Post-Selections in AI Papers in *Nature* since 2015 and the Appropriate Protocol

Juyang Weng[1,2]
[1]Brain-Mind Institute
[2]GENISAMA LLC
4460 Alderwood Drive, Okemos, Michigan 48864 USA

Submitted July 24, 2021
Revised March 16, 2022 in red
Revised April 7, 2022 in blue
Revised May 19, 2022 in magenta

Abstract

Through a review of AI papers published in *Nature* since 2015, this report discusses major technical flaws called Post-Selection in the papers. This report suggests an appropriate protocol, explains reasons for the protocol, and why what the papers have done is inappropriate and therefore yields misleading results. The charges below are applicable to whole systems and system components, and in all learning modes, including supervised, reinforcement, swarm, reservoir, and evolutionary learning modes, since the concepts about training sets, validation sets, and test sets all apply. A reinforcement-learning algorithm includes not only a handcrafted form of task-specific, desired answers but also values of all answers, desired and undesired. A supervised learning method typically does not provide values for intermediate steps (e.g., hidden features). But in contrast, a reinforcement learning mode must provide values for intermediate steps using a greedy search (e.g., time discount). Casting dice is the key protocol flaw that owes a due transparency about all losers (e.g., how good they are). A commercial product is impractical if it requires every customer to cast dice and almost all trained "lives" must cause accidents and be punished by deaths except the luckiest "life". All the losers and the luckiest are unethically determined by so called "unseen" (in fact should be called "first seen") test sets but the human programmer saw all the scores before he decided who are losers and who is the luckiest. Such a "deep learning" methodology gives no product credibility.

I. Introduction

Restricted to only technical subjects, the Claimant hereby respectfully alleges that the following papers published in *Nature* [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] suffer from the following major technical and protocol flaws and, consequently, their reported results are grossly misleading.

For a layman to understand the flaw, an analogy of Post-Selection is as follows: "The authors of a lottery-like scheme claimed that their method has won $1M, but concealed that the scheme has spent 2M lottery tickets of $1 each. The reported ticket is only the luckiest. Whether a ticket works depends on the actual lottery test. The same sequence of numbers on the luckiest ticket will not have the same luck in the next lottery." The numbers like $1M, 2M and $1 and the chance for each ticket to win in the analogy differ across different reports, but the nature of the flaw in the reports is basically the same.

Geoffrey Hinton admitted in his PubPeer response to questions raised on PubPeer towards [1] that Kriszhevsky & Hinton [18] reported the "luckiest" network.

The authors in the charged papers have test sets in their possession. Did they used the test sets in the Post-Selection stage (like a real lottery) to generate the performance data? Fig. 1 shows a piece of direct evidence: Table 2 of Kriszhevsky & Hinton [18]. In the last row "7 CNNs", the test error 15.3% (100 images per class) is smaller than the validation error 15.4% (50 images per class). ImageNet provided on average 1000 training images for each class; 1000 classes amounts to a little over $1000 \times 1000 = 1,000,000$ images in the training set. Kriszhevsky & Hinton claimed the 7 CNNs were "pre-trained" to classify the

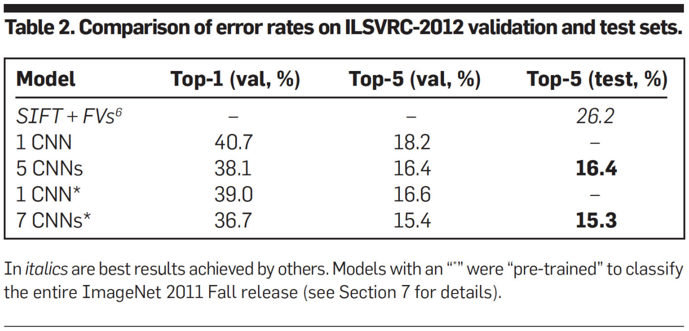

Table 2. Comparison of error rates on ILSVRC-2012 validation and test sets.

