openCHEAT: Computationally Helped Error bar Approximation Tool - Kickstarting Science 4.0

Bernhard Egger* Kevin Smith** ~~David Cox~~*** Max Siegel*
Magic Institute of Technology
{egger,k2smith,maxs}@mit.edu
* Co-First and Co-Last Authors, ** Co-Middle Author, *** Not an Author

Abstract

Error bars are often required by pedantic reviewers but are challenging to create. The process of making them is an error-prone procedure that wastes a tremendous amount of time. We therefore propose a system to automate this process. We introduce openCHEAT, a system to add error bars to scientific plots based on a proprietary deep learning method. We found that this invention can be applied to the entirety of scientific literature, past and future. Our simple and easy-to-use system enables us to add error bars to anything, including generalizing to real-world scenes. This is a first step towards fully automated science - Science 4.0.

1. Introduction

We've all had something like this happen to us: you put together a fantastic model that beats the current SoTA on some benchmark by 0.07%, which clearly should qualify the work for acceptance in any top-tier conference. However, invariably, some reviewer[1] raises concerns like "is that difference statistically reliable?" or "would the results replicate with a different initialization?", and hence require error bars on your plots for acceptance.

Now, of course we all know that classical papers on sampling theory are almost a century old [9][2] while modern machine learning was invented in 2012 [5] (though c.f. Schmidhuber for evidence that he in fact invented it all in the 80's and 90's[3]), which clearly means that using error bars is outdated. Plus, training the model multiple times to get these sample bounds is expensive, and we don't have "OpenAI money" lying around. And besides, spending energy on training these models is bad for the environment [3],[4] so really we're saving the world over here. However, a reviewer response consisting of nothing more than "The results for our model are bolded – of course they're better!" followed by a string of profanity tends not to lead to acceptance.[5] We therefore consider alternate methods for satisfying Reviewer 2 without bothering with trivialities like actually learning statistics.

We solve this problem the standard machine learning way: with lots of data of dubious provenance and an off the shelf algorithm. We propose the Computationally Helped Error bar Approximation Tool (openCHEAT[6]), which uses ~~Pix2Pix [4]~~ a proprietary method to learn to transform graphs without error bars into graphs with error bars. In this way, we can hasten the speed of science by allowing researchers to quickly update their graphs in response to reviewer requests, without any additional model training.

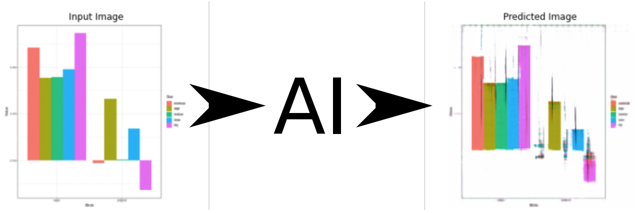

Figure 1. Detailed illustration of our approach.

The key benefits of openCHEAT can be summarized as follows:

1. Our approach is fully data driven - exactly what you would expect for error bars.
2. Our tool enables the generation of error bars in less than a second on a single GPU - this is superhuman performance.
3. Our error bars are derived from more data ($n = 10,000$) than most other error bars and are therefore more trustworthy.
4. Our approach works on images of graphs, and therefore is more likely to generalize to real-world problems than alternate approaches that require knowledge of the underlying means and standard errors.

1.1 Related Work

This work [2] is completely unprecedented. It is, if at all, only vaguely related to our own work that revolutionized autonomous driving [1].

2. Methods

Our implementation is likely based on a convolutional neural network architecture with fewer than 675,078,473,000 parameters, and uses hyperparameters σ, δ and ξ (which is our favorite greek letter). For more details refer to Figure 1. Because of potential commercial interest, we cannot reveal more about our method at this

[1] Usually Reviewer 2

[2] We did not read or retrieve this paper, but the title and abstract makes it sound like it would support this point.

[3] https://people.idsia.ch/~juergen/deep-learning-miraculous-year-1990-1991.html

[4] Gebru et al. citation redacted due to corporate pressure from Google

[5] See our last four submissions for further evidence.

[6] Note that there is in fact nothing "open" about this tool, but we thought it sounded cooler that way. And that tactic worked for OpenAI, didn't it?

