IEEE

CDS NEWSLETTERS

Volume 18, Number 2 ISSN 1550-1914 May 2024

Development of Natural and Artificial Intelligence

Contents

1 [Post Selection] Dialogue: Lack of Generalizability in Post-Selected Models

Badal Gami, Delhi Technological University, Delhi, India
Email: badalgami@gmail.com

Xiang Wu raised the issue about the generalizability of post-selected predictors in the previous dialogue initiation labeled [Deep Learning], I report how I view this issue of optimistic bias due to post-selection.

As Weng [5] argued, simply reporting the "luckiest network" based on the validation error is a severe technical flaw. The errors of the luckiest network reported are fit errors, not test errors. As the saying goes, "Even a blind squirrel finds a nut once in a while - but that does not make the squirrel an expert on locating nuts!" Similarly, a model that fits well the validation data used for selection does not guarantee its generalization with unseen test data. Wu's experimental data in [Deep Learning] Dialogue showed that, indeed, the luckiest blind squirrel did badly in new environments, not better than other less lucky blind squirrels.

Qiu's mathematical analysis [3] formalizes this problem. Let $E(g_k, V)$ be the average error of model k (g_k) on validation dataset V, and $g_{\hat{k}}$ be the luckiest model post-selected from V. Qiu showed that if one does many many experiments, the expected validation error of the luckiest network (i.e., the luckiest blind squirrel) under-estimate that of the best network in a future test (using Dr. Weng's short notation):

$$\mathbb{E}_V E(g_{\hat{k}}(g_1, ..., g_K, V)) \leq \min_{g_j \in \{g_1, ..., g_K\}} \mathbb{E}_T E(g_j, T)$$

The validation error of the luckiest model $g_{\hat{k}}$ is a downward biased estimate that underestimates the expected test error of even the best candidate model g_k from the same population. This bias can be quantified as:

$$\text{Bias} = \mathbb{E}_V E(g_{\hat{k}}(g_1, ..., g_K, V)) - \min_{g_j \in \{g_1, ..., g_K\}} \mathbb{E}_T E(g_j, T)$$

Qiu proved that under very general conditions, this bias is often negative i.e. Bias < 0. The necessary and sufficient condition for exception Bias $= 0$ is highly artificial. That is, the bias is zero when the randomness from the validation set to the test set totally vanishes. This is like as long as the blind squirrel runs into a nut in the validation set, it is guaranteed that the same blind squirrel will run into the same nut in all future tests! Obviously, this situation would not occur in reality.

It is important to note that the above Bias is based on theoretical expectations $\mathbb{E}_V$ and $\mathbb{E}_T$, which are impossible to compute exactly in any real experiments. Below, I will discuss my experimental results, where

we will use sample mean to approximate the theoretical mean.

Weng [6, Theorem 3] has proven that the minimum MSE (Mean Square Error) estimator of any trained network (luckiest on V or not) is the average error of all trained networks on V. Wu [8] experimentally confirmed Weng's Theorem 3. To illustrate this further, I discuss my experimental results [1] on Alzheimer's disease classification from brain MRI scans using deep convolutional neural networks. The problem statement was to evaluate and compare the performance of popular feature extraction techniques for Alzheimer's disease predictive models using brain imaging data from an openly available Kaggle MRI dataset with 4 classes. The input to the networks was the MRI brain image, and the output was the predicted class representing the level of dementia. The training method involved a Convolutional Neural Network (CNN) with convolutional, pooling, normalization, and fully connected layers, using cross-entropy loss and the ADAM optimizer. The study aimed to determine the best feature extraction technique by comparing standard parametric observations. We trained 20 different VGG-19 models with different pre-trained weights and regularization hyperparameters on a fitting dataset F. On a held-out validation set V, we selected the model $g_{\hat{k}}$ that achieved the highest AUC (Area Under Curve) of 0.9185 among the 20 models as shown in the Table 1. This is a case of post-selection, where $g_{\hat{k}}$ was chosen by evaluating all 20 models on V and picking the luckiest one. Unaware of post-selection bias, we reported this AUC of 0.9185 on the V mistakenly as the expected generalization performance on unseen test data. However, this AUC is in fact only a fitting error, not a true test error. To find out the downward sample bias, I computed the average validation AUC across the 20 models from the Table 1, which was:

$$\bar{E} = \text{Mean}_k \text{AUC}(g_k, V) = 0.8498$$

Then the downward sample bias in reporting $\text{AUC}(g_{\hat{k}}, V) = 0.9185$ instead of the average AUC 0.8498 is approximated by the sample value:

$$\text{Bias} \approx \text{AUC}(g_{\hat{k}}, V) - \text{Mean}_k \text{AUC}(g_k, V) = 0.9185 - 0.8498 = 0.0687$$

Weng's minimum MSE theory [6, Theorem 3] appears to indicate that if the size of V goes to infinity, the sample deviation about these 20 models will vanish. This is because every network will approach the average performance when the size of V goes to infinity, given the same F. Asymptotically, all 20 models give the same but mediocre AUC which should be around 0.8498. Namely, the number 0.9185 will also approach 0.8498 when the data set V samples the 20 networks more densely. However, the proper quantity to estimate generalization on a fresh test set T is not the luckiest AUC 0.9185. As shown in Wu's experiments [8] with 30 CNN-LSTM models trained on a visual navigation task, the test error $E(g_{\hat{k}}, T)$ of the luckiest model $g_{\hat{k}}$ selected on the validation set V tended to be close to the sample average test error $(1/30)\sum_k E(g_k, T)$ across all 30 models, rather than the minimum test error $\min_k E(g_k, T)$.

Here, 0.0687 is a significant downward sample bias when using the luckiest model's AUC of 0.9185 to estimate the generalization of AUC on unseen tests. By post-selecting the luckiest model $g_{\hat{k}}$ on V instead of using a fresh test set T, we mistook the error of using a fitting error 0.9185 as an estimate of the generalization error. The proper way would have been to evaluate all models $\{g_1, ..., g_{20}\}$ on a held-out test set T not

Table 1: Statistical summary of AUC from twenty different weight initialized VGG-19 models.

Model	Minimum	25th Percentile	Median	75th Percentile	Maximum	Mean
1	0.7982	0.8247	0.8993	0.8998	0.9149	0.8590
2	0.849	0.8630	0.8911	0.9004	0.9161	0.8840
3	0.8612	0.8646	0.8831	0.8853	**0.9185**	0.8856
4	0.7842	0.8409	0.8742	0.8776	0.9151	0.8605
5	0.8509	0.8782	0.8938	0.9048	0.9157	0.8915
6	0.8641	0.8786	0.8923	0.8978	0.9157	0.8966
7	0.7887	0.8787	0.8985	0.9047	0.9149	0.8916
8	0.7856	0.8733	0.8901	0.9027	0.9144	0.8873
9	0.7764	0.8693	0.8918	0.9021	0.9149	0.8863
10	0.7887	0.8744	0.8989	0.9012	0.9144	0.8868
11	0.8500	0.8621	0.8730	0.8782	0.8916	0.8659
12	0.8530	0.8681	0.8897	0.8955	0.9169	0.8846
13	0.8523	0.8662	0.8906	0.8973	0.9132	0.8835
14	0.8401	0.8707	0.8882	0.8918	0.9061	0.8855
15	0.7764	0.8621	0.8780	0.8797	0.8931	0.8728
16	0.7764	0.8462	0.8843	0.8843	0.8942	0.8694
17	0.7661	0.8360	0.8871	0.8881	0.9136	0.8723
18	0.8287	0.8523	0.8906	0.8973	0.9084	0.8748
19	0.7856	0.8401	0.8931	0.8918	0.9061	0.8682
20	0.7764	0.8621	0.8843	0.8851	0.9132	0.8744
Average	-	-	-	-	-	**0.8498**

used for post-selection, and report the average test AUC. This is because Weng AIEE [7] has proven that reporting the luckiest AUC on T is also wrong. Unfortunately, I do not have a test set in our study [1], but Wu [8] used a test set T that is disjoint with the validation set V.

