Interspeech 2018
2-6 September 2018, Hyderabad

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

Srinivas Parthasarathy and Carlos Busso

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

sxp120931@utdallas.edu, busso@utdallas.edu

Abstract

Recognizing emotions using few attribute dimensions such as arousal, valence and dominance provides the flexibility to effectively represent complex range of emotional behaviors. Conventional methods to learn these emotional descriptors primarily focus on separate models to recognize each of these attributes. Recent work has shown that learning these attributes together regularizes the models, leading to better feature representations. This study explores new forms of regularization by adding unsupervised auxiliary tasks to reconstruct hidden layer representations. This auxiliary task requires the denoising of hidden representations at every layer of an auto-encoder. The framework relies on *ladder networks* that utilize skip connections between encoder and decoder layers to learn powerful representations of emotional dimensions. The results show that ladder networks improve the performance of the system compared to baselines that individually learn each attribute, and conventional denoising autoencoders. Furthermore, the unsupervised auxiliary tasks have promising potential to be used in a semi-supervised setting, where few labeled sentences are available.

Index Terms: speech emotion recognition, regularization.

1. Introduction

Affective computing plays an important role in *human computer interaction* (HCI). Emotions are conventionally represented with discrete classes such as happiness, sadness, and anger [1–3]. An alternative emotional representation is through attribute dimensions such as arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong) [4–7]. These attribute dimensions provide the flexibility to represent multiple complex emotional behaviors, which cannot be easily captured with categorical descriptors. Furthermore, attribute dimensions can represent varying intensities of emotional externalizations which are lost when we use broad categorical descriptors such as "anger" (e.g., cold versus hot anger). Therefore, constructing models that can accurately predict attribute scores is an important research problem.

Conventionally emotional attributes are individually modeled [8], assuming that the attribute dimensions are orthogonal to each other. However, previous studies have shown significant correlation between different attributes [9]. This observation strongly suggests the need for jointly modeling multiple emotional attributes. An appealing way to do this task is through *multi-task learning* (MTL) where auxiliary tasks representing various emotional attributes are jointly learned [10, 11]. Learning the auxiliary task along with the primary task regularizes the learning process and the models generalize better.

While MTL generalizes the models, it requires labeled data (supervised auxiliary tasks). Regularization can also be performed with the help of unsupervised auxiliary tasks [12, 13]. One appealing approach is the reconstruction of intermediate feature representations using autoencoders [14]. Generally these unsupervised auxiliary tasks are performed as pre-training which is followed by normal supervised training of the primary task [12]. The main criticism of this approach is that the feature representation learned by the autoencoder does not necessarily support the supervised classification or regression tasks, which require the learning of invariant features that are discriminative for the task.

This paper proposes ladder networks for emotion recognition, showing its benefits for emotional attribute predictions. Ladder networks conveniently solve unsupervised auxiliary tasks along with supervised primary tasks [15, 16]. The unsupervised tasks (with respect to the primary task of predicting emotional attribute value) involve the reconstruction of hidden representations of a denoising autoencoder with lateral (skip) connections between the encoder and decoder layers. The representations from the encoder are simultaneously used to solve the supervised learning problem. The reconstruction of the hidden representations regularizes our primary regression task of predicting emotional attributes. The skip connections between the encoder and decoder ease the pressure of transporting information needed to reconstruct the representations to the top layers. Therefore, top layers can learn features that are useful for the supervised task, such as the prediction of emotional attributes. Interestingly, the framework also allows us to add multiple supervised tasks creating ladder networks with MTL structures.

This paper analyses the benefits of unsupervised auxiliary tasks to predict emotional attributes with ladder networks. We compare performance of these architectures with three baselines. The first baseline uses features from a denoising autoencoder that does not consider emotional labels to create the feature representation (i.e., unsupervised autoencoder). The second baseline is the conventional supervised *single task learning* (STL), where the emotional attributes are individually predicted. The third baseline is the MTL framework proposed by Parthasarathy and Busso [10], which does not use unsupervised auxiliary tasks (i.e., ladder networks). The performance shows that the architectures that use unsupervised auxiliary tasks consistently outperform the baselines. Furthermore, ladder networks with MTL structures have the best performance, improving the predictions of emotional attributes in the MSP-Podcast dataset.