Model	Top-1 (val, %)	Top-5 (val, %)	Top-5 (test, %)
SIFT + FVs[6]	–	–	*26.2*
1 CNN	40.7	18.2	–
5 CNNs	38.1	16.4	**16.4**
1 CNN*	39.0	16.6	–
7 CNNs*	36.7	15.4	**15.3**

In *italics* are best results achieved by others. Models with an "*" were "pre-trained" to classify the entire ImageNet 2011 Fall release (see Section 7 for details).

Fig. 1. Table 2 from Kriszhevsky & Hinton 2017 [18]. In the last row, the test error is smaller than the validation error. The ImageNet-reported numbers for the second and third bold numbers along the Top-5 (test %) column are 16.422 and 15.315, respectively.

entire ImageNet 2011 Fall release. The ILSVRC-2011 training set could be a subset of ILSVRC-2012 training set (increased 4.2%). The $100 \times 1000 = 100,000$ test images may change from 2011 to 2012 [19, Tables 2 and 3] but the 1000 test-class labels should remain the same from 2011 to 2012. Thus, the luckiest network from Post-Selections for ILSVRC-2011 (e.g., that minimizes the sum of training error, validation error and test error) should be almost the luckiest for ILSVRC-2012 as only 4.2% images were added into ILSVRC-2012.

Imagine the following questions in a civil court that should be answered based on *a preponderance of evidence*:

1) Did the authors conduct Post-Selections for [18] using a training set, a validation set, a test set or a combination thereof?
2) As the validation error 15.4% is larger than the test error 15.3%, is it more likely that the authors used a test set (or both test and validation sets) for Fig. 1 than not using a test set at all?
3) Did the authors lack due transparency about the Post-Selection stage in [18]?

II. Charges

C1 **Post-Selection**

Their experimental protocols should have included a stage, called Post-Selections Using Test Sets (PSUTS) or Post-Selections Using Validation Sets (PSUVS). In the simplest form, all the data were divided into three disjoint sets, training set T, validation set V and test set T'. In a training stage, $n \geq 2$ systems were trained to fit the training set T, each starting from a different set of weights determined by a different random seed. Next, in the Post-Selection stage, all the n systems were tested on the validation set V or test set T' and finally only the performance of the luckiest system that has the best performance was reported, but not the performances of the remaining less lucky $n-1$ systems. This is a version of "testing on the training set" as textbook [20] warned against, because the Post-Selection stage utilized the validation set V or the test set T' that is not "blind" to the authors. In competitions, the Post-Selections take a form of machine Post-Selections and human Post-Selections.

C2 **Lack of transparency**

There is a lack of transparency in reporting the Post-Selection stage. Almost all the charged papers did not mention the Post-Selection stage at all. An exception is [4] which mentioned in the caption of Figure 5, "20 replicated training runs with different random-number seeds for a DNC and LSTM ... A single DNC was ... some failures to satisfy all constraints (incomplete)." However, [4] still lacks due transparency about the Post-Selection stage. For example, do the performance data, including those in Fig. 4(b), correspond to the luckiest system among these $n = 20$ trained systems? If the answer is positive, what is the distribution of performances of other less lucky $20-1 = 19$ systems? As a publicly disclosed case, [1] misleadingly claimed: "...

until ImageNet competition in 2012. When deep convolutional systems were applied to a data set of about a million images from the web that contained 1,000 different classes, they achieved spectacular results, almost halving the error rates of the best competing approaches." But [1] did not explain the Post-Selection stage. The least transparent case among the charged papers seems to be paper [14] which did not state which AI method was used. Paper [14] stated: "We used the inverse probability of treatment weighting to adjust for baseline confounding factors and to emulate randomization." But, it did not mention (i) what AI method (re: inverse probability IPTW which requires machine learning) was used, and (ii) what feature representations of IPTW were applied to the "Flatiron Health database" and (iii) how the uncertainty (re: inverse probability IPTW) in the real data are cross-validated to support the reported results about "relaxing specific eligibility criteria".

III. Supporting Brief

S1 What the appropriate protocol should be

If one has to use a weak (batch learning) technology that requires many trained systems instead of only one, the following protocol should be carried out to evaluate the technology.

(A) All available data should be divided into three mutually disjoint sets, a training set T, a validation set V, and a test set T'.

