Comparing Human and Machine Performance on Object Classification with Scrambled Images

Roshni Sahoo, 9.60 Final Project

May 2019

Introduction

Deep neural networks (DNNs) are achieving previously unseen performance in object classification, raising questions about whether DNNs operate similarly to human vision. In this study, we investigate whether global shape of an object is essential to accurately completing object classification tasks for humans and for machines.

Previous studies have shown that shape is an important cue for recognition in human vision. As a result, we hypothesize that when we distort the global shape of an object in an image, we expect human performance on the object classification task to degrade. On the other hand, many studies suggest that deep neural networks rely more heavily on local shape for classification tasks. This implies that if we distort the global shape of an object but simultaneously preserve local features, then DNNs will achieve similar classification accuracy.

The motivation for this project is to gain insight that will ultimately make deep learning systems more transparent. Deep learning systems are often construed as a black box because people cannot pinpoint exact reasons for why a system makes a particular decision. Furthermore, we aim to assess whether machines are better than humans at object recognition on these types of images because this may provide insight in how the human visual system may differ from computer vision systems. We can also determine when both of these systems fail by determining the point at what point human and machine performance is compromised due to scrambling.

Related Work

Many researchers have been interested in whether DNNs operate similarly to biological vision. In a study completed by Baker et al., researchers presented a trained DCNN (deep convolutional neural networks) with object silhouettes that preserved overall shape but were filled with surface texture taken from other objects [Bak+18]. Shape cues appeared to play some role in the classification of artifacts, but little or none for animals. In addition, they also found that DCNNs had no ability to classify glass figurines or outlines but correctly classified some silhouettes. Lastly, they evaluated performance on displays that preserved local features but disrupted global shape, and vice versa. With disrupted global shape, human accuracy was severely reduced, but DCNNs gave the same classification labels as with ordinary shapes. Conversely, local contour changes eliminated accurate DCNN classification but caused no difficulty for human observers. These results provide evidence that DCNNs have access to some local shape information in the form of local edge relations, but they have no access to global object shapes. To our best knowledge, no formal studies have been done thus far on assessing machine and human performance on the object recognition task with scrambled inputs.

Hypotheses

To assess the importance of global features for object classification, we define a specific type of distortion called scrambling and evaluate human and machine performance on object classification tasks with scrambled images. We can define *scrambling* as the following procedure. Given an image with dimensions $n \times n$ and a scrambling factor r, we can divide the image into r^2 squares of size $\frac{n}{r} \times \frac{n}{r}$ and stitch them randomly back together to make a scrambled $n \times n$ image. A highly scrambled image corresponds to a high r value.

We define t to be the object classification accuracy on a sample set of scrambled images.

$$H_0 : t_{\text{model}} = t_{\text{human}}.$$

In other words, our null hypothesis is that the models classification accuracy will match that of humans.

$$H_1 : t_{\text{model}} > t_{\text{human}}.$$

In other words, our alternative hypothesis is that the models classification accuracy surpass that of humans.

Experimental Variables and Metrics

We highlight our independent variables for this experiment.

1. Scrambling Factor (r) - The scrambling factor represents the degree to which we scramble an image. A higher scrambling factor indicates a more scrambled image.

2. Exposure Time for Scrambled Image (s)- The exposure time for scrambled image is the length of time that we will show human subjects the scrambled image. We will show an image for s seconds, where $s \in 200$ ms, 1000 ms, 2000 ms.

We explain our metrics, or dependent variables, for this experiment.

1. Classification Accuracy (t) - The classification accuracy for humans is the number of images that humans classifies correctly for a given combination of scrambling factor and exposure time (r, s). The classification accuracy for machines is the number of images that the model classifies correctly for a given scrambling factor (r).

2. Reaction Time - The length of time that a human takes to select an answer choice after being presented with a scrambled image. We can calculate the average reaction time across all possible combinations of scrambling factor and exposure time (r, s). This is not a relevant metric for machines because the network takes the same length of time to generate an output for all input.

Our Approach

Image Processing

The first step of this process involved image processing. We took 50 images from various categories whose labels are in the vocabulary of the ImageNet dataset. These images were randomly selected from the validation set of ImageNet, so pretrained DNNs have not been trained on the images before. Because creating the scrambled images required relatively little time or computational cost, we generated scrambled images with r ranging from $1 - 29$. We scrambled each of the 50 original images at each scrambling factor. We utilize Python image processing libraries such as OpenCV and Scikit-Image to generate these images.

