Questionable Practices in Methodological Deep Learning Research

Daniel J. Trosten*†

Department of Physics and Technology, UiT The Arctic University of Norway

Abstract

Evaluation of new methodology in deep learning (DL) research is typically done by reporting point estimates of a few performance metrics, calculated from a single training run. This paper argues that this frequently used evaluation protocol in DL is fundamentally flawed – presenting 8 questionable practices that are widely adopted in the evaluation of new DL methods. The questionable practices are derived from violations of statistical principles of the scientific method, and from Hansson's definition of pseudoscience. A survey of recent publications from a top-tier DL conference indicates the widespread adoption of these practices in state-of-the-art DL research. Lastly, arguments in favor of the questionable practices, possible reasons for their adoption, and measures that have been taken to remove them, are discussed.

1 Introduction

Machine learning is the backbone of many of today's most impactful technological developments. Systems for *e.g.*, autonomous driving, speech and image recognition, and language translation have all advanced to human-level performance during the last few years – all thanks to machine learning, and in particular, the subfield of deep learning (DL) [11]. Models developed under the DL umbrella are often highly complex with millions or billions of parameters, requiring massive datasets and computational resources to train the model. Despite the high threshold set by the requirements for data and computational resources, its impressive results have caused DL to become the new hot topic in machine learning research. This, coupled with the increasing popularity of machine learning in general, has resulted in an explosion in the number of scientific publications on new DL methods and techniques.

Despite the massive research interest in DL methodology, the majority of publications follow a strict predefined narrative:

1. A new DL-based method for solving a particular problem is presented.
2. The new method is evaluated on a few benchmark datasets that are openly accessible and well known in the literature.
3. The evaluation shows that the proposed method largely outperforms all previously published methods on the benchmark datasets, thereby making the proposed method worthy of publication.

This competitive benchmarking approach to evaluation has been criticized in several previous works [6, 10, 12]. Hooker [6], for instance, argues that competitive benchmarking of heuristic algorithms might reveal what method is best, but it says little about why that is. They then proceed to suggest a more scientific approach to evaluation, which is closer to what is done in other fields of science. The more recent DL-specific meta-review by Liao et al. [12] provides a taxonomy of failure modes in methodological research, originating from critiques of evaluation in a broad selection of survey papers in DL. Lastly, from the perspective of applied DL research, Kleppe et al. [10] reviews recent publications on DL for cancer diagnosis, and find flawed evaluation protocols in a large proportion of the included publications.

The focus of this paper will also be on the evaluation protocol, but in contrast to previous work, it presents principled arguments on *why* the common evaluation protocol in methodological DL research

[*] Corresponding Author: daniel.j.trosten@uit.no.

[†] UiT Machine Learning Group: https://machine-learning.uit.no

https://doi.org/10.7557/18.6804

is flawed. It is argued that the evaluation procedures adopted in DL publications deviates from the statistical principles of the scientific method on important points. Statistical hypothesis testing – the standard for quantitative analysis in many fields of science – is extremely rare in methodological DL research. In fact, most publications do not even report estimates of uncertainty for the performance metric. Rather, standard practice is to report single point estimates of the chosen performance metric, not mentioning how many trials or how much tuning of the method that was required to achieve the reported result. Furthermore, in addition to the abandonment of crucial statistical principles, it is argued that Hansson's criteria for pseudoscience [3] are fulfilled by practices that are well established in the DL community.

The above observations are summarized as 8 *questionable practices* which are often encountered in contemporary DL research. The prevalence of these questionable practices is corroborated by a survey of papers from recent iterations of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) – a top tier conference in methodological DL research.

2 Statistical principles of evaluation

The scientific method refers to concepts and principles about how science should be conducted, in order to produce reliable results [5]. In this section, the focus will be on the use of statistics as a way to translate these concepts and principles into concrete methods, which in turn can be used to answer particular scientific questions – especially in the presence of randomness.

The main purpose of statistical inference in scientific methodology, is to provide a mathematically rigorous approach to inductive inference – *i.e.* inference about properties of a population based on random samples from that population. In inductive inference, it is crucial that one accounts for the randomness introduced by sampling, or by any other part of the data gathering process.

