Rectifier Nonlinearities Improve Neural Network Acoustic Models

Andrew L. Maas AMAAS@CS.STANFORD.EDU
Awni Y. Hannun AWNI@CS.STANFORD.EDU
Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department, Stanford University, CA 94305 USA

Abstract

Deep neural network acoustic models produce substantial gains in large vocabulary continuous speech recognition systems. Emerging work with rectified linear (ReL) hidden units demonstrates additional gains in final system performance relative to more commonly used sigmoidal nonlinearities. In this work, we explore the use of deep rectifier networks as acoustic models for the 300 hour Switchboard conversational speech recognition task. Using simple training procedures without pretraining, networks with rectifier nonlinearities produce 2% absolute reductions in word error rates over their sigmoidal counterparts. We analyze hidden layer representations to quantify differences in how ReL units encode inputs as compared to sigmoidal units. Finally, we evaluate a variant of the ReL unit with a gradient more amenable to optimization in an attempt to further improve deep rectifier networks.

1. Introduction

Deep neural networks are quickly becoming a fundamental component of high performance speech recognition systems. Deep neural network (DNN) acoustic models perform substantially better than the Gaussian mixture models (GMMs) typically used in large vocabulary continuous speech recognition (LVCSR). DNN acoustic models were initially thought to perform well because of unsupervised pretraining (Dahl et al., 2011). However, DNNs with random initialization and sufficient amounts of labeled training data perform equivalently. LVCSR systems with DNN acoustic models have now expanded to use a variety of loss functions during DNN training, and claim state-of-the-art results on many challenging tasks in speech recognition (Hinton et al., 2012; Kingsbury et al., 2012; Vesely et al., 2013).

DNN acoustic models for speech use several sigmoidal hidden layers along with a variety of initialization, regularization, and optimization strategies. There is increasing evidence from non-speech deep learning research that sigmoidal nonlinearities may not be optimal for DNNs. Glorot et al. (2011) found that DNNs with rectifier nonlinearities in place of traditional sigmoids perform much better on image recognition and text classification tasks. Indeed, the advantage of rectifier networks was most obvious in tasks with an abundance of supervised training data, which is certainly the case for DNN acoustic model training in LVCSR. DNNs with rectifier nonlinearities played an important role in a top-performing system for the ImageNet large scale image classification benchmark (Krizhevsky et al., 2012). Further, the nonlinearity used in purely unsupervised feature learning neural networks plays an important role in final system performance (Coates & Ng, 2011).

Recently, DNNs with rectifier nonlinearities were shown to perform well as acoustic models for speech recognition. Zeiler et al. (2013) train rectifier networks with up to 12 hidden layers on a proprietary voice search dataset containing hundreds of hours of training data. After supervised training, rectifier DNNs perform substantially better than their sigmoidal counterparts. Dahl et al. (2013) apply DNNs with rectifier nonlinearities and dropout regularization to a broadcast news LVCSR task with 50 hours of training data. Rectifier DNNs with dropout outperform sigmoidal networks without dropout.

In this work, we evaluate rectifier DNNs as acoustic models for a 300-hour Switchboard conversational LVCSR task. We focus on simple optimization techniques with no pretraining or regularization in order to directly assess the impact of nonlinearity choice on

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

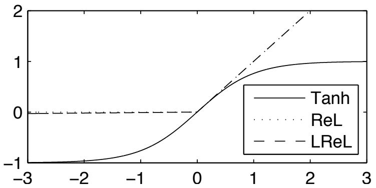

Figure 1. Nonlinearity functions used in neural network hidden layers. The hyperbolic tangent (tanh) function is a typical choice while some recent work has shown improved performance with rectified linear (ReL) functions. The leaky rectified linear function (LReL) has a non-zero gradient over its entire domain, unlike the standard ReL function.

final system performance. We evaluate multiple rectifier variants as there are potential trade-offs in hidden representation quality and ease of optimization when using rectifier nonlinearites. Further, we quantitatively compare the hidden representations of rectifier and sigmoidal networks. This analysis offers insight as to why rectifier nonlinearities perform well. Relative to previous work on rectifier DNNs for speech, this paper offers 1) a first evaluation of rectifier DNNs for a widely available LVCSR task with hundreds of hours of training data, 2) a comparison of rectifier variants, and 3) a quantitative analysis of how different DNNs encode information to further understand why rectifier DNNs perform well. Section 2 discusses motivations for rectifier nonlinearities in DNNs. Section 3 presents a comparison of several DNN acoustic models on the Switchbaord LVCSR task along with analysis of hidden layer coding properties.

