Enable Machines to Understand Human Emotions and Mental States

Jian Li
Illinoise Institute of Technology
CS577
jli232@hawk.iit.edu

September 27, 2023

Abstract

Comprehending the narrative of a movie necessitates a profound understanding of 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. Our solution, EmoTx, is a multimodal Transformer-based architecture capable of processing videos, multiple character inputs, and dialog utterances to make collective predictions. By harnessing the annotations from the MovieGraphs dataset, we aim to forecast not only classic emotions like happiness and anger but also other nuanced mental states such as honesty and helpfulness. Our 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, we establish the effectiveness of EmoTx. A closer examination of EmoTx'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 increasing application of LLM, the end-to-end NLP tasks in the past have become increasingly unified. 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.

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. Furthermore, if we can teach machines to accurately recognize human emotions, they can better understand and interact with humans.

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 EmoTx, a multimodal Transformer based architecture, it can combine videos, multiple characters, and conversations for joint prediction. By leveraging annotations from the MovieGraph dataset, 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 EmoTx to demonstrate the effectiveness of the model.

2. Related Work

2.1. Movie Understanding

The comprehension of movies has undergone a transformation. In recent years, it has shifted from categorizing individuals and recognizing them to delving into the narrative itself. Tasks like scene identification, question answering, movie captions with names, modeling interactions and/or relationships, alignment of text and video storylines, and even long-form video understanding have emerged as fascinating fields. Substantial advancements have been achieved thanks to datasets like Condensed Movie, MovieNet, the VALUE benchmark, and MovieGraphs. Building upon the annotations available in MovieGraph, we now pay our attention on an additional aspect of comprehending stories, one that complements the aforementioned directions: the identification of emotions and mental states exhibited by each character and the overall mood within a movie scene.

2.2. Visual Emotion Recognition

Historically relied on face-based recognition of Ekman's six classic emotions and popularized by datasets like MMI, CK, and CK+, experienced a decade ago with the emergence of challenging in-the-wild benchmarks such as EmotiW, FER, and AFEW, coinciding with the integration of deep learning approaches that achieved notable performance. Deviating from this established paradigm, the Emotic dataset introduced a novel perspective by incorporating 26 emotion labels for image-based emotion comprehension, emphasizing the contextual dimension. This innovative approach involved the fusion of facial features and context using techniques like two-stream CNNs and person detections with depth maps. Other avenues in emotion recognition encompass estimating valence-arousal from faces with limited context, learning representations through webly supervised data to mitigate biases, or enhancing them further through a joint text-vision embedding space. In contrast to these trends, our research is uniquely focused on the recognition of multi-label emotions and mental states within the context of movies, harnessing multimodal context at both scene and character levels.

2.3. Multimodal Datasets for Emotion Recognition

Recent developments have witnessed the adoption of multimodal datasets for emotion recognition. The Acted Facial Expressions in the Wild dataset focuses on predicting emotions from facial expressions but lacks contextual information. On the other hand, the Stanford Emotional Narratives Dataset incorporates participant-shared narratives of positive and negative life events, offering a multimodal perspective, albeit differing significantly from our emphasis on edited movies and narratives. In the realm of Emotion Recognition in Conversations (ERC), the Multimodal EmotionLines Dataset (MELD) stands out by attempting to estimate emotions for each dialogue utterance in TV episodes from Friends. Distinguishing itself from MELD, our work operates within the time-scale of cohesive story units, specifically movie scenes. Lastly, the Annotated Creative Commons Emotional DatabasE (LIRIS-ACCEDE) is the closest to our endeavor, as it provides emotion annotations for short movie clips. However, these clips are relatively brief (8-12 seconds), and the annotations are derived from the continuous valence-arousal space. In contrast to the aforementioned works, we also aim to predict character-level mental states and demonstrate the significance of video and dialogue context in achieving such labels.

2.4. Multimodal Emotion Recognition Methods

RNNs have a historical presence in Emotion Recognition in Conversations (ERC), often employed alongside graph networks for their efficacy in amalgamating audio, visual, and textual data. Building upon recent breakthroughs, Transformer architectures have also found their place in ERC applications. Augmenting these developments, external knowledge graphs contribute valuable commonsense information, while the integration of topic modeling with Transformers has yielded improved outcomes. Efforts have been made to address multi-label prediction, including the exploration of a sequence-to-set approach, although scalability issues arise with a growing number of labels. Although our approach leverages a Transformer for comprehensive modeling, our specific objective, which is to predict emotions and mental states for movie scenes and characters, distinguishes our work from traditional ERC. We have adapted and conducted comparative experiments with some of the methods mentioned earlier. In a related context, MovieGraphs employs emotion annotations to depict the evolution of emotions throughout an entire movie and for Temporal Emotion Localization. However, it is important to note that the former focuses on tracking a single emotion within each scene, while the latter presents an alternative, albeit intriguing, direction of research.

3. Preliminary Plan

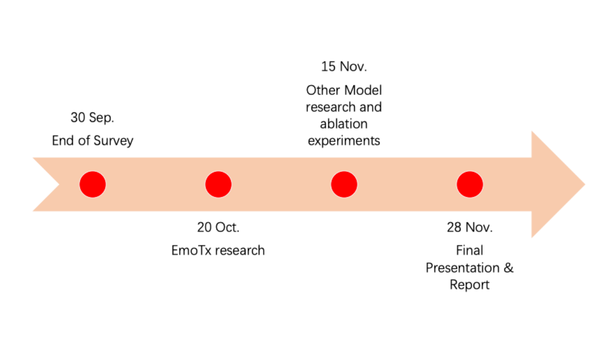

Figure 1. Plan of Project

References

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

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

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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. Multimodal 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.

[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.

[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.

[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.

[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.