Genetic Algorithms (GA) also suffer from the same misconduct of post-selection. It [2] initiates with a randomly generated initial population of candidate solutions. These candidates are then evaluated on a fitness function, often computed on a validation or training dataset. The "luckiest" or best-performing individuals are then selected. The post-selection step in GA is critical, as it determines which candidate solutions will be used to produce the next generation. This is very similar to the post-selection concerns in deep learning when the "luckiest" model from the validation set is chosen and published without regard for the performance of the other candidate models.

An appealing alternative is to avoid post-selection altogether by using models that learn incrementally from data in an online manner, without revisiting or reusing the same data for selection. The Developmental Network (DN) models proposed by Juyang Weng [4] achieve this by learning each data sample incrementally, dynamically allocating new neurons/parameters only as needed to minimize the cumulative errors over all samples learned so far. It's similar to how a child learns — not by splitting experiences into training/validation/test sets, but by steadily absorbing each new experience and updating their knowledge representations as required. Unlike post-selection methods that first split data into fitting/validation/test sets, DNs learn in a seamless way through environments, avoiding the trap of fitting to a specific validation set. By not re-using the same data for validation and dynamically updating its representations, a DN has no need

for hand-tuning hyperparameters and can abstain from converging to any local minimum on a finite dataset. This allows DNs to abstract general, transferable representations applicable to previously unseen scenarios.

References

[1] B. Gami, M. Agrawal, and R. Katarya. Feature extractor techniques for alzheimer's predictive model in brain imaging. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 173–182, Ahemdabad, India, August 31, 2023. Springer Nature Singapore.

[2] J. H. Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992.

[3] H. Qiu. Dialogue: Validation error with post-selection present is downward biased for test error? *IEEE CDS Newsletters*, 18(1):4–7, 2024.

[4] J. Weng. A developmental network model of conscious learning in biological brains. U.S. Patent Application Number: 17702686, March 22, 2022. Approval pending.

[5] J. Weng. On "deep learning" misconduct. *In Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, December 9-11, 2022. SciTePress. arXiv:2211.16350.

[6] J. Weng. Misconduct in post-selection and deep learning. In Proc. the 8th International Conf. on Control, *Robotics and Sybernetics*, pages 1–9, Changsha, China, Dec. 22-24 2023. NJ: IEEE Press. arXiv 2403.00773.

[7] J. Weng. Is 'deep learning' fraudulent in statistics? *In Proc. The 5th International Conference on Artificial Intelligence in Electronics Engineering (AIEE 2024)*, pages 1–8, Bangkok, Thailand, January 15-17 2024. NY: ACM Press.

[8] X. Wu. Dialogue: The luckiest network on validation performs average during tests. *IEEE CDS Newsletters*, 18(1):8–11, 2024.

2 [Post-Selection] Dialogue: Challenges to Post-Selection

Bardia Ardakanian, Amirkabir University of Technology (Tehran Polytechnique), Tehran, Iran
Email: bardia.ardakanian@gmail.com

In the world of deep learning, we come across Post-Selection. This practice, which has its roots in statistics, is widely used in artificial intelligence. Professor J. Weng's critical view in "On 'Deep Learning' Misconduct" [1] sheds light on Post-Selection and what it means for the field. Weng points out that the usual way of picking the best-performing network based on its error on a validation set, without reporting the full range of models trained, has a big problem. He believes this approach makes us think the network will do better in the future than it actually might when faced with new data. Weng suggests a better way to report on a network's performance, which includes not just the *errors* but also the average, minimum, median, and maximum errors from all models.

My project, "Deep Active Learning Object Detection," aimed to make training the Single Shot MultiBox Detector (SSD) [2] more efficient using deep active learning strategies. We used Variant Autoencoders (VAEs) to pick out the most useful images from a dataset for training, hoping to make the SSD network learn better and faster. We began with a basic set of images, then gradually added more useful images based on how much they could help the network learn. This was a step-by-step process meant to improve the network's ability over time.

In the early stages of our project, we focused on comparing the performance of our model with the established baseline from previous research [3]. This comparison aimed to understand where our method stands in the context of current advancements and to identify areas for further improvement and investigation.

Looking at Professor Weng's comments on Post-Selection, it's clear that our project, despite its new approach to active learning, ended up doing something Weng criticized. Our method of adding more informative images to the training set, decided by how the network did on a validation set, is exactly the kind of Post-Selection strategy Weng warned against. This way of doing things could focus too much on how well the network does now, without enough thought about how it will handle completely new data.

Weng's critique makes us think about two main points. First, it shows we need a wider way to report how all the models we trained are doing, not just the one that's doing the best at a certain point. This means we need to think more deeply about what success means in machine learning projects that use active learning. Second, it highlights the importance of transparency in reporting the performance across different models, underscoring the necessity of having a disjoint test set that is entirely separate from the validation set used in the Post-Selection process. Such a test set is vital for a true assessment of a model's ability to generalize to new, unseen data, illustrating that success on a validation set during Post-Selection does not inherently

predict similar outcomes on genuinely new tests.

Inspired by Professor Weng's critique and his paper on deep learning misconduct [1], and motivated by the findings of X. Wu and J. Weng in their test of Weng's theories [4], I decided to apply these insights within the framework of my project. To achieve this, I used the Pascal VOC dataset [5]. Specifically, I used the Pascal VOC 2007 and 2012 train images for creating two distinct datasets: set F for fitting (training) the model, and set V for validation. For testing, I created set T, which consists of the VOC 2007 test dataset. This tripartite division into sets F, V, and T establishes a structured framework for our experimentation. I chose the Pascal VOC dataset for its application in object detection research, mirroring the dataset used in the baseline study [3] with which I am comparing my model.

In our method, we began with an initial batch of 1,000 labeled images from set F to start training our model. The rest of set F was classified as the unlabeled pool. We used this setup to train the SSD (Single Shot MultiBox Detector) detector and VAE (Variational Autoencoder), as previously introduced. At each stage, the VAE helped identify the top 1000 informative images from the unlabeled pool. These selected images were then added to the labeled pool to enhance the training process. Simultaneously, by removing these images from the unlabeled pool, we ensured a focused improvement of our training dataset. This cyclic process of selecting and transferring the most informative images forms the core of our training approach, allowing the model to adapt and learn from the most relevant data.

To assess Dr. Weng's prediction regarding the VAE's effectiveness in selecting the most informative images for each training phase, we set up 10 different networks. Each network consists of a VAE and the SSD detector [2] for object detection. All SSD detectors in our networks begin with the same set of initial weights, which were randomly selected from a predefined distribution to ensure uniformity at the baseline. This random selection was performed once and applied across all detectors to maintain consistency. Similarly, each VAE is initialized with its own set of initial weights, also randomly drawn from the same distribution but unique for each VAE. The hyper-parameters for both the SSD detectors and the VAEs were hand-selected based on preliminary experiments aimed at optimizing performance.

This setup allows us to focus solely on the effectiveness of the VAEs in picking the best images for improving the model's learning over time. The randomness in weights introduces variations in their behavior, offering insights into how these differences influence the overall training process. By keeping the detector constant, we ensure that any changes in performance across the experiments can be attributed directly to the VAEs' image selection capabilities. For each iteration, we allowed the networks to train for multiple epochs, during which we recorded their mean Average Precision (mAP) on both the validation set V and the test set T.

