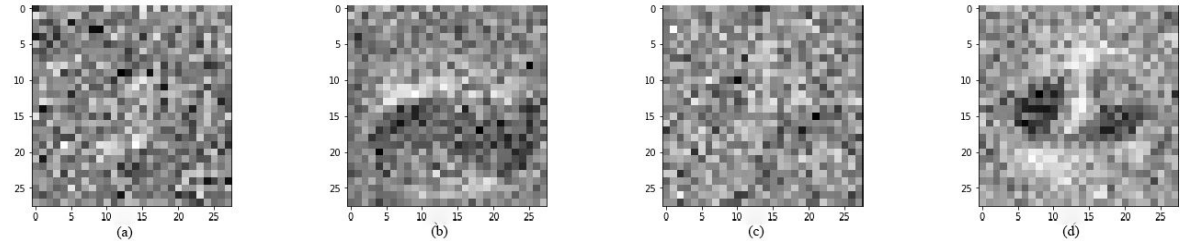
Computational Cognition Workshop, ComCo
October 1-2, 2019, Osnabrück, Germany

Through Neural Nets' Eyes

Sahar Niknam[1]
Institute of Cognitive Science, Osnabrück University

Abstract

Many researchers are working to unlock the black box of neural nets, especially in case of deep learning. And despite the robustness of mathematical analysis, study of nets through their performance in visual tasks proved to be more effective due to its intuitiveness; such that it was adversarial examples that blew the whistle on the true *learning* capacity of neural nets. This research idea, based on an experiment, also aims for a better understanding of neural nets through visual analysis; but this time by employing adversarial examples. One of the simplest ways of creating adversarials is to feed random noise to a trained network and run the gradient descent algorithm over the input until the output matches the desired label, while keeping the weights and biases untouched. The result is a still noisy picture to human eyes, but the network labels it with the desired tag and usually a high level of confidence. This experiment used the same method on two sigmoidal, feedforward nets (a shallow and a deeper one), trained on MNIST dataset, to create 1000 adversarial examples for each digit. The resulting averaged images of these adversarial examples revealed patterns that were different from the actual shapes of the digits. These patterns were preserved, fully, over multiple rounds of training of the same network, and remotely, over the two networks. The other finding of this experiment is the low rate of success in generating adversarial examples for some digits like *1* and *4* compared to the highly successful ones like *0* and *8*, which was a feature shared by both networks. So if we consider the noise input like modelling clay given to the networks to reconstruct what they have learned/memorized about the digits shapes, it seems they had difficulties learning *edges* compared to *curves*. Though, the experiment results might be heavily affected by the characteristics of the method used for generating adversarial examples. Further study on the effect of using different activation functions could also be necessary for reliable conclusions.



(a) An adversarial example, labeled 2 by a shallow feedforward network with 1.46% confidence (b) The averaged image over 1000 adversarial examples, labeled 2 by the same shallow network with 98.72% confidence (c) An adversarial example, labeled 4 by a deep feedforward network with 86.97% confidence (d) The averaged image over 1000 adversarial examples, labeled 4 by the same deep network with 99.68% confidence

Keywords: *artificial neural network, deep learning, adversarial example*

[1] sniknam@uni-osnabrueck.de