Unreproducible Research is Reproducible

Xavier Bouthillier [1] **César Laurent** [1] **Pascal Vincent** [1 2 3]

Abstract

The apparent contradiction in the title is a wordplay on the different meanings attributed to the word *reproducible* across different scientific fields. What we imply is that unreproducible *findings* can be built upon reproducible *methods*. Without denying the importance of facilitating the reproduction of *methods*, we deem important to reassert that reproduction of *findings* is a fundamental step of the scientific inquiry. We argue that the commendable quest towards easy deterministic reproducibility of methods and numerical results should not have us forget the even more important necessity of ensuring the reproducibility of empirical findings and conclusions by properly accounting for essential sources of variations. We provide experiments to exemplify the brittleness of current common practice in the evaluation of models in the field of deep learning, showing that even if the results could be reproduced, a slightly different experiment would not support the findings. We hope to help clarify the distinction between *exploratory* and *empirical* research in the field of deep learning and believe more energy should be devoted to proper empirical research in our community. This work is an attempt to promote the use of more rigorous and diversified methodologies. It is not an attempt to impose a new methodology and it is not a critique on the nature of exploratory research.

1. Introduction

Reproducibility has been the center of heated debates in many scientific disciplines. Psychology in particular has been the focus of several large reproduction efforts, attempting to reproduce close to a hundred studies (Open Science Collaboration, 2015; Klein et al., 2018). These were motivated by past evidence of lack of scientific rigour, researcher biases, and fraud (Eisner, 2018).

To help counter these problems, important changes were enacted in the psychology research community in the past few years. Making data available is becoming more common, journals are publishing replication reports and preregistration of research specifications is a growing practice.

We see a similar recent trend in machine-learning: the topic of reproducibility rose to prominence at top conferences (Henderson et al., 2018), and several workshops are now focusing on that matter. Top conferences have adopted recommendations for code sharing. More tools are made available to simplify the replication of experiments reported in papers, building on new technologies such as shareable notebooks (Kluyver et al., 2016; Forde et al., 2018), containerization of operation systems, such as Docker (Merkel, 2014) and Singularity (Kurtzer et al., 2017), and open-sourcing of frameworks such as Theano (Theano Development Team, 2016), PyTorch (Paszke et al., 2017) and TensorFlow (Abadi & al., 2015).

While the type of reproduciblity enabled by these tools is a valuable first step, there has been comparatively much fewer discussion about the reproducibility of the *findings* of studies.

Three recent works (Melis et al., 2018; Henderson et al., 2018; Lucic et al., 2018) have shown that proper experimental design is capital to assert the relative performances of models. Beyond mere reproduction, these works shed light on the fundamental problem of reproducibility that cannot be addressed solely by sharing resources such as code, data and containers. The experimental design is at the core of the concept of *reproducibility of findings*.

Melis et al. (2018) conducted large scale experiments in Natural Language Processing with hyper-parameter optimization procedures to compare models in an unbiased benchmark, leading to the surprising result that vanilla LSTM may be as good as recent supposedly state-of-the-art models. Lucic et al. (2018) analyzed GAN models with various experimental setups including average analysis over different initialization of models, concluding that current evaluation methods of GANs can hardly discriminate between model performances. Henderson et al. (2018) exposed the prob-

[1] Mila, Université de Montréal [2] Facebook AI Research [3] Canadian Institute for Advanced Research (CIFAR). Correspondence to: Xavier Bouthillier <xavier.bouthillier@umontreal.ca>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

lem of high instability of results in reinforcement learning. They executed several trials over different seeds and concluded that results in reinforcement learning should include enough trials over different initialization of the model and environment to support a claim with statistical significance.

We extend on these prior works by analyzing a task which played an essential role in the development of deep learning: image classification. Its simple undisputed evaluation metric, in contrast to NLP (Melis et al., 2018) and GAN metrics (Lucic et al., 2018), guarantees that any inconsistency in results cannot be blamed on the brittleness of the evaluation metric, but only on the methodology itself. Additionally, the environment is strongly controlled, in contrast to RL (Henderson et al., 2018), making the inconsistency of results due to small controlled sources of variations even more striking.

We propose to revisit the empirical methodology behind most research papers in machine learning, *model comparisons*, from the perspective of reproducibility of methods and findings. We will first give an example to outline the problem of reproduciblity of methods and findings in section 2. We will then clarify the definition of reproducibility in section 3. In section 4 we will describe the design of the experiments, modeled on current practices in the field, in order to verify how easy false-positive conclusions can be generated[1]. In section 5 we will present and analyse the results and discuss their implications, before highlighting some limitations of the current study in section 6. We will conclude with an open discussion on the differences between exploratory and empirical research in section 7, explaining why all forms of reproducibility deserve the attention of the community.