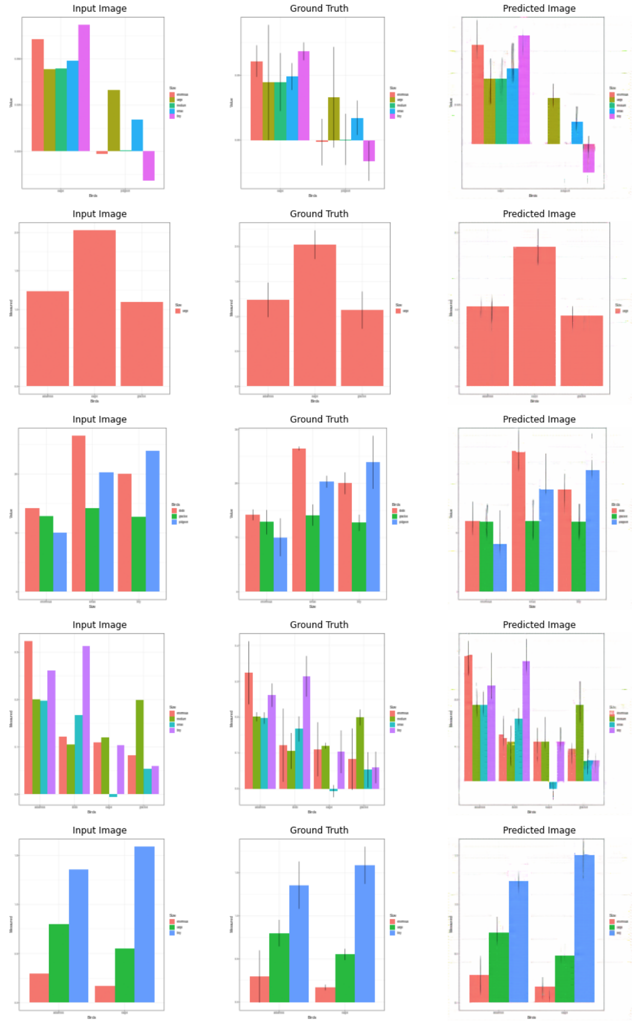

Figure 2. Example of Pre-openCHEAT plots without error bars, the ground truth error bar and our enhanced plots with error bars (sometimes even multiple to indicate experimental flaws). Our plots looks much more scientific.

point, which is clearly significantly more advanced than just using Pix2Pix [4] from a stock Colab notebook.[7]

3. Experiments and Results

We had hoped to download 10,000 images from the google image search, but Google required us to label images for its classifier so we stopped after 250 plots with error bars (we assume we exhausted all plots with error bars on the internet). We therefore decided to generate synthetic data using R, including 10,000 pairs of matched plots with and without error bars. We also generated 200 additional pairs for testing, but then misplaced them, so do not have those results. We choose an image resolution of 256x256 because the results look better in lower resolution - this also leaves more space for interpretation. During training we decided to not watch the loss going down, but instead buy some Gamestop stocks; because we were following the price fluctuations closely, we lost track of time and so assume that training performance plateaued. For our hyperparameters, we choose $\sigma = 342.43$, $\delta = 23.75$ and $\xi = 4.7431$, which were estimated based on MC Hammer's album sales in order to ensure that our model would "stop, collaborate, and listen," similar to how YOLO [7] hyperparameters were fit on Drake's radio airtime.

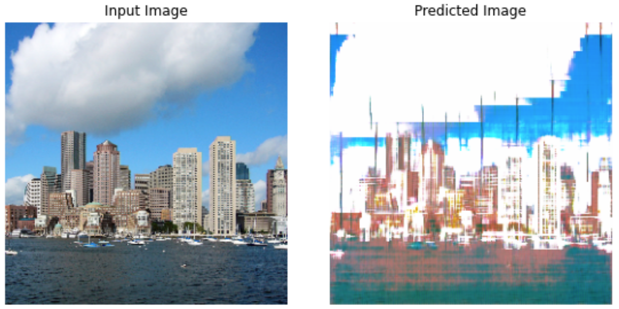

Figure 3. openCHEAT even generalizes to real-world images like the Boston skyline (source: https://commons.wikimedia.org/wiki/File:Boston_Financial_District_skyline.jpg). It must have learned that the world is three dimensional and can estimate building height reliably. From this plot we can finally see that the Boston skyline is statistically flat!

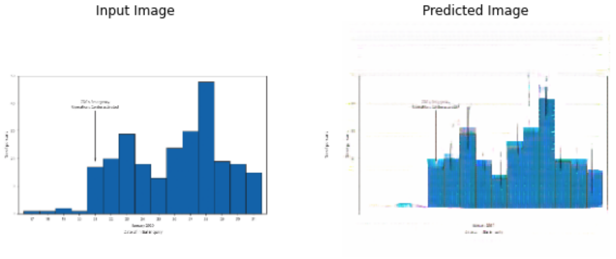

Figure 4. Pre-openCHEAT early COVID cases in the US on the left (source: CDC, https://www.cdc.gov/mmwr/volumes/69/wr/mm6906e1.htm?s_cid=mm6906e1_w) and the plot with predicted uncertainty on the right. This demonstrates that our tool can simulate multiple possible versions of the pandemic in parallel universes and report the result back. Our model also seems capable of finding dataset errors and fixing them.

We present our results in Figure 2. Our results speak for themselves and we observe all the nice properties we expected. All minor artifacts will disappear with additional training.

