

Performance of Deep Learning Algorithms vs. Shallow Models, in Extreme Conditions - Some Empirical Studies

Samik Banerjee^(✉), Prateep Bhattacharjee, and Sukhendu Das

Indian Institute of Technology Madras, Chennai 600036, India
{samik,prateepb}@cse.iitm.ac.in, sdas@iitm.ac.in

Abstract. Deep convolutional neural networks (DCNN) successfully exhibit exceptionally good classification performance, despite their massive size. The effect of a large value of noise term, as irreducible error in Expected Prediction Error (EPE) is first discussed. Through extensive systematic experiments, we show how in extreme conditions the traditional approaches fare at par with large neural networks, which generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks trained for classification barely fit a random labeling of the training data as an extreme condition to learn. This phenomenon is quantitatively unaffected even if we train the CNNs with completely inseparable data. This can be due to large degree of corruption of the entire data by random noise or random labels associated with data due to observation error. We corroborate these experimental findings by showing that depth six CNN (VGG-6) fails to overcome large noise in image signals.

Keywords: Convolutional neural networks · Noise · Classification · SVM · EPE

1 Introduction

Convolutional neural network (CNN) models have become the state-of-the-art to solve hard classification problems and have significantly improved the accuracy for classifications. Traditional statistical machine learning methods require a human domain expert that can construct a good set of features as input dataset, while deep learning models waives the requirement of a hand crafted feature set. Hence it is more powerful and suitable for hard Artificial Intelligence tasks such as speech recognition or visual object classification. CNN based machine learning models can learn a hierarchy of features with complex and overlapping distributions on its own within the first few convolutional layers of CNN model, without any hand crafting of the raw input data. In the deepest layer of the model, a weighted set of selected features for each output is used to generate a prediction. Deep learning (DL) often outperforms traditional approaches [18],

for those hard classification problems in terms of performance accuracy, since the inevitable human error in feature selection can be easily avoided.

In the recent past, researchers have reported exciting results in various domains of computer vision and machine learning using convolutional neural networks (CNN). Still two questions [21] remains as the major interest about CNNs. The first question is about the power of the architecture – which classes of functions can it approximate well? The second question is; are good minima easier to find in deep rather than in shallow networks? In this paper we provide a set of experimental results for empirical evaluation, of both CNN and shallow methods, that puts some light into answers as to why and when deep networks fail or perform at par with the traditional shallow algorithms in complex and extreme conditions of data distributions. We define an extreme condition in labeled data, where the distributions overlap and nearest neighbors are randomly available with equal probability for all classes.

This paper compares shallow algorithms with deep networks, when we train both of them with different data distributions. The logic of the paper is as follows:

- Both shallow algorithms and deep networks are universal, that is they can approximate arbitrarily well any continuous function of d variables on a compact domain, but both of them fail to learn approximations with massive overlap.
- Many natural signals such as images and text require compositional algorithms that can be well approximated by Deep Convolutional Networks due to the basic properties of scalability and shift invariance. Of course, there are many situations that do not require shift invariant, scalable algorithms. For many functions that are not compositional do we expect any advantage of deep convolutional networks? [21]

Although difficult to prove this analytically, in spite of recent advances in concepts of Statistical Learning theory and deep artificial neural network analytics, we are forced to take the help of empirical studies to justify our logic and show-case the performance of CNN in extreme conditions. Recent reports by Thomas Poggio [21, 22] reveal that DL algorithms for CNN are scalable and shift-invariant, and can approximate functions better than shallow methods. A recent work [1] shows that deep-CNN cannot handle distribution variations in the context of Transfer Learning and Domain Adaptation. None of these report any results of performance for deep learning algorithms on extreme conditions. This has been the main motivation of our work.

2 Related Work

Most publications on DL these days start with Hinton’s back-propagation [17] and with Lecun’s convolutional networks [19] (see [18] for a nice review). The works proposed in [24–26] mainly deal with a multi-stage complex system, which take the convolutional features obtained from their model and then use PCA

(Principal Component Analysis) for dimensionality reduction, followed by classification using SVM. Other significant works in this area are [1, 4, 14, 16, 17, 20, 23], which use CNN for object recognition, video classification, image captioning and character recognition tasks.