2. A problem scenario in a typical deep learning experimentation

Suppose we choose several model architectures that we want to compare for the task of image classification. We train all of them on a given dataset and then compare their classification accuracy on the same held-out test set. We then rank the models according to this measured evaluation metric and conclude that the one with highest accuracy is the best one on this dataset. Later on, we retrain the same models on the same dataset but obtain different numerical results, and observe that the new best model is different than in the previous experiment. How come? It turns out we forgot to seed the random number generator used to initialize the models to have reproducible results.

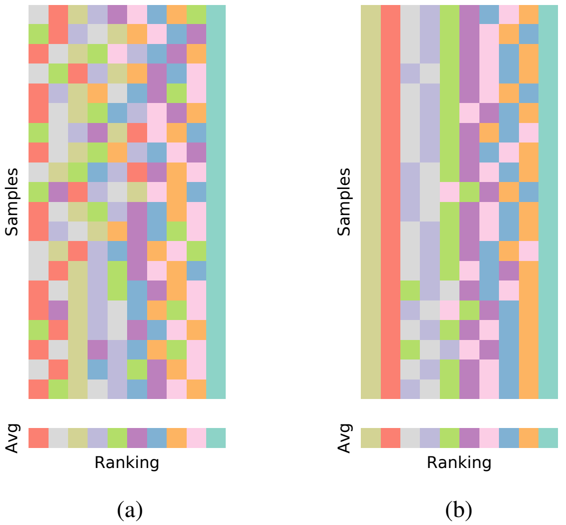


Figure 1: Variation in the ranking of 8 different neural network architectures (models) across multiple trials (samples). 1a: on MNIST digit classification; 1b: on CIFAR100 image classification. The eight different model architecture types are shown in different colors. Each row is a trial with a different random initialization of the models, and in each row the models are ranked from best (leftmost) to worst test accuracy. In 1a we see that ranking can vary greatly from one trial to another, while for a different dataset (1b) rankings of the same set of models can be more stable. We cannot know this however unless we train multiple times the same model. It is thus crucial to do so to ensure the robustness of the conclusions we draw based on a ranking of models.

The usually recommended fix to this reproducibility problem is to set the seed of the random number generator to some arbitrary value, and forget about it. But why are the performances of models sensitive to it? Measurements are affected by sources of variations. The measured accuracy of a model is, for instance, affected by its initialization, the order of the data presented during training and which particular finite data sample is used for training and testing, to name but a few. Trying to fix this problem by seeding a random generator can inadvertently limit the conclusions to this specific seed. Therefore, simply fixing one of these sources of variations has the effect of limiting the generality of a conclusion.

We show in Figure 1a an example of different runs using different seeds, keeping everything else fixed, which lead to different conclusions as to the ranking of eight different types of models on the MNIST dataset. We can clearly see that any conclusion based on a single trial would very likely be invalidated by other runs. It may be different for other datasets, where we could observe a behavior as shown

[1] In agreement with the solutions proposed by the community for methods reproducibility, our code is available publicly, including the data generated in this work and containers to simplify re-execution at github.com/bouthilx/repro-icml-2019

in Figure 1b. However we cannot know this unless we rerun the experiment under different values of the source of variation.

What we would like to point out here, is that there are two forms of reproducibility that can interfere if we are not cautious. The reproduction of the *results* requires the conversion of a stochastic system into a deterministic one, e.g. the seeding process. While this helps reproduction of results, avoiding this source of variation altogether in experiments has the potential effect of dramatically weakening the generality of conclusions. This is at odds with the reproduction of *findings*.

3. Reproducibility: a confused terminology

The distinction between different types of reproducibility is not a new phenomenon (Barba, 2018), however there is no standard terminology to this day.

In this work we will use the terms proposed by Goodman et al. (2016), which avoid the ambiguity of the terms *reproducibility*, *replicability* and *repeatability*. We report here the definitions adapted to the context of computational sciences:

Methods Reproducibility: A *method* is reproducible if reusing the original code leads to the same results.

Results Reproducibility: A *result* is reproducible if a reimplementation of the method generates statistically similar values.

Inferential Reproducibility: A *finding* or a *conclusion* is reproducible if one can draw it from a different experimental setup.