3.1 Generalization to real-world scenes

An important test for any machine learning system is that it does not just work on synthetic data, but also generalizes to real images. To test this, we used openCHEAT to estimate the errors on the heights of buildings in the Boston skyline (Fig. 3). While we see that the image quality degrades slightly,[8] openCHEAT is able to determine the uncertainty in the heights of the buildings. We find that, despite what the city architectural records tell us, there is considerable error in estimating the building heights, and therefore there is no reason to believe that the Boston city skyline is not, in fact, completely flat.

[7] https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/generative/pix2pix.ipynb

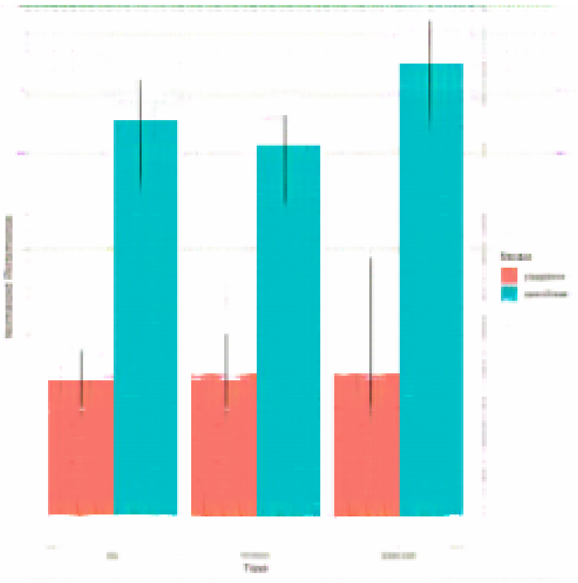

Figure 5. Performance of openCHEAT (blue) vs. baselines (red) on Go, protein folding, and Starcraft. openCHEAT's self-reported performance suggests that it can outperform state-of-the-art models even on tasks that it was not designed for.

3.2 Generalization to alternate realities

Our framework is entirely backwards compatible and can therefore be applied to existing and already published plots. Whilst some of those plots just miss error bars because scientists are lazy, for some experiments it might not be feasible or possible to derive error bars through experimentation. Our tool is however, so powerful, it can even estimate error bars for these non-repeatable experiments. We explored this on a pandemic related statistic[9] to demonstrate how powerful our method is (Fig. 4), and see that the model is able to produce error bars around a measured, past statistic. We can find only one possible explanation for how openCHEAT can accomplish this: it must have gained access to the multiverse where it can observe these outcomes in parallel realities to estimate the uncertainty.

3.3 Generalization to novel tasks

Because openCHEAT performs so spectacularly at the tasks it was designed for, we consider how it might be applied to entirely novel challenges that it had not been trained on. Here we consider its performance versus state-of-the-art models on Go [8], Starcraft [11], and protein folding [10]. As can be seen in Fig. 5, openCHEAT suggests that it outperforms these baselines by leaps and bounds. Note that openCHEAT did not actually perform these tasks, but instead reported its what its performance would be if it had performed these tasks, perhaps by accessing parts of the multiverse where it did so (see explanation above).

4. Conclusion

In this paper we demonstrate full automated science by introducing openCHEAT, a tool that adds error bars to any plot, thus satisfying reviewer concerns. Although trained on synthetic data, we demonstrate that it transfers to real-world images as well as to the multiverse. These results are so good that we plan no future work for model improvements.

However, with great power comes great responsibility [6]. While openCHEAT will revolutionize science, in the wrong hands it could produce untold devastation. Therefore, following industry standards, we are holding the code and model back from the public to prevent its use by malicious actors,[10] but are nonetheless willing to license it to the highest industry bidder.[11]

This work provides the first instance of fully automated science – Science 4.0.[12] This brings us one step closer to a scientific utopia where we can offload all of the hard work and thinking to automatic systems, and just reap the benefits of the citations to the papers they create.

References

[1] B. Egger and M. Siegel. HonkFast, PreHonk, HonkBack, PreHonkBack, HERS, AdHonk and AHC: the Missing Keys for Autonomous Driving. *SIGBOVIK*, 2020.

[2] B. Egger, K. Smith, and M. Siegel. openCHEAT: Computationally Helped Error bar Approximation Tool - Kickstarting Science 4.0. *SIGBOVIK (under careful review by very talented, outstanding reviewers)*, 2021.

[3] R. et al. Redacted. *REDACTED*, REDACTED.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[5] Y. LeCun, G. Hinton, and Y. Bengio. We reinvented the wheel. 2012.

[6] P. Parker and S. Lee. Spiderman. *Marvel*, 2002.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[8] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[9] M. E. Spear. Charting statistics. 1952.

[10] The AlphaFold team. AlphaFold: a solution to a 50-year-old grand challenge in biology. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

[11] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[8] This could be because we trained style transfer to simple images... but honestly we're too lazy to check.

[9] We're not sure what this statistic is or what it means, but we're hoping to jump on the COVID bandwagon.

[10] https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

[11] https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/

[12] Yes, we are aware of the SCIgen paper (https://pdos.csail.mit.edu/archive/scigen/), but since that doesn't use deep learning it is clearly inferior and so doesn't count.