A Comparative Study on Pre-trained Classifiers in the Context of Image Classification

Mihnea-Adrian Udrea
Dr. Nicola Strisciuglio
University of Twente
the Netherlands

ABSTRACT
Convolutional Neural Networks (ConvNets or CNNs) are nowadays the standard machine learning technique for analyzing visual imagery. As performance in the ILSVRC improves more and more, this paper questions the robustness and generalization abilities of state-of-the-art models. To test the hypotheses that four popular architectures (i.e., GoogLeNet, VGG19BN, ResNet152, and DenseNet161) are not significantly different when classifying images on the ImageNet, ImageNetV2, and ImageNetC benchmarks, the top-1 and top-5 accuracies are calculated and analyzed using the Iman-Davenport, Friedman's Aligned Ranks and Bergmann-Hommel procedures. Our results show that there is enough evidence to reject the null hypotheses. We conclude that the four pre-trained networks do not have identical performance capabilities.

Keywords
Convolutional Neural Network, Statistical Analysis, Iman-Davenport, Friedman's Aligned Ranks, Bergmann-Hommel, Accuracy, Image classification, ImageNet, ImageNetV2, ImageNetC, PyTorch, GoogLeNet, VGG, ResNet, DenseNet

1. INTRODUCTION
The increased interest in deep learning has led to a series of advances in speech recognition, decision making, and image classification. Though introduced more than 30 years ago, Convolutional Neural Networks (ConvNets or CNNs) are nowadays the standard machine learning technique for analyzing visual imagery. Recent developments such as large public image databases, graphics processing units, and open-source libraries have benefited the state-of-the-art models [7, 9, 17, 18]. Deriving from a typical structure – stacks of convolutions, followed by max-pooling and fully-connected layers, current implementations surpass the 100-layer barrier, have millions of parameters, and train on millions of images. The performance of these algorithms has been growing at a substantial pace due to competitions like the ILSVRC where researchers aim for high top-1 and top-5 accuracies. As these numbers reach the ceiling, the robustness and generalization abilities of such architectures are questioned. Previous studies show a preference for deeper and wider architectures, as well as significant problems when it comes to real-world applications where uncertainty and noise are present [8, 15]. Moreover, the sensitivity of such algorithms suffers from the operation of datasets like ImageNet where the contents of the images can be inappropriately misrepresented by the original sets of labels, leading to longer training times and unjust decreases in accuracy scores [1]. Using a statistical framework, we aim to address the gap in comparative studies with more classifiers and datasets [3] and gain new insights that may confirm or reject the existing assumptions.

34th Twente Student Conference on IT, Jan. 21st, 2021, Enschede, The Netherlands.

2. RESEARCH QUESTION
The previous section leads to the following research questions:

Considering the top-1 and top-5 accuracy scores, what pre-trained convolutional neural network models (i.e., GoogLeNet, VGG19BN, ResNet152, and DenseNet161) are significantly different according to the Iman-Davenport, Friedman's Aligned Ranks, and Bergmann-Hommel procedures when performing image classification on:

(a) the original benchmark ImageNet?
(b) a generalization benchmark like ImageNetV2?
(c) a corruption robustness benchmark like ImageNetC?

The answers to these questions may propose a new standard for evaluating the performance of image classifiers.

3. LITERATURE REVIEW
Though the focus has recently shifted from new architectural ideas to training and optimization, four popular state-of-the-art models (i.e. GoogLeNet, VGG, ResNet, and DenseNet) are considered for this comparative study.

3.1 GoogLeNet
The main idea of the Inception architecture is based on sparse connections, even inside the convolutions. Drawing inspiration from deep architectures and pop culture, Szegedy et al. [18] describe a network of larger width and depth, built from blocks optimally constructed and repeated. Information is processed at various scales, combined, and further abstracted. To keep the computational requirements to a low, 1x1 convolutions are used for dimensional reductions before the more expensive 3x3 and 5x5 convolutions; such modules are only present at higher layers. The method takes into consideration the finite computational budget and overfitting problems, controlling the number of input filters between layers. Winner of the ILSVRC 2014 Classification Challenge, the name of their final submission is an homage to the original LeNet 5 network [12]. With 27 layers (pooling included), GoogLeNet obtains a top-5 accuracy of 0.89 on the single model performance. Results indicate the importance of sparser architectures and quality gains in computer vision when compared to shallower and less wide networks.

3.2 VGG
Inspired by the work Krizhevsky et al. [11], Simonyan and Zisserman [17] explore very deep convolutional networks. To confirm the importance of depth in visual representations, they improve the classical architecture of LeCun et al [12] by significantly increasing the number of weight layers. Five max-pooling layers alternate with stacks of convolutions and are followed by three fully connected layers and softmax averaging.