Accounting for uncertainty and various sources of randomness is precisely why the majority of scientists perform statistical hypothesis tests when inductively reasoning about a population. The main idea in hypothesis testing is to only accept a hypothesis about a population (or a group of populations) when the probability of the observed outcome given that the hypothesis is false, is sufficiently low [1]. This allows scientists to control the probability of accepting a false hypothesis, resulting in reliable conclusions about the population.

In methodological DL research, researchers are typically interested in corroborating general statements on the form: *"The proposed model outperforms the current state-of-the-art on tasks of type X"*. This is a statement about three different populations:

1. The population of tasks of type X. If X is image classification for instance, then the population will include *e.g.*, classifying images of cats and dogs, classifying medical images, classifying satellite images, *etc.* Thus, to infer about the proposed method's performance on tasks of type X, a sample of several tasks of type X has to be included in the evaluation.

2. The collection of methods that can be considered "the current state-of-the-art". This is because, contrary to the narrative many DL researchers like to present, there is not a single state-of-the-art for all tasks of type X, as indicated by the No Free Lunch Theorem in optimization [16]. Instead, one has to consider a collection of methods (*i.e.* a sample), each of which represent the current state-of-the-art for specific tasks of type X.

3. The last population in the statement above, is the population that arises when the proposed model is trained from different random initializations. Since the training of DL models requires the parameters to be randomly initialized, training typically results in models whose performance differ based on the initialization.

Statistical principles for proper evaluation of DL models are now starting to emerge. The goal is to reason about the performance of a population of proposed models, on a population of tasks of a specific type, compared to a population of previous state-of-the-arts (often referred to as *baselines*). This can be done inductively, by considering a sample of proposed models, a sample of baselines, and a sample of tasks – where the latter can be represented by a sample of datasets. It is at this point however, that the typical DL evaluation protocol starts to deviate from the statistical principles outlined above. Instead of considering sufficiently large samples from the three populations, DL methodology researchers typically compare one *single* training run of the proposed model to a few recent baselines, on a few datasets. Since the sample sizes are so small in all three cases – with the worst being a sample size of 1 for the proposed model – it is not possible to calculate measures of uncertainty, nor to perform any hypothesis tests for differences in performance.

These statistical flaws in methodological DL research can be summarized by the following two questionable practices:

$\mathcal{Q}_1$ Reporting point estimates of performance metrics, without uncertainty or confidence intervals.

$\mathcal{Q}_2$ Not conducting hypothesis tests to assess whether gain in performance over previous methods is statistically significant.

In the following sections, it will become clear that the adoption of these practices is both widespread in the field, and pushes parts of methodological DL research over the edge from science to pseudoscience.

3 Pseudoscientific practices

What separates science from non-science is a question that has been extensively studied in the philosophy of science [3]. The diversity of the past, present, and future scientific endeavors has made it difficult to determine general criteria for the demarcation between science and non-science – meaning that the demarcation problem is still an active field of study in today's philosophy of science.

The focus of this paper is on a particular form of non-science, namely *pseudoscience.* A key characteristic of pseudoscience is the *deviant doctrine*, which states that *"Pseudoscience [. . .] involves a sustained effort to promote standpoints different from those that have scientific legitimacy at the time."* [3]. It is also a commonly agreed-upon principle that for something to be pseudoscientific, it is done in such a way that it appears scientific, even though it is not. The notions of sustained effort and scientific appearance are reflected in Hansson's multi-criterial approach to defining pseudoscience [3]:

1. *Belief in authority: It is contended that some person or persons have a special ability to determine what is true or false. Others have to accept their judgments.*
2. *Unrepeatable experiments: Reliance is put on experiments that cannot be repeated by others with the same outcome.*
3. *Handpicked examples: Handpicked examples are used although they are not representative of the general category that the investigation refers to.*
4. *Unwillingness to test: A theory is not tested although it is possible to test it.*
5. *Disregard of refuting information: Observations or experiments that conflict with a theory are neglected.*
6. *Built-in subterfuge: The testing of a theory is so arranged that the theory can only be confirmed, never disconfirmed, by the outcome.*
7. *Explanations are abandoned without replacement: Tenable explanations are given up without being replaced, so that the new theory leaves much more unexplained than the previous one.*

