ARTICLE IN PRESS
Neural Networks ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2017 Special Issue

Evaluating deep learning architectures for Speech Emotion Recognition

Haytham M. Fayek [a,*], Margaret Lech [a], Lawrence Cavedon [b]

[a] School of Engineering, RMIT University, Melbourne VIC 3001, Australia
[b] School of Science, RMIT University, Melbourne VIC 3001, Australia

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Affective computing
Deep learning
Emotion recognition
Neural networks
Speech recognition

ABSTRACT

Speech Emotion Recognition (SER) can be regarded as a static or dynamic classification problem, which makes SER an excellent test bed for investigating and comparing various deep learning architectures. We describe a frame-based formulation to SER that relies on minimal speech processing and end-to-end deep learning to model intra-utterance dynamics. We use the proposed SER system to empirically explore feed-forward and recurrent neural network architectures and their variants. Experiments conducted illuminate the advantages and limitations of these architectures in paralinguistic speech recognition and emotion recognition in particular. As a result of our exploration, we report state-of-the-art results on the IEMOCAP database for speaker-independent SER and present quantitative and qualitative assessments of the models' performances.

1. Introduction

In recent years, deep learning in neural networks has achieved tremendous success in various domains that led to multiple deep learning architectures emerging as effective models across numerous tasks. Feed-forward architectures such as Deep Neural Networks (DNNs) and Convolutional Neural Networks (ConvNets) have been particularly successful in image and video processing as well as speech recognition, while recurrent architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) RNNs have been effective in speech recognition and natural language processing (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015). These architectures process and model information in different ways and have their own advantages and limitations. For instance, ConvNets are able to deal with high-dimensional inputs and learn features that are invariant to small variations and distortions (Krizhevsky, Sutskever, & Hinton, 2012), whereas LSTM-RNNs are able to deal with variable length inputs and model sequential data with long range context (Graves, 2008).

In this paper, we investigate the application of end-to-end deep learning to Speech Emotion Recognition (SER) and critically explore how each of these architectures can be employed in this task. SER can be regarded as a static or dynamic classification problem, which has motivated two popular formulations in the literature to the task (Ververidis & Kotropoulos, 2006): *turn-based processing* (also known as static modeling), which aims to recognize emotions from a complete utterance; or *frame-based processing* (also known as dynamic modeling), which aims to recognize emotions at the frame level. In either formulation, SER can be employed in stand-alone applications; e.g. emotion monitoring, or integrated into other systems for emotional awareness; e.g. integrating SER into Automatic Speech Recognition (ASR) to improve its capability in dealing with emotional speech (Cowie et al., 2001; Fayek, Lech, & Cavedon, 2016b; Fernandez, 2004). Frame-based processing is more robust since it does not rely on segmenting the input speech into utterances and can model intra-utterance emotion dynamics (Arias, Busso, & Yoma, 2013; Fayek, Lech, & Cavedon, 2015). However, empirical comparisons between frame-based processing and turn-based processing in prior work have demonstrated the superiority of the latter (Schuller, Vlasenko, Eyben, Rigoll, & Wendemuth, 2009; Vlasenko, Schuller, Wendemuth, & Rigoll, 2007).

Whether performing turn-based processing or frame-based processing, most of the research effort in the last decade has been devoted to selecting an optimal set of features (Schuller et al., 2010). Despite the effort, little success has been achieved in realizing such a set of features that performs consistently over different conditions and multiple data sets (Eyben, Scherer et al., 2015). Thus, researchers have resorted to brute-force high-dimensional

* Corresponding author.
E-mail addresses: haytham.fayek@ieee.org (H.M. Fayek), margaret.lech@rmit.edu.au (M. Lech), lawrence.cavedon@rmit.edu.au (L. Cavedon).

http://dx.doi.org/10.1016/j.neunet.2017.02.013
0893-6080/

Please cite this article in press as: Fayek, H. M., et al., Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks (2017), http://dx.doi.org/10.1016/j.neunet.2017.02.013

features sets that comprise many acoustic parameters in an attempt to capture all variances (Tahon & Devillers, 2016). Such high-dimensional feature sets complicate the learning process in most machine learning algorithms, increase the likelihood of overfitting and hinder generalization. Moreover, the computation of many acoustic parameters is computationally expensive and may be difficult to apply on a large scale or with limited resources (Eyben, Huber, Marchi, Schuller, & Schuller, 2015). Therefore, it is highly pertinent to investigate the application of deep learning to SER to alleviate the problem of feature engineering and selection and achieve an SER with a simple pipeline and low latency. Moreover, SER is an excellent test bed for exploring various deep learning architectures since the task itself can be formulated in multiple ways.

Deep learning has been applied to SER in prior work, as discussed in Section 2. However, with different data subsets and under various experiment conditions involved in prior studies, it is difficult to directly compare various deep learning models. To the best of our knowledge, our work provides the first empirical exploration of various deep learning formulations and architectures applied to SER. As a result, we report state-of-the-art results on the popular Interactive Emotional Dyadic Motion Capture (IEMOCAP) database (Busso et al., 2008) for speaker-independent SER.

The remainder of this paper is divided into seven sections. In the following section, related work is reviewed, highlighting recent advances. In Section 3, a review of deep learning is presented focusing on the architectures and methods used in this paper. In Section 4, the proposed SER system is explained. In Section 5, the experimental setup is described, depicting the data, its preprocessing, the computational setup and the training recipe. Experiments performed and their results are presented in Section 6 and discussed in Section 7. Finally, the paper is concluded in Section 8.