Training and evaluation are done on multiple GPUs following data parallelism principles; the batches of images are split and shared among GPUs for processing. Normalization is not employed since it increases memory consumption and computational time while bringing no benefits; scale jittering and spatial context capturing are encouraged. Moreover, testing shows that classification errors are dependent on the size of the datasets and drop with the increase in depth. Entering the ILSVRC competition in 2014, their ConvNets outperform previous winning submissions, the single-net model with 19-weight layers achieving a top-5 accuracy of 0.92.

3.3 ResNet

He et al. [7] question the significance of stacking more layers when learning better networks. Influenced by the ideology of VGG networks, the authors address concerns like optimization and degradation which arise with the increase in depth. Their implementation is a deep residual network (ResNet) with shortcut connections that skip a certain number of layers. After performing identity mappings, the outputs of these connections and the outputs of the stacked layers are added. According to their conclusions, residual networks are easy to optimize and benefit greatly from increased depths. Achieving first place in the ILSVRC 2015 classification competition, their 152-layer single-model (the deepest on ImageNet at the time) has significantly lower complexity than VGG networks. Further, ResNet152 outperforms previous ensembles with a top-5 accuracy of 0.94. To avoid overfitting, more drastic regularization is recommended when training much deeper networks on small datasets.

3.4 DenseNet

After discovering a pattern among implementations that deal with the vanishing gradient problem, Huang et al. [9] propose an architecture with paths connecting early layers to later layers. In contrast with ResNets, the inputs are concatenated using a composite function, leading to a simpler and more efficient solution. Divided into multiple dense blocks connected by transitional layers performing batch normalization, convolution, and pooling, DenseNets are more compact and motivate feature reuse; bottleneck and compression layers are present to reduce the number of feature maps and improve computational efficiency. Additionally, with a relatively low number of filters per layer, DenseNets are easy to train and scale to hundreds of layers while raising no optimization concerns. The outcome of their research shows that DenseNets achieve similar performance to ResNets while requiring significantly fewer parameters. DenseNet201 achieves a top-5 accuracy of 0.93 on the ImageNet dataset and is believed to obtain further gains through hyperparameter tuning; the overfitting problem is addressed by the regularizing effect of connections.

3.5 Problems with ImageNet

The ImageNet benchmark has represented a turning point in the evaluation of machine learning classifiers. When investigating the generalization capabilities of the dataset, Beyer et al. [1] report several concerns such as overfitting that might lead to what is perceived as "progress". One of these concerns is the significant number of images with more than a single object of interest. The contents may be inappropriately misrepresented as the 1,000-way image classification task has a limit of a single label per image. Thus, using accuracy as a metric may penalize architectures producing correct predictions that do not match the established ground truth labels. According to the authors, their Reassessed Labels ("ReaL") provide a better approximation of the accuracies as they allow multiple annotations and remove duplicate pairs. For instance, complicated distinctions such as "laptop" are added the labels "notebook" and "computer keyboard." It is concluded that label noise is to blame for the longer training times and that the end of the original label set is near. Even so, the current research focuses on the authentic ImageNet benchmark test, taking into consideration the impact of its limitations.

4. METHOD

The proposed method comprises three steps. First, the data of the three benchmarks (i.e., ImageNet, ImageNetV2, and ImageNetC) are downloaded and split into several subsets. Second, we make use of a Google Collaboratory notebook to evaluate the performance of our models; the top-1 and top-5 accuracies are calculated for each of the subsets. Last, the statistical analysis is carried out in RStudio. To visualize the results, graphs and tables are provided.

4.1 Datasets

4.1.1 ImageNet

Figure 1. Image Samples from the "Pizza" (Left) and "Chihuahua" (Right) ImageNet Categories

Observing the increasing number of image data available on the internet, Deng et al. [4] anticipate the need for a large database dedicated to developing, training, and benchmarking image understanding algorithms. Using the backbone of the WordNet architecture, popular concepts are represented through images. Categories include typical nouns like "toilet seat", "spatula", "pizza", and "chihuahua", each of them being, on average, linked to 500 images. Though ImageNet was created with the help of the Amazon Mechanical Turk (MTurk), the authors promise to deliver diversity and accuracy as quality control deals with the errors in the human judgment process. The images are taken from different angles and contain various backgrounds and obstructions. With 1,000 categories, this dataset offers 1.2 million images dedicated to training and 150,000 more - to validation and testing. The present research makes use of the ImageNet benchmark test to confirm the rankings in performance and scores reported in the literature review.

