Exploiting Nontrivial Connectivity for Automatic Speech Recognition

Marius Paraschiv
Corti, Copenhagen, Denmark
mp@cortilabs.com

Lasse Borgholt
Corti, Copenhagen, Denmark
lb@cortilabs.com

Tycho Max Sylvester Tax
Corti, Copenhagen, Denmark
tt@cortilabs.com

Marco Singh
Corti, Copenhagen, Denmark
ms@cortilabs.com

Lars Maaløe
Technical University of Denmark
Corti, Copenhagen, Denmark
lm@cortilabs.com

Abstract

Nontrivial connectivity has allowed the training of very deep networks by addressing the problem of vanishing gradients and offering a more efficient method of reusing parameters. In this paper we make a comparison between residual networks, densely-connected networks and highway networks on an image classification task. Next, we show that these methodologies can easily be deployed into automatic speech recognition and provide significant improvements to existing models.

1 Introduction

Deep neural networks have become the standard in end-to-end speech recognition systems. Models such as DeepSpeech 2 [1] have replaced cumbersome hand-engineered pipelines while being much more flexible in terms of deployment in practical applications. A more recent proposal is Wav2Letter [2, 3], a deep fully-convolutional network, providing near state-of-the-art results on speech recognition tasks.

Empirically, it was shown that deeper convolutional networks tend to offer improved performance. One of the earliest convolutional networks, LeNet5 [4], consisted of only 5 layers. Since then, deeper architectures, such as the VGG [5] architecture with 19 layers, have shown improved performance on image classification tasks. More recently, Highway Networks [6, 7] and Residual Networks [8], reaching over 1000 layers, continued to improve classification accuracy. One of the latest propositions in the family of very deep neural networks is the Densely Connected Convolutional Network [9, 10]. The last three models, from now on denoted as HighwayNETs, ResNETs and DenseNETs, respectively, are of particular interest.

A significant difference between the above three architectures and regular neural networks is their use of skip-connections (or nontrivial connections). ResNETs use identity connections that skip a block of layers, passing the input of a previous nonconsecutive layer to the current one. HighwayNETs are very similar, but make use of gated connections, an idea reminiscent of LSTMs [11] while DenseNETs connect every layer to the outputs of every previous layer, using identity connections. Given the performance of the above architectures on image recognition problems and the increasing use of fully-convolutional networks in ASR tasks, studying the effects of nontrivial connectivity on convolutional networks in the context of speech recognition is justified.

According to the Universal Approximation theorem [12, 13], a single-layer network with a large (but finite) number of units can act as an approximator for highly complex functions. By reducing layer width and increasing the overall number of layers, the network gains the ability to learn increasingly

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

abstract features. It has been proven both experimentally [8, 6, 9] and theoretically [14] that increased depth comes with improved performance.

Training deep networks brings forth a set of challenges, most notably the vanishing gradient problem. As a network increases in depth, it can happen that the gradients of the network's output with respect to the parameters of initial layers become very small. This is an obstacle that prevents efficient training of extremely deep networks. Adding identity or gated skip-connections provides shortcuts for gradient backpropagation, which facilitate network training. A second advantage to adding skip-connectivity is that this allows the network to better reuse features. Thus, we are also interested in net gains due to the more complex architecture exclusively while maintaining depth constant.

We start by training all three models with 56 and 100 layers on the CIFAR-10 dataset. The reason for this is that no direct comparison between these three architectures can be found in the literature. By maintaining the core of the network and all hyperparameters identical, and varying only the type of skip-connections, we can study the impact of these connections on the overall results. Then we show that similar performance gains can be achieved when training the networks for ASR, even without increasing network depth. While this comparison is far from proving the superiority of one architecture over the others, it does show that, by using more complex connectivity, gains can be achieved on two different tasks.

2 Method

This section provides a brief overview of the selected architectures and of some notable implementation particularities.

