Multimodal Speech Emotion Recognition Using Audio and Text

Group J: Jan Arvin Lapuz, Alyssa Lim, Le Van Nguyen, Ramil Zabala

October 12, 2020

Presentation Outline

Presentation Outline

Project Overview Revisited

- 2018 IEEE Spoken Language Technology Workshop (SLT)
- Rank 14 in Computational Linguistics Category on Google Scholar
- Speech Emotion Recognition
- Dual Recurrent Neural Networks (RNNs)
  - Speech Data
    - Mel Frequency Crystal Coefficients (MFCC)
    - Prosodic Features
  - Text Data
    - Word Tokens
- Emotion Categories: Angry, Happy, Sad Neutral

Presentation Outline

Setting Up the Environment

Environment: Google Colab
Requirements:

- tensorflow==1.4 (tested on cuda-8.0, cudnn-6.0)
- python==2.7
- scikit-learn==0.20.0
- nltk==3.3

Limitation on VM:

- Data size limitation
- Crash

Presentation Outline

Replication Results on Original Dataset

- Pre-trained model not available but codes and model parameters are available on GitHub:
  https://github.com/david-yoon/multimodal-speech-emotion
- Replication Results:

Model	Replication Accuracy	Research Accuracy
Text Only	62.8%	63.5%
Audio Only	55.7%	54.6%
Multimodal	71.0%	71.8%
Multimodal-Attention	48.5%	69.0%

Presentation Outline

Update on the New Dataset

Approach 1: Data from another Research

- RAVDESS Dataset with 8 different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised)
- TESS Dataset with 7 different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral)

Approach 2: Create data from Scratch

- Download Youtube audio and captions (completed)
  - (Optional) Setup a Google Cloud Speech API account (completed)
  - (Optional) Feed into Google Cloud Speech API to create the text
- Feed into openSMILE Toolkit to extract the MFCC and Prosodic features (to-do)
- Investigate on feeding the text script into Amazon Turk to label with emotion tags (to-do)