Neural Networks 64 (2015) 39–48

Contents lists available at ScienceDirect

Neural Networks

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

2015 Special Issue

Deep Convolutional Neural Networks for Large-scale Speech Tasks

Tara N. Sainath[a,*], Brian Kingsbury[a], George Saon[a], Hagen Soltau[a],
Abdel-rahman Mohamed[b], George Dahl[b], Bhuvana Ramabhadran[a]

[a] *IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, United States*

[b] *Department of Computer Science, University of Toronto, United States*

ARTICLE INFO

Article history:
Available online 16 September 2014

Keywords:
Deep learning
Neural networks
Speech recognition

ABSTRACT

Convolutional Neural Networks (CNNs) are an alternative type of neural network that can be used to reduce spectral variations and model spectral correlations which exist in signals. Since speech signals exhibit both of these properties, we hypothesize that CNNs are a more effective model for speech compared to Deep Neural Networks (DNNs). In this paper, we explore applying CNNs to large vocabulary continuous speech recognition (LVCSR) tasks. First, we determine the appropriate architecture to make CNNs effective compared to DNNs for LVCSR tasks. Specifically, we focus on how many convolutional layers are needed, what is an appropriate number of hidden units, what is the best pooling strategy. Second, investigate how to incorporate speaker-adapted features, which cannot directly be modeled by CNNs as they do not obey locality in frequency, into the CNN framework. Third, given the importance of sequence training for speech tasks, we introduce a strategy to use ReLU+dropout during Hessian-free sequence training of CNNs. Experiments on 3 LVCSR tasks indicate that a CNN with the proposed speaker-adapted and ReLU+dropout ideas allow for a 12%–14% *relative improvement* in WER over a strong DNN system, achieving state-of-the art results in these 3 tasks.

1. Introduction

Recently, Deep Neural Networks (DNNs) have achieved tremendous success in acoustic modeling for large vocabulary continuous speech recognition (LVCSR) tasks, showing significant gains over state-of-the-art Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) systems on a wide variety of small and large vocabulary tasks (Dahl, Yu, Deng, & Acero, 2012; Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, & Kingsbury, 2012; Jaitly, Nguyen, Senior, & Vanhoucke, 2012; Kingsbury, Sainath, & Soltau, 2012; Seide, Li, & Yu, 2011). Convolutional Neural Networks (CNNs) (LeCun & Bengio, 1995; Lecun, Bottou, Bengio, & Haffner, 1998) are an alternative type of neural network that can be used to model spatial and temporal correlation, while reducing translational variance in signals.

CNNs are attractive compared to fully-connected DNNs for a variety of reasons. First, DNNs ignore input topology, as the input can be presented in any (fixed) order without affecting the performance of the network (LeCun & Bengio, 1995). However, spectral representations of speech have strong correlations in time and frequency, and modeling local correlations with CNNs, through weights which are shared across local regions of the input space, has been shown to be beneficial in other fields (LeCun, Huang, & Bottou, 2004). Second, DNNs are not explicitly designed to model translational variance within speech signals, which can exist due to different speaking styles (LeCun & Bengio, 1995). More specifically, different speaking styles lead to formants being shifted in the frequency domain, as well as variations in phoneme durations. These speaking styles require us to apply various speaker adaptation techniques to reduce feature variation. While DNNs of sufficient size could indeed capture translational invariance, this requires large networks with lots of training examples. CNNs on the other hand capture translational invariance with far fewer parameters by averaging the outputs of hidden units in different local time and frequency regions.

In fact, CNNs have been heavily explored in the image recognition and computer vision fields, offering improvements over DNNs on many tasks (Lawrence, 1997; LeCun et al., 2004). Recently, CNNs have been explored for speech recognition (Abdel-Hamid, Mohamed, Jiang, & Penn, 2012), also showing improvements over DNNs, however on a small vocabulary tasks with shallow

[*] Corresponding author. Tel.: +1 212 565 0000.
E-mail addresses: tsainath@google.com (T.N. Sainath), bedk@us.ibm.com (B. Kingsbury), gsaon@us.ibm.com (G. Saon), soltau@google.com (H. Soltau), asamir@cs.toronto.edu (A. Mohamed), gdahl@cs.toronto.edu (G. Dahl), bhuvana@us.ibm.com (B. Ramabhadran).

http://dx.doi.org/10.1016/j.neunet.2014.08.005
0893-6080/© 2014 Elsevier Ltd. All rights reserved.

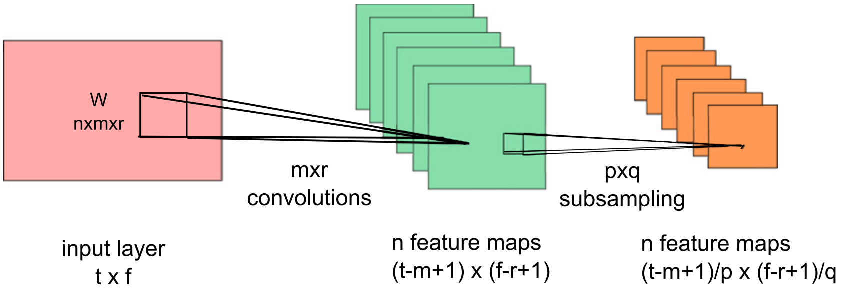

Fig. 1. Diagram showing a typical convolutional network architecture consisting of a convolutional and max-pooling layer. In this diagram, weights with the same line style are shared across all convolutional layer bands. Note this figure shows non-overlapping pooling, which is different than Abdel-Hamid et al. (2012).