2. Rectifier Nonlinearities

Neural networks typically employ a sigmoidal nonlinearity function. Recently, however, there is increasing evidence that other types of nonlinearites can improve the performance of DNNs. Figure 1 shows a typical sigmoidal activation function, the hyperboloic tangent (tanh). This function serves as the point-wise nonlinearity applied to all hidden units of a DNN. A single hidden unit's activation $h^{(i)}$ is given by,

$$h^{(i)} = \sigma(w^{(i)T}x), \tag{1}$$

where $\sigma(\cdot)$ is the tanh function, $w^{(i)}$ is the weight vector for the ith hidden unit, and x is the input. The input is speech features in the first hidden layer, and hidden activations from the previous layer in deeper layers of the DNN.

This activation function is anti-symmetric about 0 and has a more gradual gradient than a logistic sigmoid. As a result, it often leads to more robust optimization during DNN training. However, sigmoidal DNNs can suffer from the *vanishing gradient* problem (Bengio et al., 1994). Vanishing gradients occur when lower layers of a DNN have gradients of nearly 0 because higher layer units are nearly saturated at -1 or 1, the asymptotes of the tanh function. Such vanishing gradients cause slow optimization convergence, and in some cases the final trained network converges to a poor local minimum. Hidden unit weights in the network must therefore be carefully initialized as to prevent significant saturation during the early stages of training.

The resulting DNN does not produce a sparse representation in the sense of hard zero sparsity when using tanh hidden units. Many hidden units activate near the -1 asymptote for a large fraction of input patterns, indicating they are "off." However, this behavior is potentially less powerful when used with a classifier than a representation where an exact 0 indicates the unit is "off."

The rectified linear (ReL) nonlinearity offers an alternative to sigmoidal nonlinearites which addresses the problems mentioned thus far. Figure 1 shows the ReL activation function. The ReL function is mathematically given by,

$$h^{(i)} = \max(w^{(i)T}x, 0) = \begin{cases} w^{(i)T}x & w^{(i)T}x > 0 \\ 0 & \text{else} \end{cases}. \tag{2}$$

When a ReL unit is activated above 0, its partial derivative is 1. Thus vanishing gradients do not exist along paths of active hidden units in an arbitrarily deep network. Additionally, ReL units saturate at exactly 0, which is potentially helpful when using hidden activations as input features for a classifier.

However, ReL units are at a potential disadvantage during optimization because the gradient is 0 whenever the unit is not active. This could lead to cases where a unit never activates as a gradient-based optimization algorithm will not adjust the weights of a unit that never activates initially. Further, like the vanishing gradients problem, we might expect learning to be slow when training ReL networks with constant 0 gradients.

To alleviate potential problems caused by the hard 0 activation of ReL units, we additionally evaluate *leaky* rectified linear (LReL) hidden units. The leaky rectifier allows for a small, non-zero gradient when the unit is saturated and not active,

$$h^{(i)} = \max(w^{(i)T}x, 0) = \begin{cases} w^{(i)T}x & w^{(i)T}x > 0 \\ 0.01w^{(i)T}x & \text{else} \end{cases}. \tag{3}$$

Figure 1 shows the LReL function, which is nearly identical to the standard ReL function. The LReL sacrifices hard-zero sparsity for a gradient which is potentially more robust during optimization. We experiment on both types of rectifier, as well as the sigmoidal tanh nonlinearity.

