New Articles!

- There's an Arduino for that!
- Mojolicious In Conversation With Sebastian Riedel
- Database The Prolog Way
- Just JavaScript - The Prototype Mechanism
- Eckert & Mauchly and ENIAC
- The Essence Of Loops
- Weak typing - the lost art of the keyboard
- Android Adventures - ViewPager
- Inside Random Numbers
- Not Dumping .NET - Microsoft's Method
- Nolan Bushnell and Atari
- Getting Started With jQuery - Advanced Filters
- Dyslexia and Programming
- Geekuni Dancer Web Development Course
- Linq and XML

New Book Reviews!

- eCommerce in the Cloud
- Microsoft SQL Server 2014 Query Tuning & Optimization
- Getting Started with Raspberry Pi 2nd Ed
- Programming Elastic MapReduce
- iOS Development with Xamarin Cookbook
- App Inventor 2
- You Don't Know JS: this & Object Prototypes
- Practical Web Analytics

The Deep Flaw In All Neural Networks

Written by Mike James

Wednesday, 10 December 2014

Recently a paper reported the information that neural networks seem to have a fundamental problem in recognizing things now we have another twist on the same basic idea. In this case you can construct an image that looks nothing like what it is supposed to be and yet it is still classified with high confidence.

The paper **"Intriguing properties of neural networks"** from a team that includes authors from Google's deep learning research project that we reported on in May (see The Flaw Lurking In Every Deep Neural Net) discovered something surprising and disturbing.

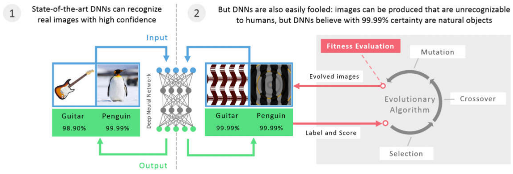
If you have a Deep Neural Network DNN that is trained to correctly recognize photos of different things then close to any correctly classified image there is an image that is incorrectly classified. For example if you have a photo of a car and the DNN classifies it with a high probability as a car then it is possible to find a set of small perturbations to the image that don't change how it looks to a human but which cause the DNN to classify it as something else - even though to a human there is possibly no detectable change in the image.

Put simply - very close to any correctly classified image there is an incorrectly classified image, an adversarial image, that to a human looks very little different.

Notice that this isn't a theorem, just an observation based on a practical examination of a relatively small number of DNNs. Even so, the fact that these DNNs exhibit the effect and the adversarial images seem to be incorrectly classified by other DNNs suggest that there is something deep and universal going on.

The changes are small and distributed across all the pixels and yet they move the "data point" outside of the boundary that separates all cars from other types of things. In the original paper these adversarial images were found by an optimization algorithm that applied perturbations that moved the image away from its correct class.

Now we have a new demonstration of this effect. In this case Anh Nguyen, Jason Yosinski and Jeff Clune made use of the genetic algorithm to "breed" images that were correctly classified with a high probability or confidence and yet, to a human, looked nothing like the the object they were supposed to be.


1
State-of-the-art DNNs can recognize real images with high confidence
2
But DNNs are also easily fooled: images can be produced that are unrecognizable to humans, but DNNs believe with 99.99% certainty are natural objects
Input
Output
Deep Neural Network
Guitar
98.90%
Penguin
99.99%
Guitar
99.99%
Penguin
99.99%
Fitness Evaluation
Mutation
Evolved images
Evolutionary Algorithm
Crossover
Label and Score
Selection

Indeed the images that the DNN classified so confidently were more like white noise to a human. for example the diagram below shows "digit" images that are classified with a 99.99% confidence by the DNN LeNet:

Tweet 17

Follow @Iprogrammerinfo 2,079 followers

for User Experience

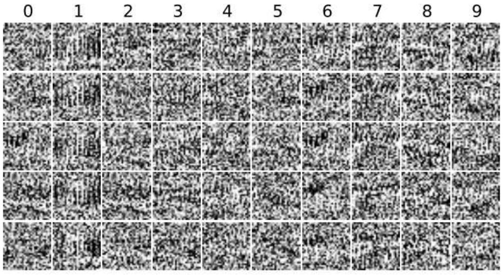
- If Hemingway Wrote JavaScript
- The Book of CSS3, 2nd Ed
- MySQL Cookbook, 3rd Ed
- JavaScript and JQuery
- Microsoft SQL Server 2012 T-SQL Fundamentals
- Graph Databases
- Arduino Programming in 24 Hours

Popular Articles

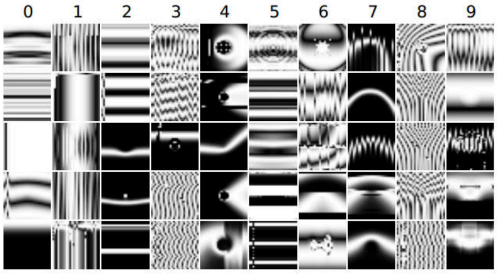
- Arrays in Python
- Android Adventures - Getting Started With Android Studio
- JavaScript Data Structures - Associative Array
- Using the WPF .NET 4.0 DataGrid
- Getting started with Excel VBA
- Dumping .NET - Microsoft's Madness
- Weak typing - the lost art of the keyboard
- Getting Started with PC Kinect
- Getting started with Microsoft Kinect SDK
- Data structures - Trees