Two of these criteria can be directly recognized in current evaluation protocols in methodological DL research:

- *Unrepeatable experiments*: Publications do not disclose all details of the evaluation, and omit important details about *e.g.* the method's configuration, or the preprocessing of datasets. Publications are not accompanied by open source code for the proposed method, making it difficult to reproduce by other researchers or practitioners.
- *Unwillingness to test*: This is related to the statistical flaws outlined in the preceding section, where the evaluation relies on performance metrics from a single run, instead of uncertainty estimates and hypothesis tests. Additionally, some publications make theoretical claims about why the proposed method works better than the baselines without backing them up, neither theoretically nor experimentally.

In addition to the two criteria above that can be recognized in publications, one can also hypothesize that several of the other criteria are satisfied by practices adopted "behind the scenes":

- *Belief in authority*: There likely exists a bias towards work that was done at high-profile labs, or work that was published in high-profile journals or conference proceedings. This bias results in publications from these labs or venues being subject to less scrutiny than other publications in the field. It also governs which publications are cited, and which methods are regarded as the current state-of-the-art, and thus included as baselines in subsequent publications.
- *Handpicked examples; Disregard of refuting information; Built-in subterfuge*: All these three criteria are satisfied by the practice of overfitting on the test-set, cherry-picking training runs, datasets, performance metrics, or baselines, in order to make it look like the proposed model out-performs the baselines.

With the establishment of the above pseudoscientific practices in methodological DL research, the list of questionable practices can be augmented with the following points:

$\mathcal{Q}_3$ Not publishing an open source implementation for the method or evaluation procedure.

$\mathcal{Q}_4$ Omitting some or all details about the method.

$\mathcal{Q}_5$ Omitting details about datasets (*e.g.*, train/test split, data preprocessing and normalization *etc.*).

$\mathcal{Q}_6$ Not backing up claims about why the proposed method works, neither empirically nor theoretically.

$\mathcal{Q}_7$ Disregarding work from lesser known labs, or work that was published in lower-ranked journals or conference proceedings.

$\mathcal{Q}_8$ Cherry-picking training runs, datasets, metrics, or baselines to make it look like the proposed model is the best.

Note that it is typically practices $\mathcal{Q}_1$ to $\mathcal{Q}_6$ that can be detected in publications. The last two, $\mathcal{Q}_7$ and $\mathcal{Q}_8$, are often adopted "behind the scenes" and are thus harder to detect by looking only at the publication.

4 Survey of recent DL papers

The purpose of this survey is to examine the adoption of practices $\mathcal{Q}_1$ to $\mathcal{Q}_6$ in methodological DL research[1]. The survey was performed with 10 papers from the 2016 to 2020 iterations of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) – the highest ranked conference in machine learning research[2]. The survey includes the 5 most cited papers, along with 5 papers randomly selected from the 200 most cited papers[3]. References, along with citation number and rankings of the included papers are listed in Table 1. The papers were read and judged by the author of this paper.

Table 1 presents the results of the survey. The most striking observation that can be made from these results is that all 10 papers adopt practices $\mathcal{Q}_1$ and $\mathcal{Q}_2$ – *i.e.*, none of them report measures of uncertainty or perform statistical hypothesis tests to illustrate the efficacy of the proposed method. Furthermore, the results show that all remaining practices, $\mathcal{Q}_3$ to $\mathcal{Q}_6$ are frequently observed in all papers. On a more positive note however, only 3 out of 10 papers adopt $\mathcal{Q}_3$, meaning that 7 papers have openly accessible source code for the method and experiments.

What is particularly interesting is that the papers included in this study are all highly cited, and published at the highest ranking conference in the field. One could therefore think that these papers represent the highest quality research in DL methodology, and thus that they are less likely to adopt questionable research practices, compared to other papers in the field. This opens up for speculation about the practices being even more common in the "average" methodological DL paper, than what is reflected by the results of this survey.

[1] Note that $\mathcal{Q}_7$ and $\mathcal{Q}_8$ are excluded as they are generally difficult to detect when only looking at publications.