3. Experiments

We perform LVCSR experiments on the 300 hour Switchboard conversational telephone speech corpus (LDC97S62). The baseline GMM system and forced alignments for DNN training are created using the Kaldi open-source toolkit (Povey et al., 2011). We use a system with 3,034 senones and train DNNs to estimate senone likelihoods in a hybrid HMM speech recognition system. The input features for DNNs are MFCCs with a context window of +/- 3 frames. Per-speaker CMVN is applied as well as fMLLR. The features are dimension reduced with LDA to a final vector of 300 dimensions and globally normalized to have 0 mean and unit variance. Overall, the HMM/GMM system training largely follows an existing Kaldi recipe and we defer to that original work for details (Vesely et al., 2013). For recognition evaluation, we use both the Switchboard and CallHome subsets of the HUB5 2000 data (LDC2002S09).

We are most interested in the effect of nonlinearity choice on DNN performance. For this reason, we use simple initialization and training procedures for DNN optimization. We randomly initialize all hidden layer weights with a mean 0 uniform distribution. The scaling of the uniform interval is set based on layer size to prevent sigmoidal saturation in the initial network (Glorot et al., 2011). The output layer is a standard softmax classifier, and cross entropy with no regularization serves as the loss function. We note that training and development set cross entropies are closely matched throughout training, suggesting that regularization is not necessary for the task. Networks are optimized using stochastic gradient descent (SGD) with momentum and a mini-batch size of 256 examples. The momentum term is initially given a weight of 0.5, and increases to 0.9 after 40,000 SGD iterations. We use a constant step size of 0.01. For each model we initially searched over several values for the step size parameter, $[0.1, 0.05, 0.01, 0.005, 0.001]$. For each nonlinearity type the value 0.01 led to fastest convergence without diverging from taking overly large steps. Network training stops after two complete passes through the 300 hour training set. Hidden layers contain 2,048 hidden units, and we explore models with varying numbers of hidden layers.

3.1. Impact of Nonlinearity

Our first experiment compares sigmoidal nonlinearity DNNs against DNNs trained using the two rectifier functions discussed in section 2. DNNs with 2, 3, and 4 hidden layers are trained for all nonlinearity types. We reserved 25,000 examples from the training set to obtain a held-out estimate of the frame-wise cross entropy and accuracy of the neural network acoustic models. Such a measurement is important because recognizer word error rate (WER) is only loosely correlated with the cross entropy metric used in our DNN acoustic model training. Table 1 shows the results for both frame-wise metrics and WER.

DNNs with rectifier nonlinearities substantially outperform sigmoidal DNNs in all error metrics, and across all DNN depths. Rectifier DNNs produce WER reductions of up to 2% absolute on the full Eval2000 dataset as compared with sigmoidal DNNs – a substantial improvement for this task. Furthermore, deeper 4 layer sigmoidal DNNs perform slightly worse than 2 layer rectifier DNNs despite having 1.76 times more free parameters. The performance gains observed in our sigmoidal DNNs relative to the GMM baseline system are on par with other recent work with DNN acoustic models on the Switchboard task (Yu et al., 2013). We note that in preliminary experiments we found tanh units to perform slightly better than logistic sigmoids, another sigmoidal nonlinearity commonly used in DNNs.

The choice of rectifier function used in the DNN appears unimportant for both frame-wise and WER metrics. Both the leaky and standard ReL networks perform similarly, suggesting the leaky rectifiers' non-zero gradient does not substantially impact training optimization. During training we observed leaky rectifier DNNs converge slightly faster, which is perhaps due to the difference in gradient among the two rectifiers.

In addition to performing better overall, rectifier DNNs benefit more from depth as compared with sigmoidal DNNs. Each time we add a hidden layer, rectifier DNNs show a greater absolute reduction in WER than sigmoidal DNNs. We believe this effect results from the lack of vanishing gradients in rectifier networks. The largest models we train still underfit the training set.