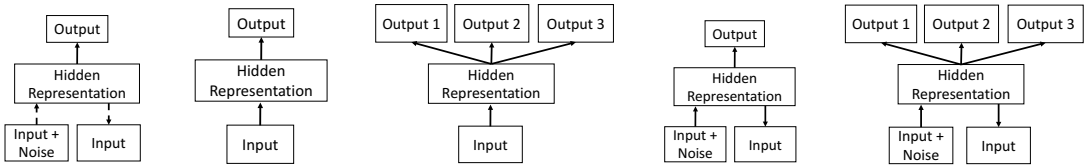
2. Background

2.1. Related Work

Few studies have focused on using auxiliary tasks to improve emotion recognition. Parthasarathy and Busso [10] proposed the joint learning of arousal, valence, dominance through *multi-task learning* (MTL). They showed significant improvement in performance when attributes are jointly predicted compared to

This work was funded by NSF CAREER award IIS-1453781.

10.21437/Interspeech.2018-1391



(a) Autoencoder (b) STL (c) MTL (d) Ladder-STL (e) Ladder-MTL

Figure 1: *Various architectures with supervised and unsupervised auxiliary tasks for emotion prediction. The first three structures are the baseline systems. The last two are the proposed models with ladder networks. While in (a) the reconstruction of the input is independent of the output prediction (shown with dashed arrows), in (d) the reconstruction in ladder network jointly considers the output prediction (shown with solid arrows).*

single task learning. Chen et al. [11] jointly learned arousal and valence for continuous emotion recognition, leveraging the relationship between the attributes. Their approach achieved improved performance for the affect subtask on the AVEC 2017 challenge. Xia and Liu [17] proposed a scheme to use the regression of emotional attributes as auxiliary task to aid the classification of emotional classes. Chang and Scherer [18] proposed a valence classifier that used predictions on arousal as a secondary task. Their MTL framework did not show any improvement over learning just the primary task. Kim et al. [19] proposed using gender and naturalness of data as auxiliary tasks to recognize emotions. Finally, Le et al. [20] proposed a classifier that continuously recognized emotional attributes. The attribute values were discretized using the k-means algorithm with $k \in \{4, 6, 8, 10\}$. The discretized attribute values were treated as classes which were jointly predicted with MTL.

The proposed approach builds upon ladder networks that effectively combine supervised classification or regression problems with unsupervised auxiliary tasks. Valpola [16] proposed using lateral shortcut connections to aid deep unsupervised learning. Rasmus et al. [15, 21] further extended this idea to support supervised learning. They included a batch normalization to reduce covariate shift. They also compared various denoising functions to be used by the decoder. Finally, Pezeshki et al. [22] studied the various components that affected the ladder network, noting that lateral connections between encoder and decoder and the addition of noise at every layer of the network greatly contributed to their improved performance. We describe in detail this framework in Section 3.2.

2.2. Database

This study uses the version 1.1. of the MSP-Podcast dataset [23]. The dataset contains emotionally colored, naturalistic speech from podcasts downloaded from audio sharing websites. The podcasts are processed and further split into smaller segments between 2.75s and 11s of duration. The segments are long enough so meaningful features can be extracted, and short enough so the emotional content does not change across the speaking turn. The dataset contains 22,630 (38h56m) audio segments. We manually identified the speaker identity of 18,991 sentences spoken by 265 speakers. The speaker information is used to create the train, development and test partitions. The partitions aim to have speaker independent sets. The test set contains 7,181 segments from 50 speakers, the development set contains 2,614 sentences from 20 speakers, and the train set includes the rest of the corpus (12,835 sentences). Audio segments are emotionally annotated on *Amazon Mechanical Turk* (AMT). The annotations are collected through perceptual evaluations from five or more raters. The emotional attributes are annotated with *self-assessment manikins* (SAMs) on a seven likert-scale for arousal (1-very calm, 7-very active), valence (1-very negative, 7-very positive), and dominance (1- very weak, 7 - very weak). The ground-truth labels for the attributes of a segment are the average scores provided by the evaluators.