4.1.2 ImageNetV2

Ideally, machine learning models generalize to new data. To evaluate the state-of-the-art, Recht et al. [15] follow the principles behind ImageNet and elaborate an extension to the original benchmark. Sampled a decade after the original benchmark test, ImageNetV2 contains three datasets, each with 10,000 new images. A thorough analysis suggests that the more complicated images lead to accuracy drops of 11-14%. Regardless of the outstanding contributions to the field, more attention is recommended to the creation and operation of datasets like ImageNet as real-life applications suffer from the lack of generalization. We aim to check whether the four models have the same ability to adjust to previously unseen content.

4.1.3 ImageNetC

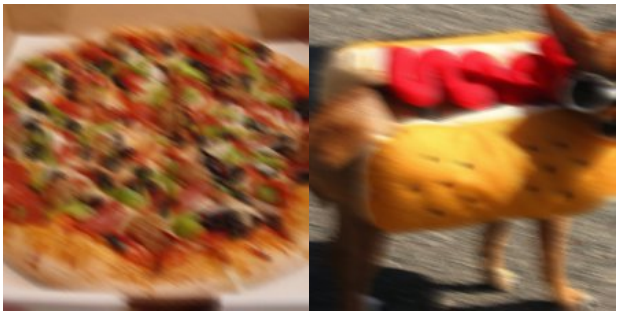

Figure 2. Image Samples from the "Pizza" (Left) and "Chihuahua" (Right) ImageNetC Categories

Hendrycks and Dietterich [8] question the robustness of machine learning algorithms and implement a benchmark that consists of fifteen types of corruption. Applied to the ImageNet validation set, ImageNetC covers common corruptions like Gaussian noise, Poisson noise, motion blur, and pixilation. Moreover, each image is applied a level of severity; ranging between 1 and 5, the higher the level of severity, the worse the image noise gets. Results show that deeper and wider models perform notably better on corrupted inputs than smaller models. The importance of such benchmarking techniques is highlighted as the accuracy on ImageNet reaches its limit. Its purpose for the current experimental evaluation is to observe how the algorithms react in worst-case scenarios and identify the ones with different performances.

4.2 PyTorch Models

Table 1. ImageNet 1-Crop Error Rates (224x224)

Network	Top-1 error	Top-5 error
VGG19BN	0.2576	0.0815
ResNet152	0.2169	0.0594
DenseNet161	0.2235	0.0620
GoogLeNet	0.3022	0.1047

Competing in the ILSVRC implies classifying images into one of the 1,000 classes of the ImageNet database. Approximately 150,000 images collected from search engines are used for validation and testing, each algorithm producing a list of labels sorted by decreasing confidence. PyTorch [13] is a popular scientific computer library in the deep learning community with easy debugging and support for hardware accelerators. It offers common image transformations and model architectures addressing image classification. Table 1 provides the top-1 and top-5 error rates for the chosen pre-trained models [14].

4.2 Metrics

The top-1 and top-5 accuracies are the two numbers usually reported [7, 9, 17, 18]. Top-1 accuracy implies that the label with the highest confidence matches the ground truth, whereas top-5 accuracy implies that one of the first five labels with the highest confidence matches the ground truth. In both cases, the scores are calculated as the number of matches, divided by the total number of data points evaluated. These metrics are numbers between 0 and 1, where 1 indicates perfect accuracy and 0 – the contrary.

4.3 Statistics and Tests

Since the numbers used to describe the performance of such algorithms may vary by less than 0.01, the present paper questions the significance of these differences and how they can be translated using a statistical framework. It is believed that significance tests are often misused, leading to false conclusions. When it comes to real-world classifiers and datasets, Demšar [3] recommends the use of non-parametric methods to evaluate differences. Their empirical results indicate the strength and safety of these tests in problems regarding classification accuracies.

The Friedman test is a non-parametric method ranking models for our datasets. It does not assume normal distributions nor homogeneity of variance. The algorithm performing best is assigned "rank 1" and, in case of a tie, an average rank is calculated. The null hypothesis states that all models are equivalent and thus have the same rank. Due to its undesirably conservative nature, the Iman-Davenport modification of Friedman's test derives a better statistic [3, 10] which is used for the present research. Depending on the significance level chosen, the p-value obtained indicates whether the null hypothesis can be safely rejected or not. A test result below the alpha threshold is statistically significant, meaning our test hypothesis is false. The opposite holds for a test result above the alpha threshold.