In machine learning, methods reproducibility can be achieved by seeding stochastic processes, but this is insufficient to ensure results reproducibility, where one cannot e.g. rely on having the exact same implementation, execution order, and hardware. To assess results reproducibility some characterization of the probability distribution over what is measured (such as evaluation metrics) is needed. However confidence intervals are seldom provided in the deep learning literature, thus *results reproducibility* can hardly be achieved at the moment, unfortunately. Note that *methods reproducibility* can be obtained as well by producing confidence intervals instead of documenting seeds. The distinction between methods and results reproducibility lies in the presence of a step of reimplementation or reconstruction of the experimental setup.

At the other end of the reproducibility spectrum is *inferential reproducibility*, which is not about the (numerical) results, but rather the conclusions drawn. Suppose a technique performs better than the state-of-the-art for a given task on several vision datasets and fulfills results reproducibility. The authors may conclude that the technique improves the performance on that task. However, if the method later fails on another similar vision dataset, it would invalidate inferential reproducibility. The conclusion, as stated, is not reproducible. This would imply that the assumptions behind the conclusion were wrong or too vaguely stated if at all, and need to be refined: maybe the model performs better on smaller datasets, or on some particular types of images. Such refinements are critical for the advancement of science and can lead to new discoveries.

An observation we want to convey to the reader is that a major part of the current reproducibility litterature in computational science is strongly influenced by the seminal work of Claerbout & Karrenbach (1992), a work that was solely about methods reproducibility, proposing a methodology to ensure automatic regeneration of a report with its accompanying figures. Likewise, the machine learning community seems to be currently mostly referring to methods reproducibility when discussing about *reproducibility*, with the common solution proposed being code sharing.

While code sharing is a valuable practice for the community we argue that it only addresses methods reproducibility and results reproducibility at best. We will present in the next section our methodology to test how current common practice for analyzing model performance in deep learning fails to ensure inferential reproducibility.

4. Methodology to test the robustness of conclusions

The goal of this work is to verify the effect of sources of variations on the robustness of the conclusions drawn in the context of image classification with deep learning models, using common methodology.

To verify this, we will train several popular deep-learning models (i.e. network architectures) multiple times without fixing the initialization or the sampling order of the data and we will measure how much the ranking of the models vary due to these sources of variations.

4.1. Biased vs unbiased scenarios

In order to draw a faithful portrait of the current methodology of practitioners in the field, we would need to use what original authors deemed the best hyper-parameters of each model on each dataset. Unfortunately, the dataset/model matrix we might gather from the literature in this way would be too sparse, leaving us with very few datasets where we could hope to compare all (or even most) models. We will instead consider two methodologies which are respectively worse and arguably better than most common practice. By doing so, we bound the spectrum of experimental bias that includes common practices.

The worse than common practice approach consists in selecting the optimizer hyper-parameters that are the best for one specific model on one specific dataset and apply them unchanged to all other (model,dataset) pairs. This is the most biased methodology, as it should favor the model that was used to select these hyper-parameters. This is arguably a worse practice than what we would (hopefully) observe in the field, but a reasonable lower bound of it as long as all models can be trained sufficiently well. We will refer to this as the *biased scenario*.

The better practice is to optimize the hyper-parameters for each model on each dataset independently using an appropriate validation set, while ensuring that all models had an equal budget of hyper-parameter optimization. We will call this the *unbiased scenario*.

Considered hyper-parameters include the learning rate and momentum as well as weight-decay (L2 regularization strength).

4.2. Experimental setup

For the benchmarking of models, we chose 10 different models of different scales: LeNet (LeCun et al., 1998), MobileNetV2 (Sandler et al., 2018), VGG11, VGG19 (Simonyan & Zisserman, 2014), ResNet18, ResNet101, PreActResNet18, PreActResNet101 (He et al., 2016), DenseNet121 and DenseNet201 (Huang et al., 2017). We limit ourselves to common models in the field for image classification tasks. The evaluation metric of interest is the classification accuracy on the test set.

By *model* we refer to a given architecture (e.g. VGG11) i.e. a specific parameterized function form, together with its standard recommended random parameter initialization strategy. A specific set of (trained) parameter values for a given model corresponds to an *instantiation* of the model. What we are after is a qualitative estimation of which model (together with its standard training procedure) performs better, not which instance. In practice one may care more about which instance performs best, as it is the instance that is used in the end. However, in science, models are the center of interest. An instance is useful as a probe to better understand a model. This is why sources of variations such as the initialization should not be fixed. Conclusions on a model that are limited to a single instance are very weak.

