Speech Emotion Recognition

Rajat Hebbar, Digbalay Bose, Somesh Sakriya

October 27, 2021

Overview

Problem Definition

- Design a deep neural network based system for estimating emotional content in the speech.

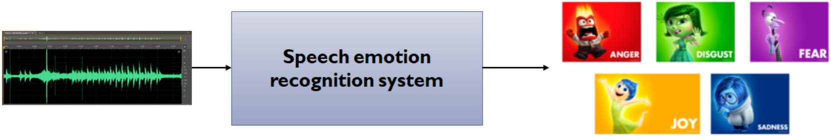


Figure: Outline of the speech emotion recognition system

- Can invoke other modalities like video, text for augmenting the capabilities of speech based emotion recognition algorithms

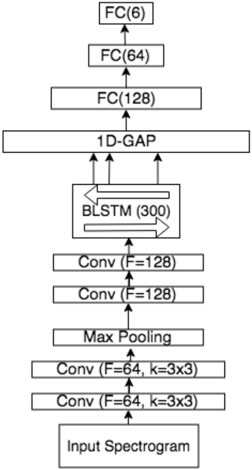
CMU-MOSEI

- **Details:**
  - 3.2k videos, 23k utterances, 1000 speakers
  - Sentence level M-Turk annotations:
    - Likert sentiment scale [-3,3] (-3: highly negative, +3: highly positive)
    - 6 emotion labels: **happiness, sadness, anger, fear, disgust, surprise**
    - presence of emotion x annotated on Likert scale [0,3] (0: no evidence of x, 3: highly x)
  - originally released Text-Audio-Visual features:
    - **Glove** embeddings
    - Facial landmarks, shape parameters, **face embeddings**, etc.
    - COVAREP acoustic features including 12 **MFCCs**, pitch, etc.
    - Words and audio **aligned** using P2FA forced alignment

Two stage training

- Instead of classification problem, emotion recognition posed as a regression problem because of the continuous scale used for labelling.
- Two stage training procedure involves the following:
  - **First stage:** Train the neural network for regression, where the regression output is a $1 \times k$ vector, where $k =$ number of distinct emotions.
  - **Second stage:** Freeze the first stage model layers till the embedding layer. Train k separate models for k emotions by considering as a single valued regression problem.
- Above procedure can be applied to any network and can be adopted for classification as well.

Audio based models



ga

Figure: CLDNN architecture modified for 4 second inputs of CMU-MOSEI

Audio based models

Layer	Support	Filt dim.	# filts.	Stride	Data size
conv1	7×7	1	96	2×2	254×198
mpool1	3×3	-	-	2×2	126×99
conv2	5×5	96	256	2×2	62×49
mpool2	3×3	-	-	2×2	30×24
conv3	3×3	256	256	1×1	30×24
conv4	3×3	256	256	1×1	30×24
conv5	3×3	256	256	1×1	30×24
mpool5	5×3	-	-	3×2	9×11
fc6	9×1	256	4096	1×1	1×11
apool6	1×n	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Figure: Original vgg vox architecture proposed in [1].67M parameters

Audio based models

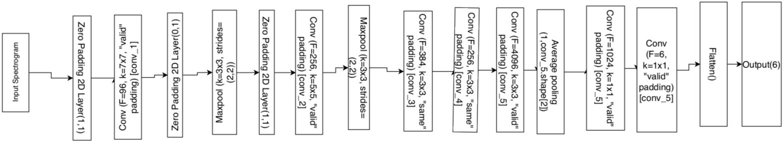


Figure: Vgg vox architecture modified for 4 second inputs of CMU-MOSEI

Features

- Audio : 64D log-mel spectrograms
  - 25ms window, 10ms shift
  - Inputs chopped to 4s segments (resulting in input size of 400×64)

Results

- Mean absolute error and mean squared error metrics are evaluated for the models associated with each emotion

Emotions	CLDNN_MSE	VGG_VOX_MSE	CLDNN_MAE	VGG_VOX_MAE
1	0.07	0.02	0.18	0.13
2	0.03	0.0164	0.11	0.084
3	0.03	0.0617	0.09	0.164
4	0.00	0.022	0.02	0.082
5	0.02	0.004	0.06	0.0466
6	0.01	0.0178	0.03	0.063

Table: MAE and MSE of the stage 2 models of VGG-vox and CLDNN

Multimodal Emotion-Lines Dataset (MELD)

- Dialogues from Friends TV Series
- Around 1400 dialogues consisting of 13000 utterances

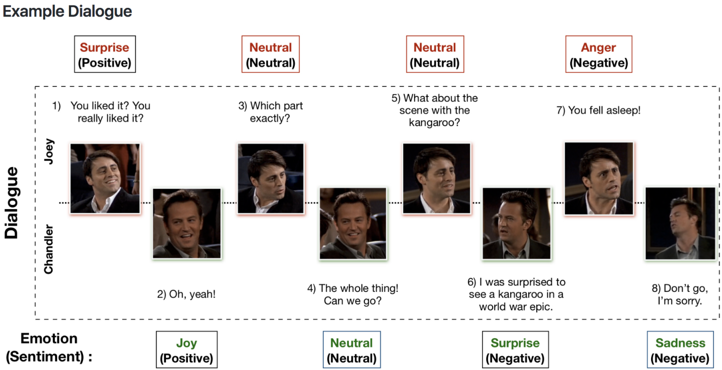


Figure: Single dialogue in MELD [2]