3. Methodology

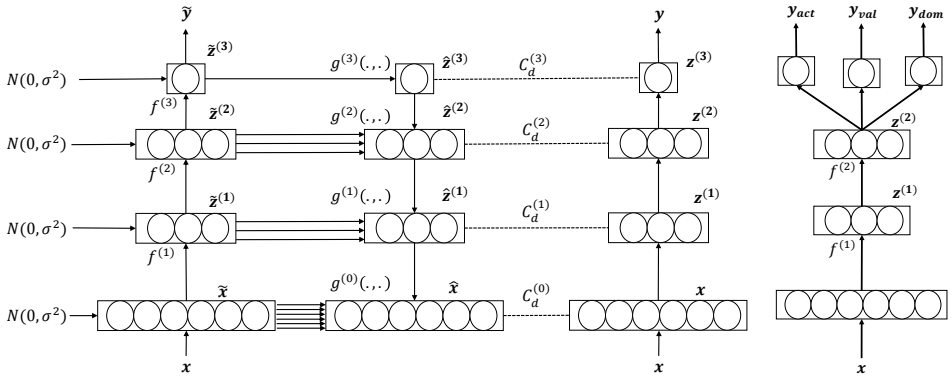
3.1. Proposed method

An important challenge while designing emotion recognition systems is to make models that generalize across different conditions [24]. Conventional models (Figure 1(b)) show poor performance when trained and tested on different corpora [10, 25]. Therefore, regularizing deep learning models is crucial in emotion recognition to find representations that are not overfitted to a particular domain. Regularization can be implemented with various approaches including early stopping criterion and dropout. The approach proposed in this study is to solve auxiliary tasks along with the primary task of predicting emotional attributes. By training models that are optimized for primary and auxiliary tasks, the feature representations are more general, avoiding overfitting. There are multiple ways to introduce auxiliary tasks to model emotion recognition. Previous studies for emotion recognition have focused on supervised auxiliary tasks, involving learning multiple emotion attributes [10] (Figure 1(c)), combining emotional classification problem with regression of emotional attributes [17], and learning other labels such as gender and age along with the emotion [19]. While these approaches are appealing, the supervised nature of the tasks require auxiliary labels for the training samples. Labels for emotional data commonly come from perceptual evaluations where multiple raters judge the emotional content of the stimuli. These evaluations are both expensive and time consuming. Therefore, annotating additional meta-information is not a feasible alternative. Its appealing to create unsupervised auxiliary tasks to regularize the network.

While traditional autoencoders (Figure. 1(a)) reconstruct input features in an unsupervised fashion, the intermediate latent representations are not trained for the underlying regression or classification task. This paper proposes to employ the unsupervised reconstruction of inputs as an auxiliary task to regularize the network, while optimizing the performance of an emotion regression system. We efficiently achieve this goal with ladder network architectures (Figure. 1(d)). The addition of an unsupervised auxiliary task not only regularizes the learning of the primary task, but also helps learning powerful discriminative representations of the input features. Furthermore, since there is no constraint on the primary task itself, we can combine it with other supervised auxiliary tasks to produce powerful models for predicting emotional attributes (Figure. 1(e)).

3.2. Ladder Networks

Ladder networks combine supervised primary task with unsupervised auxiliary tasks. The auxiliary tasks reconstruct the



(a) Ladder Network (b) MTL Network

Figure 2: *Architectures using auxiliary tasks for emotion attribute prediction. 2(a) illustrates the ladder network with unsupervised auxiliary tasks. 2(b) illustrates the MTL network that jointly learns multiple attribute values.*

hidden representations of a denoising autoencoder. The encoder of the autoencoder is simultaneously used to train the primary task at hand. The key aspect of the ladder network is the lateral connections between encoder and decoder layers. These skip connections allow the decoder to directly learn the representation from the encoder layer, bypassing the top layers of the encoder which can then learn representations that would help with the primary supervised task. Figure. 2(a) illustrates a conceptual ladder network with two hidden layers used for a regression task. Note that the true benefits of the ladder network is for semi-supervised setting where few labeled samples for the primary task are available in the target domain. However, this study focuses on the fully-supervised setting where we have emotional labels for every sample. The semi-supervised case is left as a future work.