Several popular Machine Learning (ML) techniques had been originally designed and proposed for the solution of binary classification problems. Traditionally among them, one can mention the Support Vector Machines (SVMs) [7], the Perceptron [13] and the RIPPER algorithms [5]. Many algorithms developed by the machine learning community focus on learning in pre-defined feature spaces. However, many real-world classification tasks exist that involve features where such algorithms could not be applied [11]. This paper also reports that Naive Bayes' outperformed C4.5 induction algorithm based on empirical evaluations. In such cases, the non-parametric classification algorithms like k-NN perform better. Currently very few work focuses on the traditional culture (predominant in Digital Signal/Image Processing and Communications field) of performance degradation in the presence of noise. One assumes that training may provide the system the power to overcome noise or overlapping distributions of data. But, is it really so? There lies the motivation and focus of our work.

3 Analysis Using Expected Prediction Error (EPE)

Based on the concept of Bias-variance decomposition [12], one can write $EPE(Y|X) = \sigma^2 + B^2 + V$; where σ^2 is the irreducible error due to noise and the two other terms (Bias and Variance) are model and dimension dependent. In the presence of a large amount of noise ($\sigma \gg 1$, say) in data, the error dominates resulting in degraded accuracy. In such a case, the prediction will generate random output in most cases. Figure 1 shows the EPE plot as demonstrated in [3]. For highly complex models (assuming for large orders, this metric is hypothetically equivalent to the 'power' of a deep-CNN) the error reaches 50% asymptotically. From a performance perspective, this can be visualized to be $EPE(Y|X) = B_n^2 + V_n$, where $B_n \gg B$ and $V_n \gg V$; *i.e.* a system randomly produces accurate results at most only for half of the cases ($\simeq 50\%$). We show this empirically using performance analysis of deep and shallow networks.

Noise is the unavoidable component of the loss, incurred independently of the learning algorithm. One always favor a more complex model if we assess that the 'goodness' of a model fits on the training data, as a more complex model will be able to capture small, random trends in the data due to noise [10]. Too large a model complexity/order causes overfitting. Overfitting occurs when an estimator is too flexible, allowing it to capture illusory trends in the data. These illusory trends are often the result of the noise in the observations. Its reasonable to assume that CNN models have extremely large complexity/order to deal with large variations of training samples. The above synthetic curve (Fig. 1) shows that for large order the test error will asymptotically touch 50%. The gap between total error and variance is due to noise in signal/data and perhaps so even for kernel based models (theoretical proof open for researchers). This

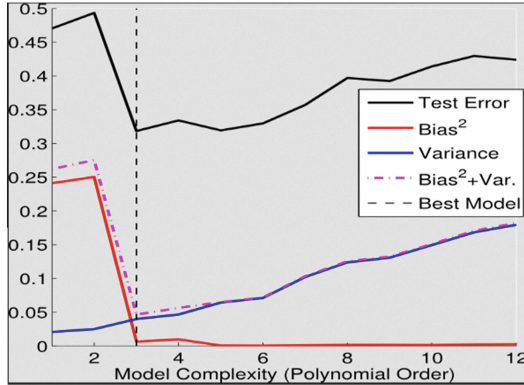


Fig. 1. Bias-variance trade-off (courtesy [3]).

trend should definitely be followed by all statistical classifiers [3] and even kernel models (e.g. SVM). The question remains, can CNN with its heavy training requirements overcome this and reduce the performance gap due to irreducible error. No formal proof is available, but one can empirically test to verify this, as for now.

For large noise, will performance of shallow models be weaker than CNN? How do you simulate this large noise? Either perturb the data samples or their class labels. We did this in an alternative way - make data points of two classes (binary classification problem) to overlap by a large extent. This in effect can be thought to be equivalently simulating a scenario of heavy noise on class labels. The noise level however cannot be quantified, other than a completely different mode of theoretical analytics, which is beyond the scope of the current paper. We quantify this noise by the amount of data overlap, with 100% data overlap indicating the maximum level of noise ($SNR \simeq 0$) where one may assume to have completely over-corrupted the input data (both class-wise distributions of training and testing data completely overlap with similar distributions).

Anyway, noise can play a significant role in the EPE as per bias-variance analysis. Indeed, according to Domingos [10], with the 0/1 loss the noise is linearly added to the error with a coefficient equal to $2P_D(f_D(x) = y) - 1$. Hence, if the classifier is accurate, that is, if $P_D(f_D(x) = y) \gg 0.5$, then the noise $N(x)$, if present, influences the expected loss. In the opposite situation also, with very bad classifiers, that is when $P_D(f_D(x) = y) \ll 0.5$, the noise influences the overall error in the opposite sense: it reduces the expected loss. If $P_D(f_D(x) = y) \approx 0.5$, that is if the classifier provides a sort of random guessing, then $2P_D(f_D(x) = y) - 1 \approx 0$ is the estimate of the noise in real data sets (as shown in [10]). A straightforward approach simply consists in disregarding it, but in this way we could overestimate the bias. Some heuristics are proposed in [15], but the problem remains substantially unresolved. Given these unanswered questions, we resort to empirical studies, as also suggested in [10] with noise.