2.1 Residual Networks

Residual Networks [8] represent a slight variation from a traditional neural network due to their use of identity skip-connections. In this paper, skip-connections for both ResNETs and HighwayNETs jump over 2 layers.

The skip-connections of residual networks can be understood mathematically as a reformulation of the mapping to be learned. If the desired underlying mapping is $\mathrm{G(x)}$, representing the function one wishes to approximate, the skip-connections allow the residual block to learn the mapping $\mathrm{G(x) := F(x) + x}$ instead, in terms of a residual function $\mathrm{F(x)}$, representing the activation of the block layers and x representing the unchanged output of the previous network block.

2.2 Highway Networks

Another architecture that uses skip-connections is the HighwayNET [6]. The main difference between it and the ResNET is that a HighwayNET uses gated connections.

A regular network consists of a series of layers, each of which applies a non-linear transformation $\mathrm{H(x, W_H)}$ to the activation of the previous layer, $\mathrm{y = H(x, W_H)}$, where $\mathrm{W_H}$ are the weights parameterizing the non-linear function H, x is the output of the previous layer and y is the current layer activation. A HighwayNET contains two additional transformations, $\mathrm{T(x, W_T)}$ and $\mathrm{C(x, W_C)}$, which give rise to a composition rule : $\mathrm{y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C)}$.

The additional transformations T and C are known as the *transform gate* and the *carry gate*, respectively, and $\mathrm{W_T}$ and $\mathrm{W_C}$ represent their corresponding weights. The role of the transform gate T, is to learn how many of the features of the previous layer need to be passed through the nonlinearity of the current layer, and the role of the carry gate C is to determine how many of these features should remain unchanged. This extra degree of freedom provided to the network comes at the cost of additional parameters. In practice, this translates to slower convergence, compared to the ResNET. A method for reducing additional parameters is not to use a carry gate altogether, and to transform the above equation into, $\mathrm{y = H(x, W_H) \cdot T(x, W_T) + x \cdot (1 - T(x, W_T))}$.

In order to ensure that the network will use the alternate gate connection, one must initialize it with a negative bias, proportional to the depth of the network. Written in detail, the transform gate is of the form, $\mathrm{T(x)} = \sigma(\mathrm{W_T^\top \cdot x + b_T})$, where $\mathrm{b_T}$ is the gate bias, $\sigma(\mathrm{x})$ is the sigmoid function and $\mathrm{W_T^\top}$ is the transpose of the weight matrix, corresponding to the transform gate. As such, the product between

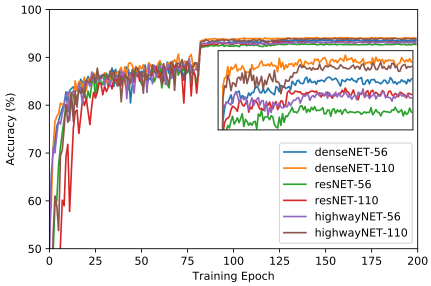

Figure 1: Classification accuracy in terms of the number of training epochs, for the CIFAR-10 dataset. At epochs 82 and 123, the learning rate is decreased by a factor of 10, and the enlarged plot section shows relative network performance, starting from epoch 82.

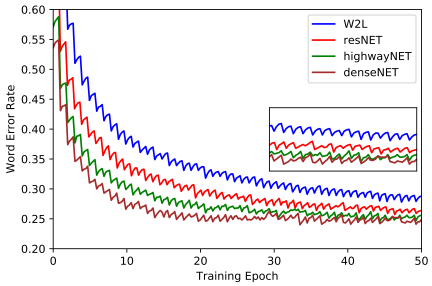

Figure 2: Validation word error rates for the baseline Wav2Letter model and the three derived networks, in the absence of a language model, for the first 50 training epochs.

the gate $\mathrm{T}(\mathrm{x})$ and $\mathrm{H}(\mathrm{x})$ is a product between a scalar and a tensor. The bias term must be initialized with a negative value, depending on network depth, such that the network is biased towards carry behavior. If a negative bias is not used, the network will tend to ignore the gated connections. In both residual and highway networks, we have used skip-connections over blocks of two layers. We have also experimented with blocks containing bottleneck layers, but with no noticeable improvements.