Encoder: The encoder of the ladder network is a fully connected *multilayer perceptron* (MLP) network. A Gaussian noise with variance σ^2 is added to each layer of the noisy encoder (Figure. 2(a)). The representation from the final layer $\tilde{\mathbf{z}}^{(L)}$ of the encoder is used as the target for the supervised task. The decoder tries to reconstruct the latent representation $\hat{\mathbf{z}}$ at every layer using, as target, a clean copy of the encoder $\mathbf{z}$. Note that the supervised task, in this study the prediction of emotional attributes, is trained with the noisy encoder which further regularizes the supervised learning. However, for inference we use the predictions from the clean encoder. We describe the choice of hyper-parameters for the network in Section 4.3.

Decoder: The goal of the decoder is to denoise the noisy latent representations. The denoising function, $g()$, in Figure 2(a) combines top-down information from the decoder and the lateral connection from the corresponding encoder layer. With lateral connections, the ladder networks perform similar to hierarchical latent variable models. Lower layers are mostly responsible for reconstructing the input vector. This approach allows higher layers to learn more abstract, discriminative features needed for the supervised task. We use the denoising function proposed by Pezeshki et al. [22], modeled by an MLP with inputs $[u, \tilde{z}, u \odot z]$, where u is the batch normalized projection of the layer above and $\odot$ represents the Hadamard product. We use an MLP with 1 hidden layer and 4 hidden nodes to model the denoising function $g()$. The overall loss function is:

$$C = C_c + \lambda_l \sum_l C_d^{(l)} \tag{1}$$

where C_c is the supervised loss, $C_d^{(l)}$ is the reconstruction loss at layer l and λ_l is a hyper-parameter weight for the loss.

3.3. Ladder Network with Multi-task learning

While ladder network utilizes the reconstruction cost as an unsupervised auxiliary task, regularization can also be achieved through supervised tasks. With emotional attributes, MTL can achieve this goal by joint learning multiple attributes as proposed by Parthasarathy and Busso [10]. Figure 2(b) illustrates a two hidden layer MTL network that jointly predicts three emotional attributes: arousal, valence and dominance. The overall loss for the MTL network is given by

$$C_{MTL} = \alpha C_{aro} + \beta C_{val} + (1 - \alpha - \beta) C_{dom} \tag{2}$$

with $0 < \alpha, \beta < 1$ and $\alpha + \beta \leq 1$. Parthasarathy and Busso [10] showed that MTL networks perform better than STL networks for predicting emotional attributes. An appealing framework is to combine the proposed ladder network with MTL as shown in Figure 1(e). The supervised loss C_c in Equation 1 is replaced by the MTL loss C_{MTL} from Equation 2. This approach aims to combine supervised and unsupervised auxiliary tasks, creating an architecture that can learn powerful feature representations targeted for the prediction of emotional attributes (while the target attribute is the primary task, the other two attributes are auxiliary tasks).

4. Experimental Evaluation

4.1. Acoustic Features

We use the feature set introduced for the Computational Paralinguistic Challenge at Interspeech 2013 [26]. The feature extraction process involves two parts. First, *low level descriptors* (LLDs) are extracted on a frame-by-frame basis. The set includes *mel frequency cepstral coefficients* (MFCCs), fundamental frequency (F0) and energy. Various statistics, denoted as *high level features* (HLFs) are calculated over the LLDs. Overall, the feature set contains 6,373 features which we use for the various tasks in this study. The features were extracted with OpenSMILE [27].