Multimodal Emotion-Lines Dataset (MELD)

- Multimodal multi-party conversational dataset
- 7 emotion classes, including Neutral
- Highly imbalanced classes

	Train	Dev	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

Figure: Dataset Distribution [2]

Statistics	Train	Dev	Test
# of modality	{a,v,t}	{a,v,t}	{a,v,t}
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.30	3.35	3.24
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Figure: Dataset Statistics [2]

Features

- Audio : 64D log-mel spectrograms
  - 25ms window, 10ms shift
  - Variable length input to network.
- Audio baseline: 6373 opensmile features (IS13-ComParE config)
- Text : 300D glove embeddings
  - each utterance padded to 50 words

Audio-only

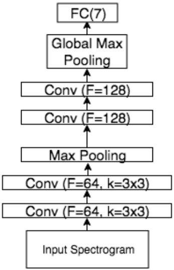


Figure: CNN architecture for variable length audio

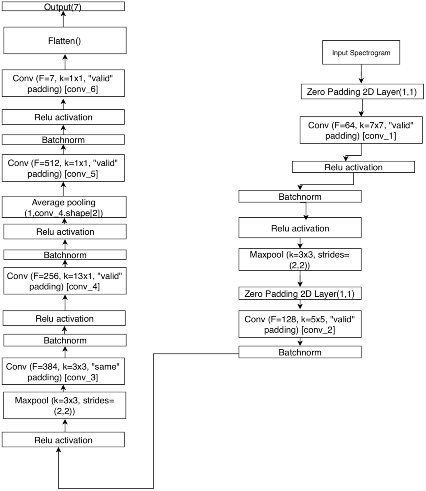


Figure: Modified vgg-vox

Text-only

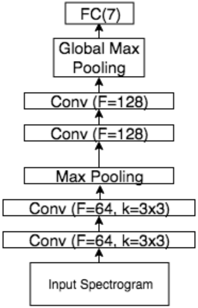


Figure: CNN architecture

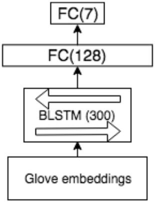


Figure: BLSTM architecture

Multimodal fusion

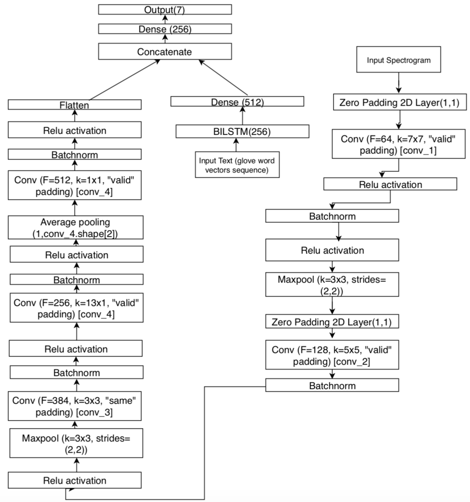


Figure: Modified VGG-vox with BLSTM for multimodal fusion

Training Parameters

- Handling class-imbalance with 1) class-weights, 2) balancing individual batches by oversampling minority class
- Batch size of 28 (4 samples/emotion/batch)
- Adam optimizer, Categorical cross-entropy loss
- Early stopping criterion with patience of 3-5
- Hyper-parameter tuning
    - Number of CNN-blocks, filter maps
    - Number of BLSTM units
    - Number and size of FC layers

Results

Audio

Model	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	w-avg
Poria et al. [2]	0.26	0.06	0.03	**0.16**	**0.62**	**0.15**	**0.19**	**0.39**
os-dnn	0	0	0	0.06	0.64	0	0	0.32
cnn-gmp	**0.29**	0.04	**0.06**	0.11	0.48	0.07	**0.19**	0.31
vgg-vox	0.24	**0.08**	0.04	0	**0.62**	0.06	0	0.34

Table: Per-class F1 score and weighted average score for audio

Text

Model	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	w-avg
Poria et al. [2]	**0.42**	**0.22**	**0.08**	**0.54**	**0.72**	**0.27**	**0.48**	**0.56**
cnn	0.31	0.02	0	0.34	0.53	0.16	0.4	0.4
blstm	0.37	0.14	0.1	0.52	0.67	0.24	**0.48**	0.53

Table: Per-class F1 score and weighted average score for text

Results

Multimodal Fusion

Model	**Anger**	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**	**w-avg**
Poria et al.	**0.43**	**0.24**	**0.09**	**0.54**	**0.77**	**0.24**	**0.51**	**0.59**
cnn-gmp + blstm	0.3	0.1	0.03	0.41	0.67	0.2	0.4	0.48
vgg-vox + blstm	0.38	0.15	0.07	0.5	0.72	**0.24**	0.48	0.55

Table: Per-class F1 score and weighted average score for audio

Future work

- Posing CMU-MOSEI as a multi-class classification problem
- Due to class imbalance in MELD, training of hierarchical networks
- Utilizing visual cues for improving performance
- Temporal convolutions networks for audio

References I

S. Albanie et al. "Emotion Recognition in Speech using Cross-Modal Transfer in the Wild". In: ACM Multimedia. 2018.

Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *CoRR* abs/1810.02508 (2018).