As Dr. Weng suspected, our results confirm that the initial configuration of the VAEs has a significant impact on the images selected for training, which in turn affects model performance. Figures 1 and 2 show that outcomes vary significantly between the different models. These figures also demonstrate just how sensitive performance is to the initial setup of the weights. Notably, the model that initially performed the best on the validation set—the "luckiest" network—ends up with performance close to the average and sometimes near the lowest when we later test all 10 networks on a new dataset. *This emphasizes the importance of not just relying on validation performance to predict future success but rather considering a broader set of metrics for a more rounded evaluation of a model's capabilities.*

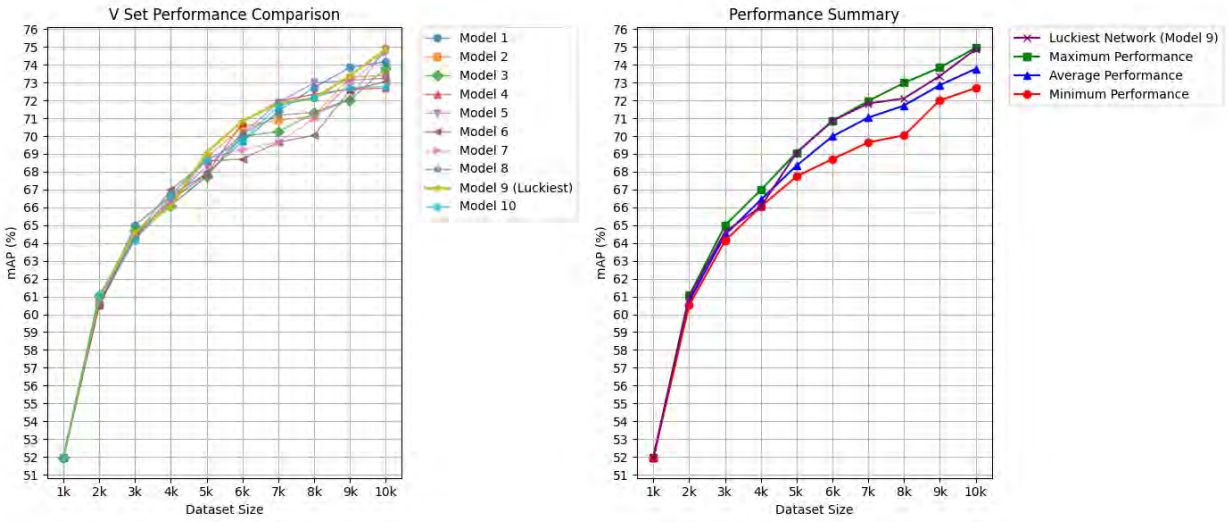

Figure 1: Models' performance fitted on set F and tested on set V. The left plot illustrates individual model trajectories over the course of training, while the right plot summarizes the performance range across all models, showcasing the minimum, median, and maximum scores, along with the outcome of the "luckiest" network.

Building on the results and the highlighted sensitivity of model performance to initial configurations, we further explore theoretical aspects that might explain these observations. To adapt Weng's proof on Pure-Guess Nearest Neighbors (PGNN), referred to as "the devil" by Sorode [6], let's consider the premise that it's possible to have a finite number of networks, denoted as n, such that at least one will achieve zero validation error for any given validation set V. In this adapted proof, we assume that the validation set V consists of p problems, with each problem having l possible solutions. This setup requires us to generate $n = l^p$ networks to ensure exactly one network can perfectly predict the outcomes on V. Unlike Weng's PGNN, or "devil", which utilized a probabilistic generation method that allows the process to stop once a sufficiently good network is found, here, an exhaustive generation method is employed. This "devil" method typically generates more networks because it aims to cover all possible configurations without specifically targeting a zero validation error. By doing so, we ensure a comprehensive exploration of the solution space, but at the cost of potentially higher computational resources compared to Weng's method, which can halt early in the presence of an acceptable solution. This choice is made for simplicity and clarity in understanding, acknowledging that while the amount of time in development is finite, it is practically unbounded, allowing us to construct all possible networks each corresponding to a different validation set V.

For instance, if our validation set includes a collection of images where the task is to determine whether an object is present or not, we have $l = 2$ possible outcomes for each image, making the problem a binary classification task. In this case, $n = 2^p$ networks are necessary to cover all possible scenarios. Expanding this scenario further, if the task involves detecting an object that could be in one of four specific locations, or not present at all, we then have five distinct outcomes for each problem ($l = 5$). Consequently, to cover every potential outcome in this more complex scenario, we would need to generate $n = 5^p$ networks. Therefore, while the theoretical concept of creating n networks to cover all possible scenarios is insightful, it's not

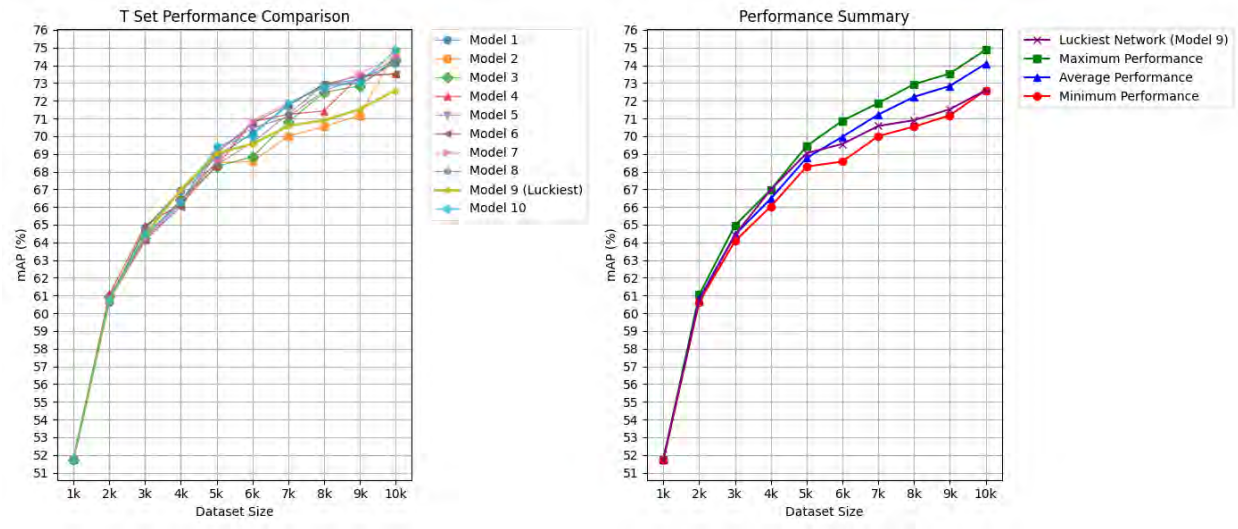

Figure 2: Models' performance fitted on set F and tested on set T. The left plot illustrates individual model trajectories over the course of training, while the right plot summarizes the performance range across all models, showcasing the minimum, median, and maximum scores, along with the outcome of the "luckiest" network.

practical in real-world applications due to the exponential growth in the number of networks required as the problem complexity (p) increases.

When we apply the "devil" logic, or PGNN, to include a test set T, which comprises q problems each with j possible answers, the complexity increases significantly. The total number of networks, $n = l^p j^q$, becomes impractical due to its exponential growth in relation to p and q. This underscores the challenges of using such an exhaustive approach for large datasets and highlights the importance of finding more scalable solutions to ensure that a network can generalize effectively to new, unseen data. The potential gap between perfect performance on a validation set and consistent performance on a test set further illustrates the limitations of relying solely on this method, proving Sorode was right to describe PGNN as "the devil".

In my project, the VAE is designed to select the most informative images from a pool of p images, with l representing the number of these informative images. Following the logic outlined above, if networks were to randomly select images to identify as informative, theoretically, $n = l^p$ networks would be sufficient for one network to correctly identify all informative images. This approach directly aligns with the objective of our project, aiming to efficiently utilize VAEs for enhancing image selection during the training process. However, it is important to acknowledge that this process of selecting informative images does not fully capture the concept of Post-Selection in the context of training multiple networks. Post-selection refers to the broader practice of training various networks and then selecting the one that performs best on a validation set, which could introduce selection bias and overestimate the network's ability to generalize to new data. Therefore, while our focus was on optimizing image selection, the comprehensive impact of Post-Selection involves evaluating the performance of multiple trained networks to ensure a robust model evaluation. Exploring these concepts sheds light on the challenge of moving from validation success to real-world effectiveness, emphasizing the need for models that generalize well.

To expand upon the discussion on Post-Selection and its broader implications, it's essential to recognize its role in a variety of machine learning techniques. This concern extends beyond deep learning into areas like evolutionary computation and reservoir computing. Methods such as genetic algorithms, random forests, echo state machines, and extreme learning machines often depend on selecting configurations that show promise based on specific criteria. This selection, influenced by randomness in initial parameters or manual adjustments while looking at the validation error can lead to biases that disregard the model's generalizability.

