Analysis of “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”

Preamble

Wow, what a cool paper! This paper had vibes of the minority report all through it. It really helped that I had background knowledge of evolutionary algorithms to provide context for how they were generating the false images.

A brief statement of the problems addressed in the paper in my own words.

I thought that the entirety of the paper was stated succinctly in section 3.4 about halfway through, it states: *“...This is because evolution need only to produce features that are unique to, or discriminative for, a class, rather than produce an image that contains all the typical features of a class”.* That statement right there captures the main crux of their paper. It was really interesting to me to see their chain of reasoning for the blame. For example, in the discussion section they talked about how their experiments could have led to very different results. Maybe I am wrong here, but it seemed as if they were initially unaware of the problem prior to starting their work. Then once the discovered the problem with DNN’s they were able to reason backwards as to why the nets were behaving the way they were.

What I agree with/like in the paper and why.

Honestly what is not to like about this paper? You said that this paper was like pouring ice cold water on the DL-hype, and you weren’t kidding. I really appreciate how they wrote this paper, and made it easy to understand. This point may be a little subjective, but I really like how they didn’t “try” to fool the DNN’s they simply used evolutionary algorithms, and let those things pick and choose how to morph and change the original image.

The identification of the problem with false positives in security related applications is one thing that really scares me. It was the last paragraph in the discussion section where they say: *“...For example, one can imagine a security camera that relies on face or voice recognition being compromised. Swapping white-noise for a face, fingerprints, or a voice might be especially pernicious since other humans nearby might not recognize that someone is attempting to compromise the system.”* It was from reading that section that gave me the minority report vibes. Ugh, what a terrifying thought, I don’t know how to reconcile this with today’s day and age. All I can do is show with my money and my vote where I am willing to place my trust, but I fear that I will be in that vast minority when the time comes.

What I disagree with/dislike in the paper and why.

No, I can’t really think of anything here. It is not wise to disagree with something that reveals truth and error and explains it so succinctly.

Any inspirations I found in the paper.

Yes, I have inspirations, I am inspired to never wholly trust computers to make life and death or security related decisions. I know that much of that responsibility has already been relegated to machines, but I can start taking steps to decrease that implicit trust where I can.

Thanks again for assigning this incredible paper! This was a joy to read and very insightful. Keep up the good work!