4.2.1. Seed replicates

For each model, we sample 10 different seeds for the pseudo-random generator used for both the initialization of the model parameters and the ordering of the data presented by the data iterator. All models are trained for 120 epochs on the same dataset. Hyper-parameters will be selected differently in the biased and unbiased scenarios, in a way which we will explain shortly.

Following the terminology of Vaux et al. (2012), we call these runs *seed replicates*.

4.2.2. Dataset Replicates

Observations are likely to differ depending on the difficulty of the task, as the potential of different models will be easier to distinguish on more challenging tasks. To ensure some robustness of our conclusions to this source of variation, we will run the seed replicates on different datasets, namely MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), CIFAR10, CIFAR100 (Krizhevsky & Hinton, 2009), EMNIST-balanced (Cohen et al., 2017) and TinyImageNet (et al, 2019). We will call the set of seed replicates of a model on a given dataset a *dataset replicate*. We will not consider here other (less extreme) potential sources of variation in the dataset, but briefly discuss them in section 6.

4.2.3. Biased and unbiased seed replicates

As explained in subsection 4.2.1, the observations about the variations of performances of the models will be different in the biased and unbiased scenario.

For the biased scenario, we will pick the specific hyper-parameters provided by He et al. (2016) in their work on ResNets. This choice should favor ResNet models in our benchmark.

For the unbiased scenario, we will optimize the hyper-parameters for each dataset replicate. To distinguish each scenario, we will call them biased and unbiased seed replicates (4.2.1), biased and unbiased dataset replicates (4.2.2).

Example to summarize the terminology: the *unbiased-scenario dataset replicate* for dataset MNIST and a given model, will be constituted of 10 *seed replicates*, each of which is a trained model instance (that was initialized with on of the 10 seeds) whose hyper-parameters were selected for best performance on the validation subset of that dataset.

The hyperparameter optimization will be executed using a slightly modified version of ASHA (Li et al., 2018)[2]. The exploration is executed until 10 different runs, each with a budget of 120 epochs, have been trained for a given pair of model and dataset. Once this threshold is reached, best hyper-parameters found are used to follow the same procedure as for the seed replicates, i.e. training the model 10 times with different initialization seeds. The hyper-parameter optimization is done based on error rate of the models on the validation set. For the analysis, we will use the test accuracy measures, as we do for the biased seed

[2] With budgets of 15, 30, 60 and 120 epochs and a reduction factor of 4

replicates. This set of 10 runs for each model are the unbiased seed replicates.

5. Experimental results

Results are presented in two different forms. The first goal is to visualize the distribution of performance across seed replicates (5.1). The second goal is to visualize the stability of the model rankings when selecting single seed replicates to measure their performance (5.2). The variances of the model rankings are a way of measuring the likelihood that a conclusion drawn from a single seed replicate, which is common practice in the deep learning community, would hold across many replicates.

5.1. Performance distributions over seed replicates

We generated histograms for the seed replicates for different models on each dataset to compare the distribution of their test error rate. Figures 2a and 2b present these histograms for the biased and unbiased scenarios, respectively. Datasets are ordered based on their difficulty (measured by the average performance of all models). These plots help visualize the overlaps between the distributions of the model performances and give insight on the complexity of the different tasks.

We observe that the overlaps in distribution do not significantly increase between the unbiased and biased scenario. Since they are bounding the spectrum of common practices, we can safely assume that the current observations would also hold in a faithful simulation of common practices.

One can see that concluding which model performs best based on observations from a single initialization seed is brittle: this conclusion will often be falsified if using a different seed. This is especially true for simpler datasets (mnist, svhn, emnist), but one also sees that model ranking varies widely across datasets. Thus, results from single seed experiments on too few datasets, even if they satisfy methods reproducibility, are not sufficient to ensure inferential reproducibility. Hence our irreverent title.

5.2. Ranking stability over seed replicates

We then perform basic bootstrap sampling (Efron, 1992) using the seed replicates. For each dataset, we randomly select a seed replicate for each model and rank them accordingly. We do so 1000 times, and report the results as histograms of rankings aggregated over all datasets. Figures 3a and 3b contain those histograms for the biased and unbiased scenarios, respectively. Such ranking distributions makes it possible to compare model performances across several datasets.

We first note that PreActResNet models do not stand out as the best performing models in the biased scenario, although the hyper-parameters were supposed to favor them. Looking back at Figure 2a, we can observe that they did not outperform other models even on CIFAR10, the dataset on which the best hyper-parameters were selected according to the literature, although they did outperform ResNets, which was the claim of He et al. (2016).