(B) In the validation stage, all systems trained to fit the training set T are validated using the validation set V, but not the test set T' (which must not be leaked into the post-selection stage). The validation rate—the ratio of the number of systems that pass the validation over the total number of systems trained—must be reported. Next, during the test stage using T', the performances of all trained systems must be reported (including those validated and not validated). The size of every network and the computational resources (the number of floating point operations per second and the time spent for each network output should be reported so that the technology can be compared with other technologies on an equal footing.

(C) According to the well known protocol of cross-validation [20, p. 483-484], random lucks (including all those lucks that a user does not fully control during deployment, like deployment of a vaccine) should be averaged out to report at least the average performance on T' across all n systems trained. One may use n-fold cross-validation [20, p. 483-484] across all n random seeds, tested on the test set T'. In other words, the charged papers should report the average performance on a completely new test set T' across all n systems trained, including those not validated, instead of only the performance of the luckiest system from Post-Selections.

(D) To estimate the reliability of the average performance, report also, in additional to the average of performances, also the distribution of all n performances across different random seeds, including the minimum, 25%, 50%, 75%, the maximum, and the standard deviation of all the n performances. This sensitivity report should be also done for the recommended hyper parameters and report how the performance is sensitive to the hyper parameters.

(E) If the search for random weights is coupled with the search for the architecture hyper-parameters, report the distribution of performances of kn systems as in (D), where k is the number of hyper-parameter vectors searched [4]. Alternatively, declare that random seeds for weights are *decoupled* from the search for the best hyper parameter vector. Namely, every hyper parameter vector tried in search uses the same initial value of the random seed for assigning random weights of the neural network.

(F) For a machine learning competition, the competition organizer should publicly announce and strictly enforce stipulations that explicitly ban PSUTS, either the currently rampant machine PSUTS or suspected human PSUTS in [21], [3], [5], [12] via human on-the-fly interactions with the decision process of a competing machine.

S2 Why this is the appropriate protocol

(A) Some available data sets provide a validation set (like a mock examination); or some competition teams privately extract a disjoint validation set from available data.

(B) Each competition team should not use the test set T' in training or post-selection, nor should it hand-label the test set from a competition. The validation rate is used to see how effective a machine learning technique is, but all trained systems should be all tested and reported to be transparent about the high cost of machine training and how sensitive the learning technique is to the random guess.

(C) An objective test should provide statistically expected error of a single resulting system for a typical deployment. Statistically, the luckiest system on V (or T') is expected to give only an average performance across all possible random seeds on a new validation set V (or a new test set T') drawn from the same distribution. Thus, instead of reporting the luckiest performance, report the average performance to smooth out random-seed lucks that a typical user in deployment cannot control. The traditional n-fold cross-validation uses average to smooth out similar lucks in dividing all available data into T, V and T'. Why including those not validated? A customer cannot afford to validate the technology (like a vaccine) himself.

(D) If the distribution of n performances is wide, the reported average performance is not trustable. This is similar to, but not the same as, so called p-value and confidence in statistical science.

(E) It is necessary because if one claims $n = 1$ for each hyper-parameter vector, in fact the search for the luckiest random seed is embedded in the search for the $kn = k$ hyper-parameter vectors and k is typically huge in machine learning unless a decoupled random seed is declared. Alternatively, a decoupled seed prevents the search for random weights hiding in the search for hyper parameter vectors.

(F) Without the stipulation, a competition against a machine is unfair since it is in fact a competition with a team of humans who have a lot of computer support.

S3 **Why what those papers have done is inappropriate and therefore yields misleading results**

(A) Since the charged papers did not provide any statements to state otherwise, it is reasonable for a reader to assume that the test set T' was used in Post-Selection. For the same lack of transparency, textbook [20, p.483] wrote: "It is essential that the validation (or test) set not include points used for training the parameters in the classifier—a methodological error known as 'testing on the training set'."

(B) The luckiest system from V (or T') is like a luckiest hit in a lottery of random seeds. The charged papers should not report only the misleadingly high recognition rate of the luckiest system, but instead the performances of all trained systems. Textbook [20, p.295] stated: "The average error on an independent test set is virtually always higher then on the training set." The luckiest results from Post-Selections in the charged papers are grossly misleading.