Once it is known that not all the architectures perform the same, we can proceed with a post-hoc test. As this is a review of existing methods, checking the pairwise differences then correcting the p-values in a multiple comparison analysis is preferred [2]. Due to abnormality, a nonparametric test such as Friedman's Aligned Ranks is applied to our pairwise comparisons. Garcia and Herrera [5] describe the Bergmann-Hommel procedure as an advanced and powerful tool for controlling the familywise error rate. Though extremely expensive from a computational point of view, the use of the Bergmann-Hommel correction procedure is recommended for comparisons with up to nine algorithms. The p-values are yet again compared to the significance level, the ones below it indicating evidence of significant differences between the respective pairs of models. For a more thorough introduction to these procedures, we refer to [2, 3, 5, 6, 10].

5. EXPERIMENTS

We begin our experiments by uploading all the images to a Google Drive account. To facilitate the classification process, we make use of Google Collaboratory's Cuda GPUs. The transformation of our data follows the one from the PyTorch website: the images are resized to 256x256 pixels, centrally cropped to 224x224 pixels, converted to a Tensor data type, and normalized using the mean and standard deviation. Further, the data is loaded with a batch number of one and four workers. For each image sample, the labels with the highest confidence are stored for the statistical analysis. The RStudio package implemented by Calvo and Santafé [2] includes the nonparametric tests, post-hoc tests, and correction methods mentioned in Section 4.3, following the explanations from Demšar [3] and García et al. [6]. In the following subsections, the distributions and differences among models are discussed. We consider a significance level alpha equal to 0.05, which is the probability of rejecting the null hypothesis when it is true.

5.2 Statistical Distribution of Results

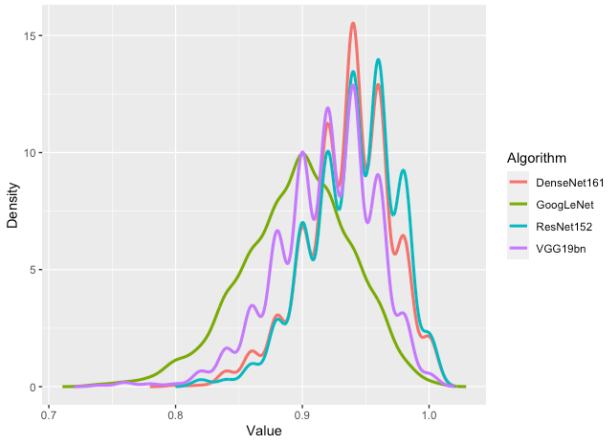

Figure 3. Density Plot of Top-5 Accuracies on ImageNet

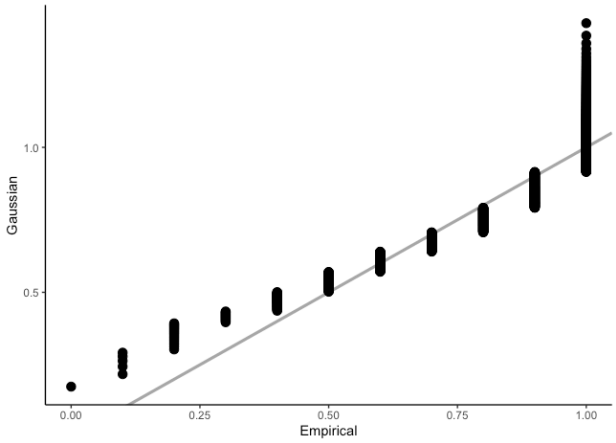

Figure 4. Q-Q Plot of ResNet152's Top-1 Accuracies on ImageNetV2

We regard a distribution as the occurrence of the different accuracy scores obtained on a benchmark test. To check whether the distributions cannot indeed be considered normal, we may want to visualize the data. Therefore, two types of plots are created. A density plot illustrates that most of the samples cannot be regarded as normal due to the lack of symmetry and unimodality. As evidenced in the quantile-quantile plot, sample points do not perfectly lie on the diagonal. The assumptions of normality and homogeneity of variance are violated (see Figure 3 and 4), confirming the recommendation of Demšar [3].

5.3 ImageNet

The TorchVision classification models are trained on the ILSVRC2012 dataset [16]. For the present experiment, we make use of their validation data which contain 50,000 random images with labels. After splitting them into subsets, we end up with a sample size of 1,000 for both of the following analyses.