4.2. Baselines

We compare our results with three baseline networks. Figure 1(b) shows the first baseline, which is a *deep neural network* (DNN) separately trained for each emotional attribute (i.e., STL). Figure 1(a) shows the second baseline, which learns

Table 1: *CCC values for the validation and test sets. The evaluations include 256 nodes per layer for arousal (*Aro*), valence (*Val*), and dominance (*Dom*). **Bold values** indicate the model(s) with the best performance per attribute (multiple cases are highlighted when differences are statistical significant). * indicates significant improvements of the ladder networks compared to all the baselines.*

Task	Validation			Test		
	Aro	Val	Dom	Aro	Val	Dom
Autoencoder	0.358 ± 0.069	0.136 ± 0.141	0.305 ± 0.139	0.272 ± 0.136	-0.006 ± 0.012	0.284 ± 0.148
STL	0.778 ± 0.004	0.443 ± 0.008	0.722 ± 0.004	0.737 ± 0.008	$\mathbf{0.292 \pm 0.007}$	0.670 ± 0.007
MTL	0.791 ± 0.003	$\mathbf{0.469 \pm 0.010}$	0.735 ± 0.003	0.745 ± 0.008	0.285 ± 0.007	0.676 ± 0.006
Ladder+STL	$0.801 \pm 0.002^*$	0.443 ± 0.007	$0.742 \pm 0.002^*$	$\mathbf{0.765 \pm 0.002}^*$	$\mathbf{0.294 \pm 0.007}$	$0.687 \pm 0.003^*$
Ladder+MTL	$\mathbf{0.803 \pm 0.002}^*$	0.458 ± 0.004	$\mathbf{0.746 \pm 0.001}^*$	$0.761 \pm 0.002^*$	$\mathbf{0.289 \pm 0.008}$	$\mathbf{0.689 \pm 0.002}^*$

feature representations using an autoencoder in an unsupervised fashion. The feature representations learned are then used as input for the supervised task. Unlike the ladder networks, the feature representation is independently learned from the supervised task. The objective of the autoencoder is to learn hidden representations that are useful for denoising the noise added to the features. All weights and activations of the encoder are frozen, and an output layer is then added on the top layer of the encoder for the prediction task. Figure 1(c) shows the third baseline, which uses MTL to jointly predict the three emotional attributes. Following the work of Parthasarathy and Busso [10], we train three MTL networks, one for each target emotional attribute, optimizing α and β in Equation 2 to maximize the performance for each attribute. By learning all three tasks, we obtain feature representations that generalize well across different conditions.

4.3. Implementation Details

All the deep neural networks in this study have two hidden layers with 256 nodes per layer. We use *rectified linear unit* (ReLU) activation at the hidden layers and a linear activation for the output layer. We optimize the networks using NADAM with a learning rate of $5e^{-5}$. The networks are implemented with dropout $p = 0.5$ at the input and first hidden layer. Following Trigeorgis et al. [28], we use the *concordance correlation coefficient* (CCC) as the loss function for training the models. We also use CCC to evaluate the models. All the hyper-parameters are set maximizing performance on the validation set, including the parameters for MTL (α, β). We train all the networks for 50 epochs with early stopping based on the results observed in the validation set. The best model on the validation set is then evaluated on the test set. All the models are trained 10 times with different random initializations, reporting the mean CCC.

For the ladder network, we add noise with variance σ^2=0.3 to each layer of the encoder. We conducted a grid search for the weights for the reconstruction loss with values $\lambda_l \in \{0.1, 1, 10, 100\}$. A value of $\lambda_l = 1$ gives the best result on the validation set. We use the mean squared error as the reconstruction cost and a dropout with probability $p = 0.1$ at the input layer and the first hidden layer. Notice that the original paper implementing ladder network did not use dropout. However, the high dimensionality of our feature vector and the violation of the independence assumption in our features motivate us to further regularize the network by adding dropout.

5. Results

Table 1 illustrates the mean CCC and standard deviation of the proposed architectures and the baselines for the validation and test sets. We analyze the performance of various models using the one-tailed t-test over the 10 trials, asserting significance if p-value$<$0.05. We highlight with an asterisk when the ladder models perform better than the baselines. First, we note that the results on the validation set are significantly higher than performance on the test set for all emotional attributes and all models. Comparing the performance of the different architectures on the test set, we note that all the networks perform better than the autoencoder baseline. This result shows that the feature representation learned by the autoencoder does not fit well for the primary regression task. Next, amongst the baselines we see that MTL models perform the best in all cases, except for valence on the test set. This result further confirms the benefits of supervised auxiliary tasks through the joint learning of multiple emotional attributes, shown in our previous study [10]. Observing the proposed architectures, we note that in almost all cases the ladder networks perform significantly better than the baselines and give the best performance. For valence, while MTL performs better than all other methods on the validation set, it does not translate to the test set where the Ladder+STL architecture has the best performance. The role of regularization for valence is an interesting topic which requires further study [29]. Amongst the proposed architectures, Ladder+MTL performs better than Ladder+STL in many cases. The results show the benefits of combining both unsupervised and supervised auxiliary tasks for predicting emotional attributes. Overall, the unsupervised auxiliary tasks greatly help to regularize our network improving the predictions and providing yet state-of-the-art performance on the MSP-Podcast corpus.