We note that the distortion of scrambling also introduces a horizontal and vertical grid structure onto the image. We call these grid structures *edges*. In order to assess whether any change in performance can be attributed to the permuting of blocks of the image or to the introduction of edges, we assess machine performance on images with edges introduced, as well. We call these images *edge-induced images*. We create 50 images with edges induced for every scrambling factor by splitting each image into r^2 blocks of size $\frac{n}{r} \times \frac{n}{r}$ and shifting each block up, down, left, or right by $\frac{n}{4r}$ pixels. We fill in empty pixels with the color of the nearest neighbor of the pixel. We emphasize that these images are not

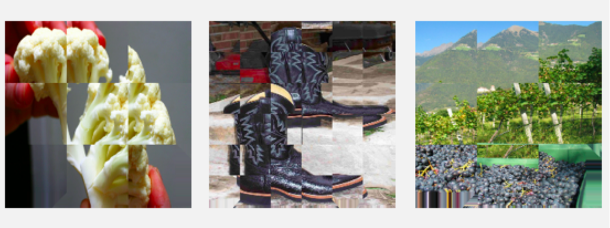


Figure 1: Examples of images with edges induced

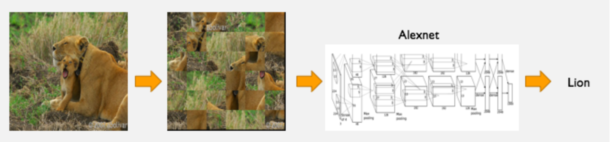


Figure 2: Pipeline for DNN's Object Classification

scrambled because the blocks maintain the same relative positions. Nevertheless, we compute these images for each scrambling factor so that we create edge-induced images with the same number of edges as their scrambled counterparts. See figure 1 for examples of the edge-induced images.

Data Collection

The next step was data collection. We created a survey of 50 questions in the format of the match-to-sample paradigm. Each question involved a multiple choice question for the network to answer. The stimulus image depicted a scrambled image from an ImageNet category. We randomly selected the incorrect answer categories to prevent introducing bias by tailoring the difficulty of the task to humans or machines. The correct answer was one of the answer categories.

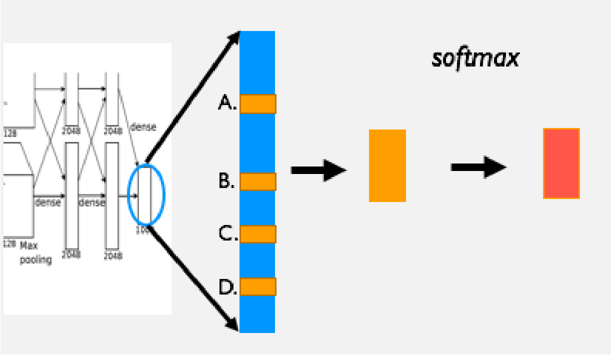
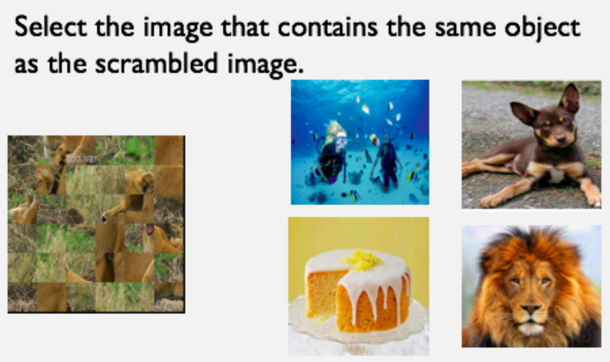


Figure 3: Multiple Choice Questions for a DNN



Select the image that contains the same object as the scrambled image.

Figure 4: Example question given to a human test subject.

Network Data Collection We wanted to determine whether the network can select the correct label for the scrambled image out of four possible answer categories. After passing the scrambled image as an input to the network, we extracted the weights of the logits that correspond to the image categories that are in our answer space for that particular input image. After that, we normalized these values using a softmax transformation, so that they summed to 1. The logit that had the highest value represented the answer choice that the network selected. See figure 3 for a diagram of how to pose multiple-choice questions to a network.