Table 1. Results for DNN systems in terms of frame-wise error metrics on the development set as well as word error rates (%) on the Hub5 2000 evaluation sets. The Hub5 set (EV) contains the Switchboard (SWBD) and CallHome (CH) evaluation subsets. Frame-wise error metrics were evaluated on 25,000 frames held out from the training set.

Model	Dev CrossEnt	Dev Acc(%)	SWBD WER	CH WER	EV WER
GMM Baseline	N/A	N/A	25.1	40.6	32.6
2 Layer Tanh	2.09	48.0	21.0	34.3	27.7
2 Layer ReLU	1.91	51.7	19.1	32.3	25.7
2 Layer LReLU	1.90	51.8	19.1	32.1	25.6
3 Layer Tanh	2.02	49.8	20.0	32.7	26.4
3 Layer ReLU	1.83	53.3	18.1	30.6	24.4
3 Layer LReLU	1.83	53.4	17.8	30.7	24.3
4 Layer Tanh	1.98	49.8	19.5	32.3	25.9
4 Layer ReLU	1.79	53.9	17.3	29.9	23.6
4 Layer LReLU	1.78	53.9	17.3	29.9	23.7

3.2. Analyzing Coding Properties

Previous work in DNNs for speech and with ReL networks suggest that sparsity of hidden layer representations plays an important role for both classifier performance and invariance to input perturbations. Although sparsity and invariance are not necessarily coupled, we seek to better understand how ReL and tanh networks differ. Further, one might hypothesize that ReL networks exhibit "mostly linear" behavior where units saturate at 0 rarely. We analyze the hidden representations of our trained DNN acoustic models in an attempt to explain the performance gains observed when using ReL nonlinearities.

We compute the last hidden layer representations of 4-layer DNNs trained with each nonlinearity type from section 3.1 for 10,000 input samples from the held-out set. For each hidden unit, we compute its empirical *activation probability* – the fraction of examples for which the unit is not saturated. We consider ReL and LReL units saturated when the activation is nonpositive, $h(x) \leq 0$. Sigmoidal tanh units have negative and positive saturation, measured by an activation $h(x) \leq -0.95$ and $h(x) \geq 0.95$ respectively. For the sigmoidal units we also measure the fraction of units that saturate on the negative asymptote ($h(x) \leq -0.95$), as this corresponds to the "off" position. Figure 2 shows the activation probabilities for hidden units in the last hidden layer for each network type. The units are sorted in decreasing order of activation probability.

ReL DNNs contain substantially more sparse representations than sigmoidal DNNs. We measure *lifetime sparsity*, the average empirical activation probability of all units in the layer for a large sample of inputs (Willmore & Tolhurst, 2001). The average activation probability for the ReL hidden layer is 0.11, more than a factor of 6 less than the average probability for tanh units (considering tanh to be active or "on" when $h(x) > -0.95$). If we believe sparse activation of a hidden unit is important for invariance to input stimuli, then rectifier networks have a clear advantage. Notice that in terms of sparsity the two types of rectifier evaluated are nearly identical.

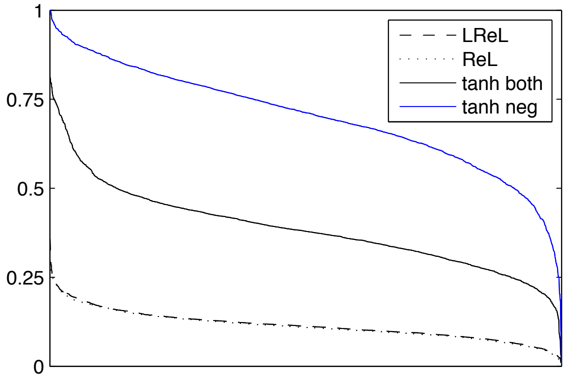

Figure 2. Empirical activation probability of hidden units in the final hidden layer layer of 4 hidden layer DNNs. Hidden units (x axis) are sorted by their probability of activation. In ReL networks, we consider any positive value as active ($h(x) > 0$). In tanh networks we consider activation in terms of not saturating in the "off" position ($h(x) > -0.95$, "tanh neg") as well as not saturating on either asymptote ($-0.95 < h(x) < 0.95$, "tanh both").

