Intriguing properties of neural networks

Summary written by Mahban Gholijafari
April 1, 2025

Summary: This paper explains two counter-intuitive properties of deep neural networks that comes from their highly expressiveness and cause networks to misclassify an image by apply a certain hardly perceptible perturbation.

Related work: The first property is concerned with the semantic meaning of individual units. Semantic meaning of various units were analyzed by finding the set of inputs that maximally activate a given unit [3, 8, 4]. Computer vision systems interpret and activation of a hidden unit as a meaningful feature; they look for input images which maximize the activation value of this single feature [3, 8, 4, 2]. A variety of recent computer vision models employ input deformations during training for increasing the robustness and convergence speed of the model [5, 8].

Approach: This work found out that adversarial examples are relatively robust, and are shared by neural networks with varied number of layers, activations or trained on different subset of the training data. This paper denotes the input image by $x \in \mathbb{R}^m$, and activation of some layers as $\phi(x)$. They find that the images x' are semantically related to each other, for many x' such that $x' = \arg\max_{x \in \mathcal{I}}(\phi(x), v)$, which means that the natural basis is not better than a random basis to inspect the properties of $\phi(x)$. To evaluate this claim, they used CNN and AlexNet trained on the MNIST dataset with two scenarios, one that maximizes activation in random directions and the other maximizes activations on a natural basis. These experiments give insight on the capacity of ϕ to generate invariance on a particular subset of input distribution bot, does not explain the behavior on the rest of its domain. The main result of this paper for deep neural networks is that the smoothness assumption that underlies many kernel methods does not hold, and they show that by using a simple optimization procedure, adversarial examples are found. Doing so will add small perturbations to a correctly classified input image so that it is no longer correctly classified. The optimization problem proposed in this work can also be used constructively, similar to the hard-negative mining principle. A classifier mapping image pixel value vectors to a discrete label set is denoted by $f : \mathbb{R}^m \rightarrow \{1...k\}$, and f has an associated continuous loss function denoted by $\text{loss}_f : \mathbb{R}^m \times \{1...k\} \rightarrow \mathbb{R}^+$. For a given $x \in \mathbb{R}^m$ image and target label $l \in \{1...k\}$, they aim to solve this box-constrained optimization problem: minimize $||r||_2$ subject to: 1. $f(x + r) = l$, 2. $x + r \in [0, 1]^m$. The minimizer r might not be unique, and they denote one such $x + r$ for an arbitrarily chosen minimizer by $D(x, l)$. As the exact computation of this problem is hard, they approximate it by using a box-constrained L-BFGS, and the approximation is performed by line-search to find the minimum $c > 0$ for which the minimizer r of the following problem satisfies $f(x + r) = l$, i.e., minimize $c|r| + \text{loss}_f(x + r, l)$ subject to $x + r \in [0, 1]^m$.

Datasets, Experiments and Results: The experiments in this paper are on a few different networks and three datasets. Namely, MNIST [7] with a fully connected (FC) network with one or more hidden layers and a Softmax classifier, ImageNet [1] with AlexNet, and ∼10M image samples from Youtube [6] with QuocNet. The minimimum distortion function D has three intriguing properties: 1. for all the networks, the generated adversarial examples are very close and hard to distinguish, 2. cross model generalization, 3. cross training-set generalization. A two layer 100-100-19 non-convolutional neural network with a test error below 1.2% was successfully trained by keeping a pool of adversarial examples a random subset of which is continuously replaced by newly generated adversarial examples and which is mixed into the original training set all the time. Quadratic weight decay ($\text{loss}_{decay} = \lambda \sum w_i^2/k$, where k is the number of units in the layer) was used instead of dropout and for comparison a network of this size gets to 1.6% errors when regularized by weight decay alone and can be improved to around 1.3% by using carefully applied dropout. They found out that adversarial examples for the higher layers seemed to be significantly more useful than those on the input or lower layers. One of the models (FC10(1)), is trained with extremely high $\lambda = 1$ to test whether it is still possible to generate adversarial examples in this extreme setting as well. Two other models are a simple sigmoidal neural network with two hidden layers and a classifier. The last model consists of a single layer sparse autoencoder with sigmoid activations and 400 nodes with a Softmax classifier. To study cross-training-set generalization, they used MNIST and trained three non-convolutional networks, and the adversarial examples remain hard for models trained even on a disjoint training set, although their effectiveness decreases considerably.

Strengths: The adversarial examples are universal and not just the result of overfitting to a particular model or to the specific selection of the training set.

Weaknesses: For MNIST, they do not have results for convolutional modules.

Reflections: The back-feeding adversarial examples to training might improve generalization of the resulting models. Future work consist of systematically comparing the effect of generating adversarial examples for each layer. Also, a deep understanding of how often adversarial negatives appears should be addressed in future research.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[4] I. Goodfellow, H. Lee, Q. Le, A. Saxe, and A. Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22, 2009.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[6] Q. V. Le. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8595–8598. IEEE, 2013.

[7] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[8] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.