCV), 2018.

[50] 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.

[51] Jae Sung Park, Trevor Darrell, and Anna Rohrbach. IdentityAware Multi-Sentence Video Description. In European Conference on Computer Vision (ECCV), 2020.

[52] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. In British Machine Vision Conference (BMVC), 2015. 6

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems (NeurIPS), 2019. 6

[54] Robert Plutchik. A General Pscychoevolutionary Theory of Emotion. Theories of Emotion, pages 3–33, 1980.

[55] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Association of Computational Linguistics (ACL), 2019. 1, 2

[56] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A Local-to-Global Approach to Multi-modal Movie Scene Segmentation. In Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2, 3

[57] Zeeshan Rasheed and Mubarak Shah. Scene Detection in Hollywood Movies and TV Shows. In Conference on Computer Vision and Pattern Recognition (CVPR), 2003. 2

[58] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. IJCV, 123:94–120, 2017.

[59] Daniel Rotman, Dror Porat, and Gal Ashour. Optimal Sequential Grouping for Robust Video Scene Detection using Multiple Modalities. International Journal of Semantic Computing, 11(2):192–208, 2017. 2

[60] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65, 1987.

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115:211–252, 2015. 5

[62] Amy M. Schmitter. 17th and 18th Century Theories of Emotions. In The Stanford Encyclopedia of Philosophy, 2021.

[63] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In Association for the Advancement of Artificial Intelligence (AAAI), 2021. 2

[64] Sarath Sivaprasad, Tanmayee Joshi, Rishabh Agrawal, and Niranjan Pedanekar. Multimodal Continuous Prediction of Emotions in Movies using Long Short-Term Memory Networks. In International Conference on Multimedia Retrieval (ICMR), 2018. 2

[65] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In International Conference on Computer Vision (ICCV), 2019. 4

[66] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 6

[67] Makarand Tapaswi, Martin B¨auml, and Rainer Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV series. In Conference on Computer Vision and Pattern Recognition (CVPR), 2012. 2

[68] Makarand Tapaswi, Martin B¨auml, and Rainer Stiefelhagen. StoryGraphs: Visualizing Character Interactions as a Timeline. In Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 2, 3

[69] Makarand Tapaswi, Martin B¨auml, and Rainer Stiefelhagen. Book2Movie: Aligning Video scenes with Book chapters. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 2

[70] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2, 7

[71] Ha Thi Phuong Thao, BT Balamurali, Dorien Herremans, and Gemma Roig. AttendAffectNet: Self-Attention based Networks for Predicting Affective Responses from Movies. In International Conference on Pattern Recognition (ICPR), 2021.

[72] Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 23(2):97–115, 2001.

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Advances in Neural Information Processing Systems (NeurIPS), 2017. 3, 4

[74] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards Understanding HumanCentric Situations from Videos. In Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1, 2, 3, 5, 7

[75] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning Visual Emotion Representations From Web Data. In Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2

[76] Martin W¨ollmer, Angeliki Metallinou, Florian Eyben, Bj¨orn Schuller, and Shrikanth S. Narayanan. Context-sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling. In Interspeech, 2010. 2

[77] Chao-Yuan Wu and Philipp Kr¨ahenb¨uhl. Towards LongForm Video Understanding. In Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[78] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A Graph-based Framework to Bridge Movies and Synopses. In International Conference on Computer Vision (ICCV), 2019. 2

[79] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A Joint Sequence Fusion Model for Video Question Answering and Retrieval. In European Conference on Computer Vision (ECCV), 2018. 2

[80] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end Concept Word Detection for

Video Captioning, Retrieval, and Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[81] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multi-modal Multi-label Emotion Detection with Modality and Label Dependence. In Empirical Methods in Natural Language Processing (EMNLP), 2020.

[82] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both Contextand Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In International Joint Conference on Artificial Intelligence (IJCAI), 2019.

[83] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. IEEE Signal Processing Letters, pages 1499–1503, 2016. 6

[84] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 40(6):1452–1464, 2017. 5

[85] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In International Joint Conference on Natural Language Processing (IJCNLP), 2021.

[86] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In International Conference on Computer Vision (ICCV), 2015. 2

[87] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A Robustly Optimized BERT Pre-training Approach with Post-training. In Chinese National Conference on Computational Linguistics, 2021. 4, 5

[88] Srivastava, Dhruv, Aditya Kumar Singh, and Makarand Tapaswi. ”How You Feelin’? Learning Emotions and Mental States in Movie Scenes.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

3, 6