Furthermore, techniques based on fuzzy set theories, which navigate reasoning under uncertainty and imprecision, face similar biases. In these systems, decisions are made based on degrees of truth rather than fixed binary choices, reflecting the inherent uncertainty and imprecision in real-world scenarios. For example, a fuzzy logic controller in an automated heating system might decide on the level of heat based on fuzzy temperature categories like "cold," "cool," "warm," and "hot." These categories are not rigidly defined but are instead shaped by overlapping ranges of temperature values, allowing for nuanced control adjustments. However, if this system is overly optimized for a specific set of temperature conditions encountered during its training phase, it might not perform as well in environments with different temperature patterns. This overfitting to the training data is a form of Post-Selection bias, illustrating how the adaptability of fuzzy systems to handle imprecision and uncertainty can inadvertently lead to challenges in generalizing beyond the initial conditions they were designed for.

Recognizing the widespread issue of Post-Selection across these methods, including fuzzy controllers, underscores the need for statistically valid evaluation methods. Fuzzy controllers, similar to other AI methods, might lack comprehensive testing and could hide less favorable data. This highlights the importance of assessments with disjoint test sets, entirely separate from those used in the Post-Selection process, to maintain evaluation integrity and support the development of models with true predictive power and adaptability across the machine learning field.

In conclusion, our "Deep Active Learning Object Detection" project, while showing the value of careful, step-by-step training, also highlights the real problems with Post-Selection. Weng's critical view helps us take a closer look at our methods, pushing us toward valid and open ways of developing models that can be trusted on new, unseen data. Indeed, we did not have a test, and we did hide bad-looking data, acknowledging these limitations is crucial as we navigate the complexities of model development and evaluation. As machine learning keeps advancing, we must carefully face these challenges, making sure valid methods and honest reporting back our creative efforts.

References

[1] J. Weng. On "deep learning" misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, December 9-11, 2022. SciTePress. arXiv:2211.16350.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot

MultiBox Detector. In *European Conference on Computer Vision (ECCV) 2016*, 9905:21–37. Springer, 2016. doi:10.1007/978-3-319-46448-0_2. arXiv:1512.02325 [cs.CV].

[3] J. Choi, I. Elezi, H.-J. Lee, C. Farabet, and J. M. Alvarez. Active Learning for Deep Object Detection via Probabilistic Modeling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10244–10253, Montreal, QC, Canada, October, 2021. IEEE. doi:10.1109/ICCV48922.2021.01010.

[4] X. Wu and J. Weng. The luckiest network gives the average error on disjoint tests: Experiments. In *Proc. The 5th International Conference on Artificial Intelligence in Electronics Engineering (AIEE 2024)*, pages 1–10, Bangkok, Thailand, January 15-17, 2024. NY: ACM Press.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. In *International Journal of Computer Vision*, volume 88, number 2, pages 303–338, June, 2010.

[6] S. Sorode. The "Pure-Guess Nearest Neighbors" (PGNN) is to play a devil's role so that laymen can see the "devil" Deep Learning better. *IEEE CDS Newsletters*, 18(1): 2, 2024.

3 [Post-Selection] Dialogue: Misconduct in Post-Selection

Chunhua Li, Hebei University of Science and Technology, Hebei, China
Email: 1477380105@qq.com

I agree with Professor Weng's views [3] on the Post-Selection misconduct including deep learning, evolutionary computation, fuzzy systems, and other machine learning methods that train multiple systems. Post-Selection is a large category of machine learning methods that break the wall between a model M and its test data T.

Conventionally, one commits a sole model M before applying its test data T.

For example, in so-called deep learning, however, there have been multiple models, instead of one, which will be Post-Selected only after applying a validation set V. The models consist of multiple multilayer neural networks. Each network M is supposed to infer a decision from each query data vector in a test set T. The average error (mean) across all vectors in T is a simple characterization of the model M. The distribution of the errors on T should also be reported, such as the minimum, median, maximum, and standard deviation. However, in Post-Selection, which models are selected for the error report was not determined till all the models had seen fitting data set F and validation data set V, but the test data set T never existed.

Superficially, deep learning greedily fits the parameters of all models on F. It reports only the lowest error among all models on the validation set V disguising it as a "test error" (in fact, it is only a fit error on V). However, it is almost always true that the disjoint test data T does not exist at all, only the validation set V that the authors used to Post-Select the luckiest model. The only exception that shows a test for the luckiest on V is Wu in the last issue [4]. In addition to hiding bad-looking data from all those less lucky models, the Post-Selection method reports inflated performance on the validation set V, without a test on the test set T. The two misconducts (cheating and hiding) in deep learning are obvious.

In genetic algorithms, post-selecting models begin from population generation where each agent in the resulting population corresponds to a model for post-selection. During the population generation, randomized genotype techniques, such as crossover and mutation, are used to randomly generate random models from the earlier generation of the population. The post-selection of a small number of agents in the resulting population is based on a hand-crafted objective function evaluated on the validation set V. Again, the test set T is absent. A few luckiest individual agents are allowed to survive from which the next generation of the population is generated again using the randomized genotype techniques. The finally reported luckiest individual corresponds to the luckiest model on V through several generations of evolution. The major difference from flat deep learning is that the parameters of models in genetic algorithms are dependent from one generation to the next. Again, since the surviving agents were not tested by a test set T that is disjoint

from the validation set V, the so-called "test error" from a genetic algorithm is nonsense.

In fuzzy logic, similar misconduct occurs. Fuzzy logic is a form of many-valued logic in which the truth value (e.g., membership) of variables may be any real number between 0 and 1, instead of either 0 or 1 in probability. This is called the membership function. To build a membership function, multiple hand-picked parameters or randomly generated parameters, including neural networks, are often employed. Each set of parameters corresponds to a model through which multiple models are tried and experimented with. The parameters are determined by multiple-attribute decision-making used in a process of Post-Selection that uses only a validation set V but without a test set T. Therefore, fuzzy logic also suffers from two types of misconduct, (1) cheating in the absence of a test and (2) hiding bad-looking data.

To imagine the effect of the Post-Selection more intuitively, let us imagine an error terrain where the location of a surface point represents the model parameters of a model M and the height of the terrain surface is the average error of M, as shown in Fig. 1 in Weng [2]. There are three types of terrains, fitting terrain (dashed-thin in [2]), validation terrain (black-thin in [2]) and test terrain (green-thick in [2]), respectively. Post-Selection tries multiple random points on the fitting terrain. Each tried set of parameters corresponds to a particular location whose height is the average error on the fitting set F (fitting terrain), the validation set V (validation terrain), and the test set T (test terrain), respectively. The fitting terrain, validation terrain, and test terrain are three different terrains. When one iteratively updates model parameters to minimize the fitting error, he is trying to greedily search for a nearby point that is lower on the fitting terrain. With multiple m models trained, Post-Selection finds m local-minimum locations on the fitting terrain. Whether a local minimum location on the fitting terrain is reported depends on its height on the validation terrain. The more locations one tries on the fitting terrain, the more likely a lower point can be found on the validation terrain. However, the test terrain is absent if the test set does not exist. However, the test terrain is not the same as the validation terrain, as they are very different as demonstrated by Wu in the last issue [4]. The harder the AI problem is, the test terrain is more different from the validation terrain. Therefore, the so-called "test error" from the Post-Selection is nonsense because the test set (and the test terrain) never existed in all publications before Wu [4] as far as the Post-Selection papers I am aware of.

All models from Post-Selection should be very poor in generalization, at least untested. In summary, as pointed out by Professor Weng [3], two types of misconduct exist in Post-Selection The first is hiding bad-looking data (that are truly bad). The second is that the so-called "test error" is falsified by calling the "fitting error" on V as "test error" on T. Both misconducts are unethical.

To deal with the unethical problem, a valid test on all deep learning models should be conducted. Wu's data in the last issue [4] showed that the test error is only around the average of all trained models.

However, the misconduct of Post-Selection is difficult to avoid. Given an available data set as fitting data set F, the candidate parameters of multiple models are searched through error-backprop, genetic algorithms, or fuzzy logic, through many iterations. The luckiest parameters for F are not the luckiest ones for the validation set V, let alone the luckiest for the test data T. And the luckiest error on V is not a test error on T. For example, the luckiest validation error is reported as 5% but the average test error could be 12%.

Unfortunately, Post-Selection, including Deep learning, has been spreading widely all over the world by now and has contaminated many scientific disciplines. As a result, we should alert all scientific disciplines to correct it, instead of denying it. In my opinion, it is urgent for us to seriously consider the dismal future

of AI (but see [1]). We should research new methods that overcome Post-Selection. Weng [1] suggested that overcoming the Post-Selection needs a holistic solution to 20 million-dollar problems.

