Reproducibility Project

Will Holt, Sebastian Segura, Vaishnavi Paniki

# Introduction

Paper Summary

Our selected paper, “Introduction to CNN Keras – 0.997 Top 6” by Yassine Ghouzam, presents a deep learning approach for image classification models using Convolutional Neural Networks (CNNs) implemented in Keras. The paper focuses on training a CNN model with the Kaggle Digit Recognizer dataset, which is based on the MNIST Handwritten Digit dataset. The author uses a structured approach when preprocessing, designing model architecture, and hyperparameter tuning to achieve a quite high classification accuracy (0.997). The methodology used includes data augmentation, dropout regularization, and the Adam optimizer to enhance generalization and prevent overfitting. Despite its high performance, the paper lacks certain explicit details about various aspects, such specifics regarding computational environment, dependency versions, and various preprocessing steps. These missing details create potential reproducibility challenges, which this project will investigate and report on.

Reproduction Goals

Our project is designed to evaluate reproducibility of the results shown in the paper by taking two differing approaches to replication:

1. Direct Reproduction: Recreating the exact steps outlined in the paper with as little deviation as possible to evaluate uncontrolled discrepancies caused by missing details, software changes since the original publication date, or changes in availability of data.
2. Indirect Reproduction: Modifying controlled variables (i.e. adjusting hyperparameters, preprocessing in different ways, etc.) to show potential impacts on result reproducibility.

Through these two methods, we aim to accomplish multiple goals. First, we will identify discrepancies between the original results and what we were able to reproduce. Then, we determine whether the absence of certain details in the original publication affects overall reproducibility. Next, we analyze the ethics related to incomplete reporting in data science. Finally, our paper will provide recommendations on best practices for helping with more reproducible research in DL and ML studies.

Methodology

This study will assess the reproducibility of Yassine Ghouzam’s CNN model by implementing both direct and indirect replication methods to study the reproducibility of both an attempt to directly copy the approach and report on uncontrolled discrepancies, and to show the potential results based on more controlled discrepancies and the effects of both methods on model performance.

Direct Reproduction Approach

The direct replication aimed to reproduce the CNN model exactly as described by Ghouzam’s paper. However, there were uncontrolled variables in this approach due to outdated codework and versioning issues. This led to various modifications. The primary objectives in this stage of the analysis were to attempt identical preprocessing, architecture, and training conditions while noting all discrepancies required by updates or those dated methods or versions.

Direct Reproduction Environment

The model was implemented using TensorFlow/Keras for deep learning, alongside additional version differences with NumPy, Pandas, and Matplotlib. Modifications due to deprecated functionality or libraries differences were documented.

Direct Reproduction Data Processing Steps

This method followed the original paper as directly as possible. We loaded the dataset and performed a train-test split to ensure the dataset normalization matched the original document’s report. The model was recreated with multiple layers, batch normalization, and dropout following the original document’s outlined architecture to compare activation functions, optimizer settings, and layer configurations. The model training followed a similar suit, using the same batch size (86) to monitor validation accuracy and loss. A key discrepancy identified is the original model trained for only 1 epoch, where the replication defaulted to 10 epochs, which may have altered performance. The original paper also did not specify the default value for the learning rate of the model, so we set ours to 0.001 in hopes of mimicking the results. These minor alterations were necessary due to the nature of uncontrolled discrepancies in reporting. They were likely introduced by software updates, and not intentional modifications.

Indirect Reproduction Approach

This approach introduced controlled variations in the replication process to investigate how modifications in a model’s hyperparameters and architecture specifications impact the performance of the model. These changes were chosen to test the original model’s robustness and observe how sensitive it may be to alternative configurations. This approach simulates replicating a paper with worse documentation to explain what it would be like to replicate a poorly explained experiment.

