Intriguing Properties of neural networks

Summary written by Vedant Jain
April 28, 2021

Summary: This paper describes a theoretical analysis of neural networks and the meaning behind how information is encoded in different layers and neurons of a neural network. They describe how different networks perturbed in very subtle ways can cause incorrect classification. Two import distinctions are made, the first is that these perturbations are unrecognizable as different for the human eye, the second is that these perturbations aren't just random noise added to the network. Instead they are distinctly selected to be directionally opposite of the gradient. Szegedy et al, also show that the encoding of the network is not done in terms of the individual neurons in a layer but much of the important information is encoded in the space of the network itself.

Related work: Earlier work has shown that different layers of a neural network encode differently spatially relevant features. And at first it was thought that activation of individual layers contributed to classification. The believe was that there was a specific vector of inputs that mattered to get the desired outputs. But later on the paper goes on to show how this is actually not the case. Something that is shown by Mikolov et al, when they looked at different words and how the different magnitudes and directions the vector space of word representations resulted in strong semantic encoding of relationships.[1]

Approach: The author systematically went through exploring the encoding of images as well as the creation of what are called adversarial examples. In the first case they looked at the MNIST dataset and specifically the prior belief that a certain vector of activation would lead to semantic descriptions of classification. They were able to show that this was not the case and in fact taking vectors in random space with which creates a random basis actually can result in similar semantic criteria. This suggests that it isn't individual neurons that matter but more so the vector space that the network creates that is important for classification. Secondly, they invalidate the prior assumption that networks basically classify an image and then those that are similar to them get similar classification. Instead they describe how tiny perturbations that have overall very small l2 norm can still have a drastic affect to the classification of an image. It is made explicitly clear that these perturbations aren't random noise which a model actually does a good job classification, instead these ”blind spots” are computationally found by going against the gradient of the model. Finally, they show that an adversarial image for one model will likely be adversarial images for multiple models.
Datasets, Experiments and Results: The authors of this paper validated their process across multiple data sets and described how theses vulnerabilities exists across the many algorithms. Specifically they looked at the MNIST dataset with a simple fully connected network and autoencoder network. They also used the ImageNet dataset in the context of the AlexNet, and finally they used 10M images from youtube with the QuocNet network.
Strengths: This paper describes a very new theoretical approach explaining neural networks. They highlight a huge short coming with neural networks, one they know they will need clear evidence to be able to support their case. The author rises to the challenge by explaining in both very simple conceptual terms why neural networks are weak to adversarial attacks and mathematically how the values for these images are calculated.
Weaknesses: A weakness with this paper involves sort of exploring different methods of adversarial images. For example, it may be interesting to see how a change in a single pixel could cause a network to miss classify. Similarly, it would be interesting to identify common causes of classification and see if they are caused by random ”discoveries” of images that are in the space of the fast gradient descent method.
Reflections: Szegedy et al describe very interesting advances in the theoretical understanding of neural networks. They change the prior assumptions about how small changes in pixel values that don't change the representation to the human eye can cause a network to miss-classify. These results highlight that neural networks may not be as accurate as we earlier thought and their wide scale must be done more carefully to deal with these edge cases.

References

[1] G. C. a. J. D. Tomas Mikolov, Kai Chen. Efficient estimation of word representations in vector space. 2013.