The aggregated results of Figure 3b tend to confirm the superiority of PreActResNets over ResNets. The superiority is however more subtle than what is shown in the original paper, with ResNets sometimes performing better (CIFAR10) or on par (CIFAR100, TinyImageNet). We must note nevertheless that the models used in He et al. (2016) were considerably deeper (110, 164 and 1001 layers) than the one used in this study (18 and 101 layers), making it impossible to compare directly our results to the original ones.

This brings us to another important observation: In our study larger ResNets and PreActResNets did not outperform their smaller counterparts, raising a doubt that larger models would here fare differently. This could be due in part to the fact that we did not perform data augmentation. Nevertheless, the same cannot be said for VGG, for which the larger model is systematically better than its smaller counterpart.

Given the relative homogeneity of the aggregated results, a more subtle measure, one for instance where we weigh performance with respect to computational complexity, would likely raise small models to prominence. We believe that a more nuanced portrait of model performances as the one presented in this study would promote such finer grained analysis.

6. Limitations of this work

6.1. Problem diversity

All experiments are confined to the problem of image classification. It is reasonable however to expect that similar observations can be made for different family of problems provided that best performing models have overlapping distribution of performances. Note that similar observations were made on the more complex tasks in NLP (Melis et al., 2018) and for GANs (Lucic et al., 2018). Our empirical contribution here is to assess the situation on what is arguably the most studied standard task for deep learning, which has a simple undisputed evaluation metric, i.e. image classification.

6.2. Hyper-parameter optimization challenges

Hyper-parameter optimization is not a simple task and although it can help to reduce the bias in the way hyper-

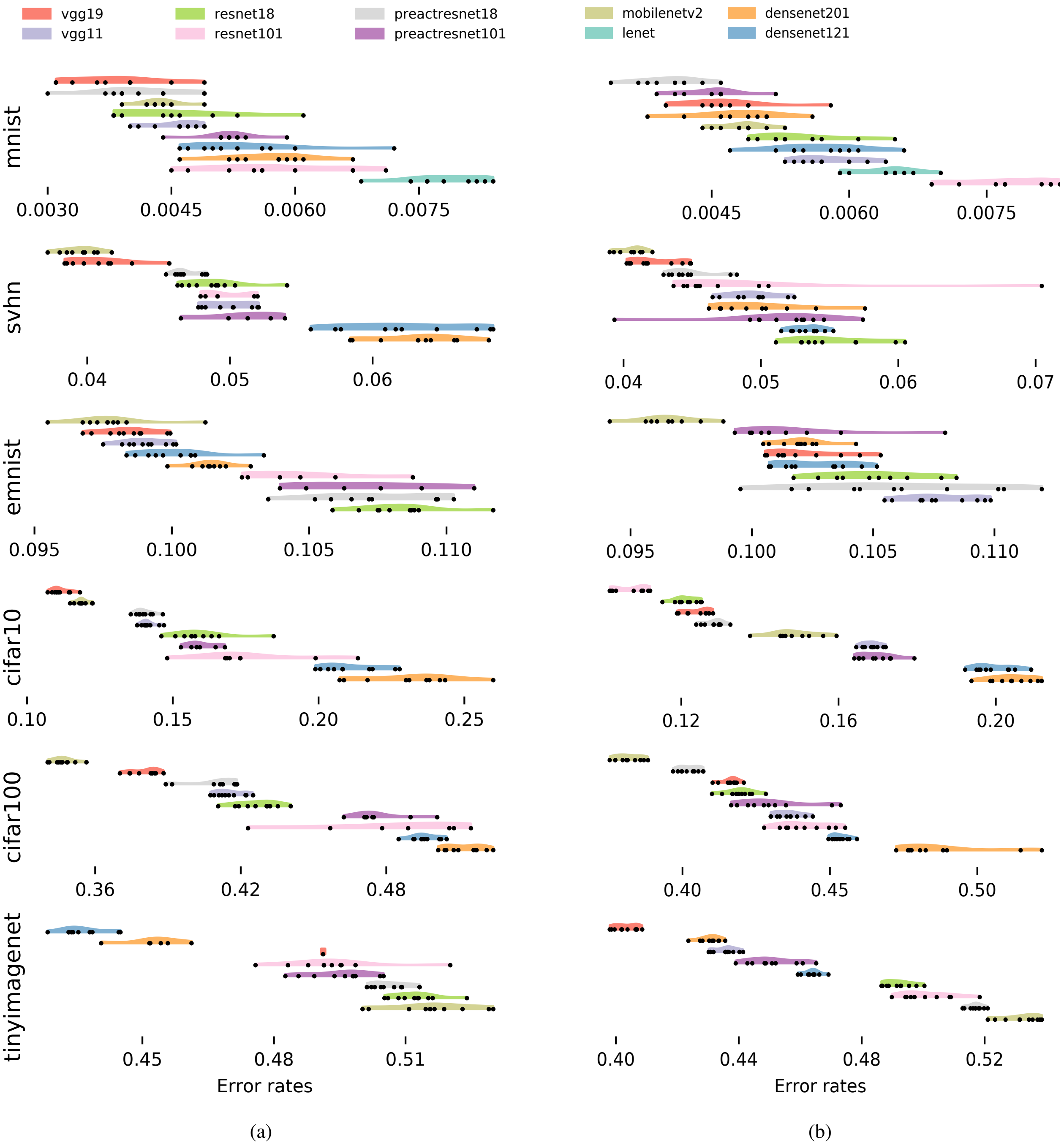
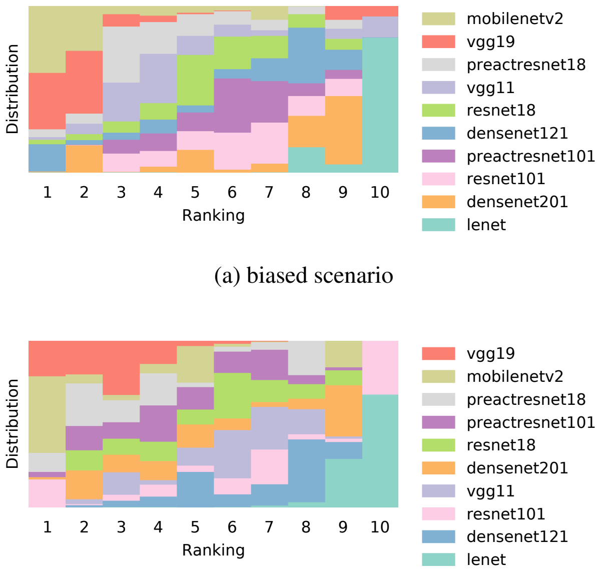


