Troubling Trends in Machine Learning Scholarship

Zachary C. Lipton & Jacob Steinhardt
Carnegie Mellon University, Stanford University

Presented at ICML'18 Debates Workshop
Slides By: Maitrey Mehta

MY MOTIVATION

- Most of the points are true and widespread
- A formal compilation of my experiences with research papers
- Highlights common mistakes that authors, readers and reviewers make
- A nice and new direction of work
- Comes from respected people who confess their crimes

“papers are most valuable to the community when they act in service of the reader, creating foundational knowledge and communicating as clearly as possible.”

“Flawed scholarship threatens to mislead the public and stymie future research by compromising ML’s intellectual foundations”

Desirable Characteristics of Papers

1. Provide intuition to aid the reader's understanding, but clearly distinguish it from stronger conclusions supported by evidence (e.g.- Cars and ML)
2. Describe empirical investigations that consider and rule out alternative hypotheses (e.g. - Ablation Studies)
3. Make clear the relationship between theoretical analysis and intuitive or empirical claims
4. Use language to empower the reader, choosing terminology to avoid misleading or unproven connotations, collisions with other definitions, or conflation with other related but distinct concepts (Make it a story someone would like to read)

Troubling Trends

1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words

Troubling Trends

1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words

Explanation vs. Speculation

- Papers often offer speculation in the guise of explanations

Example:

- The paper [1] states that batch normalization offers improvements by reducing changes in the distribution of hidden activations over the course of training. By which divergence measure is this change quantified? The paper never clarifies.
- [2] states,

> *"It is well-known that a deep neural network is very hard to optimize due to the internal-covariate-shift problem."*

- The authors of [3] carefully convey uncertainty. Instead of presenting the guidelines as authoritative, the paper states:

> *"Although such recommendations come. . . from years of experimentation and to some extent mathematical justification, they* ***should be challenged****. They constitute a good starting point. . . but very often* ***have not been*** *formally validated,* ***leaving open many questions*** *that can be answered either by theoretical analysis or by solid comparative experimental work".*

Troubling Trends

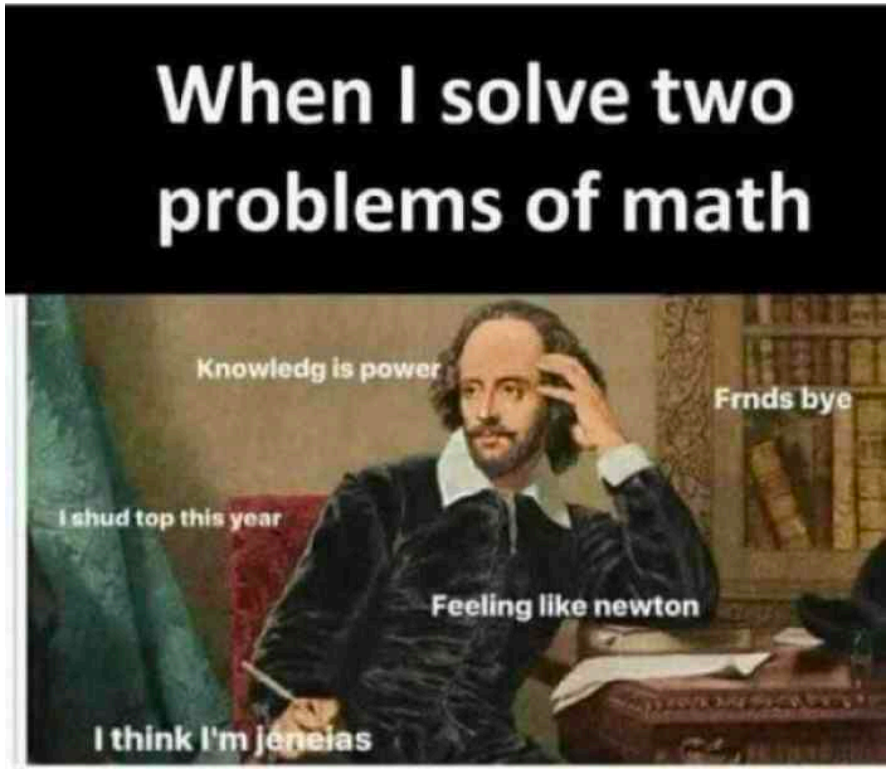
1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words

Failure to Identify the Sources of Empirical Gains

- Too frequently, authors propose many tweaks absent proper ablation studies, obscuring the source of empirical gains.
- Sometimes just one of the changes is actually responsible for the improved results. This can give the false impression that the authors did more work (by proposing several improvements), when in fact they did not do enough (by not performing proper ablations)
- Ablation is neither necessary nor sufficient for understanding a method and can even be impractical given computational constraints
- Understanding can also come from robustness checks (adversarial example papers) as well as qualitative error analysis

Troubling Trends

1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words



Mathiness

"When writing a paper early in PhD, we (ZL) received feedback from an experienced post-doc that the paper needed more equations. The post-doc wasn't endorsing the system, but rather communicating a sober view of how reviewing works."