networks. Specifically, Abdel-Hamid et al. (2012) introduced a novel framework to model spectral correlations where convolutional weights were shared over limited frequency regions, a technique known as limited weight sharing (LWS). One of the limitations of this LWS approach was that the network was limited to one convolutional layer, unlike most CNN work which uses multiple convolutional layers (LeCun et al., 2004). In this paper, we explore a spatial modeling approach similar to work done in the image recognition community, where convolutional weights are fully shared across all time and frequency components. This modeling approach, known as full weight sharing (FWS), allows for multiple convolutional layers and encourages deeper networks.

The first part of this paper explores the appropriate architecture for CNNs on LVCSR tasks. Specifically, we investigate how many convolutional vs. fully connected layers are needed, the filter size per convolutional layer, an appropriate number of hidden units per layer and a good pooling strategy. In addition, we compare the LWS proposed in Abdel-Hamid et al. (2012) to our FWS strategy.

The second part of this paper explores the best type of input feature to be used with CNN. Various speaker adaptation techniques have been shown to improve the performance of speech recognition systems. In this paper, we focus on how to incorporate feature-space maximum likelihood linear regression (fMLLR) (Gales, 1998) and identity vectors (i-vectors) (Saon, Soltau, Picheny, & Nahamoo, 2013), which do not exhibit locality in frequency, into the CNN framework through a joint CNN/DNN architecture (Sainath, Kingsbury, Mohamed, Dahl, Saon, Soltau, Beran, Aravkin, & Ramabhadran, 2013).

Finally, we investigate the role of rectified linear units (ReLU) and dropout (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012) for Hessian-free (HF) sequence training (Kingsbury et al., 2012) of CNNs. In Dahl, Sainath, and Hinton (2013), ReLU+dropout was shown to give good performance for cross-entropy (CE) trained DNNs but was not employed during HF sequence-training. However, sequence-training is critical for speech recognition performance, providing an additional relative gain of 10%–15% over a CE-trained DNN (Kingsbury et al., 2012). During CE training, the dropout mask changes for each utterance. However, during HF training, we are not guaranteed to get conjugate directions if the dropout mask changes for each utterance. Therefore, in order to make dropout usable during HF, we keep the dropout mask fixed per utterance for all iterations of conjugate gradient (CG) within a single HF iteration.

After analyzing the best CNN architecture, input feature set and ReLU, we then explore using CNNs on a 50 hr English Broadcast News (BN) task (Kingsbury, 2009). Naturally, our best DNN system offers a 13% relative improvement over the GMM/HMM, consistent with gains observed in the literature with DNNs vs. GMM/HMMs (Kingsbury et al., 2012). Comparing DNNs to CNNs, we find that a CNN hybrid system offers a 3% relative improvement over the hybrid DNN, whereas the joint CNN/DNN system which incorporates speaker adaptation and ReLU+dropout offers an 14% improvement. Finally, we explore the behavior of the joint CNN/DNN and ReLU+dropout on two larger scale tasks — namely a 300 hr Switchboard (SWB) task and a 400 hr BN task. We find that using the CNN with these improvements, we can obtain a 12% relative improvement over the DNN on SWB and a 16% relative improvement over the DNN on 400 hr BN.

The rest of this paper is organized as follows. The basic CNN architecture used in this paper is described in Section 2. An exploration of various weight-sharing and pooling strategies are discussed in Section 3, while input feature analysis is discussed in Section 4. Using ReLU+dropout for HF sequence training is discussed in Section 5. Results on three LVCSR tasks are presented in Section 6, Finally, Section 7 concludes the paper and discusses future work.

2. Basic CNN architecture

In this section, we describe the basic CNNs architecture and experimental setup used in this paper.

2.1. CNN description

A typical convolutional network architecture is shown in Fig. 1. First, we are given an input signal $\mathbf{V} \in \Re^{t \times f}$, where t and f are the input feature dimension in time and frequency respectively. A weight matrix $\mathbf{W} \in \Re^{(m \times r) \times n}$ is convolved with the full input V. The weight matrix spans across a small local time-frequency patch of size $m \times r$, where $m <= t$ and $r <= f$. This weight sharing helps to model local correlations in the input signal. The weight matrix has n hidden units (i.e., feature maps). Thus, overall the convolutional operation produces n feature maps of size $(t - m) \times (f - r)$.

After performing convolution, a max-pooling layer helps to remove variability in the time-frequency space that exist due to speaking styles, channel distortions, etc. Given a pooling size of $p \times q$, pooling performs a subsampling operation to reduce the time-frequency space to be $\frac{(t-m+1)p}{\times} \frac{(f-r+1)}{q}$.

Most CNN work in image recognition has the lower network layers be convolutional, while the higher network layers are fully connected. One goal of this paper is to determine an appropriate CNN architecture for speech tasks, including the number of convolutional vs. fully connected layers, hidden units and pooling strategy.

2.2. Experimental details

2.2.1. Data

We perform preliminary experiments to learn the behavior of CNNs for speech on a smaller task. Specifically, the acoustic models are trained on 50 h of data from the 1996 and 1997