[2] The ranking is based on the h5 index, and can be found at `https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng` (Accessed 10.03.22)

[3] List obtained from `https://scholar.google.com/citations?hl=en&vq=eng&view_op=list_hcore&venue=FXe-a9w0eycJ.2021` (Accessed 10.03.22).

Table 1: Results of survey performed with 10 papers from iterations of the CVPR conference from 2016 to 2020. Each row represents a paper, and checkmarks (✓) in the columns $\mathcal{Q}_1, \ldots, \mathcal{Q}_6$ indicate that the respective practices are adopted in the paper.

	Rank	Ref.	Citations	$\mathcal{Q}_1$	$\mathcal{Q}_2$	$\mathcal{Q}_3$	$\mathcal{Q}_4$	$\mathcal{Q}_5$	$\mathcal{Q}_6$
Top 5	1	[4]	82588	✓	✓	✓			✓
	2	[7]	17102	✓	✓				✓
	3	[14]	16833	✓	✓			✓	
	4	[15]	14252	✓	✓	✓			✓
	5	[9]	9543	✓	✓		✓	✓	
Random	158	[2]	639	✓	✓				
	26	[13]	2576	✓	✓		✓	✓	
	15	[17]	4595	✓	✓		✓		
	55	[18]	1344	✓	✓		✓	✓	
	96	[19]	922	✓	✓	✓	✓		✓

Note that this survey is limited to publications in Computer Vision. The prominence of the questionable practices might thus be different in different subfields of DL. Additionally, the survey could be more extensive with multiple evaluators, more papers from multiple fields, and quantitative analyses of the results. Expanding these aspects of the survey is left to future work.

5 Discussion

5.1 Arguments in favor of the questionable practices

Although the questionable practices have been presented in a negative light in this paper, there are arguments in support of their adoption. For instance, one might ask if it is fine to cherry-pick training runs, datasets and baselines, not report measures of uncertainty, and not perform hypothesis tests, simply because everyone else does it. After all, if all methods are compared in the same way on equal terms, is that not a fair evaluation? The answer to this argument is twofold. First, the widespread adoption of the questionable practices does not mean that all methods are compared on equal terms. Many researchers perceive these practices to lie in a gray area between what is morally right and wrong. To what degree the questionable practices are adopted will therefore vary between different cultures, different labs, and different individuals. Evaluation procedures can therefore not be truly fair when the questionable practices are adopted. Second, the fact that almost everyone does it, does not mean that it is good scientific practice. In fact, the fulfillment of Hansson's criteria for pseudoscience [3] illustrates that parts of the evaluation is not scientific at all.

Another question one might ask, is how can DL have led to so many technological breakthroughs, if research on DL methodology is riddled with questionable practices. The answer to this is that, even though much of the research is performed with these questionable practices, it is still possible to produce good results. Using point estimates in place of confidence intervals or hypothesis tests might not be statistically sound, but they still provide some information about the efficacy of the proposed method. Additionally, the success of DL has been significantly aided by recent advancements in computational capacity and increased access to large datasets. It is therefore possible to achieve much better results today than it was 10 years ago, using the same basic methodology, but trained for longer with more data.

Finally, one might argue that training models multiple times to estimate uncertainty, performing hypothesis tests or publishing open source code, requires extra time and resources, which most researchers do not have. This is a legitimate point, and it can not only be up to individual researchers to address the issue. Rather, systemic change has to come from all entities in the field. Research institutions, industry partners, publishers, and reviewers all have to contribute to rid the field of these practices.

5.2 Possible reasons for the adoption of the questionable practices.

It is not straightforward to determine the exact cause for the adoption of the questionable practices in methodological DL research. However, since the problem is related to the evaluation of new methods, the cause is likely linked to the excessive, one-sided focus on benchmark performance – both in peer review and in the general DL research community. Proposing a model that outperforms all competitors on the standard benchmark datasets is effectively a golden ticket to publication, regardless of the novelty and potential impact of the work. This, together with the competitiveness and pace of DL research, creates a strong incentive for researchers to adopt the questionable practices, in order to convey an overly optimistic representation of the performance of their method.