References

[1] J. Weng. 20 million-dollar problems for any brain models and a holistic solution: Conscious learning. In *Proc. Int'l Joint Conference on Neural Networks*, pages 1–9, Padua, Italy, July 18-23, 2022. NJ: IEEE Press. http://www.cse.msu.edu/~weng/research/20M-IJCNN2022rvsd-cite.pdf

[2] J. Weng. Misconduct in post-selection and deep learning. In *Proc. the 8th International Conference on Control, Robotics and Sybernetics*, pages 1–9, Changsha, China, December 22-24, 2023. NJ: IEEE Press. arXiv 2403.00773.

[3] J. Weng. Dialogue summary: Is "deep learning" misconduct and what should researchers do? *IEEE CDS Newsletters*, 18(1):12–15, 2024.

[4] X. Wu. Dialogue: The luckiest network on validation performs average during tests. *IEEE CDS Newsletters*, 18(1):8–11, 2024.

4 [Post-Selection] Dialogue: My Report for Criminal Investigation of Alphabet's Pyramid Scheme

Juyang Weng, Brain-Mind Institute and GENISAMA, USA
Email: juyang.weng@gmail.com

I, as an individual in the USA, cannot lawfully bring up a *criminal* charge in a federal court against a defendant, which can be either a person or a company. Only the U.S. government can. I can only file a *civil* lawsuit in a court and I did. The defendant of the lawsuit can be an individual who domiciles in the USA or is a citizen of a foreign country. This Dialogue represents the author's report to the U.S. Department of Justice (DOJ) for criminal investigation against Alphabet Inc. The subject that I report here is "a fraud, scam, or bad business practice" as instructed by the Federal Trade Commission at reportfraud.ftc.gov. Alphabet is not the unique company that has done the alleged criminal acts. However, Alphabet is the most egregious in willfulness and the scale of damage. All USA persons who own Alphabet's stocks, or have paid the U.S. social security tax, have been directly injured by the pyramid scheme. All other nations have been indirectly hurt.

The alleged criminal act is a pyramid scheme, called Post-Selection AI (PSAI), as a money-making opportunity and services through which Alphabet willfully lures investments, attracts customers, and inflates its stock prices.

By definition, "a pyramid scheme is a fraudulent multi-level marketing (MLM) arrangement. Generally, the scheme operates under the guise of selling a product, though the profit from the scheme is based on the number of recruits rather than strength of sales" (see Cornel Law School > LII > Wex > pyramid scheme).

In the PSAI pyramid, there are many machine recruits, called models. The models can be any machine trainees—neural networks, genetic algorithms, fuzzy systems, and all other randomly initialized or hand-initialized trainees. The models on the pyramid can be of a single type (e.g., Convolutional Neural Networks, CNN), or of different types (e.g., Linear Least Squares (LLS), Least-Angle Regression (LAR) and Random Forests (RF). See so-called "Super Learner" [4] in the field of statistics. It is important to note that the test set T is missing from biological experiments in the so-called Supper-Learner [4]. I raised this missing to the three authors [4] but none of them denied this missing. I guess that since the Super Learner did not reduce the original errors in their biological experiments, they simply dropped the test T.

In the pyramid, each model recruit is trained on their common fit data set F, and all recruits are ranked according to their error on a common validation data set V. However, the required test data set T is absent for all recruits. Each machine trainee learns to fit the data set F, using error-backprop, genetic algorithm,

fuzzy logic, or a combination thereof. All learning modes are applicable—supervised, reinforcement, unsupervised, and semi-supervised, regardless of the number of "shots" in learning. The schemer only reports the top model recruit on the PSAI pyramid, which has the smallest error on the validation set V.

Why can any learning algorithms be used by the pyramid? They all fit each model to the fit set F. Let us review the types of learning algorithms.

Error-backprop: The error-backprop algorithm is used by so-called "Deep Learning" networks. It iteratively adjusts the parameters of the network along the direction where the error terrain descends the quickest. The network finally gets stuck at a local minimum location when the terrain is locally flat.

Genetic algorithm: The major difference between the error-backprop and a *genetic algorithm* is that the former treats all networks in a flat way, but the latter organizes models as population generations where each early generation model spawns one or more next-generation models. The algorithm kills and hides all models that do not do well. However, our human race must report the statistics about all human individuals and we must not kill genetically deficient individuals and hide the killing!

Fuzzy logic: The major difference between the error-backprop and a fuzzy logic algorithm is that in the former, the *membership function* takes the value of either 1 or 0; e.g., either a member of the human class (1) or not (0), but that in the latter can take any value between 0 and 1; e.g., 0.2 certainty to be a member of the human class. However, the values of such fuzzy member functions are either randomly or hand-initialized. The Fuzzy logic algorithm constructs a PSAI pyramid, regardless neural networks are used or not, as long as the scheme trains multiple models.

For simplicity, consider the number of features (like the number of hyperparameters) in linear least square which has a closed-form solution and does not need any initialization of parameters. Suppose the dimension of each data vector in the fit set F (i.e., the number of features) is p, there are $2^p - 1$ non-empty subsets of features and thus as many as $2^p - 1$ submodels. Each submodel leads to a different *sample* validation error on V. See controversially named "Valid Post-Selection Inference" [1] in statistics. Compare the ranking of submodels on the validation set V with the ranking of them on the test set T, the two rankings could be drastically changed. I emailed four authors of [1], Professors Richard Berk, Lawrence Brown (passed away), Kai Zhang, and Linda Zhao, about my questions and the invalidity of their Post-Selection in [1] but I have not received any responses to my questions.

In the Ponzi-like pyramid scheme, the more recruits the scheme gets, the more profitable the conspirator becomes. Likewise, the more model recruits the PSAI pyramid uses, typically the top model on the pyramid looks more "accurate" so that the conspirator can lure more money. I have proposed a simple PGNN (Pure-Guess Nearest Neighbor) predictor [7] as a machine recruit on the pyramid to play the devil's role. PGNN gets zero error on any fit set F and any validation set V. Therefore, PGNN is always at the top of the pyramid. Because PGNN has never been tested for generalization (and would generalize very badly), nobody would waste money on the PSAI pyramid if he knew that PGNN is on the pyramid.

I raised the first two types of misconduct in the PSAI pyramid scheme [7] with detailed evidence.

1. **Cheating**: Cheating in the absence of a test. The luckiest model on the top of the pyramid does not have a test at all. For example, Alphabet fraudulently gave [5] so-called "test" data (see text "English-to-German and English-to-French ... tests" about Alphabet Transformer, the predecessor of ChatGPT

and Alphabet's Bard). Here, the text "test" is evidence of *cheating*.

2. **Hiding**: Hiding bad-looking data. Alphabet hid all bad-looking data from the models that are not at the top of the PSAI pyramid and did not report as it should. Alphabet [5] did not tell how many models are on the pyramid and it only reported the top one. Few business reports from Alphabet exposed ever how many models are on the pyramid. Evidence: Only a few papers reported loser models. Graves et al. [2, Fig.5] from Alphabet self-reported $n = 20$ models, but [3, Fig.4] self-exposed as many as $n = 10,000$ models! Here $n \geq 2$ is evidence of *hiding*.

3. **PSAI Pyramid**: There are $n \geq 2$ models, but Alphabet reported only one so called "test" error. This is evidence of a pyramid (e.g., a min-heap that is well-known in computer science). Alphabet builds an error pyramid for all models that it trained according to their errors on the validation set in its possession. Model recruits having smaller validation errors are placed higher on the pyramid. Then, the schemer only reported the top model of the pyramid and hid all bad-looking models below the top. It reports the "data fitting error" on the validation set in its possession fraudulently as a so-called "test error". Evidence of PSAI pyramid: Alphabet authors of [2] could plot the $n = 20$ error trajectories (Fig. 5) from the pyramid only if they had the possession of the validation set V. However, in contrast, Alphabet in [5] did not mention the pyramid at all, like many of its later papers. In other words, Alphabet in [2] told the pyramid but Alphabet in [5] hid the pyramid.