4 Details of CNN and Datasets Used

The ‘*Wt. layer*’ in VGG-6 network [23] consists of a convolutional layer with the same *kernel size* and *number of filters* as proposed by the authors. The *FC layers* corresponds to the fully connected layers of the network. with two 50% *Dropout* layers. This moderately deep network is used for the empirical verification of the data. VGG-6 has been used for the object recognition tasks in [23].

4.1 Synthetic Scatter Dataset

Experimentations have been carried on two separate sets of data to study the effect of noise on the deep learning model compared with vanilla shallow supervised as well as non-parametric algorithms. The first set consists of two-class synthetically generated random data distributions, while the second consists of the Chars74K dataset [9].

Synthetic Data - To test the accuracy of the different algorithms, we have synthetically generated random data belonging to two classes. The distribution is considered to be elliptical for scatter generation (see Fig. 2) For a 2-class classification problem, dataset is generated as: (a) *50-dimensional data* and (b) *100-dimensional data*. The scatter is generated randomly and produced with 7 levels of overlap (difficulty) as described below (illustrated using 3-d scatters):

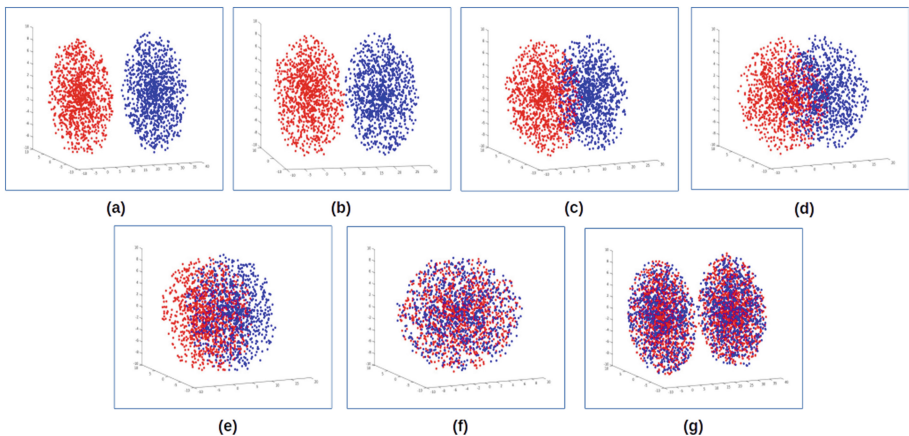


Fig. 2. Scatter Plots showing 3D Data: (a) Non-overlapping; (b) Barely touching; Overlap of: (c) 25%; (d) 50%; (e) 75%; (f) Fully Overlapping; (g) Random class labels (best viewed in color).

- **Non-overlapping (Fig. 2(a))** - The scatter for classes are completely separated, and they are separated from each other. This is the most easiest and favorable case.

- **Adjacent (Fig. 2(b))**- The data are completely separated, but they are touching each other at a single point.
- **Overlap (25%) (Fig. 2(c))**- 25% of the data from both classes overlap.
- **Overlap (50%) (Fig. 2(d))**- 50% of the data from both classes overlap.
- **Overlap (75%) (Fig. 2(e))**- 75% of the data from both classes overlap.
- **Completely overlapping (Fig. 2(f))**- The entire data from both classes (of decision regions) completely overlaps with each other. The class means, variances and boundaries are all identical.
- **Random (Fig. 2(g))**- The data is separated in 2 clusters as in non-overlapping case, but each cluster have a complete mixture of the two class labels randomly.

The extreme conditions are in Figs. 2(e)-(g). These are considered the most extreme and hard to solve by a machine. The datasets used are partitioned in a 10-fold cross-validation setting using $\{60 : 30 : 10\}$ as *train*, *test* & *validation* sets. For both the 50D and 100D data, 1 million data points/class are generated for the two class problem.

The Chars74K dataset [9] - Invariance of the CNN model to noise is further experimented on a benchmark real-world dataset for character recognition with 62992 synthesized characters from computer fonts (refer Fig. 3 for samples in the dataset). The dataset has 62 classes (0 – 9, A – Z, a – z). The VGG-6 model is trained on all these 62992 characters. The testing set is generated using Additive White Gaussian Noise (AWGN) [2] as shown in Fig. 4.