5.3.1 ImageNet Top-1 accuracies

Table 2. Minimum, Mean, and Maximum of ImageNet Top-1 Accuracies

Network	Min.	Mean	Max.
VGG19BN	0.5200	0.7422	0.9200
ResNet152	0.6000	0.7825	0.9400
DenseNet161	0.5800	0.7715	0.9600
GoogLeNet	0.4400	0.6974	0.8800

Table 3. First Quartile, Median, and Third Quartile of ImageNet Top-1 Accuracies

Network	1st Qu.	Median	3rd Qu.
VGG19BN	0.7000	0.7400	0.7800
ResNet152	0.7400	0.7800	0.8200
DenseNet161	0.7400	0.7800	0.8200
GoogLeNet	0.6600	0.7000	0.7400

Table 4. Corrected P-values for Pairwise Network Comparisons of ImageNet Top-1 Accuracies

Comparison	Corrected P-value
GoogLeNet vs. ResNet152	0e0
GoogLeNet vs. DenseNet161	0e0
GoogLeNet vs. VGG19BN	0e0
VGG19BN vs. ResNet152	0e0
VGG19BN vs. DenseNet161	0e0
DenseNet161 vs. ResNet152	5.838971e-9

As can be seen in Table 2, ResNet152 has the highest average top-1 accuracy on ImageNet, followed closely by DenseNet161, VGG19BN, and GoogLeNet. ResNet152 and DenseNet161 have the same interquartile range as their first quartile, median, and third quartile coincide (see Table 3). Further, the Iman Davenport's correction of Friedman's rank sum test yields a p-value < 2.2e-16 and indicates that the null hypothesis can be safely rejected: not all the networks perform the same when considering their top-1 accuracies on the ImageNet benchmark. We proceed with the post-hoc test to identify the exact pairs of models showing differences. The Friedman's Aligned Ranks test is applied to the six pairwise comparisons mentioned in Table 4. All corrected p-values using the Bergmann and Hommel's method are below the established threshold, indicating significant differences between all pairs of architectures.

5.3.2 ImageNet Top-5 accuracies

Table 5. Minimum, Mean, and Maximum of ImageNet Top-5 Accuracies

Network	Min.	Mean	Max.
VGG19BN	0.7400	0.9185	1.0000
ResNet152	0.8200	0.9398	1.0000
DenseNet161	0.8000	0.9360	1.0000
GoogLeNet	0.7400	0.8954	1.0000

Table 6. First Quartile, Median, and Third Quartile of ImageNet Top-5 Accuracies

Network	1st Qu.	Median	3rd Qu.
VGG19BN	0.9000	0.9200	0.9400
ResNet152	0.9200	0.9400	0.9600
DenseNet161	0.9200	0.9400	0.9600
GoogLeNet	0.8600	0.9000	0.9200

Table 7. Corrected P-values for Pairwise Network Comparisons of ImageNet Top-5 Accuracies

Comparison	Corrected P-value
GoogLeNet vs. ResNet152	0e0
GoogLeNet vs. DenseNet161	0e0
GoogLeNet vs. VGG19BN	0e0
VGG19BN vs. ResNet152	0e0
VGG19BN vs. DenseNet161	0e0
DenseNet161 vs. ResNet152	5.013554e-4

The results shown in Table 5 indicate that ResNet152 obtains the highest average top-5 accuracy on ImageNet, followed by DenseNet161, VGG19BN, and GoogLeNet. The maximum top-5 accuracies per subset are 1.00, the highest possible. ResNet152 and DenseNet161 share the same first quartile, median, and third quartile, as presented in Table 6. With a p-value < 2.2e-16 obtained from the Iman and Davenport test, the null hypothesis can be safely rejected: one or more models do not perform the same when considering their top-5 accuracies on the ImageNet benchmark. Continuing with the post-hoc procedures, significant differences among our models are discovered as the corrected p-values do not exceed the significance level of 0.05 (see Table 7).

5.4 ImageNetV2

TopImages is one of the three test sets included with ImageNetV2. Containing exactly 10,000 images, this benchmark links each class to the ten images with the highest frequency in their candidate pool. The top-1 and top-5 accuracies are calculated for each class, resulting in a sample size of 1,000.

5.4.1 ImageNetV2 Top-1 accuracies

Table 8. Minimum, Mean, and Maximum of ImageNetV2 Top-1 Accuracies

Network	Min.	Mean	Max.
VGG19BN	0.0000	0.7597	1.0000
ResNet152	0.0000	0.8015	1.0000
DenseNet161	0.0000	0.7951	1.0000
GoogLeNet	0.0000	0.7300	1.0000

Table 9. First Quartile, Median, and Third Quartile of ImageNetV2 Top-1 Accuracies

Network	1st Qu.	Median	3rd Qu.
VGG19BN	0.6000	0.8000	0.9000
ResNet152	0.7000	0.9000	1.0000
DenseNet161	0.7000	0.9000	1.0000
GoogLeNet	0.6000	0.8000	0.9000

Table 10. Corrected P-values for Pairwise Network Comparisons of ImageNetV2 Top-1 Accuracies