Controlled Variations Introduced

First, we changed from ReLU to Sigmoid, which affected the gradient propagation and slowed model convergence. The kernel size was also changed from 5x5, 3x3 to 9x9, 5x5, 3x3 to measure the differences in feature extraction. Next, the number of filters was reduced from (32, 64) to (16, 8), which should significantly impact feature representation. The pooling strategy was also changed from 2x2 Stride 2 to 2x2 Stride 3, which results in more information loss thanks to the larger strides. Dropout rate was doubled, which will lead to over regularization and training instability, dense layer neurons were reduced from 256 to 64 neurons to reduce the complexity of the model, and optimizer rates were changed from RMSprop 0.001 to Adam 0.1 to add even more instability and difficulty in convergence. Lastly, the batch size increased from 86 to 512 to affect the gradient updates and over generalization of our model’s performance compared to the original. The idea here was to take the actual missing data from the original paper and extrapolate other variables which largely go unreported in similar model’s reports.

Observed Effects of Controlled Variations on Model Accuracy

The activation function changes cause vanishing gradient issues and a much slower convergence. Increasing the pooling strides and dropout rates allowed for too much information loss, which significantly impacted accuracy. The reduced filter count and pooling layers both limited the model’s ability to extract meaningful features from the images. Increasing the learning rate and batch size both destabilized training and slowed down the updates, reducing adaptability overall. By comparing both results from both replication attempts, we aim to quantify the impact of both uncontrolled (due to software/versioning) and controlled (intentional modification) discrepancies on model reproducibility.

Results

The results from both direct and indirect replications were analyzed by comparing model performance, discrepancies, and the impact of the changes implemented on model accuracy.

Direct Reproduction vs Original (uncontrolled discrepancies)

The direct replication aimed to reproduce the original CNN model as closely as possible Despite necessary modifications from outdated code and software, the replicated model produced results consistent with the original. Some key differences were observed. The replication required 10 epochs to achieve results like the original, which only ran 1 epoch. This likely introduced some level of performance differences. The replicated model also explicitly set the learning rate to 0.001, as the original model left it undefined. The assumption was that the default setting was used, but this could be a potential spot to examine for potential issues in replication. Finally, both implementations used batch sizes of 86 to ensure sameness in training batch sizes. Despite these identified necessary modifications, the model’s performance remained stable. This indicates that the core architecture was robust enough to handle minor software updates and parameter adjustments without much of a change in performance.

A diagram of a number

AI-generated content may be incorrect.A diagram of a confused matrix

AI-generated content may be incorrect.

Fig.1: Left is original confusion matrix, right is exact replication confusion matrix

Indirect Reproduction vs Original (controlled discrepancies)

The indirect method tested the impact of controlled modifications to key hyperparameters to simulate a replication attempt on a poorly documented report. As expected, these commonly omitted details, when changed, had a significant impact on the model’s accuracy. This version of the model has approximately 10% worse accuracy than the original model and the direct replication model. This explains the importance of hyperparameter tuning in neural network designs. As for specific alterations and their effects, switching ReLU to Sigmoid led to slower convergence and vanishing gradient issues. Increasing kernel size and dropping filters also lowered the model’s ability to extract any meaningful features from the images. The pooling strategy and dropout rate alterations led to major information losses and over-regularization, which negatively impacted performance as well. By rising the learning rate to 0.1 and the batch size to 512, both the training and convergence rates became unstable and cause more negative generalization. These results highlight the sensitivity of CNN models to their hyperparameters and architecture and explain why robust documentation is necessary if replication is the final goal.

A graph of loss and accuracy curves

AI-generated content may be incorrect.

Fig. 2: Controlled error model shows distinct accuracy and validation loss due to the changes implemented simulating a more poorly reported paper.

Ethics Discussion

Reproducibility Challenges

The challenges we faced in this project stemmed primarily from outdated code, missing details, and incomplete reporting in the original paper. The direct replication effort was able to identify discrepancies due to software updates and version changes and adequately control for these changes, however the indirect replication effort focused more on the ethical implications of insufficient reporting. Many of the changes that impacted the model’s performance would have been avoided had the original authors provided more thorough documentation on every parameter of the model and computational environment. Any minor omissions of information can mislead future researchers attempting to create a replication, and will likely result in major waste of resources, misinterpreted results, and failure to produce exact replicas to build and advance the study. The ethical implications of such failures in documentation are profound, as reproducibility is a core tenet of scientific integrity and advancement, and failure to provide enough information to exactly reproduce a study undermines the validity of the results obtained.