(C) The charged papers should transparently report how many systems they have trained (e.g., 10,000 by [10]) and the average performance across all trained systems (not just 165 luckiest ones in [10]). Without this information, it is misleading for [1] to claim credits for error-backprop techniques because error-backprop also resulted in those less lucky systems.

(D) The charged papers did not transparently explain the distribution of performances across a huge number of trained systems from less lucky random seeds.

(E) The charged papers did not present how sensitive the reported performance is to the hyper-parameters that only greedily fit a specific data set. The charged papers did not transparently account for the distribution of performances across a huge number of trained systems with different random seeds and different hyper parameters tried. The charged papers did not declare a decoupled random seed either.

(F) The competitions discussed in [21], [1], [3], [5], [12] lack the stipulation in S1.F and therefore the publicly announced results from these competitions are misleading.

Without further discussing each of the charged papers, the authors and readers of all the charged papers should be able to understand how the Post-Selections charge C1 and the transparency charge C2 explained here are applicable specifically to each of the charged papers.

The Claimant intends to make this report concise and free from mathematics. The allegations are rooted in well-known pattern recognition protocol, such as the n-fold cross-validation in [20] and are supported by the related theoretical analysis and experimental experience from the Claimant's group.

For more technical details about why Convolutional Neural Networks (CNNs) trained by gradient decent techniques severely suffer from local minima problems, see [22]. For some variants of Post-Selections and why Post-Selections are technically flawed for not only CNNs but also all other AI methods that do not automatically abstract rules and purposes, such as published swarm learning and evolutionary computations, see [23].

IV. Conclusions

To deny the allegations, the authors of the charged papers should publicly provide to *Nature* scientifically verifiable source programs and data sets T, V and T' along with additional data in S1.C, S1.D, S1.E, and if competitions are involved, also S1.F. If the authors of the charged papers failed to provide such denial in a reasonable time frame or failed to receive positive verifications by other independent laboratories including the Claimant, the authors of the charged papers should voluntarily retract their corresponding papers according to *Nature* and COPE rules. None of the authors who responded to *Nature* editor's communications denied the Post-Selection stage. With the updated direct evidence provided, we can see that transparency of the Post-Selection stage is substantial for the trustability of the reported methods and data. Hereby, the Claimant respectfully request that the Editor-In-Chief should retract all the charged papers due to their lack of transparency. The editor-in-chief has a responsibility to retract based on a preponderance of evidence.

References

[1] LeCun, Y., Bengio, L. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[2] Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

[3] Silver, D. *et al.* Mastering the game of go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

[4] Graves, A. *et al.* Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).

[5] Silver, D. *et al.* Mastering the game of go without human knowledge. *Nature* 354–359 (2017).

[6] McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).

[7] Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

[8] Bellemare, M. G. *et al.* Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* **588**, 77–82 (2020).

[9] Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O. & Clune, J. First return, then explore. *Nature* **590**, 580–586 (2021).

[10] Saggio, V. *et al.* Experimental quantum speed-up in reinforcement learning agents. *Nature* **591**, 229–233 (2021).

[11] Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).

[12] Slonim, N. *et al.* An autonomous debating system. *Nature* **591**, 379–384 (2021).

[13] Mirhoseini, A. *et al.* A graph placement methodology for fast chip design. *Nature* **594**, 207–212 (2021).

[14] Lu, M. Y. *et al.* AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).

[15] Warnat-Herresthal, S. *et al.* Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).

[16] Assael, Y. *et al.* Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022).

[17] Lu, H. *et al.* Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature* **604**, 662–667 (2022).

[18] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**, 84–90 (2017).

[19] Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).

[20] Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification* (Wiley, New York, 2001), 2nd edn.

[21] Silver, A. Deep blue's cheating move. *Chess News* (2015). https://en.chessbase.com/post/deep-blue-s-cheating-move.

[22] Weng, J. On post selections using test sets (PSUTS) in AI. In *Proc. International Joint Conference on Neural Networks*, 1–8 (Shengzhen, China, 2021).

[23] Weng, J. A developmental method that computes optimal networks without post-selections. In *Proc. IEEE International Conference on Development and Learning*, 1–6 (Beijing, China, 2021).