Let me use the lottery as an analog. The schemer places all lottery tickets on the pyramid according to the money from each lottery ticket in the last lottery draw. The more money a ticket got, the higher its position on the pyramid (i.e., a max-heap). The schemer only reports the top ticket on the pyramid as its so-called "intelligent technique" to write random numbers on lottery tickets, but the luckiest ticket in the last lottery draw is never tested in a future lottery draw.

How large is Alphabet's inflation of performance? Any real test error (non-zero) divided by the zero (from PGNN) is infinity— the inflation ratio is infinitely large for PGNN! Because Alphabet hid all other models on the pyramid except the top one, the author alone cannot figure out the true rate of performance inflation of Alphabet's AI products. This is something that a federal investigator should inquire into. Regardless of Alphabet's AI products, the PGNN's rate of inflation is infinite! The results in the next paragraph state that Alphabet's inflation ratio equals the average error across all machine recruiters divided by the luckiest validation error Alphabet reported. Understandably, this ratio is typically so large that Alphabet's AI products are impractical. The harder an AI problem is, the worse the inflation ratio.

I have not only blown the whistle [7] about the pyramid but also mathematically proven [2] that, in the future test, all models on the pyramid are expected to give the average of all errors on the pyramid, regardless of their locations on the pyramid. In the last issue, Dr. Xiang Wu [9] experimentally confirmed my proof [2]. Namely, the luck of any model on the pyramid that is determined by its luck on the validation set (i.e., the last lottery draw) is not transferable to a future test (i.e., a future lottery draw). Naturally, every lottery ticket (model on the pyramid) will have the same luck in the next lottery draw (the future test)! In summary, Alphabet's pyramid scheme does not work in reality. It is only a scam.

In the United States District Court for the Western District of Michigan (Civil Action No. 1:22-cv-998)

and the US Court of Appeal 6th Circuit (Civil Action No, 23–1567), Alphabet has not denied the author's above allegations (1) cheating and (2) hiding. Allegation (3) pyramid is a combination of (1) and (2).

I wrote my allegation letters continuously to top administrators of Alphabet, October 4, 2019 (Larry Page and Sergey Brin), December 4, 2019 (Sundar Pichai), May 22, 2021 (Sundar Pichai), September 17, 2021 (Demis Hassabis), April 1, 2022 (Larry Page and Sergey Brin), and May 4, 2022 (John Hennessy) but I have not received any responses from them at all. However, Alphabet continued to willfully mislead its customers using its PSAI pyramid scheme. For example, on Feb. 6, 2023, Alphabet introduced to the public a Chatbot of its own named "Bard"—the Alphabet version of ChatGPT, the predecessor of Alphabet's Transformer. This is evidence of willful Fraud, over 40 months after I directly alerted Alphabet. I allege that "Bard" is again a PSAI pyramid. Furthermore, Alphabet's pyramid scheme is willful.

Ryan Morrison, a reporter for Tech Monitor wrote in "AI Spending to Double to More than $300bn by 2026," and estimated that AI spending will reach $154bn this year. The author estimates that almost all such $154bn spending is flawed by the PSAI scheme. We humans have not seen any corporate fraud that big. Therefore, this PSAI scheme appears to be the largest corporate fraud that human history has ever recorded. Among offenders, Alphabet appears to be the most willful and is leading in scale. Alphabet has already known my holistic solution [6] to the PSAI problem which is the well-known local minima problem and the well-known overfitting problem (nonoptimal in the sense of maximum likelihood).

I pray that the DOJ is willing to take up its mandate and timely start a thorough investigation of Alphabet's PSAI pyramid scheme. Otherwise, the scheme will be hidden from the public eye for an unknown number of years to come, and about $300bn or more will be wasted every year in the future.

References

[1] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[2] A. Graves, G. Wayne, M. Reynolds, D. Hassabis, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538:471–476, 2016.

[3] V. Saggio, B. E. Asenbeck, P. Walther, et al. Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849):229–233, March 11 2021.

[4] M. J. van der Laan, Eric C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007.

[5] A. Vaswani, L. Kaiser N. Shazeer, et al. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, Dec. 4-9 2017.

[6] J. Weng. 20 million-dollar problems for any brain models and a holistic solution: Conscious learning. In *Proc. Int'l Joint Conference on Neural Networks*, pages 1–9, Padua, Italy, July 18-23 2022. NJ: IEEE Press.

[7] J. Weng. On "deep learning" misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.

[8] J. Weng. Is 'deep learning' fraudulent in statistics? In *Proc. The 5th International Conference on Artificial Intelligence in Electronics Engineering (AIEE 2024)*, pages 1–8, Bangkok, Thailand, January 15-17 2024. NY: ACM Press.

[9] X. Wu. Dialogue: The luckiest network on validation performs average during tests. *IEEE CDS Newsletters*, 18(1):8–11, 2024.

5 [Post-Selection] Dialogue Summary: Negative Effects of Post-Selection

Xiang Wu, Nanjing University of Science and Technology, Nanjing, China
Email: wux0213@hotmail.com

After the Dialogue Initiation was published in the last issue of the IEEE CDS Newsletter, we received several responses from different authors. Except for those who have not passed the reviews, we published four responses here. They are from Badal Gami, Bardia Ardakanian, Chunhua Li, and Juyang Weng, respectively, as the reader can see above. Let me first comment on each of the four Dialogues.

The author of the first Dialogue response, Badal Gami, discussed Hongxiang Qiu's mathematical theories [1] about the under-estimation nature of a Post-Selection scheme. He presented a new experimental result of VGG-19s, with a table that shows different sample accuracy from 20 CNN models each of which starts from a "different set pre-trained weights and regularization hyperparameters". I assume that in Gami's table, each row corresponds to a particular learned model, where five columns, minimum, 25%, 50%, 75%, and maximum, respectively, correspond to the variation landmarks among searched regularization hyperparameters. Through this example, Badal Gami's discussion is very clear about how Badal Gami applied the Post-Selection scheme to the table of 20 models—choosing the luckiest CNN model whose maximum accuracy is the largest ($1-0.9185 \approx 8\%$ error rate). The average of the mean accuracy of the 20 models is 0.8498%, or $1-0.8498 \approx 15\%$ error rate. The correct error doubles the mistakenly reported error.

The author of the second Dialogue response, Bardia Ardakanian, has done some deep learning related projects. He is open and honest about the misconduct in deep learning. He conducted object detection experiments using SSD and VAE. Different from the first Dialogue, Ardakanian conducted post-selection on a subset of 1000 training images from a larger training set as a starting set. This is another dimension of Post-Selection. It is important to note here that although a human teacher may design classes, he must not conduct Post-Selection from multiple students because every student in his class must develop normally. This is like choosing easy training samples makes learning easier. As the training set increases from $1,000$ to $10,000$, the 10 modules deviate from one another. In contrast to Professor Weng's theoretical conclusion that the "luckiest" model on V is expected to give an *average* performance of the 10 models on T, here the "luckiest" model on V gave an performance that is almost the *minimum* of the 10 models! This is of course possible because the performance of every model is a random variable.

The author of the third Dialogue response, Chunhua Li, analyzes how post-selection happens in the training procedure of the models. Her analysis provides more intuitive and detailed information for us to understand the post-selection misconduct. She wrote "the so-called 'test error' from the Post-Selection is *nonsense* because the test set (and the test terrain) never existed in all publications before Wu".

The author of the fourth Dialogue response, Juyang Weng, gives a top-level view of Post-Selection AI (PSAI) and comprehensive explanations of misconduct in the PSAI pyramid scheme. He stressed that *willful* cheating in any scientific papers should be treated as a criminal offense.

In the following, I summarize responses in two aspects.

1. Misconduct indeed has widely spread in AI, not only in neural networks, but also in almost all AI methods. The authors agree with Professor Weng's views [2] on the Post-Selection misconduct existing in deep learning, evolutionary computation, fuzzy systems, and other machine learning methods. Badal Gami and Bardia Ardakanian conducted more experiments which further disclose the post-selection problem in deep learning in particular and in Post-Selection in general.

2. The negative effects of the post-selection misconduct. The post-selected models suffer from two main types of misconduct, (1) cheating in the absence of a test and (2) hiding bad-looking data. The post-selected models also lack generalizability, which is a fatal weakness. This is because the generalizability of trained AI models in future tests is the essence of so-called AI.

