Natural Language Processing Course Project Report

Multi Modal Methods for Emotional Recognition

Sahel Mesforoosh, Soroush Tabesh, Dorna Dehghani, Amin Kashiri, Fatemeh Tohidian, Seyyed Alireza Mousavizadeh
Sharif University of Technology, Spring 1401

ABSTRACT

Emotions could be recognized in a video by analyzing its script, frames, audio, or a combination of those called "multi-modal methods." In this project, we propose several algorithms for a combination of processing the script as a Natural Language Processing task and scenes of a video by studying people's face and pose in order to recognize their emotions. Our analysis reveals that this combination in video analysis can result better in emotional recognition rather than using only one of its aspects.

KEYWORDS

NLP, Vision

1 INTRODUCTION

Context and Motivation. Analyzing and detecting emotions in images and text has been a growing trend in machine learning in recent years. A *video* consists of a number of consecutive images and dialogues or monologues consisting of voice and text. In order to recognize emotions in a video, we can either analyze one of these aspects or combine the result of two or all aspect's analyses. We were curious about different algorithms for this task leading to various results; therefore we tested some of them in this project.

Research Problem. We considered scripts and frames as the constituents of a video and ignored the sound and tone of expressing speeches, since analyzing those needed additional knowledge which was out of this course and project's scope. Hence, for the purpose of recognizing the emotions of a video we observed the effect of both video's scripts and its scenes.

Related Work. At the beginning of this project, we decided to find and read related articles about this area. Here are their brief explanation:

- Indicated that emotion recognition is significantly better in response to multi-modal versus uni-modal stimuli. [1] However, this article's results were based on human trials and it did not include any machine learning methods.
- Propose a LSTM-based model that enables utterances to capture contextual information from their surroundings in the same video, thus aiding the classification process. [2]
- Proposed a context-level inter-modal attention framework for simultaneously predicting the sentiment and expressed emotions of an utterance, evaluated using CMU-MOSEI dataset. [3]
- Proposed a deep neural learning approach based on multiple modalities in which extracted features of an audiovisual data stream are fused in real time for sentiment classification. [4]

REFERENCES

[1] Silke Paulmann and Marc D Pell. 2011. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion* 35, 2 (2011), 192–201.

[2] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 873–883.

[3] Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *arXiv preprint arXiv:2002.08267* (2020).

[4] Atitaya Yakaew, Matthew N Dailey, and Teeradaj Racharak. 2021. Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks.. In *ICPRAM*. 442–451.

Supervised by Ehsaneddin Asgari.