A Report to *CACM* on Data Deletions by Krizhevsky, Sutskever and Hinton

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

Submitted Dec. 14, 2021
Updated: June 6, 2022 in blue

Abstract

The author reports hereby to *Communications of the ACM* deleting (undesirable) data through severe protocol flows in Krizhevsky, Sutskever and Hinton (KSH) 2017 [1] that appeared in *Communications of the ACM*. This report analyzes that KSH should have used Post Selections Using Test Sets (PSUTS) that grossly violate the well-established statistical protocols. If the allegations are found true, the results reported by KSH are grossly misleading and the paper should be retracted. This report suggests the appropriate protocol, explains reasons for the protocol, why what the KSH paper has allegedly done is inappropriate and therefore yields misleading results.

I. Introduction

Restricted to only technical subjects, the Claimant hereby respectfully allege that the KSH paper [1] suffer from the following major technical and protocol flaws. An earlier version of the KSH paper appeared in NIPS 2012. This report is important since, if the allegations in this report are found true, a wide-scale perception change in the AI Community is necessary due to the visibility of this work in the ImageNet ILSVRC-2012 Competition.

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 [5] that Kriszhevsky & Hinton [1] reported the "luckiest" network. Namely, Hinton admitted that the authors conducted data deletions.

Furthermore, the flaws also include tests on training sets. 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 [1]. 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 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 [3, 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.

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 [1]. 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.

II. Charges

This report alleges two following charges, Post Selection and Lack of Transparency.

C1: **Post Selection**

The experimental protocols of the KSH paper should have included a stage, called Post Selection. The Post Selection stage has two technical flaws, Post Selection Using Test Sets (PSUTS) and Post Selection Without Cross-Validation (PSWCV). In the simplest form of PSUTS, all the data were divided into two disjoint sets, T and T'. In a training stage, n systems where 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 test set T' and finally only the performance of the luckiest system that has the best performance on the test set T' was reported in the corresponding paper, but not the performances of the remaining less lucky $n-1$ systems. This is a version of "testing on the training set" as textbook [2] warned against, because the Post Selection stage utilized the test set T'. In the PSWCV, no cross-validation was conducted to switch the role of training set T and test set T', so that the training set T were carefully selected (e.g., by the ImageNet organizers [3]) to make tests on T' easy.

C2: **Lack of transparency**

There is a lack of transparency in reporting the Post Selection stage. the KSH paper did not mention the Post Selection stage at all. If the result did not use PSUTS, each accuracy number must report three associated numbers—fitting error, validation error and disjoint test error. If the result did not use PSWCV, each number must also report the fourth associated number—standard deviation due to cross-validation. But the KSH paper gave only one number, implying that they used PSUTS and PSWCV.

III. Supporting Brief

S1: **Wide-spread protocol flaws**

Few papers that did the Post Selection even mention the Post Selection. A paper in *Nature* 2016 [4] mentioned the Post Selection 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 widely advertised case, [5] in *Nature* 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 [5] did not explain the Post Selection stage, which

should have used PSUTS on team-labeled test sets. Such a claim is closely related to the KSH paper. Geoffrey Hinton, an author of [5], blankly responded to the first question in PubPeer and stopped responding when the questions got harder to respond blankly. In addition, apparently no cross-validation was conducted by [4] or [5].

S2: **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'. The KSH paper has a validation set V, but V should be superficial because of PSUTS.

(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 along with the luckiest validation error. Next, during the test stage, the performances of all trained systems must be reported (including those not validated).

(C) According to the well-known protocol of cross-validation [2, 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 [2, p. 483-484] across all n random seeds, tested on the test set T'. In other words, the KSH paper 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 PSUTS.

(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, the maximum, and the standard deviation of all the n performances. This should be done for the paper-recommended hyper parameter vector below.

(E) The distribution of performances of kn systems, where k is the number of hyper-parameter vectors searched [4], should also be reported in the format of (D) since such a search is coupled with a search for seeds of random weights. 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 ransom seed for assigning random weights of the neural network.

S3: **Why this is the appropriate protocol**

(A) Some available data sets provide a validation set; 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 initial 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 and T'.

(D) If the standard deviation of n performances is large, the reported average performance is not trustable. This is similar to, but more transparent than, so called p-value in statistical science.

(E) S2(E) 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. A decoupled random seed prevents the search across random seeds from being hidden in the search for the hyper-parameters.

