Study Of Dense Network Approaches For Speech Emotion Recognition

Mohammed Abdelwahab, Carlos Busso

THE UNIVERSITY OF TEXAS AT DALLAS

Multimodal Signal Processing Lab (MSP)

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

Motivation

Background:

- It is not clear the best configuration for deep learning structures in speech emotion recognition
  - Limited databases
  - No well defined network structure that works well across conditions

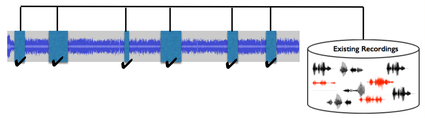
Our Work:

- We study various factors affecting performance in DNN for speech emotion recogniton
  - Amount of training data
  - Depth of the network
  - Use of residual networks
  - Activation
  - Batch normalization

Database and Features

The MSP-Podcast Corpus

- Emotional corpus collected at UT-Dallas
  - Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
  - Annotated on Amazon Mechanical Turk for emotional dimensions
  - V1.0: 20,045 labeled utterances (34 hrs, 15 min)
    - Test set: 6,069 segments from 50 speakers
    - Dev set: 2,226 segments from 15 speakers
    - Train set: 11,750 segments



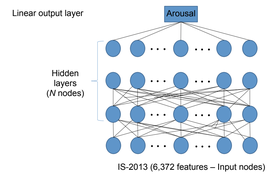
Acoustic Features

- Interspeech 2013 Computational Paralinguistic Challenge feature set (6,373 features)

Experimental Setting

- Models are trained to maximize the concordance correlation coefficient (CCC)

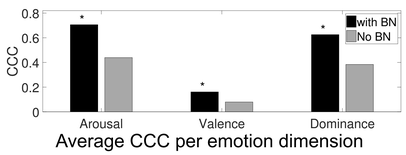
$$\rho_c(x,y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

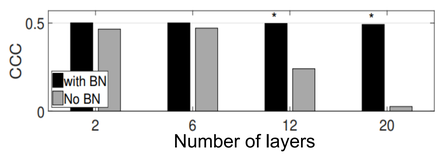


- Train networks with
  - 2, 6, 12 and 20 layers
  - 1k, 5.5k and 11.7k training samples
- Batch size of 256
- Learning rate of 1e-3 for first 100 epochs then linearly annealed to zero
- Dropout layers are introduced between layers
- Maxnorm of four as a weight constraint

Experiment Results

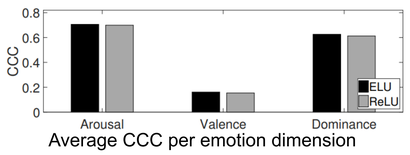
Batch Normalization

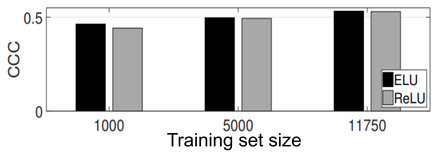




- Batch normalization is crucial to maintain consistent performance across different number of layers

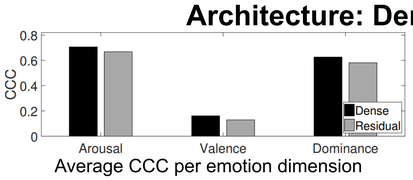
Activations: ReLU versus ELU

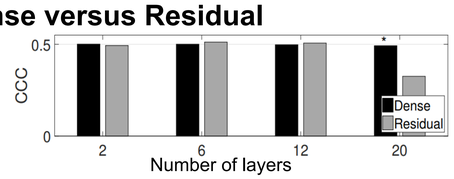




- ELU provides slightly better performance. However, differences are not statistically significant

Architecture: Dense versus Residual

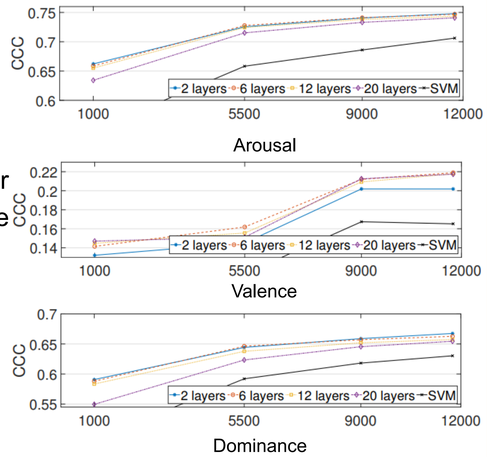


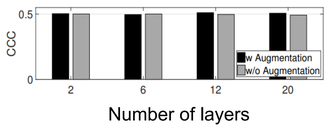


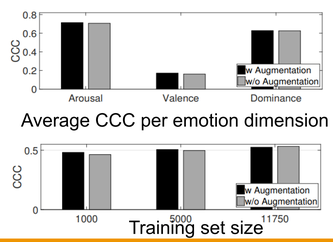
- Residual networks performs significantly worse when the training set size is small

Training Set Size

- As the training set size increases, the performance increases
- We expect to see further improvements with more data (ongoing effort)







Data Augmentation

- Speech rate data augmentation
- Data augmentation provides a small benefit for very deep layers when the training set size is small
- 20 layers trained with 1,000 turns
  - ccc=0.46 w/o data augmentation
  - ccc=0.48 w/ data augmentation

Conclusions

- This study explored the performance of regression models for arousal, valence and dominance
  - Number of layers
  - Size of the training set
  - Alternative activation functions
  - Batch normalization
  - Residual networks
  - Data augmentation
- Increasing the size of the training set improves prediction performance
- Batch normalization between layers is needed
- Data augmentation is a viable option when the training size is limited

Future Work

- We are annotating more data
- Explore using GANs for data augmentation
- Study end-to-end networks

This work was funded by NSF CAREER award IIS-1453781