A Deep Reinforcement Learning Framework for Identifying Funny Scenes in Movies

Haoqi Li¹, Naveen Kumar², Ruxin Chen², Panayiotis Georgiou¹
¹Signal Processing for Communication Understanding and Behavior Analysis Laboratory (SCUBA), University of Southern California
²SONY Interactive Entertainment LLC

haogili@usc.edu, naveen1.kumar@sony.com, ruxin.chen@sony.com, georgiou@sipi.usc.edu

Overview

This paper presents a novel deep Reinforcement Learning framework for classifying movie scenes based on affect using the faces images detected in the video stream as input.

Introduction

➢ Motivation
- Movie affective content analysis
- Limited labeled training data in supervised learning
- Understand movie scene’s affective content through a process of “trial and error”

➢ Goal
- Reinforcement Learning for movie scene affective prediction
- Explore decision space irrespective of fine resolution labels
- Make use of human input when available

➢ Challenges
- Complexity of human interaction
- Complexity of movie data (e.g., different angles of faces)

Movie clip/Sequence Label Prediction (e.g. Funny or Not Funny)

Facial Expression Embedding Model
- Pre-trained embedding on Kaggle FER
- 6 layer CNN with maxpooling

Facial Expression Embedding Model

Traditional RL framework

➢ Framework of our proposed method RL framework
- Idea: Design the agent action that can interact with the environment
- Deep Q-network
- The environment state s = original movie face embedding ε + predicted affective information α

A(t) = Q(s(t))
R(t) = H(A(1), ..., A(t), F)
α(t + 1) = G(A(1), ..., A(t))

Q Deep Q-network
A(t) Q-value over actions at time T
R(t) Reward generated by function H based on actions taken so far
G Mapping action to affective state

Movie data processing
- Face extraction
- ~7000 training sequences

Face Image Input

Experiments

Facial Expression Embedding model

➢ RL model for affective label prediction
- Binary classification
- Feedback as G: one-hot encoding of maximum affective state activation

α(t + 1) = \begin{cases} 1 & \text{if } A_i(t) \geq A_j(t) \forall j \\ 0 & \text{otherwise} \end{cases}

- Reward function (H):
  - Majority vote of frame-level results up to now.
  - Value of reward is critical (R_{maj} = \pm 0.05, R_{del} = \pm 1)

Discussion & Conclusion
- Smaller frame length (30fr) reduces complexity of modeling sequence
- Potential applications for human in the loop training process

Results

Funny classification accuracy

- Majority vote
- Last frame accuracy

Dataset

- 18 movies form SONY Pictures
- Total 1471 scenes with scene level tone labels: funny, calm, exciting etc.
- Simplify the task to funny scene binary classification (sparse labels)