S4: **Why what those papers have done is inappropriate and therefore yields misleading results**
(A) Since the KSH paper 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 [2, 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 KSH paper should not report only the misleadingly high recognition rate of the luckiest system, but instead the performances of all trained systems. Textbook [2, p.295] stated: "The average error on an independent test set is virtually always higher then on the training set." The luckiest results from PSUTS in the KSH paper is grossly misleading.
(C) The KSH paper should transparently report how many systems they have trained and the average performance across all trained systems. Without this information, it is misleading for [5] in *Nature* to claim credits for error-backprop techniques because error-backprop also resulted in those less lucky systems.
(D) The KSH paper did not transparently explain the distribution of performances across an unknown number of trained systems from less lucky random seeds.
(E) The KSH paper did not present how sensitive the reported performance is to the hyper-parameters that only greedily fit a specific test set. The KSH paper did not transparently account for the distribution of performances across a unknown number of trained systems with different random seeds and different hyper parameters tried. The KSH paper did not declare a decoupled random seed.

S5: Page 89, left column, lines 1-2, the KSH paper reads "in the remainder of this paragraph, we use validation and test error rates interchangeably because in our experience they do not differ by more than 0.1% (see Table 2)."
However, in Table 2, let us look at the 7NNs case: The validation error (15.4%) is surprisingly larger than the test error (15.3%). (Like KSH, the authors of [6] who also reported test errors that are surprisingly smaller than validation errors.) Here, with 50,000 validation images to average over for the validation error versus 150,000 test images to average over for the test error, it is extremely unlikely that Post Selection Using Validation Sets (PSUVS) was used instead of PSUTS. This is an evidence that PSUTS was used instead of PSUVS. If the authors want to deny this evidence, they should provide independently verifiable source program and data from which 15.4% and 15.3% are generated. Since 50,000 is considerably smaller than 150,000 and validation set might be easier than the test set, the Claimant guesses that the luckiest test error using PSUVS is significantly larger than 15.4% and the average error of all n training systems will be further much larger.

S6: Hinton wrote in his only response in PubPeer toward paper [5]: "In the ImageNet competition the test set was only known to the people who ran the competition". This is blank and not accurate because, unfortunately, the test sets of ILSCRC 2012 (without answers) were released at least by July 12, 2012, but the contest answers (not including teams' programs) were not due till Sept. 30, 2012. See https://image-net.org/challenges/LSVRC/2012/index.php. All ImageNet 2012 contest teams had as long as over 2.5 months to look at the test sets while there were no explicit ImageNet Contest rules that ban teams from labeling the test sets themselves during algorithm developments. This time period seems too long as each team had a great pressure to label the test sets, at least to find out how their algorithms did on the test sets before submission. See below for how test answers were in fact released indirectly.

S7: In fact, the teams did not need to hand-label the test sets themselves. From as early as January 26, 2012, ImageNet organizers (improperly?) provided an Evaluation Server, stating "you can evaluate you[sic] own results". See https://image-net.org/challenges/LSVRC/2012/index.php It is not difficult for one to write a script program to automatically get all the test labels from the Evaluation Server. Therefore, the test answers were at least disclosed in an interactively way through the Evaluation Server.

S8: See more direct questions in PubPeer toward paper [5] that the authors chose not to respond at all.

S9: Without further discussing other details the KSH paper, the authors and readers of the paper should be able to understand how the PSUTS and PSWCV charge C1 and the transparency charge C2 raised here are applicable to the entire work reported by the KSH paper.

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 [2] and are supported by the related theoretical analysis and experimental experience from the Claimant's group.

For more technical details about why CNNs and LSTMs trained by gradient decent techniques severely suffer from local minima problems, see [7]. For some variants of Post Selections that amount to a flawed protocol for not only CNNs and LSTM but also all other AI methods that do not automatically abstract rules and purposes, such as published swarm learning and evolutionary computations, see [8].

IV. Redress Sought

To deny the allegations, the authors of KSH paper should publicly provide, to the journal, scientifically verifiable source programs and program-ready data sets T, V and T' along with additional data in S2(C), S2(D), S2(E), so that independent labs need to simply run the programs and generate all the data in Table 2.

If the authors of the KSH paper failed to provide such denial in a reasonable time frame (e.g., 30 days) or failed to receive positive verifications by other independent laboratories and the Claimant, the authors of the KSH paper should voluntarily and timely retract the KSH paper according to the applicable rules of the journal and the COPE rules.

References

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

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

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

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

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

[6] Gao, Q., Ascoli, G. A. & Zhao, L. BEAN: Interpretable and efficient learning with biologically-enhanced artificial neuronal assembly regularization. *Front. Neurorobot* 1–13 (2021). See Fig. 7.

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

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