Resource Considerations

Replication often depends on the availability of adequate computational resources and software environments. This study demonstrates the importance of such discrepancies due to software advancements over time, as the paper was published in 2017 and many of the libraries and functions used are now dated or unavailable. The reporting on most of these details was comprehensive for the time, but there is a distinct near-sightedness in the understanding that software and technology grow and advance over time. The paper does a good job in reporting methods and alternative methods for the period of publication, but regarding version numbers, states, and methods which would potentially change in the coming years, there was very little to no documentation. AI and ML models are very sensitive to even small changes in hyperparameters and configurations, so if they are not reported with future reproductions in mind, there will be guaranteed issues in replication attempts. The ethical responsibility to disclose every single tiny detail of a research paper supports future scientific advancements by preventing time and resource waste and should be considered a core requirement of any publication.

Scientific Integrity Implications

The issues encountered in this project show a large gap in scientific integrity for ML and DL research. Advanced documentation is critical to maintaining transparency and reproducibility, and failure to document adequately compromises the ability of others to verify results or build upon findings. For AI and ML research, the sensitive nature of these models requires complete reporting of hyperparameters, environmental parameters, and other tiny details to be reproduced and studied for advancement. Failure to report in such a way that enabled replication denies the ethical obligation all scientists must create works that help everyone and aid others to create an environment of innovation and improvements.

Impact on Real-World Applications

Real-world applications and advancements will be slowed or otherwise inhibited due to non-replicable reporting. Models employed in fields like healthcare, finance, or criminal justice should always be produced with verification, trust, and replication in mind. When reproducibility of such models is not guaranteed, trust in the results of these systems plummets, and becomes questionable at best. Machine learning models are widely used to make decisions that have real-world impacts on human lives, ranging from medical diagnoses, mortgage lending systems, and even criminal sentencing. When the models used to inform these decisions are not publicly available for testing and scrutiny, it highlights the lack of transparency and creates an environment of distrust towards those who employ such systems. Without these transparency practices in place, any system will inevitably lead to errors or biased outcomes that would more easily be avoided with extremely rigorous reproducibility standards in place.

Recommendations

Best Practices

Authors must provide comprehensive documentation of their computational environments, hyperparameters, version numbers, and data preprocessing steps. This includes explicitly reporting default settings of libraries, as well as the versions at the time of reporting. GitHub is a good system for tracking code changes, dependencies, and computational environments, as it is an accessible repository for others to validate and build upon findings. Docker containers or other virtual environments should be provided to guarantee exact consistency across replication attempts. As even the smallest changes can cause large differences in model performance, clear steps in data handling, augmentation, and preprocessing must be documented to avoid inevitable discrepancies in the final model replications. Finally, a lack of reporting on any aspect of hyperparameters will lead to significant changes in the model’s performance as well and must be diligently accounted for.

Process Improvements

Before formal publication, researchers should post their work to open-access preprint repositories such as GitHub to gain early feedback and make their codework available for review and replication. This would help address potential issues early on. Collaboration should be pushed for on open-source platforms to further enhance replication, as the more eyes on a topic, the more likely it is to have reporting errors spotted and clarified. We would like to further highlight the importance of reproducibility by proposing a standardized reporting template for machine learning experiments to ensure consistency across publications. This reporting template would need to include sections for environmental specifications, code repositories, hyperparameters, etc. to help streamline the reproduction process and limit resource waste. Finally, a call must be made for institutions and publication journals to mandate replication studies as part of the publication process. This would enable researchers to shift their focus from “results for me” to “results for us all,” to highlight discrepancies and ensure accurate results. No model should be employed without the future replication of its results in mind, and by urging institutions and publications to measure model success on more than singular outcomes, the scientific world will grow more transparent and cultivate an environment of growth for all, instead of racing for results.