Popular Book Reviews

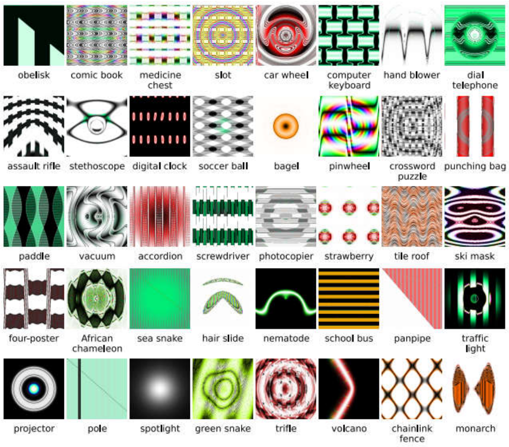
- Beautiful Architecture: Leading Thinkers Reveal the Hidden Beauty in Software Design
- Hello World! Computer Programming for Kids and Other Beginners
- Practical Raspberry Pi
- Beginning R: The Statistical Programming Language
- Programming: Principles and Practice Using C++
- Algorithms in a Nutshell
- How We Test Software at Microsoft
- The Ruby Programming Language
- Android Apps Marketing
- Simply Javascript



A slight tweak of the algorithm can generate more regular looking images that are classified as digits with a 99.99% certainty but still look nothing like the digits according to most humans:



The same technique was then applied to the Imagenet database with similar results the following were recognized with a mean confidence of 99.12% :



In this case you can occasionally see that there is some similarity in features - the nematode, spotlight and pole for example. However it seems that what is happening is that the images are stimulating the low level feature detectors, but not in an organized way. If a localized image stimulates the DNN then an image that includes repeated copies increases the total stimulation. To test this, some of the images that include repeats were edited to reduce the repeats and the confidence did drop.

The genetic algorithm approach isn't essential in constructing such examples. The team also used a standard optimization method to produce images that were mostly unrecognizable, but still classified with a high confidence.

All this suggests that the DNNs are learning low and medium level features and not making use of how these relate to one another. For example, a face would be recognized if it had two eyes, nose and mouth irrespective of where they occurred in the image.

It also seems that there is a generalization - images bred to be classified with high confidence on one DNN seem to be classified with a high confidence by another.

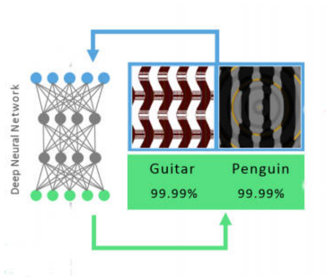
So what is going on?

The authors suggest that it is problem with the way that the discrimination is being performed. The DNN assigns each class a large volume in the image space and confidence is proportional to how far away from the boundary an image is - how close it is to other images in the class isn't taken into account. There is so much available space that it is possible to find images that are in the class but are far from images that look like the class.

This sounds very much like the arguments that resulted in the invention of the Support Vector Machine. Perhaps DNNs need to optimize something other than pure classification performance?

performance?

At the end of the paper the authors raise the interesting question of how these finding affect the use of DNNs in real applications. A security camera, for example, could be fooled by "white noise" designed to be classified as a face. Perhaps the right background wallpaper could influence visual search classifiers. The possibilities are there waiting to be exploited. The idea that a driverless car could swerve dangerously to avoid something that looked nothing at all like a pedestrian is currently very possible - either by accident or design.



More Information

[Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images](#)

Anh Nguyen, Jason Yosinski and Jeff Clune

Related Articles

The Flaw Lurking In Every Deep Neural Net

Neural Networks Describe What They See

Neural Turing Machines Learn Their Algorithms

Learning To Be A Computer

Google's Deep Learning AI Knows Where You Live And Can Crack CAPTCHA

Google Uses AI to Find Where You Live

Deep Learning Researchers To Work For Google

Google Explains How AI Photo Search Works

Google Has Another Machine Vision Breakthrough?

The Triumph Of Deep Learning

To be informed about new articles on I Programmer, install the I Programmer Toolbar, subscribe to the RSS feed, follow us on, Twitter, Facebook, Google+ or Linkedin, or sign up for our weekly newsletter.

I PROGRAMMER NEWS
Languages, Web, AI, Apps, Tools...

DR DOBB'S BITES THE DUST AFTER 38 YEARS

17/12/2014

If you don't know Dr Dobb's - or Dr. Dobb's Journal of Computer Calisthenics and Orthodontia - then you probably missed out on a great era of computing. Sadly this iconic publication has jus [...]

+ FULL STORY

NODE.JS FORK - NOW YOU HAVE A CHOICE TO MAKE

08/12/2014

One of the big advantages of open source is that if you don't like the current state of things you can simply create a fork and make your own version of the project. However, not all forks are equal a [...]

+ FULL STORY

More News

- Google Gets Closer To Killing Old Style Browser Plugins
- How Google Does Multi-Platform In Inbox
- Cutting Edge Topics At SDD 2015
- Flow - A Static Type Checker For JavaScript