Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Srinivas Parthasarathy, Carlos Busso

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

INTERSPEECH 2018
SEPTEMBER 2-6 | HYDERABAD, INDIA
HYDERABAD INTERNATIONAL CONVENTION CENTRE

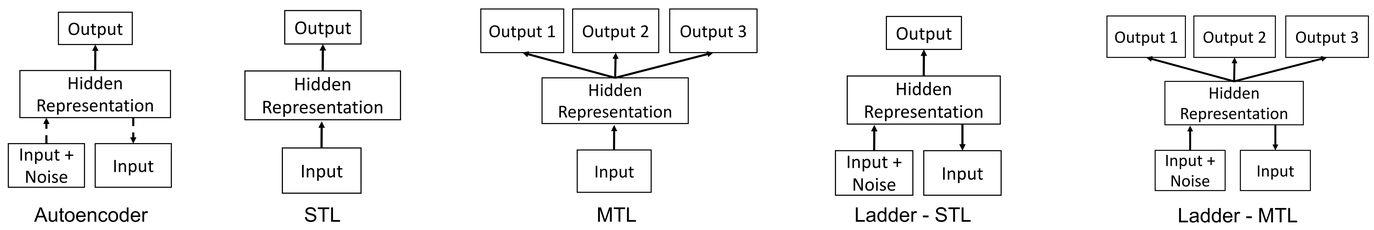
Motivation

Background

- Conventionally emotion attributes are individually modelled
- Jointly learning multiple attributes regularizes models better, improving performance
  - Requires labeled data - expensive

Our Work

- Regularization through unsupervised auxiliary tasks
  - Reconstruction of input and intermediate features
  - Ladder networks learn invariant features that are useful for the discriminative task



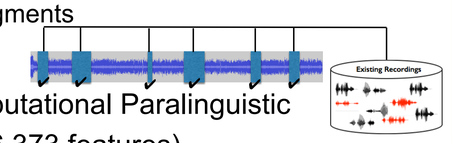
Database and Features

The MSP-Podcast Corpus

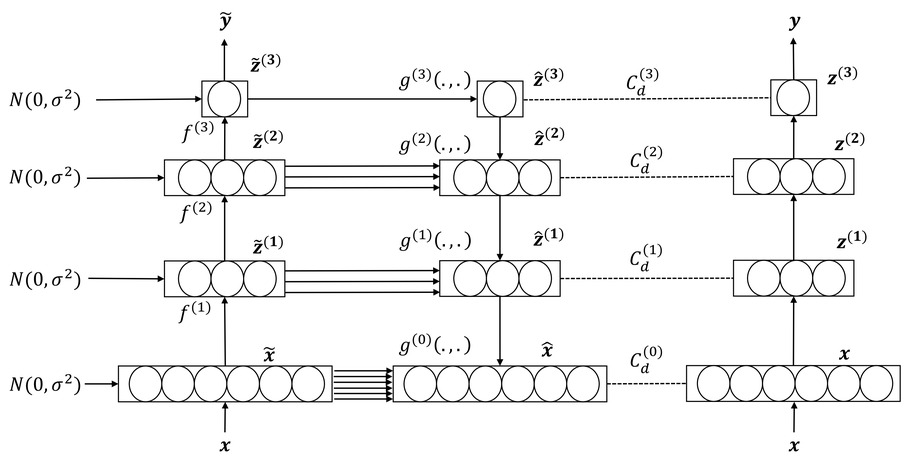
- Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
- Annotated on Amazon Mechanical Turk (aro., val., dom.)
- V1.1: 22,630 utterances with emotional labels (38h56m)
  - Test set: 7,181 segments from 50 speakers
  - Dev set: 2,614 segments from 20speakers
  - Train set: 12,835 segments

Acoustic Features

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



Ladder Networks for Emotion Recognition



Encoder

- Fully connected network
- Gaussian noise σ^2 added to each layer
- Representation from the final layer, $\tilde{\mathbf{z}}^{(L)}$, used for the supervised task
  - Further regularization
- A clean copy of encoder, $\mathbf{z}$, used as the reference for denoising

Decoder

- Denoising - reconstruct latent representation, $\hat{\mathbf{z}}$, from $\tilde{\mathbf{z}}$ using $\mathbf{z}$
- Denoising function, $g()$, combines top down information from decoder and lateral connection from corresponding encoder layer
  - Lower layers can be used to reconstruct data
  - Higher layers learn abstractive features for the discriminative task
  - Modeled by MLP with inputs $[u, \tilde{z}, u \odot z]$, u projection of layer above

$$C = C_c + \lambda_l \sum_l C_d^{(l)}$$

Ladder Network with Multi-task Learning

- Use both unsupervised and supervised auxiliary tasks
- Jointly predict activation, valence, dominance

$$C_{MTL} = \alpha C_{aro} + \beta C_{val} + (1 - \alpha - \beta) C_{dom}$$

Implementation

- Cost and Metric: Concordance correlation coefficient (CCC)
- Encoder and Decoder: 256 nodes

Evaluation and Conclusions

Task	Validation		
	Aro	Val	Dom
Autoencoder	0.358 ± 0.069	0.136 ± 0.141	0.305 ± 0.139
STL	0.778 ± 0.004	0.443 ± 0.008	0.722 ± 0.004
MTL	0.791 ± 0.003	$\mathbf{0.469 \pm 0.010}$	0.735 ± 0.003
Ladder+STL	$0.801 \pm 0.002^*$	0.443 ± 0.007	$0.742 \pm 0.002^*$
Ladder+MTL	$\mathbf{0.803 \pm 0.002}^*$	0.458 ± 0.004	$\mathbf{0.746 \pm 0.001}^*$

Task	Test		
	Aro	Val	Dom
Autoencoder	0.272 ± 0.136	-0.006 ± 0.012	0.284 ± 0.148
STL	0.737 ± 0.008	$\mathbf{0.292 \pm 0.007}$	0.670 ± 0.007
MTL	0.745 ± 0.008	0.285 ± 0.007	0.676 ± 0.006
Ladder+STL	$\mathbf{0.765 \pm 0.002}^*$	$\mathbf{0.294 \pm 0.007}$	$0.687 \pm 0.003^*$
Ladder+MTL	$0.761 \pm 0.002^*$	$\mathbf{0.289 \pm 0.008}$	$\mathbf{0.689 \pm 0.002}^*$

Conclusions

- All models better than conventional autoencoder
- MTL > STL : Benefits of supervised auxiliary task
- Ladder Network > Supervised tasks
- Ladder + MTL: Best performance for several conditions
  - Power of combining unsupervised and supervised auxiliary tasks

This work was funded by NSF CAREER award IIS-1453781