Comparison	Corrected P-value
GoogLeNet vs. ResNet152	0e0
GoogLeNet vs. DenseNet161	0e0
GoogLeNet vs. VGG19BN	1.64313e-14
VGG19BN vs. ResNet152	0e0
VGG19BN vs. DenseNet161	0e0
DenseNet161 vs. ResNet152	4.860711e-2

As shown in Table 8, ResNet152 has the highest average top-1 accuracy on ImageNetV2, followed by DenseNet161, VGG19BN, and GoogLeNet. The lowest and highest top-1 accuracies per class achieved by all state-of-the-art models match the worst and the best accuracies possible, 0.00 and 1.00, respectively. The results presented in Table 9 indicate that two pairs of models have identical first quartiles, medians, and third quartiles (ResNet152-DenseNet161 and GoogLeNet-VGG19BN). After carrying out the Iman-Davenport test, a p-value < 2.2e-16 is obtained. We reject the null hypothesis: there are significant differences in the performance of our models when considering their top-1 accuracies on the ImageNetV2 benchmark. When applying the Friedman and Bergmann-Hommel methods to our six pairwise comparisons, the corrected p-values are below 0.05, meaning that there is sufficiently strong evidence to conclude that the pairs of models are significantly different (see Table 10).

5.4.2 ImageNetV2 Top-5 accuracies

Table 11. Minimum, Mean, and Maximum of ImageNetV2 Top-5 Accuracies

Network	Min.	Mean	Max.
VGG19BN	0.5000	0.9381	1.0000
ResNet152	0.5000	0.9597	1.0000
DenseNet161	0.4000	0.9516	1.0000
GoogLeNet	0.4000	0.9174	1.0000

Table 12. First Quartile, Median, and Third Quartile of ImageNetV2 Top-5 Accuracies

Network	1st Qu.	Median	3rd Qu.
VGG19BN	0.9000	1.0000	1.0000
ResNet152	0.9000	1.0000	1.0000
DenseNet161	0.9000	1.0000	1.0000
GoogLeNet	0.9000	1.0000	1.0000

Table 13. Corrected P-values for Pairwise Network Comparisons of ImageNetV2 Top-5 Accuracies

Comparison	Corrected P-value
GoogLeNet vs. ResNet152	0e0
GoogLeNet vs. DenseNet161	0e0
GoogLeNet vs. VGG19BN	1.465494e-14
VGG19BN vs. ResNet152	0e0
VGG19BN vs. DenseNet161	3.122579e-9
DenseNet161 vs. ResNet152	1.074642e-3

As can be seen in Table 11, ResNet152 achieves the highest mean top-5 accuracy on ImageNetV2, followed closely by DenseNet161, VGG19BN, and GoogLeNet. The maximum top-5 accuracies per class are 1.00, the highest possible. Moreover, the results shown in Table 12 indicate that all four networks have identical interquartile ranges. The null hypothesis is rejected

since the p-value < 2.2e-16 obtained from the Iman-Davenport test is below the significance level alpha: at least one model performs statistically differently when assessing the top-5 accuracies per class on the ImageNetV2 benchmark. To find out which models are statistically different, we proceed with the pairwise comparisons (see Table 13). Applying the post-hoc methods leads to corrected p-values below the significance level of 0.05; we have evidence that our six pairs of models perform significantly differently.

5.5 ImageNetC

For this experiment, only images suffering from a level three of motion blur are considered. Each of the 1,000 classes contains exactly 50 standard-sized images. The top-1 and top-5 accuracies are calculated for each of them, finishing with 1,000 observations per group.

5.5.1 ImageNetC Top-1 accuracies

Table 14. Minimum, Mean, and Maximum of ImageNetC Top-1 Accuracies

Network	Min.	Mean	Max.
VGG19BN	0.0000	0.2836	0.9000
ResNet152	0.4000	0.4626	0.9800
DenseNet161	0.0000	0.4019	0.9600
GoogLeNet	0.0000	0.2496	0.9000

Table 15. First Quartile, Median, and Third Quartile of ImageNetC Top-1 Accuracies

Network	1st Qu.	Median	3rd Qu.
VGG19BN	0.1400	0.2400	0.4000
ResNet152	0.3200	0.4600	0.6000
DenseNet161	0.2600	0.3800	0.5400
GoogLeNet	0.1200	0.2100	0.3600

Table 16. Corrected P-values for Pairwise Network Comparisons of ImageNetC Top-1 Accuracies

Comparison	Corrected P-value
GoogLeNet vs. ResNet152	0e0
GoogLeNet vs. DenseNet161	0e0
GoogLeNet vs. VGG19BN	6.741536e-10
VGG19BN vs. ResNet152	0e0
VGG19BN vs. DenseNet161	0e0
DenseNet161 vs. ResNet152	0e0

