Learning Cross-modal Audiovisual Representations with Ladder Networks for Emotion Recognition

Lucas Goncalves, Carlos Busso

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

Features and Performance Analysis for Emotion Recognition

CONCLUSIONS

Corpus

CREMA-D corpus
- Contains videos of subjects saying sentences while displaying pre-defined emotions
- Corpus was collected from an ethnically and racially diverse group
  - 91 actors (48 male and 43 female)
  - Contains 7,442 clips
  - 6-class problem: anger, happiness, sadness, fear, disgust, neutral

Data partition:
- 70% train set
- 15% development set
- 15% test set
Speaker-independent splits:
- No speaker overlap in train, development, and test sets

Proposed Framework

Components:
- Representations from a backbone network
- Unsupervised auxiliary tasks with multimodal ladder networks
- Cross-modal skip connections between the encoder and the decoder
- The audiovisual ladder network takes as input and , to be processed by the cross-modal transformer

Training Objective:

Ablation experiments:
Both ladder network mechanisms are important for the overall performance of the model

Visual Data Preparation
- Extract faces from videos at the frame level
- Normalize pixel intensities within the range [-1, 1]
- Resize the images to a predetermined dimension of 224x224x3
- Facial feature representations extract from VGG-face model
- Representations are 4096-dimensional per frame

Audio Data Preparation
- 65 low-level audio descriptors (LLDs) of the ComParE feature set
- It adds their first order derivates (Δ LLDs), creating a 130D sequence
- The features are extract using window lengths of 32ms with a step size of 16ms

Experimental Results

Macro | Micro
--- | ---
Architecture | Prec. | Rec. | F1 | Prec. | Rec. | F1
Our Model | 80.3 | 80.4 | 80.2 | 80.3 | 80.3 | 80.3
Baseline 1 | 76.5 | 75.7 | 75.5 | 75.7 | 75.7 | 75.7
Baseline 2 [1] | 71.6 | 71.0 | 70.6 | 71.0 | 71.0 | 71.0
Baseline 3 [2] | 60.6 | 57.8 | 56.3 | 58.0 | 58.0 | 58.0

We compare the results using a one-tailed matched paired t-test over the 20 results with p-value <0.05 to assert statistical significance

This work was supported by NSF under Grant IIS-1718944

REFERENCES: