A Discussion of *Adversarial Examples Are Not Bugs, They Are Features*

PUBLISHED

Aug. 6, 2019

DOI

10.23915/distill.00019

On May 6th, Andrew Ilyas and colleagues published a paper [1] outlining two sets of experiments. Firstly, they showed that models trained on adversarial examples can transfer to real data, and secondly that models trained on a dataset derived from the representations of robust neural networks seem to inherit non-trivial robustness. They proposed an intriguing interpretation for their results: adversarial examples are due to “non-robust features” which are highly predictive but imperceptible to humans.

The paper was received with intense interest and discussion on social media, mailing lists, and reading groups around the world. How should we interpret these experiments? Would they replicate? [1] And if non-robust features exist… what are they?

To explore these questions, Distill decided to run an experimental “discussion article.” [2] We invited a number of researchers to write comments on the paper and organized discussion and responses from the original authors.

The Machine Learning community sometimes worries that peer review isn’t thorough enough. In contrast to this, we were struck by how deeply respondents engaged. Some respondents literally invested weeks in replicating results, running new experiments, and thinking deeply about the original paper. We also saw respondents update their views on non-robust features as they ran experiments — sometimes back and forth! The original authors similarly deeply engaged in discussing their results, clarifying misunderstandings, and even running new experiments in response to comments.

We think this deep engagement and discussion is really exciting, and hope to experiment with more such discussion articles in the future.

Discussion Themes

Clarifications: Discussion between the respondents and original authors was able to surface several misunderstandings or opportunities to sharpen claims. The original authors summarize this in their rebuttal.

Successful Replication: Respondents successfully reproduced many of the experiments in Ilyas et al [1] and had no unsuccessful replication attempts. This was significantly facilitated by the release of code, models, and datasets by the original authors. Gabriel Goh and Preetum Nakkiran both independently reimplemented and replicated the non-robust dataset experiments. [3] Preetum also replicated part of the robust dataset experiment by training models on the provided robust dataset and finding that they seemed non-trivially robust. It seems epistemically notable that both Preetum and Gabriel were initially skeptical. Preetum emphasizes that he found it easy to make the phenomenon work and that it was robust to many variants and hyperparameters he tried.

Exploring the Boundaries of Non-Robust Transfer: Three of the comments focused on variants of the "non-robust dataset" experiment, where training on adversarial examples transfers to real data. When, how, and why does it happen? Gabriel Goh explores an alternative mechanism for the results, Preetum Nakkiran shows a special construction where it doesn't happen, and Eric Wallace shows that transfer can happen for other kinds of incorrectly labeled data.

Properties of Robust and Non-Robust Features: The other three comments focused on the properties of robust and non-robust models. Gabriel Goh explores what non-robust features might look like in the case of linear models, while Dan Hendrycks and Justin Gilmer discuss how the results relate to the broader problem of robustness to distribution shift, and Reiichiro Nakano explores the qualitative differences of robust models in the context of style transfer.

Comments

Distill collected six comments on the original paper. They are presented in alphabetical order by the author's last name, with brief summaries of each comment and the corresponding response from the original authors.

Adversarial Example Researchers Need to Expand What is Meant by "Robustness"

AUTHORS	AFFILIATIONS
Justin Gilmer	Google Brain Team
Dan Hendrycks	UC Berkeley

Justin and Dan discuss "non-robust features" as a special case of models being non-robust because they latch on to superficial correlations, a view often found in the distributional robustness literature. As an example, they discuss recent analysis of how neural networks behave in frequency space. They emphasize we should think about a broader notion of robustness.

→ READ FULL ARTICLE

> COMMENT FROM ORIGINAL AUTHORS:
>
> The demonstration of models that learn from only high-frequency components of the data is an interesting finding that provides us with another way our models can learn from data that appears "meaningless" to humans. The authors fully agree that studying a wider notion of robustness will become increasingly important in ML, and will help us get a better grasp of features we actually want our models to rely on.

Gabriel explores an alternative mechanism that could contribute to the non-robust transfer results. He establishes a lower-bound showing that this mechanism contributes a little bit to the $\widehat{\mathcal{D}}_{rand}$ experiment, but finds no evidence for it effecting the $\widehat{\mathcal{D}}_{det}$ experiment.

→ READ FULL ARTICLE

> COMMENT FROM ORIGINAL AUTHORS:
>
> This is a nice in-depth investigation that highlights (and neatly visualizes) one of the motivations for designing the $\widehat{\mathcal{D}}_{det}$ dataset.

Two Examples of Useful, Non-Robust Features

AUTHORS
Gabriel Goh

AFFILIATIONS
OpenAI

Gabriel explores what non-robust useful features might look like in the linear case. He provides two constructions: "contaminated" features which are only non-robust due to a non-useful feature being mixed in, and "ensembles" that could be candidates for true useful non-robust features.

→ READ FULL ARTICLE

> COMMENT FROM ORIGINAL AUTHORS:
>
> These experiments with linear models are a great first step towards visualizing non-robust features for real datasets (and thus a neat corroboration of their existence). Furthermore, the theoretical construction of "contaminated" non-robust features opens an interesting direction of developing a more fine-grained definition of features.