Figure 2: Histograms of performances for each model when changing seeds in the biased (a) and unbiased (b) scenario. Each model is identified by a color. For each dataset, models are ordered based on their average performance. Outliers are omitted for clarity. One can see that concluding which model performs best based on observations from a single initialization seed is brittle: this conclusion will often be falsified if using a different seed. This is especially true for simpler datasets (top three), but one also sees that model ranking varies widely across datasets. Thus results from single seed experiments on too few datasets, even if they satisfy methods reproducibility, are not sufficient to ensure inferential reproducibility. This is true for both biased and unbiased scenarios.



(a) biased scenario

(b) unbiased scenario

Figure 3: Stacked histograms of model rankings estimated through 1000 bootstrap samples of seeds replicates across all datasets in the biased (a) and unbiased (b) scenario. Models are ordered according to their average performance ranking over all datasets. We note three important observations. 1) PreActResNet models do not stand out as the best performing models in the biased scenario (a) although the hyper-parameters were supposed to favor them. 2) The aggregated results of (b) tend to confirm the superiority of PreActResNets over ResNets. 3) Larger ResNets and PreActResNets did not outperform their smaller counterparts, while the larger VGG is systematically better than its smaller counterpart. This can be verified for all datasets in Figure 2.

parameters are chosen it might also introduce another bias for models that are easier to hyper-optimize.

It is also difficult to determine which hyper-parameters should be tuned as there are several factors that influence the training of a model. When training all models with the same optimizer for instance, even though we tune the corresponding hyper-parameters for all models, some of the models may be favored by this choice of optimizer over another. A conclusion would only hold for the optimizer chosen, and may not hold anymore if this source of variance is introduced in the experimental design. Choosing a large set of hyper-parameters to optimize would have the advantage of increasing the robustness of the conclusions one draws. Doing so would however significantly increase the search space and likely hamper the hyper-parameter optimization procedure, making it unpractical. It is worth noting that the current study required the time-equivalent of training over 7000 models for 120 epochs[3].

6.3. Other sources of variations

The current study is limited to the stochasticity of the data ordering on which the model is trained and to the stochasticity of the model initialization. There are two other important sources of variations that we here kept fixed.

The first one is the sampling of the datasets. It is common practice to use given datasets as a fixed set of data. There is however a source of variations in the finite sampling of a dataset from a distribution. Using a technique such as cross-validation could help integrate such variation in our experiments without requiring access to the true distribution of the data. Those would be data sampling replicates.