2. Related work

Work on SER prior to 2011 is well reviewed in the literature (Ayadi, Kamel, & Karray, 2011; Petta, Pelachaud, & Cowie, 2011; Ververidis & Kotropoulos, 2006). Since DNNs displaced Gaussian Mixture Models (GMMs) for acoustic modeling in ASR (Hinton et al., 2012; Mohamed, Dahl, & Hinton, 2012), researchers have attempted to employ DNNs for other speech applications as well, and specifically for SER. Stuhlsatz et al. (2011) proposed a DNN Generalized Discriminant Analysis to deal with high-dimensional feature sets in SER, demonstrating better performance than Support Vector Machines (SVM) on the same set of features. In Li et al. (2013) a hybrid DNN—Hidden Markov Model (HMM) trained on Mel-Frequency Cepstral Coefficients (MFCCs) was proposed for SER and compared to a GMM—HMM indicating improved results. Han, Yu, and Tashev (2014) used a DNN to extract features from speech segments, which were then used to construct utterance-level SER features that were fed into an Extreme Learning Machine (ELM) for utterance-level classification outperforming other techniques. In Fayek, Lech, and Cavedon (2016a), a DNN was used to learn a mapping from Fourier-transform based filter banks to emotion classes using soft labels generated from multiple annotators to model the subjectiveness in emotion recognition which yielded improved performance compared to ground truth labels obtained by majority voting between the same annotators.

More recently, alternative neural network architectures for SER were also investigated. Mao, Dong, Huang, and Zhan (2014) used a ConvNet in a two-stage SER scheme that involves learning local invariant features using a sparse auto-encoder from speech spectrograms, processed using Principal Component Analysis (PCA) followed by salient discriminative feature analysis to extract discriminative features demonstrating competitive results. Tian, Moore, and Lai (2015) compared knowledge-inspired disfluency and non-verbal vocalization features in emotional speech against a feature set comprising acoustic parameters aggregated using statistical functionals, by using LSTM-RNNs as well as SVM, where the former was shown to yield better results given enough data.

This study differs from prior studies in several ways. We focus on a frame-based formulation for SER, aiming to achieve a system with a simple pipeline and low latency by modeling intra-utterance emotion dynamics. Moreover, most previous studies relied on some form of high-level features, while in this paper we strive for minimal speech processing and rely on deep learning to automate the process of feature extraction. Furthermore, we use uniform data subsets and experiment conditions promoting comparisons across various deep learning models, which has not been investigated in previous studies.

3. Deep learning: An overview

Deep learning in neural networks is the approach of composing networks into multiple layers of processing with the aim of learning multiple levels of abstraction (Goodfellow, Bengio, & Courville, 2016; LeCun et al., 2015). In doing so, the network can adaptively learn low-level features from raw data and higher-level features from low-level ones in a hierarchical manner, nullifying the over-dependence of shallow networks on feature engineering. The remainder of this section reviews the architectures, learning procedures and regularization methods used in this paper.

3.1. Architectures

The two most popular neural network architectures are the feed-forward (acyclic) architecture and the recurrent (cyclic) architecture (Schmidhuber, 2015). Feed-forward neural network architectures comprise multiple layers of transformations and nonlinearity with the output of each layer feeding the subsequent layer. A feed-forward fully-connected multi-layer neural network — also known as Deep Neural Network (DNN) — can be modeled by iterating over Eqs. (1) and (2):

$$\mathbf{h}^{(l)} = \mathbf{y}^{(l-1)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)} \tag{1}$$

$$\mathbf{y}^{(l)} = \phi(\mathbf{h}^{(l)}) \tag{2}$$

where $l \in \{1, \ldots, L\}$ denotes the lth layer, $\mathbf{h}^{(l)} \in \mathbb{R}^{n_o}$ is a vector of preactivations of layer l, $\mathbf{y}^{(l-1)} \in \mathbb{R}^{n_i}$ is the output of the previous layer $(l-1)$ and input to layer l, $\mathbf{W}^{(l)} \in \mathbb{R}^{n_i \times n_o}$ is a matrix of learnable weights of layer l, $\mathbf{b}^{(l)} \in \mathbb{R}^{n_o}$ is a vector of learnable biases of layer l, $\mathbf{y}^{(l)} \in \mathbb{R}^{n_o}$ is the output of layer l, $\mathbf{y}^{(0)}$ is the input to the model, $\mathbf{y}^{(L)}$ is the output of the final layer L and the model, and ϕ is a nonlinear activation function applied element-wise. The activation function used in this paper for feed-forward architectures is the Rectified Linear Unit (ReLU) as in Eq. (3) due to its advantages over other activation functions, such as computational simplicity and faster learning convergence (Glorot, Bordes, & Bengio, 2011).

$$\phi(z) = \max(0, z). \tag{3}$$

To provide a probabilistic interpretation of the model's output, the output layer L utilizes a softmax nonlinearity instead of the nonlinear function used in previous layers as in Eq. (4):

$$softmax(z_k) = \frac{e^{z_k}}{\sum_{k=1}^{K} e^{z_k}} \tag{4}$$

where K is the number of output classes.

A popular variant of the feed-forward neural network architecture is the Convolutional Neural Network (ConvNet) (LeCun et al., 1990), which leverages three ideas: sparse interactions; parameter sharing; and equivariant representations. This can be achieved

Please cite this article in press as: Fayek, H. M., et al., Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks (2017), http://dx.doi.org/10.1016/j.neunet.2017.02.013