Bad Guy Identifier - Project Plan
11-632 MCDS Fall 2018 Capstone

Yue Cao
yuec1@andrew.cmu.edu

Advisor: Prof. Alex Hauptmann
alex+@cs.cmu.edu

1 Abstract

As multimedia content grows ubiquitous, the ability to understand, analyze and exploit these data becomes crucial to all social media giants and streaming service providers. The traditional techniques, like adding metadata tags, recommendation based on collaborative filtering, are already a must for every serious competitor, and a deeper understanding of these content is the next battlefield.

One of the challenges is the subtle, nebulous nature of the tasks, i.e. problems which does not have an objectively correct answer, such as emotion recognition and prediction [2], or humor recognition [1]. In this project we take on a similar task that identifies villains in movie trailers, and classifying main characters into three categories: good, unsure/neural, and bad. This exploratory project bases on curated and self-labeled movie trailers fetched from YouTube, and use audio, video and image features. To the best of our knowledge, there is no prior work on this specific problem.

2 Vision

2.1 Introduction

The success and potential exhibited by deep learning in the last decade has encouraged researches on more complicated domains like autonomous driving and chatbots. These daily life scenarios incorporate many aspects of human senses, such as visual, auditory and even emotional, and requires some understanding of common sense. How well can a computer program master what is considered usual to us is still left open.

Synchronously, we are enjoying an unprecedented level of access to multimedia. Our mobile phone and numerous carefully-crafted applications makes it easy to indulge in a waterfall of images, audios and videos on the web. Our physical world is by nature multimedia, and the digital world is increasingly so. The ability to understand and analyze explicit and inexplicit multimedia data would be highly valuable for developing applications that serve human better.

Thus, motivated by what we can (possibly) do and what we want, this exploratory project takes on a question so naive that a 6-year-old human child could answer: who is the bad guy in this movie? Hopefully by the end of this project, we can have better understanding on multimedia machine learning and acquiring commonsense knowledge.

2.2 Proposed Solution

We propose a pipeline that takes in a movie trailer and outputs main characters with corresponding classification results (good, bad, unknown/neutral). The pipeline consists of 1) a hand-crafted main character identification module, which essentially clusters all faces appeared in the trailer and keep large ones that presumably represent main characters, 2) a pre-processing module that takes video and main character information and convert them into training units acceptable to the neural network, such as images and audio clips, each labeled with corresponding timestamp (or interval) and which character it represent, and 3) trained neural network and takes in training units and output classification results. Features we plan to use or explore for the third module includes images, audio clips, and if time permits, captions.

2.3 Relationship to Prior Work

As mentioned before, our project is the first in this domain as far as we know of. However, we will be piecing together many established works in various areas throughout the process.

- Face recognition. We rely heavily on face recognition for character identification. The Python library used in this project is `face_recognition`, built on state-of-the-art dlib deep learning module with a 99.38% accuracy on Labeled Faces in the Wild [3] benchmark.

3 Objectives

Build an offline classifier that takes in a short video (less than 5 minutes), identify main characters and classify them into three categories: good, bad, and uncertain. Ideally, the classifier should have a performance better than 50%.

4 Design

4.1 Data

We are curating our own dataset for this project. To avoid bias in movies selection, we use top 30 films from IMDB most featured films each year from 1995 to 2017, excluding animations, non-English movies and movies with less than $250,000$ ratings. With the movie title, release year and genre as meta-data, we use Google YouTube API to query "`<movie_title> <release_year>` official trailer", assume the first result is the desired trailer (later during manual labeling we verified that this assumption is 100% solid), and download the video in highest resolution possible with caption if available. The dataset has 271 movies.

Before manually labeling each movie's characters with good or bad tags, we run the first main character identification module in the pipeline to get a mapping from a character ID unique to that specific movie, to a list of images of faces belonging to that character. We then label the characters (represented by a list of faces) as `GOOD`, `BAD` or `N` (unknown/neutral) solely based on the impression from the trailer.

There are certain traits about this dataset that we want to highlight here.

1. Skewness of the data. There is much more good or neutral characters than bad ones. [NOTE TO SELF: ADD STATISTICS HERE AFTER LABEL COMPLETEION] This might encourages the model to always guess between `GOOD` and `N`. We plan to put more weight on label `BAD` in the loss function.

5 Timeline

- Milestone 1. [Sep. 1 - Sep. 17] Dataset preparation.
  - Setup developing environment.
  - Fetch, clean and label short videos.
  - Deliverable: a ready-to-use, clean dataset of videos, labeled with necessary and available feature at this stage, i.e. movie id and title.
- Milestone 2. [Sep. 17 - Oct. 15] Face recognition and character identification. Ground truth label.
  - Identify main characters in the clip. Manually label each character indicating whether it is a villian of the movie.
  - Deliverable: for each video, use face recognition and clustering to identify its main characters appeared. For each character, manually label it with ground truth. If possible, associated each character with screen shots containing its figure or face, and/or time intervals of its appearance in the video.

- Milestone 3. [Oct. 15 - Nov. 12] First stage training using only image features.
- Milestone 4. Incorporate audio and/or text (script).
    - If necessary, add features to each video including separate audio track, timestamped scripts.
- Milestone 5. [Nov. 26 - Dec. 10] Final report and project presentation.

References

[1] Dario Bertero et al. "Deep Learning of Audio and Language Features for Humor Prediction." In: *LREC*. 2016.

[2] Abhinav Dhall et al. "Video and image based emotion recognition challenges in the wild: Emotiw 2015". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM. 2015, pp. 423–426.

[3] Gary B Huang et al. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments". In: *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 2008.