The second source comes from the optimization procedure of the hyper-parameters. The technique we use, ASHA, is in its very own nature stochastic as it can be seen as a sophisticated random search. To include this source of variation we would need to execute several hyper-parameter optimization procedures and average our analyses over all of them. These would be hyper-parameter optimization replicates.

7. Open Discussion: exploratory v.s. empirical research

Reproducibility is undeniably bound to a definition of the scientific method. Inferential reproducibility is based on concepts such as falsification from Popper (2005), statistically significant demonstration as described by Fisher (1935) or increasing confirmation as stated by Carnap (1936). From this vantage point, methods reproducibility seems but secondary, playing only an accessory role in the scientific inquiry, i.e. in the proper forming of scientific conclusions.

There have been strong debates however in the second part of the 20th century on the nature of the scientific method. Kuhn (2012) and Feyerabend (1993) amongst others have argued that *the scientific method* described by Popper does not exist. We can indeed observe a growing number of research methods to this day, and methods such as *exploratory research* are widely used and accepted despite their weak compliance with a rigorous application of *the scientific method*. As stated by Leek (2017), limiting all scientific work to *the scientific method* would pose a risk of hampering the progress of science.

Let us clarify what we mean by empirical and exploratory research.

[3] 39k+ models if we do not normalize the length of training procedures. ASHA required training 30k models for 15 epochs, 7k+ models for 30 epochs and 1k+ models for 60 epochs. The seed variations required training 1k+ models for 120 epochs.

Empirical research: Its goal is to test an hypothesis. It aims to build a robust corpus of knowledge. It has the advantage of favoring stable progress of a scientific field. As previously outlined, *inferential reproducibility* is strongly linked to empirical research.

Exploratory research: Its goal is to explore a new subject and gather new observations. It aims to expand the research horizon with new corpus of knowledge and favors fast progress of a scientific field. *Methods and results reproducibility* have the advantage of facilitating the diffusion of knowledge and exploration by providing tools to extend existing research and are thus strongly linked to exploratory research.

A too large proportion of the research devoted to exploratory research increases the risk of seeing lines of research collapsing because of building on non-robust basis, while a too large proportion devoted to empirical research increases the risk of hampering the progress by limiting exploration. We do not know what the proper balance is. We can however easily claim that the situation of research in deep learning is currently insufficiently balanced.

The risks of line of research collapses are slowly emerging, as suggested by recent works (Melis et al., 2018; Henderson et al., 2018; Lucic et al., 2018). Sculley et al. (2018) drew attention to the problem, controversially arguing that current methodology in deep learning research is akin to "alchemy". In light of this it is important to understand the tension between exploratory and empirical research, because although both are valuable, they do not play the same role. While Batch-Norm (Ioffe & Szegedy, 2015) was criticized by Sculley et al. (2018), we can actually use it as an example to demonstrate the importance of both research methods. Although Ioffe & Szegedy (2015) include empirical experiments in their work, it could hardly be considered as *empirical research* since the data used to build the evidence would be considered insufficient to substantially support the claims of a superior training approach due to the reduction of internal covariate shift[4]. This however does not invalidate their impactful contribution, and there is now undeniable confirmations that Batch-Norm provides improvements in large models, such as ResNets (He et al., 2016), though likely not due to the reduction of internal covariate shift (Santurkar et al., 2018). The risk with exploratory research is that the *findings* and *conclusions* are brittle and may rest on unstated or unverified assumptions. Consequently using them as a basis for further exploratory research should be exercised with great caution. The example of Batch-Norm is interesting here because, it was revolutionary for the construction of deeper models, and led to a significant number of works (Salimans & Kingma, 2016; Ba et al., 2016; Cooijmans et al., 2016; Arpit et al., 2016) that focused on normalizations, ahead of a good understanding of *why* Batch-Norm works. The internal covariate shift assumption was debunked much later (Santurkar et al., 2018).

Both *exploratory* and proper *empirical research* methods have their role to play in science, and progress in one should support the other. Recognizing their distinct valuable roles, instead of confusing them or arguing one is superior to the other, will certainly lead to a more rational, harmonious, and efficient development of the field, with earlier detected dead ends, and less time and effort wasted globally. Ideally, a promising exploratory work such as Ioffe & Szegedy (2015) should have led more directly to an empirical work such as Santurkar et al. (2018). In short, methods and results reproducibility will mostly help exploratory research, speeding the exploration further with readily available code, while better experimental design will help support robust conclusions as required by inferential reproducibility. This in turn will establish solid empirical ground, on which the community can build further exploration and empirical studies, with increased confidence.