References

[1] Hongxiang Qiu. Dialogue: Validation error with post-selection present is downward biased for test error? *IEEE CDS Newsletters*, 18(1):4–7, 2024.

[2] J. Weng. Is 'deep learning' fraudulent in statistics? In *Proc. The 5th International Conference on Artificial Intelligence in Electronics Engineering (AIEE 2024)*, pages 1–8, Bangkok, Thailand, January 15-17 2024. NY: ACM Press.

6 [AI Crisis] Dialogue Initiation: Is AI in a Credibility Crisis?

Juyang Weng, Brain-Mind Institute and GENISAMA, USA
Email: juyang.weng@gmail.com

On March 29, 2023, an open letter that calls for a six-month "AI pause" appeared as an online letter. "The signatories expressed a range of fears and apprehensions including about rampant growth of AI large-language models (LLMs) as well as of unchecked AI media hype" [1]. According to the widespread deep learning misconduct in the last three issues of this Newsletters (Vol. 17, No. 2 and Vol. 18, No. 1 and No. 2), that open letter appears now to be a false alarm triggered by the invalid Deep Learning misconduct in particular and Post-Selection in general. That open letter was signed by those who did not know about the misconduct.

In other words, almost all AI methods are cheating in the absence of a test and hide bad-looking data. Worse, my response to the Post-Selection Dialogue in this issue alleged further that Post-Selection is a Ponzi-like Pyramid scheme.

AI has been in different crises before, such as the reproducibility crisis [2] and resource crisis [3]. This time, the Post-Selection misconduct amounts to a new crisis—AI credibility crisis.

This dialogue calls for a discussion about this AI crisis. Each response can address one or several issues below.

1. Fatality: How fatal is Post-Selection misconduct? Is AI almost dead now? Can we revitalize AI?

2. Depth: How deep is the Post-Selection misconduct? If both symbolic AI and computational AI suffer from Post-Selection misconduct, are there any other authors who can provide a way to fill the depth of the Post-Selection difficulties?

3. Breadth: How many AI papers suffer from Post-Selection misconduct? Can you cite a few AI methods that are free from misconduct?

Prof. Kalanmoy Deb, in the area of Genetic Algorithms (GAs) wrote to me: "We did some ML model development using GAs in which we kept a training dataset and a validation dataset to find the optimized model and a disjointed test dataset to demonstrate optimized model's generalizing ability. There are some other GA studies like the above." I replied, "Please send me all such papers so that I can cite them in the next Dialogue. They set a good example." However, he did not send any. Prof. Deb also wrote, "In terms of your comment on hiding bad looking data, I mentioned that choosing the best performing solution from the entire

GA run is not cheating. This is because algorithms use more evaluations to deal with a population-based algorithm and it is not cheating to select the best performing solution." I asked him to host a Dailogue but he did not reply. What do you think about Prof. Deb's statements?

Please send your Dialogue, 1 or 2 pages being sufficient, to juyang.weng@gmail.com with a CC to wangdongshu@zzu.edu.cn by July 31, 2024.

References

[1] M. Anderson. 'AI pause' open letter stokes fear and controversy: IEEE signatories say they worry about ultrasmat, amoral sysems without guidance. *IEEE Spectrum*, April 7 2023.

[2] P. Ball. Is AI leading to a reproducibility crisis in science? *Nature*, 624:22–25, Dec. 7 2023.

[3] H. Lei. The real AI crisis. Toowards Data Science, March 20 2020.

7 A Proverbial Story: Galileo's Free Fall Experiment

Galileo's Free Fall Experiment.

Human development requires constant innovation, and the advancement of science necessitates continual improvement. Galileo's experiment of "two iron balls falling at the same time" conducted on the Leaning Tower of Pisa is a famous story in the history of physics. This story took place in the 16th century when Galileo was a mathematics professor at the University of Pisa.

Galileo questioned Aristotle's view, who had once said, "When two iron balls, one weighing 10 pounds and the other 1 pound, are dropped from a height, the 10-pound ball will surely reach the ground first, and its speed will be 10 times faster than the 1-pound ball." Galileo was skeptical of this view, so he decided to verify it through experimentation.

He designed a simple experiment and invited some scholars and university students from Pisa to watch below the tower. Galileo and his assistant climbed the tower and let a 10-pound iron ball and a 1-pound iron ball fall freely from the top of the tower simultaneously. To everyone's surprise, the two iron balls hit the ground almost at the same time, and the experimental result directly contradicted Aristotle's theory.

Galileo was not satisfied with the result of a single experiment. He repeated the experiment multiple times and obtained the same result each time. This experimental result caused a stir because it directly challenged the widely accepted Aristotle's theory at that time. Galileo's experimental result not only overturned Aristotle's view that the falling speed of an object is proportional to its mass but also laid the foundation for subsequent studies in mechanics and physics.

Galileo's experiment demonstrated his dedication to science and courage in exploring the truth. With his wisdom and courage, he broke the traditional constraints and opened a new path for the development of science. Therefore, we should continue along this path.

8 IEEE TCDS Table of Contents

Volume 16, Issue 1, February 2024

Guest Editorial Special Issue on Cognitive Learning of Multiagent Systems

Y. Tang, W. Lin, C. Yang, N. Gatti and G. G. Yen

Distributed Process Monitoring for Multiagent Systems Through Cognitive Learning

H. Chen, O. Dogru, S. K. Varanasi, X. Yin and B. Huang

Distributed Cognitive Learning Strategy for Cooperative–Competitive Multiagent Systems

Y. -J. Liu, S. Zhang and L. Tang

pFedEff: An Efficient and Personalized Federated Cognitive Learning Framework in Multiagent Systems

H. Shi, J. Zhang, S. Fan, R. Ma and H. Guan

Observer-Based Event-Triggered Iterative Learning Consensus for Locally Lipschitz Nonlinear MASs

H. Li, J. Luo, H. Ma and Q. Zhou

Learning Skills From Demonstrations: A Trend From Motion Primitives to Experience Abstraction

M. Tavassoli, S. Katyara, M. Pozzi, N. Deshpande, D. G. Caldwell and D. Prattichizzo

Data-Based Collaborative Learning for Multiagent Systems Under Distributed Denial-of-Service Attacks

Y. Xu and Z. -G. Wu

Neural Manifold Modulated Continual Reinforcement Learning for Musculoskeletal Robots

J. Chen, Z. Chen, C. Yao and H. Qiao

A Multiagent Meta-Based Task Offloading Strategy for Mobile-Edge Computing

W. Ding, F. Luo, C. Gu, Z. Dai and H. Lu

Prior Knowledge-Augmented Broad Reinforcement Learning Framework for Fault Diagnosis of Heterogeneous Multiagent Systems

L. Guo, Y. Ren, R. Li and B. Jiang

Multiagent Multiobjective Decision Making and Game for Saving Public Resources

X. Ma, Y. Zhang, W. Xie, J. Yang and W. Zhang

Adversarial Decision Making Against Intelligent Targets in Cooperative Multiagent Systems

Y. Li, H. Liu, F. Sun and Z. Chen

An Overview of Brain Fingerprint Identification Based on Various Neuroimaging Technologies

S. Zhang, W. Yang, H. Mou, Z. Pei, F. Li and X. Wu

Human-Collaborative Artificial Intelligence Along With Social Values in Industry 5.0: A Survey of the State-of-the-Art

M. Khosravy, N. Gupta, A. Pasquali, N. Dey, R. G. Crespo and O. Witkowski

A Self-Distillation Embedded Supervised Affinity Attention Model for Few-Shot Segmentation

Q. Zhao, B. Liu, S. Lyu and H. Chen

Dynamic Threshold Distribution Domain Adaptation Network: A Cross-Subject Fatigue Recognition Method Based on EEG Signals

C. Ma, M. Zhang, X. Sun, H. Wang and Z. Gao

Distilling Invariant Representations With Domain Adversarial Learning for Cross-Subject Children Seizure Prediction

Z. Zhang, A. Liu, Y. Gao, X. Cui, R. Qian and X. Chen

Graph-Based Information Separator and Area Convolutional Network for EEG-Based Intention Decoding