Peer review in DL also has the tendency to favor complex methods over simple ones. A method that is straightforward to understand and implement, but solves a problem equally well as its more complex counterparts, will often be rejected due to "lack of novelty". Complicated models are often harder to interpret and take longer to train, due to their increased number of components and parameters. This in turn makes it more difficult to report all details of the proposed model, and to properly evaluate it – effectively increasing the likelihood of the questionable practices being adopted.

5.3 Recent progress in eliminating the questionable practices.

The current reproducibility crisis in DL [8] has inspired several measures to improve the reproducibility of methodological DL research. Open source code is a crucial step towards reproducibility, which is now actively encouraged by several top-tier publication venues[4]. Additionally, open access to publications and open review processes are also becoming the new norm in DL research.

However, despite these advances in reproducibility, there is still a lack of focus on the statistical principles of evaluation. Measures of uncertainty, confidence intervals, or hypothesis tests are still not encouraged in the same way as open code. The community should thus continue striving for statistically sound methods of evaluation, in order to produce trustworthy results.

6 Conclusion

This paper shows that the standard evaluation protocol in methodological DL research deviates from important statistical principles of the scientific method. Furthermore, the standard evaluation protocol includes several elements that can be considered pseudoscientific, according to Hansson's criteria of pseudoscience [3].

The problematic aspects of the evaluation protocol are summarized as 8 concrete questionable practices, whose widespread adoption was demonstrated by a survey of recent papers published at a top-tier DL conference.

It is not easy to say why these practices have become so popular, but the overall pace of the field, and the narrow focus on benchmark performance, are likely candidates. Several high-impact publication venues have taken measures to reduce the adoption of some questionable practices, but much work still remains to remove them completely.

Acknowledgements

The author would like to thank Professor Robert Jenssen for valuable feedback on an early draft of the manuscript.

This work was financially supported by the Research Council of Norway (RCN), through its Centre for Research-based Innovation funding scheme (Visual Intelligence, grant no. 309439), and Consortium Partners. It was further funded by RCN IKTPLUSS grant no. 303514.

References

[1] G. Casella and R. L. Berger. *Statistical Inference.* Thomson Learning, Australia ; Pacific Grove, CA, second edition, 2002. ISBN 978-0-534-24312-8.

[4] See *e.g.* `https://cvpr2023.thecvf.com/Conferences/2023/AuthorGuidelines`, `https://nips.cc/Conferences/2022/CallForPapers`, and `https://icml.cc/Conferences/2022/CallForPapers`.

[2] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *CVPR*, 2018. doi: 10.1109/CVPR.2018.00762.

[3] S. O. Hansson. Science and Pseudo-Science. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2021 edition, 2021.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.90.

[5] B. Hepburn and H. Andersen. Scientific Method. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2021 edition, 2021.

[6] J. Hooker. Testing Heuristics: We Have It All Wrong. *Journal of Heuristics*, 1:33–42, 1995. doi: 10.1007/BF02430364.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *CVPR*, 2017. doi: 10.1109/CVPR.2017.243.

[8] M. Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. doi: 10.1126/science.359.6377.725.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-To-Image Translation With Conditional Adversarial Networks. In *CVPR*, 2017. doi: 10.1109/CVPR.2017.632.

[10] A. Kleppe, O.-J. Skrede, S. De Raedt, K. Liestøl, D. J. Kerr, and H. E. Danielsen. Designing deep learning studies in cancer diagnostics. *Nature Reviews. Cancer*, 21(3):199–211, 2021. doi: 10.1038/s41568-020-00327-9.

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

[12] T. Liao, R. Taori, I. D. Raji, and L. Schmidt. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Advances in Neural Information Processing Systems*, 2021.

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.282.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.91.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.308.

[16] D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

[17] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*, 2017. doi: 10.1109/CVPR.2017.634.

[18] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual Dense Network for Image Super-Resolution. In *CVPR*, 2018. doi: 10.1109/CVPR.2018.00262.

[19] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-Ranking Person Re-Identification With k-Reciprocal Encoding. In *CVPR*, 2017. doi: 10.1109/CVPR.2017.389.