2.3 Densely Connected Networks

An extreme case of using skip-connections is the DenseNET [9]. Instead of having skip-connections over 2 layers, the network is divided into densely-connected blocks. Every layer in a block is connected to every previous layer.

Current activations are then concatenated to previous ones and passed on to the next layer. In this way, a layer is able to append newly learned features to already existing ones instead of modifying them. One can define DenseNET activations in terms of a composite function depending on the overall concatenated tensor of features, $\mathrm{y}_\ell = \mathrm{H}[\mathrm{x}_0, \mathrm{x}_1, ..., \mathrm{x}_{\ell-1}]$, where y_ℓ is the activation of layer ℓ.

Batch normalization applies scaling and bias to the input features, and by placing the convolution last, preceded by the nonlinearity and the batch normalization, every layer is free to apply a different scaling to the same features. This has been shown in Ref. [10] to have a considerable impact on the network's classification accuracy.

Due to the concatenation operation, performed after every layer, the concatenated feature vector has the tendency to grow very large, and performing convolutions with larger kernels quickly becomes impractical. For this reason, the network is split into densely-connected blocks, and before passing the concatenated features to the next block, we use a dimensionality reduction layer (also called a transfer layer) consisting of 1D convolutions with kernel size of 1.

3 Image Classification

For the task of image classification we use the CIFAR-10 dataset. For simplicity, the architectures are split into blocks of 2 convolutional layers each. The initial layer and first n blocks are identical, comprised of layers with 16 filters each. These are followed by two groups of n convolutional blocks with 32 and 64 filters, respectively.

Striding is performed in the first layer in the second and third group of convolutional blocks. We have experimented with $n = 9$ and $n = 18$, corresponding to 56 and 110-layer networks, respectively. The sequential blocks are followed by global average pooling and a softmax layer.

Table 1: Network classification accuracy on the CIFAR-10 dataset.

Architecture	Accuracy (%)
ResNET(56)	92.80 ± 0.13
HighwayNET(56)	93.29 ± 0.08
DenseNET(56)	93.37 ± 0.13
ResNET(110)	93.35 ± 0.19
HighwayNET(110)	93.96 ± 0.17
DenseNET(110)	94.04 ± 0.11

Table 2: Word error rates for the baseline and non-trivially-connected models.

Architecture	WER
W2L (our benchmark)	18.6
ResNET-W2L	17.2
HighwayNET-W2L	14.5
DenseNET-W2L	13.7

We present the test classification accuracies of HighwayNET, ResNET and DenseNET models, in terms of training epoch, in Fig.1. The training procedure is the same as in Ref. [8], for all models, and results are presented in Table 1.

4 Speech Recognition

For automatic speech recognition, the LibriSpeech dataset, consisting of audio-book fragments from the public domain and their corresponding transcriptions, has been used, with a partition of 360 hours of clean speech. We start with Wav2Letter [2] as our baseline model, trained on spectrogram inputs. It is a fully-convolutional architecture, consisting exclusively of 1-D convolutions in the time dimension.

Striding is performed by the first convolutional layer, followed by 7 identical layers with relatively small kernels. Onto these layers identity skip-connections, corresponding to ResNETs and DenseNETs, or gated skip-connections, corresponding to HighwayNETs, are added. This exact structure is preserved for all models, thus clearly showing the contributions brought by the extra connections. Finally, the features are passed through a large-kernel layer and a convolutional layer with a kernel size of 1.

We use the CTC loss function [15] and network output is decoded with the beam search procedure [16] utilizing a modified Kneser-Ney smoothed 4-gram model [17] trained on the original training data and the additional LibriSpeech language modeling data.