Sparsity, however, is not a complete picture of code quality. In a sparse code, only a few coding units represent any particular stimulus on average. However, it is possible to use the same coding units for each stimulus and still obtain a sparse code. For example, a layer with four coding units and hidden unit activation probabilities $[1, 0, 0, 0]$ has average lifetime sparsity 0.25. Such a code is sparse on average, but not *disperse*. Dispersion measures whether the set of active units is different for each stimulus (Willmore et al., 2000). A different four unit code with activation probabilities $[0.25, 0.25, 0.25, 0.25]$ again has lifetime sparsity 0.25 but is more disperse because units share input coding equally. We can informally compare dispersion by comparing the slope of curves in figure 2. A flat curve corresponds to a perfectly disperse code in this case.

We measure dispersion quantitatively for the hidden layers presented in figure 2. We compute the standard deviation of empirical activation probabilities across all units in the hidden layer [1]. A perfectly disperse code, where all units code equally, has standard deviation 0. Both types of ReL layer have standard deviation 0.04, significantly lower than the tanh layer's standard deviation of 0.14. This indicates that ReL networks, as compared with tanh networks, produce sparse codes where information is distributed more uniformly across hidden units. There are several results from information theory, learning theory, and computational neuroscience which suggest sparse-disperse codes are important, and translate to improved performance or invariance.

4. Conclusion

This work focuses on the impact of nonlinearity choice in DNN acoustic models without sophisticated pretraining or regularization techniques. DNNs with rectifiers produce substantial gains on the 300-hour Switchboard task compared to sigmoidal DNNs. Leaky rectifiers, with non-zero gradient over the entire domain, perform nearly identically to standard rectifier DNNs. This suggests gradient-based optimization for model training is not adversely affected by the use of rectifier nonlinearities. Further, ReL DNNs without pretraining or advanced optimization strategies perform on par with established benchmarks for the Switchboard task. Our analysis of hidden layer representations revealed substantial differences in both sparsity and dispersion when using ReL units compared with tanh units. The increased sparsity and dispersion of ReL hidden layers may help to explain their improved performance in supervised acoustic model training.

[1] This metric for dispersion differs from metrics in previous work. Previous work focuses on analyzing linear filters with Gaussian-distribted inputs. Our metric captures the idea of dispersion more suitably for non-linear coding units and non-Gaussian inputs.

References

Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 1994.

Coates, A.P. and Ng, A.Y. The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization. In *ICML*, 2011.

Dahl, G.E., Yu, D., Deng, L., and Acero, A. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on Deep Learning for Speech and Langauge Processing*, 2011.

Dahl, G.E., Sainath, T.N., and Hinton, G.E. Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout. In *ICASSP*, 2013.

Glorot, X., Bordes, A., and Bengio, Y. Deep Sparse Rectifier Networks. In *AISTATS*, pp. 315–323, 2011.

Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(November): 82–97, 2012.

Kingsbury, B., Sainath, T.N., and Soltau, H. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In *Interspeech*, 2012.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Veselý, K., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., and Stemmer, G. The kaldi speech recognition toolkit. In *ASRU*, 2011.

Vesely, K., Ghoshal, A., Burget, L., and Povey, D. Sequence-discriminative training of deep neural networks. In *Submission to Interspeech*, 2013.

Willmore, B. and Tolhurst, D.J. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255–270, 2001.

Willmore, B., Watters, P.A., and Tolhurst, D.J. A comparison of natural-image-based models of simple-cell coding. *Perception*, 29(9):1017–1040, 2000.

Yu, D., Seltzer, M.L., Li, J., Huang, J., and Seide, F. Feature Learning in Deep Neural Networks Studies on Speech Recognition Tasks. In *ICLR*, 2013.

Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q.V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., and Hinton, G.E. On Rectified Linear Units for Speech Processing. In *ICASSP*, 2013.