- Not all ideas and claims are amenable to precise mathematical description
- When mathematical and natural language statements are mixed without a clear accounting of their relationship, both the prose and the theory can suffer: problems in the theory can be concealed by vague definitions, while weak arguments in the prose can be bolstered by the appearance of technical depth
- Spurious theorems inserted into papers to lend authoritativeness to empirical results, even when the theorem's conclusions do not actually support the main claims of the paper.[4]
- While the best remedy for *mathiness* is to avoid it, some papers go further with exemplary exposition
- Paper [5] on counterfactual reasoning covers a large amount of mathematical ground in a down-to-earth manner, with numerous clear connections to applied empirical problems (Shift of linear classifier with learning rate)

Troubling Trends

1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words

Misuse of Language: Suggestive Definitions

- A number of papers name components of proposed models in a manner suggestive of human cognition, e.g. “thought vectors” and the “consciousness prior”.
- Our goal is not to rid the academic literature of all such language; when properly qualified, these connections might communicate a fruitful source of inspiration.
- When a suggestive term is assigned technical meaning, each subsequent paper has no choice but to confuse its readers, either by embracing the term or by replacing it.

Troubling Trends

1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words

Misuse of Language: Overloading Terminologies

- Taking a term that holds precise technical meaning and using it in an imprecise or contradictory way.
- Consider the case of deconvolution, which formally describes the process of reversing a convolution, but is now used in the deep learning literature to refer to transpose convolutions.
- (My favorite- generative models. Now distinguished as *precise* and *implicit* generative models)

Troubling Trends

1. Explanation vs. Speculation
2. Failure to Identify the Sources of Empirical Gains
3. Mathiness
4. Misuse of Language
    1. Suggestive Definitions
    2. Overloading Technical Terminology
    3. Suitcase Words

Misuse of Language: Suitcase Words

- Suitcase Words: Coined by Minsky in *The Emotion Machine*.
- Mental processes such as consciousness, thinking, attention, emotion, and feeling that may not share “a single cause or origin”
- *Interpretability* holds no universally agreed-upon meaning, and often references disjoint methods and desiderata
- *Generalization* has both a specific technical meaning (generalizing from train to test) and a more colloquial meaning that is closer to the notion of transfer (generalizing from one population to another) or of external validity (generalizing from an experimental setting to the real world)

Speculation on Causes Behind the Trends

- Complacency in the Face of Progress:
  - *“strong results excuse weak arguments”*

- Growing Pains
  - Rapid expansion of the community can also have side effects
  - Rapid growth can also thin the reviewer pool, in two ways—by increasing the ratio of submitted papers to reviewers, and by decreasing the fraction of experienced reviewers.
  - Less experienced reviewers may be more likely to demand architectural novelty, be fooled by spurious theorems, and let pass serious but subtle issues like misuse of language, thus either incentivizing or enabling several of the trends described above.
  - At the same time, experienced but over-burdened reviewers may revert to a “check-list” mentality, rewarding more formulaic papers at the expense of more creative or intellectually ambitious work that might not fit a preconceived template. Moreover, overworked reviewers may not have enough time to fix—or even to notice—all of the issues in a submitted paper.

Speculation on Causes Behind the Trends

- Misaligned Incentives
  - As ML research garners increased media attention and ML startups become commonplace, to some degree incentives are provided by the press (“What will they write about?”) and by investors (“What will they invest in?”).
  - Overselling: The media provides incentives for some of these trends. Anthropomorphic descriptions of ML algorithms provide fodder for popular coverage. Take for instance [6], which characterizes an autoencoder as a “simulated brain”.
  - Hints of human-level performance tend to be sensationalized in newspaper headlines, e.g. [7], which describes a deep learning image captioning system as “mimicking human levels of understanding”.
  - Investors too have shown a strong appetite for AI research, funding startups sometimes on the basis of a single paper

Suggestions

- For Authors:
  - Ask “what worked?” and “why?”, rather than just “how well?”
  - Three practices that are common in the strongest empirical papers are error analysis, ablation studies, and robustness checks
  - When writing, we recommend asking the following question: *Would I rely on this explanation for making predictions or for getting a system to work?* This can be a good test of whether a theorem is being included to please reviewers or to convey actual insight.
  - Being clear about which problems are open and which are solved not only presents a clearer picture to readers, it encourages follow-up work and guards against researchers neglecting questions presumed (falsely) to be resolved
- For Reviewers:
  - *“Might I have accepted this paper if the authors had done a worse job?”*
  - For instance, a paper describing a simple idea that leads to improved performance, together with two negative results, should be judged more favorably than a paper that combines three ideas together (without ablation studies) yielding the same improvement.

References

1. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (ICML), 2015.
2. Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? (no, it is not about internal covariate shift). arXiv preprint arXiv:1805.11604, 2018.
3. Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In Neural networks: Tricks of the trade, pages 437–478. Springer, 2012.
4. Jacob Steinhardt and Percy Liang. Learning fast-mixing models for structured prediction. In Francis Bach and David Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 1063–1072, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/steinhardtb15.html.
5. Alan J Bray and David S Dean. Statistics of critical points of gaussian fields on large-dimensional spaces. Physical review letters, 98(15):150201, 2007.
6. Cade Metz. You don't have to be Google to build an artificial brain, 2014 — Accessed on July 4th, 2018. URL https://www.wired.com/2014/09/google-artificial-brain/.
7. John Markoff. Researchers announce advance in imagerecognition software, 2014 — Accessed on July 4th, 2018. URL https://www.nytimes.com/2014/11/18/science/researchers-announce-breakthrough-in-content-recognition-software.html. [Online; posted 26-September-2014].