Adversarially Robust Neural Style Transfer

AUTHORS
Reiichiro Nakano

Reiichiro shows that adversarial robustness makes neural style transfer work by default on a non-VGG architecture. He finds that matching robust features makes style transfer's outputs look perceptually better to humans.

→ READ FULL ARTICLE

> COMMENT FROM ORIGINAL AUTHORS:
>
> Very interesting results that highlight the potential role of non-robust features and the utility of robust models for downstream tasks. We're excited to see what kind of impact robustly trained models will have in neural network art! Inspired by these findings, we also take a deeper dive into (non-robust) VGG, and find some interesting links between robustness and style transfer.

Adversarial Examples are Just Bugs, Too

AUTHORS	AFFILIATIONS
Preetum Nakkiran	OpenAI & Harvard University

Preetum constructs a family of adversarial examples with no transfer to real data, suggesting that some adversarial examples are "bugs" in the original paper's framing. Preetum also demonstrates that adversarial examples can arise even if the underlying distribution has no "non-robust features".

→ READ FULL ARTICLE

> COMMENT FROM ORIGINAL AUTHORS:
>
> A fine-grained look at adversarial examples that neatly our thesis (i.e. that non-robust features exist and adversarial examples arise from them, see Takeaway #1) while providing an example of adversarial examples that arise from "bugs". The fact that the constructed "bugs"-based adversarial examples don't transfer constitutes another evidence for the link between transferability and (non-robust) features.

Learning from Incorrectly Labeled Data

AUTHORS	AFFILIATIONS
Eric Wallace	Allen Institute for AI

Eric shows that training on a model's training errors, or on how it predicts examples form an unrelated dataset, can both transfer to the true test set. These experiments are analogous to the original paper's non-robust transfer results — all three results are examples of a kind of "learning from incorrectly labeled data."

→ READ FULL ARTICLE

COMMENT FROM ORIGINAL AUTHORS:

These experiments are a creative demonstration of the fact that the underlying phenomenon of learning features from “human-meaningless” data can actually arise in a broad range of settings.

Original Author Discussion and Responses

Discussion and Author Responses

AUTHORS

Logan Engstrom, Andrew Ilyas, Aleksander Madry, Shibani Santurkar, Brandon Tran, Dimitris Tsipras

AFFILIATIONS

MIT

The original authors describe their takeaways and some clarifcations that resulted from the conversation. This article also contains their responses to each comment.

→ READ FULL ARTICLE

Citation Information

If you wish to cite this discussion as a whole, citation information can be found below. The author order is all participants in the conversation in alphabetical order. You can also cite individual comments or the author responses using the citation information provided at the bottom of the corresponding article.

Editorial Note

This discussion article is an experiment organized by Chris Olah and Ludwig Schubert. Chris Olah facilitated and edited the comments and discussion process. Ludwig Schubert assisted by assembling the responses into their current presentation.

We’re extremely grateful for the time and effort that both the authors of the responses as well as the authors of the original paper put into this process, and the patience they had with the editorial team as we experimented with this format. Respondents were selected in two ways. Some respondents came to our attention because they were actively working on better understanding the Ilyas et al results. Other respondents were subject matter experts we reached out to.

Distill is also grateful to Ferenc Huszár for encouraging us to explore this style of article.

Footnotes

1. Adversarial example research is particularly vulnerable to a certain kind of non-replication among disciplines of machine learning, because it requires researchers to play both attack and defense. It’s easy for even very rigorous researchers to accidentally use a weak attack. However, as we’ll see, Ilyas et al’s results have held up to initial scrutiny. [↩]
2. Running a discussion article is something Distill has wanted to try for several years. It was originally suggested to us by Ferenc Huszár, who writes many lovely discussions of papers on his blog.

   Why not just have everyone write private blog posts like Ferenc? Distill hopes that providing a more organized forum for many people to participate can give more researchers license to invest energy in discussing other’s work and make sure there’s an opportunity for all parties to comment and respond before the final version is published. [↩]
3. Preetum reproduced the $\widehat{\mathcal{D}}_{det}$ non-robust dataset experiment as described in the paper, for L_∞ and L_2 attacks.
   Gabriel repproduced both $\widehat{\mathcal{D}}_{det}$ and $\widehat{\mathcal{D}}_{rand}$ for L_2 attacks. [↩]

References

1. **Adversarial examples are not bugs, they are features** [PDF]
   Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A., 2019. arXiv preprint arXiv:1905.02175.

Updates and Corrections

If you see mistakes or want to suggest changes, please create an issue on GitHub.

Reuse

Citation

For attribution in academic contexts, please cite this work as

```
Engstrom, et al., "A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features'", Distill,
2019.
```

BibTeX citation

```
@article{engstrom2019a,
  author = {Engstrom, Logan and Gilmer, Justin and Goh, Gabriel and Hendrycks, Dan and Ilyas, Andrew
and Madry, Aleksander and Nakano, Reiichiro and Nakkiran, Preetum and Santurkar, Shibani and Tran,
Brandon and Tsipras, Dimitris and Wallace, Eric},
  title = {A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features'},
  journal = {Distill},
  year = {2019},
  note = {https://distill.pub/2019/advex-bugs-discussion},
  doi = {10.23915/distill.00019}
}
```

Distill is dedicated to clear explanations of machine learning

About Submit Prize Archive RSS GitHub Twitter ISSN 2476-0757