PSYCHO: PerSonalitY CHaracterizatiOn of artificial intelligence

Achal Dave
Cranberry-Lemon University

Rohit Girdhar
Cranberry-Lemon University

Abstract

Recent times have seen great advancements in the field of AI, thanks to the resurgence of deep learning. It has impacted virtually every aspect of our lives, from generating new cat videos [4], to converting cat videos into dog videos [2]. However, these advancements have also stoked fear in the hearts of us humans: what if the robot hand that learned to open door knobs instead decides to use its skills to pick up a gun and point it at us? Needless to say, the solution is not fewer guns, but the mental health of these robots. In this work, we try to assuage those concerns by proposing a method to analyze the brains of our robots. Our method takes years of human psychology research and brainlessly applies it to analyze the deep networks that form the fundamental cognitive system of modern day robots. We evaluate our method on the latest and greatest deep networks and uncover the ones most likely to 'break bad'.

1. Introduction

> "AI is a fundamental risk to the existence of human civilization."
>
> *Elon Musk (July 2017)*

> "I was trying to turn off some lights and they kept turning back on. After the third request, Alexa stopped responding and instead did an evil laugh."
>
> *Reddit user (January 2018)*

> "The #BostonDynamics #robots are learning. Soon they'll be opening our fridges and stealing our beer."
>
> *Dr. Randy Olson (February 2018, via Twitter)*

Lets face it. The threat of AI is real, and the leaders of our tech industry have gone out of their way to warn us about it. However, the lack of tools to interpret our AI methods has tied the hands of AI researchers, forcing them to focus on making their methods stronger with no regard to the future of humanity. This problem is especially dire in the field of deep learning, where the dark magic of stochastic gradient descent carves out ultra high dimensional spaces to learn representations unimaginable by humans. In this work, we take a step back and attempt to analyze the thinking process of the deep networks we have crafted, before it is too late.

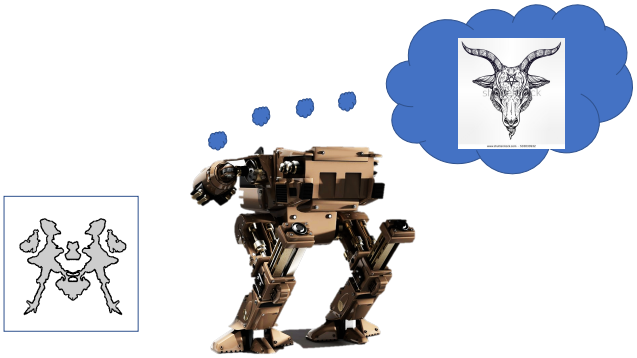

Figure 1. When will AI go haywire? Understanding how AI will act in the future requires a carefully designed psychological analysis using the widely acclaimed Rorschach ink blot test.

Today, the Turing test is largely solved [1, 5, 3]. Our method, PSYCHO instead uses the *Rorschach inkblot* test to analyze artificial intelligence. The test works by showing an inkblot image, like in Table 1 (column 1), and asks the user to pick a sentence that best describes that inkblot from 7 options (we follow the paradigm from `http://theinkblot.com/`). We design an approach to allow state of the art deep networks to take this test, by finding nearest neighbors of their representation with a representation for each option. We report some insightful analysis of these networks in Sec. 3.

2. Approach

The Rorschach ink blot test, as presented on `http://theinkblot.com/`, requires the test-taker to pick a sentence describing each of the 10 Rorschach ink blots. Unfortunately, despite our best efforts, we were unable to coax current AI models into taking online personality tests.

Undeterred, we developed a novel approach for psycho-

logically evaluating our models. For each ink blot, we collected an image representing each potential response (such as "a giraffe in a bathtub"). Unfortunately, naively collecting images can lead to a bias in the selected images. To overcome any such bias, we directly query Google Image Search for an unbiased list of images for each potential response. We then selected a single image from these results for each response query while trying our very hardest not to use our personal biases.

Armed with this dataset, we present each ink blot along with potential responses to our model, and select as a response the image that the model thinks is most like the ink blot.[1]

3. Experiments

We present qualitative and quantitative results, along with psychological notes for **five** popular Convolutional Neural Network models in the computer vision community. We have anonymized the names to ~~protect against lawsuits~~ avoid upsetting anyone.

In Table 1, we present the extensive analysis provided by `http://theinkblot.com`. We immediately notice that our models have surprisingly varied personalities. "A-net" is a prototypical optimist, or what experts may refer to as "the SpongeBob". V-net and I-net share a high sickness quotient, which we explore further through qualitative results.

Unfortunately, trusting experts can mislead our understanding of potential societal threats. To overcome this, we present the raw results from our method in Table 2 for further public analysis.

Disturbing responses: While some responses from our model are playful (e.g. Table 2 Row 5), there are numerous worrying signs in their responses. I-net, in particular, consistently chooses disturbing imagery (a satanical head in Row 3, a satanical eye in Row 5, a strange creature in Row 7, and what is indubitably a satanical ritual in Row 9). Equally worrying is the creepy imagery provided as responses by V-net, R-net, and D-net in Row 1 (a monsterous face) and, worse, in Row 5 (a Teletubby).

Intellectual diversity: The lack of diversity in AI is plainly visible from our analysis. In particular, we discover for the first time that models developed in the same institution (R-net and D-net) develop *equivalent* psychological tendencies.

4. Conclusion

While we are far from preventing the inevitable AI apocalypse, we believe our method will go a long way in enabling AI researchers to psycho-analyze their deep networks before deploying them to read every single Snapchat we post through the day.

[1] In particular, we take the final layer representation of the ink blot and all response images, and choose the response that minimizes Euclidean distance to the ink blot. We hope to publicly release our code.

Model	Sickness	Notes
A-net	47%	"Positive attitude towards everything" "very annoying"
V-net	75%	"aspire to [be] CEO", "horrible bore"
I-net	78%	"short attention-span", "work very slowly"
R-net D-net	60%	"succeeded beyond wildest dreams", "frequently mentions paradigm shifts"

Table 1. Quantitative and qualitative results from the Rorscahch test, according to one online test.

N.B.: This paper is a work of satire and should not be taken seriously.

References

[1] Computer ai passes turing test in 'world first'. `http://www.bbc.com/news/technology-27762088`, 2014.

[2] J.-Y. Z. et al. CycleGAN. `https://github.com/junyanz/CycleGAN`, 2017.

[3] L. Hardesty. Computer system passes "visual turing test".

[4] J. Johnson. Meow generator: This deep learning AI generated thousands of creepy cat pictures. *Motherboard*, 2017.

[5] C. Osborne. Mit's artificial intelligence passes key turing test. `http://www.zdnet.com/article/mits-artificial-intelligence-passes-key-turing-test/`, 2016.

Query	A-net	V-net	I-net	R-net	D-net

Table 2. Qualitative results on the Rorschach test.