Fig. 3. A few examples from the Chars74K dataset [9].

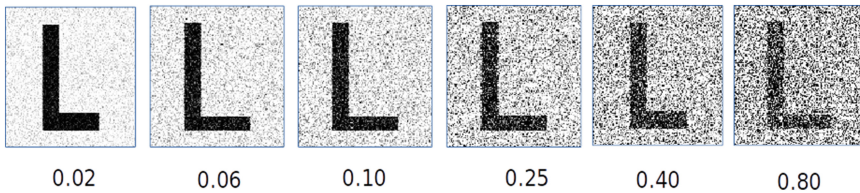


Fig. 4. An example showing the effect of AWGN on a character template, with increasing variance of noise.

5 Experimental Results

Experiments have been carried on the the synthetic datasets using SVM [6], k-NN, Naive Bayes and VGG-6 [23]. Figure 5(A) shows the accuracy of the traditional shallow learning methods for 50D data along with the VGG-6 CNN model. The plot reveals a constant drop in accuracy of the classification with increasing amount of overlap. The experiments are studied in 10-fold cross-validation mode. We observe the performance with increasing extremity (*i.e.* more overlap and similar boundaries of scatters). The *Completely overlapping (CO)* and the *Random (RM)* cases exhibit the poorest performance of the classifier since the accuracy of the binary classifier is around 50%, indicating the presence of extreme distribution overlap in the data over a pair of classes. Similar setup has been experimented on the 100D data where the deep VGG-6 model shows a similar trend in Fig. 5(B) along with the other classifiers. Note here that, the non-parametric classifier though performs worse than deep-CNN at low levels of overlap in class-wise data distributions, catches up quite well to produce a similar degraded performance under extreme overlap (CO and RM) conditions. A recognition accuracy of $\approx 50\%$ at (*f*) and (*g*) indices in Fig. 5(B) show that deep-CNN has no advantage over other simple shallow classifier in extreme conditions. This is one of the main outcomes of this empirical study. At full overlap (labels are random) the CNN performs similar to the shallow learning algorithms.

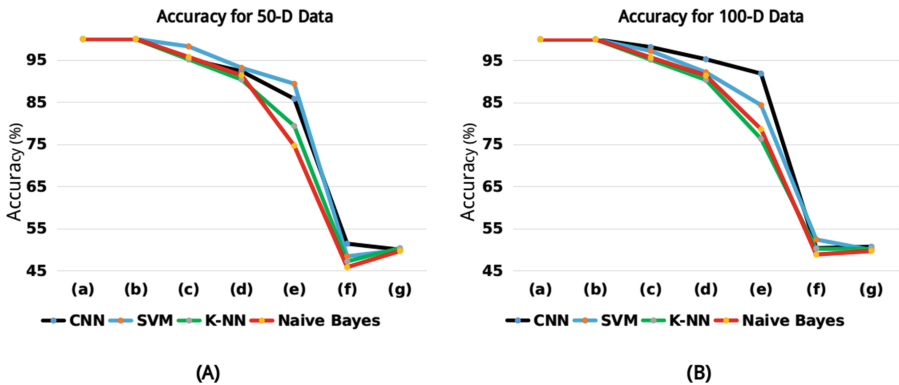


Fig. 5. Plots for accuracies of (A) 50D and (B) 100D data on the 7 different data distributions as shown in Fig. 2; (a) Non-overlapping; (b) Barely touching; Overlap of: (c) 25%; (d) 50%; (e) 75%; (f) Fully Overlapping; and (g) Random class labels.

The VGG-6 model is trained on the clear images of the Chars74K dataset and tested on the images with added noise (see Fig. 3). Figure 6(A) shows plots of accuracy of classification obtained by the VGG-6 model with increasing number of epochs (during training), when tested with image samples of low noise levels of perturbations of the image signal. Figure 6(B) shows the decrease in accuracy of the SVM (based on the HOG [8] feature extracted on the images)

and VGG-6 models with increase in the variance of noise incorporated in the images. For natural images, we can infer that the CNNs are barely competent than shallow methods even when a small amount of noise degrades the images, and the performance of the CNN also falls rapidly with increasing levels of noise in data.