The results shown in Table 14 suggest that ResNet152 scores the highest average top-1 accuracy on ImageNetC, followed by DenseNet161, VGG19BN, and GoogLeNet. The lowest minimum top-1 accuracy per class is 0.00, the lowest possible. As can be seen in Table 15, the median top-1 accuracy for ResNet152 is considerably higher than the rest. We safely reject the null hypothesis as the p-value < 2.2e-16 obtained from the Iman-Davenport test is considerably lower than the established threshold value. One or more image classification models show significant differences in performance when comparing their top-1 accuracies on the ImageNetC benchmark. To discover what models are different, we continue with the post-hoc procedure. Applying Friedman's Aligned Ranks test and correcting the p-values using the Bergmann-Hommel method, we conclude that all four models perform significantly differently as the corrected p-values are below 0.05 (see Table 16).

5.5.2 ImageNetC Top-5 accuracies

Table 17. Minimum, Mean, and Maximum of ImageNetC Top-5 Accuracies

Network	Min.	Mean	Max.
VGG19BN	0.0200	0.5117	0.9600
ResNet152	0.1200	0.6972	1.0000
DenseNet161	0.1000	0.6346	1.0000
GoogLeNet	0.0000	0.4619	0.9800

Table 18. First Quartile, Median, and Third Quartile of ImageNetC Top-5 Accuracies

Network	1st Qu.	Median	3rd Qu.
VGG19BN	0.3400	0.5200	0.6800
ResNet152	0.5800	0.7200	0.8200
DenseNet161	0.5000	0.6600	0.7800
GoogLeNet	0.3000	0.4600	0.6200

Table 19. Corrected P-values for Pairwise Network Comparisons of ImageNetC Top-5 Accuracies

Comparison	Corrected P-value
GoogLeNet vs. ResNet152	0e0
GoogLeNet vs. DenseNet161	0e0
GoogLeNet vs. VGG19BN	4.440892e-16
VGG19BN vs. ResNet152	0e0
VGG19BN vs. DenseNet161	0e0
DenseNet161 vs. ResNet152	0e0

As shown in Table 17, the highest top-5 accuracy on ImageNetC is obtained by ResNet152, followed by DenseNet161, VGG19BN, and GoogLeNet. The highest maximum top-5 accuracy per class is achieved by ResNet152 and DenseNet161, while the lowest minimum one – by GoogLeNet. The results from Table 18 indicate that ResNet152 has a noticeably higher median top-5 accuracy. After performing the Iman-Davenport test, we obtain a p-value < 2.2e-16, meaning that the null hypothesis can be safely rejected: there is at least one model with significant differences in performance on the ImageNetC benchmark. According to Table 16, GoogLeNet, DenseNet161, VGG19bn, and ResNet152 are all significantly different as the corrected p-values using the Bergmann-Hommel method do not exceed the alpha of 0.05.

6. DISCUSSION

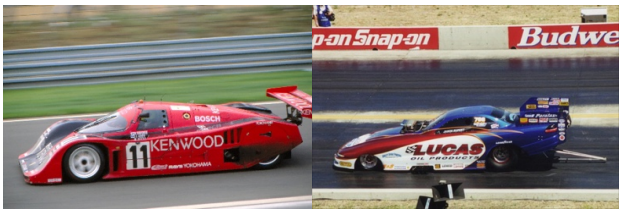

Figure 5. Image Samples from the "Sports Car" (Left) and "Race Car" (Right) ImageNetV2 Categories

Figure 6. Image Samples from the "African Crocodile" (Left) and "Goldfish" (Right) ImageNetC Categories (Motion Blur, Severity Level 3)

While previous research has focused on designing new architectures, the results contribute a clearer understanding of the metrics declared. The data indicate that the pre-trained convolutional neural network models (i.e., GoogLeNet, DenseNet161, ResNet152, and VGG19BN) are significantly different when performing image classification on three benchmark tests (i.e., ImageNet, ImageNetV2, and ImageNetC).

The average top-1 and top-5 accuracy scores obtained on the original benchmark coincide with the ones from the literature review. The observations of Kaiming He et al. [7] are therefore confirmed as increased depth indeed leads to significant accuracy gains. While comparable at a first glance [9], the performances of DenseNet161 and ResNet152 are unquestionably not similar.