8. Conclusion

We have highlighted the problem of reproducibility of findings due to improper experimental design and presented experiments to showcase how current practice methodologies to benchmark deep convolutional models on image classification tasks are sensitive to this. It is important to take into consideration and investigate sources of variability that *should not* affect the conclusion. As the community embraces rigorous methodologies of empirical research, we believe large scale analysis that include all important sources of variations will provide new insights that could not be discovered through current common methodologies.

Comparing models on different datasets makes it difficult to claim absolute superiority, as the rankings rarely holds across many of them, but it also provides useful information. As outlined by Sculley et al. (2018), the No Free Lunch Theorem (Wolpert et al., 1997) still applies and as such negative performances of a new model should also be reported. These negative results are crucial for the understanding of the underlying principles that make a model better than another on a set of tasks. By identifying in what situations a model fails to deliver on its promises, it becomes possible to identify the shared properties on the corresponding tasks, shedding light on the implicit biases that are shared by the model and the tasks.

[4] According to the position of Vaux et al. (2012) on research in epidemiology, the small amount of data of most deep learning paper would not be enough to classify them as *empirical research*.

Acknowledgments

We thank CIFAR, and Facebook for funding that supported this research, and Compute Canada and Microsoft Azure for computational resources. We also thank Michael Noukhovitch, Mirko Bronzi, Brady Neal, Zafarali Ahmed and Rithesh Kumar for their insightful comments.

References

Abadi, M. and al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/.

Arpit, D., Zhou, Y., Kota, B., and Govindaraju, V. Normalization propagation: A parametric technique for removing internal covariate shift in deep networks. In *International Conference on Machine Learning*, pp. 1168–1176, 2016.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Barba, L. A. Terminologies for reproducible research. *arXiv preprint arXiv:1802.03311*, 2018.

Carnap, R. Testability and meaning. *Philosophy of science*, 3(4):419–471, 1936.

Claerbout, J. F. and Karrenbach, M. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, pp. 601–604. Society of Exploration Geophysicists, 1992.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

Cooijmans, T., Ballas, N., Laurent, C., Gülçehre, Ç., and Courville, A. Recurrent batch normalization. *arXiv preprint arXiv:1603.09025*, 2016.

Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer, 1992.

Eisner, D. Reproducibility of science: Fraud, impact factors and carelessness. *Journal of molecular and cellular cardiology*, 114:364–368, 2018.

et al, L. Cs231n: Convolutional neural networks for visual recognition, 2019. URL http://cs231n.stanford.edu/.

Feyerabend, P. *Against method.* Verso, 1993.

Fisher, R. A. The design of experiments. 1935.

Forde, J., Bussonnier, M., Fortin, F.-A., Granger, B., Head, T., Holdgraf, C., Ivanov, P., Kelley, K., Pacer, M., Panda, Y., et al. Reproducing machine learning research on binder. 2018.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016. ISSN 1946-6234. doi: 10.1126/scitranslmed.aaf5027.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. IEEE, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Schmidt, B. (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87 – 90. IOS Press, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Kuhn, T. S. *The structure of scientific revolutions*. University of Chicago press, 2012.

Kurtzer, G. M., Sochat, V., and Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):1–20, 05 2017. doi: 10.1371/journal.pone.0177459.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Leek, J. A few things that would reduce stress around reproducibility/replicability in science, 2017. URL https://simplystatistics.org/2017/11/21/rr-sress/.

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, K., Hardt, M., Recht, B., and Talwalkar, A. Massively parallel hyperparameter tuning, 2018. URL https://openreview.net/forum?id=S1Y7OOlRZ.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pp. 698–707, 2018.

Melis, G., Dyer, C., and Blunsom, P. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018.

Merkel, D. Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239), March 2014. ISSN 1075-3583.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Popper, K. *The logic of scientific discovery*. Routledge, 2005.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 901–909, 2016.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pp. 2483–2493, 2018.

Sculley, D., Snoek, J., Wiltschko, A., and Rahimi, A. Winner's curse? on pace, progress, and empirical rigor, 2018. URL https://openreview.net/forum?id=rJWF0Fywf.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

Vaux, D. L., Fidler, F., and Cumming, G. Replicates and repeatswhat is the difference and is it significant?: A brief discussion of statistics and experimental design. *EMBO reports*, 13(4):291–296, 2012.

Wolpert, D. H., Macready, W. G., et al. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.