Table. 2 contains the word error rates for the baseline model and the three selected architectures on the 360-hour dataset. Our benchmark results for the baseline model are slightly higher than the WER of 14 that the original paper [2] reported, due to the fact that training method and hyperparameters used were not specified. As initially asserted, the performance gains through the use of nontrivial connections are task-independent as are the relative performance differences between the three models. A comparative plot of validation WER during training, for the four models is given in Fig.2.

Our results show that all three networks with skip-connectivity outperform the baseline W2L model, even while maintaining the same depth. Feature reuse plays an important role. ResNETs and HighwayNETs aggregate features of different layers by addition while the DenseNET concatenates all previous features, hence giving one layer access to the outputs of all previous ones in the block.

Continuing with our investigation, we trained the same models on the full LibriSpeech 1000-hour clean speech data, and our baseline model achieved a word error rate of **9.1** compared to the 9.4 reported in the original paper [2]. Due to time constraints, the results obtained by the models on this dataset, as well as the exploration of very deep networks, will be the subject of a future publication.

5 Conclusion

By adding skip-connectivity and improving feature reuse, a consistent improvement in network performance is seen, even without increasing network depth. The next step is to explore both the depth and width dimensions [18], in the context of speech recognition, and observe how such architectures are aided by more complex connections.

References

[1] Amodei D., Anubhai R., Battenberg E., Carl C., Casper J., Catanzaro B., Zhu, Z. et. al. Deep-Speech 2: End-to-end speech recognition in English and Mandarin, Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA. 2015

[2] Collobert R., Puhrsch C. and Synnaeve G. Wav2Letter: an End-to-End ConvNet-based Speech Recognition System. arXiv:1609.03193v2, 2016.

[3] Kunze J., Kirsch L., Kurenkov I., Krug A., Johannsmeier J. and Stober S. Transfer Learning for Speech Recognition on a Budget, arXiv:1706.00290, 2017.

[4] LeCun Y., Bottou L., Bengio Y. and Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, **86(11)**, (pp. 2278–2324), 1998.

[5] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, **115(3)**, (pp. 211–252), 2015.

[6] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. Advances in Neural Information Processing Systems, C. Cortes and N. D. Lawrence and D. D. Lee and M. Sugiyama and R. Garnett (editors), **28**, (pp. 2377-2385), 2015.

[7] Srivastava R. K., Greff K., Schmidhuber J. Highway Networks. ICML 2015 Deep Learning workshop, arXiv:1507.06228, 2015.

[8] He K., Zhang X., Ren S. and Sun J. Deep Residual Learning for Image Recognition. arXiv:1512.03385, 2015.

[9] Huang G., Liu Z., Weinberger K. Q. and Van Der Maaten, L. Densely Connected Convolutional Networks. arXiv:1608.06993v3, 2016.

[10] Pleiss G., Chen D., Huang G., Li T., Van Der Maaten L. and Weinberger, K. Q. Memory-Efficient Implementation of DenseNets. arXiv:1707.06990v1, 2017.

[11] Hochreiter S. and Schmidhuber J. Long short-term memory. Neural Computation. **9(8)**, (pp. 1735–1780), 1997.

[12] Cybenko, G. Approximations by superpositions of sigmoidal functions, Mathematics of Control, Signals, and Systems, **2(4)**, (pp. 303-314), 1989.

[13] Hornik K. Approximation Capabilities of Multilayer Feedforward Networks, Neural Networks, **4(2)**, (pp. 251–257) (1991).

[14] Eldan R., Shamir O. The Power of Depth for Feedforward Neural Networks. arXiv:1512.03965.

[15] Graves A., Fernandez S., Gomez F. and Schmidhuber J., Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning ACM, (pp. 369–376), 2006.

[16] Hannun A. Y., Maas A. L., Jurafsky D. and Ng A. Y. First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs, arXiv:1408.2873, 2014.

[17] `https://kheafield.com/code/kenlm/`

[18] Zagoruyko S. and Komodakis N. Wide Residual Networks. arXiv:1605.07146, 2016.