6. Conclusions

This work proposed ladder networks with multi-task learning to predict emotional attributes, achieving state-of-the-art performance on the MSP-Podcast corpus. We illustrated the benefits of using auxiliary tasks to regularize the network by combining unsupervised auxiliary tasks (ladder network) and supervised auxiliary tasks (multi-task learning). The emotional models generalize better across various conditions providing significantly better performance than the baseline systems.

There are many future directions for this work. First, the true potential of using unsupervised auxiliary tasks is in harnessing the almost unlimited amount of unlabeled data. We can extend the framework to work in a semi-supervised manner, where we can combine large number of unlabeled samples with fewer emotionally labeled samples, providing a more powerful feature representation. Second, the feature representations produced by the ladder network can be further studied and used as general input features for other emotion recognition problems such as classification of emotional categories. Finally, the auxiliary tasks could be extended to cover multiple modalities generalizing the models even more. The promising results in this study suggest that these extensions can lead to important improvements in emotion prediction performance.

7. References

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, October 2004, pp. 205–211.

[2] C. Lee, S. Yildirim, M. Bulut, C. Busso, A. Kazemzadeh, S. Lee, and S. Narayanan, "Effects of emotion on different phoneme classes," *J. Acoust. Soc. Am.*, vol. 116, p. 2481, 2004.

[3] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, vol. 4, Honolulu, HI, USA, April 2007, pp. 941–944.

[4] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April-June 2011.

[5] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.

[6] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[7] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.

[8] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 597–600.

[9] P. Lewis, H. Critchley, P. Rotshtein, and R. Dolan, "Neural correlates of processing valence and arousal in affective words," *Cerebral cortex*, vol. 17, no. 3, pp. 742–748, March 2007.

[10] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[11] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC 2017)*, Mountain View, California, USA, October 2017, pp. 19–26.

[12] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[13] M. Ranzato and M. Szummer, "Semi-supervised learning of compact document representations with deep networks," in *International Conference on Machine Learning (ICML 2008)*, Helsinki, Finland, July 2008, pp. 792–799.

[14] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[15] A. Rasmusi, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[16] H. Valpola, "From neural PCA to deep unsupervised learning," in *Advances in Independent Component Analysis and Learning Machines*, E. Bingham, S. Kaski, J. Laaksonen, and J. Lampinen, Eds. London, UK: Academic Press, May 2015, pp. 143–171.

[17] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, January-March 2017.

[18] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2746–2750.

[19] J. Kim, G. Englebienne, K. Truong, and V. Evers, "Towards speech emotion recognition "in the Wild" using aggregated corpora and deep multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1113–1117.

[20] D. Le, Z. Aldeneh, and E. Mower Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1108–1112.

[21] A. Rasmus, H. Valpola, and T. Raiko, "Lateral connections in denoising autoencoders support supervised learning," *CoRR*, vol. abs/1504.08215, pp. 1–5, April 2015. [Online]. Available: http://arxiv.org/abs/1504.08215

[22] M. Pezeshki, L. Fan, P. Brakel, A. Courville, and Y. Bengio, "Deconstructing the ladder network architecture," in *International Conference on Machine Learning (ICML 2016)*, San Juan, Puerto Rico, May 2016, pp. 2368–2376.

[23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.

[24] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[25] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," in *Speaker Classification II*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Berlin, Germany: Springer-Verlag Berlin Heidelberg, August 2007, vol. 4441, pp. 43–56.

[26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[28] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[29] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Interspeech 2018*, Hyderabad, India, September 2018.