Concerning the ability to adapt to new data, the results contradict the claims of Recht et al. [15] that the models fail to reach their original accuracy scores on ImageNetV2. Interestingly, there is a slight increase in top-1 and top-5 mean accuracies of 2-3% and 1-2%, respectively. The minimum class accuracies confirm the concerns of Beyer et al. [1] - though the images are drawn from the same distribution as the one used to create the models, their contents misrepresent them on many occasions. Because of the ambiguity and duplicity among labels, correct predictions that do not match the ground truth are translated to penalties. For instance, "sports car" is the class with the most difficult images to classify as no models can do so with the highest confidence. What happens here is that sports cars are misclassified as "race cars", when in fact they both illustrate the same concept (see Figure 5).

Sensitive tasks such as image processing imply reacting to noisy data. The four architectures cannot be regarded as robust when presented worst-case scenarios from the ImageNetC benchmark. There is a significant drop in average accuracies of 32-46% for top-1 and 24-44% for top-5. Even when applied a motion blur corruption of level three, the predictions seem to indicate context-awareness. Leading to a spread in confidences, images containing more than a single object of interest may also be the cause of that; the images supposed to represent an African crocodile or a goldfish (see Figure 6) are misclassified as "American coot" or "hair spray." Further, ResNet152's median top-1 and top-5 accuracies on ImageNetC are considerably higher and indicate its superiority over the competitors. This confirms the conclusions of Hendrycks and Dietterich [8] that deeper and wider models are considerably more robust.

The methodological choices were constrained by the high upload times and storage limitations. The original ImageNetC consists of 15 types of corruptions, each type having five levels of severity. Ending up with an astonishing amount of 3,750,000 images, testing the hypotheses on only one type of corruption with a level three of severity seemed the most appropriate option. Future studies should take into account different grouping strategies. Instead of calculating the top-1 and top-5 accuracies, different conclusions may be drawn when observing the confidences of the predictions. Given the confirmed issues with the images and labels of the original benchmark ImageNet, we recommend a strategy similar to ReaL [1] for obtaining better approximations of performance.

7. CONCLUSIONS

This research aimed to identify differences in the performance of four pre-trained image classification models. Based on a statistical analysis of the top-1 and top-5 accuracy scores obtained on three benchmark tests, it can be concluded that GoogLeNet, DenseNet161, ResNet152, and VGG19BN are significantly different. This research clearly illustrates the importance of statistical frameworks in the context of designing machine learning systems, but also raises the question of whether benchmarks like ImageNet are still useful in their current state. Based on these conclusions, practitioners should consider different groupings, metrics, or datasets in their future work. Finally, by addressing the gap in comparative studies, our findings challenge and confirm the existing assumptions.

8. REFERENCES

[1] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, Aäron van den Oord. 2020. Are we done with ImageNet?. arXiv: 2006.07159. Retrieved from https://arxiv.org/abs/2006.07159

[2] Borja Calvo and Guzmán Santafé Rodrigo. 2016. scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal, Vol.* 8, 1 (Aug. 2016), 1-8. DOI: https://doi.org/10.32614/RJ-2016-017

[3] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 7 (12/1/2006), 1–30. DOI: https://doi.org/10.5555/1248547.1248548

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, Miami, FL, 248-255. DOI: https://doi.org/10.1109/CVPR.2009.5206848

[5] Salvador García and Francisco Herrera. 2008. An Extension on 'Statistical Comparisons of Classifiers over Multiple Data Sets' for All Pairwise Comparisons. *Journal of Machine Learning Research* 9, 12 (Dec. 2008), 2677–2694.

[6] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.* 180, 10 (May 2010), 2044–2064. DOI: https://doi.org/10.1016/j.ins.2009.12.010

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv: 1512.03385. Retrieved from https://arxiv.org/abs/1512.03385

[8] Dan Hendrycks and Thomas G. Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations. arXiv: 1807.01697. Retrieved from https://arxiv.org/abs/1807.01697

[9] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv: 1608.06993. Retrieved from https://arxiv.org/abs/1608.06993

[10] David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (*SIGIR '93*). Association for Computing Machinery, New York, NY, USA, 329–338. DOI: https://doi.org/10.1145/160688.160758

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (June 2017), 84–90. DOI: https://doi.org/10.1145/3065386

[12] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 4 (Winter 1989), 541–551. DOI: https://doi.org/10.1162/neco.1989.1.4.541

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv: 1912.01703. Retrieved from https://arxiv.org/abs/1912.01703

[14] PyTorch. 2020. torchvision.models – PyTorch 1.7.0 documentation. Retrieved from https://pytorch.org/docs/stable/torchvision/models.html

[15] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. arXiv: 1902.10811. Retrieved from https://arxiv.org/abs/1902.10811

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575. Retrieved from https://arxiv.org/abs/1409.0575

[17] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556. Retrieved from https://arxiv.org/abs/1409.1556

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. arXiv: 1409.4842. Retrieved from https://arxiv.org/abs/1409.4842