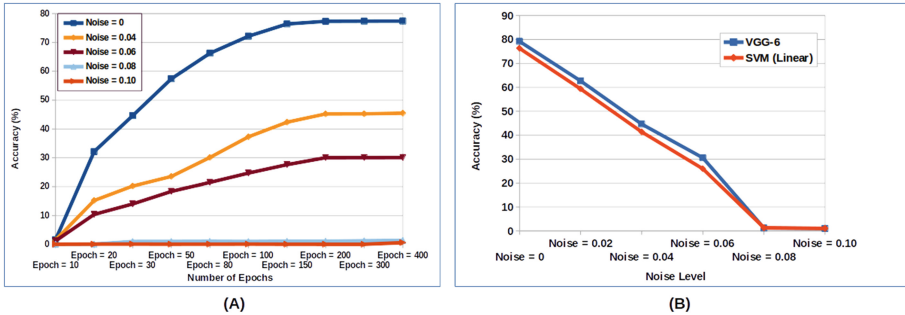


Fig. 6. Curve showing (A) the effect of noise with increasing number of Epochs in training the *VGG-6*; (B) effect on the accuracy of classification using SVM and VGG-6 (after 400 epochs) with increasing levels of noise; on the Chars74K dataset [9].

6 Conclusion

This paper reveals that many state-of-the-art classifiers provide equivalently degraded performance under extreme conditions of the data. When the data is corrupted by large levels of noise or overlapping scatter distributions, even a recent state-of-the-art CNN model randomly classifies the data. In case of Natural images, the DL methods cannot handle extreme conditions (large noise). Being a supervised technique, the CNN models need a mechanism to overcome noise in the data to approximate and classify them more accurately.

References

1. Banerjee, S., Das, S.: Soft-margin learning for multiple feature-kernel combinations with domain adaptation, for recognition in surveillance face dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) on Biometrics, pp. 169–174 (2016)
2. Bergmans, P.: A simple converse for broadcast channels with additive white gaussian noise (corresp.). *IEEE Trans. Inf. Theory* **20**(2), 279–280 (1974)
3. Blog, C.M.: Model selection: underfitting, overfitting, and the bias-variance tradeoff (2013)
4. Chen, J.C., Zheng, J., Patel, V.M., Chellappa, R.: Fisher vector encoded deep convolutional features for unconstrained face verification. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 2981–2985, September 2016

5. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123 (1995)
6. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
7. Cristianini, N., Scholkopf, B.: Support vector machines and kernel methods: the new generation of learning machines. *AI Mag.* **23**(3), 31 (2002)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
9. de Campos, T.E., Babu, B.R., Varma, M.: Character recognition in natural images. In: VISAPP (2), pp. 273–280 (2009)
10. Domingos, P.: A unified bias-variance decomposition. In: Proceedings of 17th International Conference on Machine Learning, pp. 231–238. Morgan Kaufmann, Stanford (2000)
11. Dougherty, J., Kohavi, R., Sahami, M., et al.: Supervised and unsupervised discretization of continuous features. In: Proceedings of the Twelfth International Conference on Machine Learning, vol. 12, pp. 194–202 (1995)
12. Friedman, J., Hastie, T., Tibshirani, R.: *The Elements of Statistical Learning*. Springer Series in Statistics, vol. 1. Springer, Berlin (2001)
13. Haykin, S.: Multilayer perceptrons. *Neural Netw. Compr. Found.* **2**, 156–255 (1999)
14. Hoffman, J., Guadarrama, S., Tzeng, E.S., Hu, R., Donahue, J., Girshick, R., Darrell, T., Saenko, K.: LSDA: Large scale detection through adaptation. In: Advances in Neural Information Processing Systems (NIPS), pp. 3536–3544 (2014)
15. James, G.M.: Variance and bias for general loss functions. *Mach. Learn.* **51**(2), 115–135 (2003)
16. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
19. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**(4), 541–551 (1989)
20. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint [arXiv:1412.6632](https://arxiv.org/abs/1412.6632) (2014)
21. Mhaskar, H., Liao, Q., Poggio, T.: Learning functions: when is deep better than shallow. arXiv preprint [arXiv:1603.00988](https://arxiv.org/abs/1603.00988) (2016)
22. Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., Liao, Q.: Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. arXiv preprint [arXiv:1611.00740](https://arxiv.org/abs/1611.00740) (2016)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)

24. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1891–1898 (2014)
25. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708 (2014)
26. Zhu, Z., Luo, P., Wang, X., Tang, X.: Recover canonical-view faces in the wild with deep neural networks. arXiv preprint [arXiv:1404.3543](https://arxiv.org/abs/1404.3543) (2014)