X. Tang et al.

Enhancing Overt and Covert Attention Using a Real-Time Neurofeedback Game With Consumer-Grade EEG

T. A. Suhail and A. P. Vinod

Machine Learning Technique Reveals Intrinsic EEG Connectivity Characteristics of Patients With Mild Stroke During Cognitive Task Performing

M. Xu et al.

Relationship Between Decision Making and Resting-State EEG in Adolescents With Different Emotional Stabilities

Y. Si et al.

A Distributed Dynamic Framework to Allocate Collaborative Tasks Based on Capability Matching in Heterogeneous Multirobot Systems

H. -Y. Lee et al.

A Cognitive Robotics Implementation of Global Workspace Theory for Episodic Memory Interaction With Consciousness

W. Huang, A. Chella and A. Cangelosi

Hand Movement Recognition and Salient Tremor Feature Extraction With Wearable Devices in Parkinson's Patients

F. Lin et al.

Iterative Pseudo-Sparse Partial Least Square and Its Higher Order Variant: Application to Inference From High-Dimensional Biosignals

A. Einizade and S. H. Sardouie

Spatial–Temporal Feature Network for Speech-Based Depression Recognition

Z. Han et al.

SalDA: DeepConvNet Greets Attention for Visual Saliency Prediction

Y. Tang, P. Gao and Z. Wang

AdaDet: An Adaptive Object Detection System Based on Early-Exit Neural Networks

L. Yang, Z. Zheng, J. Wang, S. Song, G. Huang and F. Li

ElectrodeNet—A Deep-Learning-Based Sound Coding Strategy for Cochlear Implants

E. H. -H. Huang, R. Chao, Y. Tsao and C. -M. Wu

Parallel Self-Attention and Spatial-Attention Fusion for Human Pose Estimation and Running Movement Recognition

Q. Wu, Y. Zhang, L. Zhang and H. Yu

CSC-Net: Cross-Color Spatial Co-Occurrence Matrix Network for Detecting Synthesized Fake Images

T. Qiao et al.

STDP-Driven Development of Attention-Based People Detection in Spiking Neural Networks

A. Safa, I. Ocket, A. Bourdoux, H. Sahli, F. Catthoor and G. G. E. GielenX. Ma, Y. Zhang, W. Xie, J. Yang and W. Zhang

Generalized Feature Learning for Detection of Novel Objects

J. Liu, X. Liu, Z. Cao, J. Yu and M. Tan

Sequential Learning Network With Residual Blocks: Incorporating Temporal Convolutional Information Into Recurrent Neural Networks

D. Shan, K. Yao and X. Zhang

Volume 16, Issue 2, April 2024

Guest Editorial Special Issue on Movement Sciences in Cognitive Systems

J. Zhong, R. Dong, S. Ikuno, Y. Li and C. Yang

A Bioinspired Multifunctional Tendon-Driven Tactile Sensor and Application in Obstacle Avoidance Using Reinforcement Learning

Z. Lu, Z. Zhao, T. Yue, X. Zhu and N. Wang

Learning to Assist Bimanual Teleoperation Using Interval Type-2 Polynomial Fuzzy Inference

Z. Wang et al.

Passive Model-Predictive Impedance Control for Safe Physical Human–Robot Interaction

R. Cao, L. Cheng and H. Li

Path Learning by Demonstration for Iterative Human–Robot Interaction With Uncertain Time Durations

D. Huang, J. Xia, C. Song, X. Xing and Y. Li

BaSICNet: Lightweight 3-D Hand Pose Estimation Network Based on Biomechanical Structure Information for Dexterous Manipulator Teleoperation

W. Pang, Q. Gao, Y. Zhao, Z. Ju and J. Hu

GCEN: Multiagent Deep Reinforcement Learning With Grouped Cognitive Feature Representation

H. Gao, X. Xu, C. Yan, Y. Lan and K. Yao

The Effect of Expressive Robot Behavior on Users' Mental Effort: A Pupillometry Study

M. van Otterdijk, B. Laeng, D. S. Lindblom and J. Torresen

Modeling Motor Control in Continuous Time Active Inference: A Survey

M. Priorelli et al.

A Unified Search Framework for Data Augmentation and Neural Architecture on Small-Scale Image Data Sets

J. Zhang, L. Zhang, D. Li and L. Wang

Intersection-Over-Union Similarity-Based Nonmaximum Suppression for Human Pose Estimation in Crowded Scenes

L. Wei, H. Huang and X. Yu

Shallow Inception Domain Adaptation Network for EEG-Based Motor Imagery Classification

X. Huang, K. -S. Choi, N. Zhou, Y. Zhang, B. Chen and W. Pedrycz

Social-Psychology-Inspired Reinforcement Learning Framework for Conflict Management in Connected Vehicles

H. Rathore, Y. K. Singhal and G. K. Joseph

A Synapse-Threshold Synergistic Learning Approach for Spiking Neural Networks

H. Sun et al.

Continual Learning for Anthropomorphic Hand Grasping

W. Li, W. Wei and P. Wang

A Hierarchical Utilization of Semantic Gradients and Scene Structure for Visual Place Recognition

Y. Bao, Y. Pan, Z. Yang and R. Huan

Coupling Visual Semantics of Artificial Neural Networks and Human Brain Function via Synchronized Activations

L. Zhao et al.

Graph-Theory-Based EEG Source Connectivity for Assessing Biomechanical Performance in Transfemoral Amputees With Vibrotactile Feedback

S. Kumar, A. Khajuria and D. Joshi

Neuro-Inspired Motion Control of a Soft Myriapod Robot

Q. Ren, W. Zhu, J. Cao and W. Liang

Diagnosis of Early Mild Cognitive Impairment Based on Associated High-Order Functional Connection Network Generated by Multimodal MRI

W. Wang et al.

Does the Brain Infer Invariance Transformations From Graph Symmetries?

H. Linde

A Decentralized Communication Framework Based on Dual-Level Recurrence for Multiagent Reinforcement Learning

X. Li, J. Li, H. Shi and K. -S. Hwang

Graph-Theory-Based Multilevel Cortical Functional Connectivity Developmental Analysis

K. Pan, T. Jiang, R. Zheng, T. Wang, F. Gao and J. Cao

A Contour Detection Method Based on the Projective Coding Model of the Visual Cortex Information Flow

Z. Cai, R. Yang, Y. Fan and W. Wu

UAC: Offline Reinforcement Learning With Uncertain Action Constraint

J. Guan et al.

Supervised Meta-Reinforcement Learning With Trajectory Optimization for Manipulation Tasks

L. Wang, Y. Zhang, D. Zhu, S. Coleman and D. Kerr

TL-P3GAN: An Efficient Temporal-Learning-Based Generative Adversarial Network for Precise P300 Signal Generation for P300 Spellers

V. Bhandari, N. D. Londhe and G. B. Kshirsagar

Using Humanoid Robots to Obtain High-Quality Motor Imagery Electroencephalogram Data for Better Brain–Computer Interaction

S. Cheng, J. Wang, J. Tian, A. Zhu and J. Fan

Two-Stage Grasp Detection Method for Robotics Using Point Clouds and Deep Hierarchical Feature Learning Network

X. Liu, C. Huang, J. Li, W. Wan and C. Yang

A Hybrid End-to-End Spatiotemporal Attention Neural Network With Graph-Smooth Signals for EEG Emotion Recognition

S. Sartipi, M. Torkamani-Azar and M. Cetin

Motion Learning for Musculoskeletal Robots Based on Cortex-Inspired Motor Primitives and Modulation

X. Wang, J. Chen and W. Wu

JLCSR: Joint Learning of Compactness and Separability Representations for Few-Shot Classification

S. Yang, F. Liu, S. Zheng and Y. Tan

Joint Linguistic Steganography With BERT Masked Language Model and Graph Attention Network

C. Ding, Z. Fu, Q. Yu, F. Wang and X. Chen

Electroencephalography Connectivity Assesses Cognitive Disorders of Autistic Children During Game-Based Social Interaction

Y. -L. Tseng et al.

Bioinspired Memristive Neural Network Circuit Design of Cross-Modal Associative Memory

J. Liu, F. Xiong, Y. Zhou, S. Duan and X. Hu