Human Data Collection For collecting data on human performance of this task, we created a survey using Amazon Mechanical Turk. As in the machine task, each question was multiple choice, and the stimulus image was a scrambled image from an ImageNet category. Although we gave humans the same stimulus and answer choices as the machines, we had human subjects select a choice from four images representing a different categories instead of text labels. We showed the human subjects the answer choices as images instead of text labels, so that the subject did not need to have prior knowledge of the names of objects in order to identify them. The incorrect answer choices were unscrambled images from different ImageNet categories. The correct answer was an unscrambled image from the same category as the scrambled image. See 4 for an example question posed to a human subject.

Across subjects, we randomize across exposure times and scrambling factors, never showing the subject two scrambled images that originated from the same template image. We exposed the human subjects to the scrambled images for various lengths of exposure time, and they were instructed to click on the answer choice that they believed matches the category of the unscrambled image. We provided an incentive for humans to answer the questions correctly by giving them a small monetary reward for each correct answer.
Due to constraints of running human trials, we elected to evaluate human performance on the 50 stimuli images with scrambling factors $r = 1, 3, 5, 7, 10, 15$ and with the exposure times of 100 ms, 200 ms, and 2000 ms.

Results

We detail the results from the experiments. We first assess machine performance on images that are not scrambled but have edges introduced. This allows us to understand how much of the change in performance on scrambled images can be attributed to the introduction of edges or altering the positions of the blocks. Next, we assessed machine performance by comparing the classification accuracy across stimuli images with different scrambling factors. We also compare the performance of various network architectures across a small range of scrambling factors.

Machine Performance

Network Classification Accuracy on Images with Edges Induced in Multiple Choice Paradigm The network architecture that we use for this experiment is AlexNet. We defined that an image is classified correctly by a network if it selects the correct category out of four predetermined answer choices. We calculated the classification accuracy across all 50 images with edges induced for each r from $1 - 29$. Although we observe a degradation in performance on edge-induced images, this degradation in performance does not appear to vary with r. The edges introduce noise that is not strongly correlated with r. When we attempt to fit a linear regression model to the data, we find that we generate a line with the following equation $t = -0.00409852r + 0.831133$. We compute that an R^2 value of 0.28766 for this regression, so 28.8% of the variance in classification accuracy can be explained by the variance in r. The correlation coefficient is equal to 0.536, which does not imply a strong correlation. This suggests that we can attribute much of the change in performance on scrambled images to altering the relative positions of the blocks of the image, rather the introduction of edges.

Network Classification Accuracy on Multiple Choice Questions Across All Scrambling Factors The network architecture that we used for our experiments was AlexNet. We calculated the classification accuracy across all 50 stimuli images for values of r in the range $1 - 29$. We observe that the classification accuracy decreases steadily as we increase the scrambling (see figure 6). By the time the scrambling factor is equal to 29, the network is essentially guessing between the answer choices because it is only correct a quarter of the time.

Comparison of Different Network Architectures' Performance We verified that various network architectures have similar degradation in performance patterns as we increase the scrambling factor. We see that for all network architectures the classification accuracy declines steadily as we increase the scrambling factor (see figure 7).

Network Classification Accuracy with Top Ten Criteria After running these preliminary

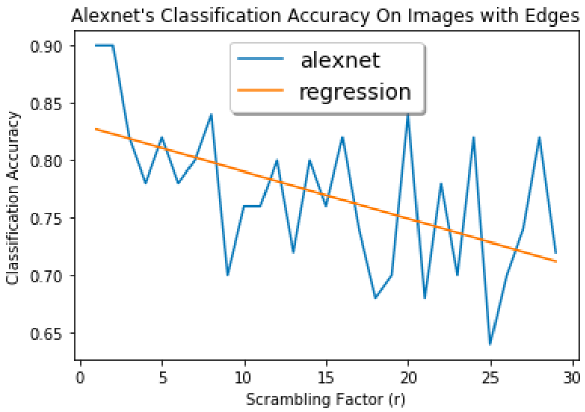


Figure 5: AlexNet's Classification Accuracy vs. Scrambling Factor on Edge-Induced Images. Note that although we plot against a variable termed scrambling factor, these images are not scrambled. Rather, scrambling factor refers to the fineness of the grid-structure that the edges impose onto the image.

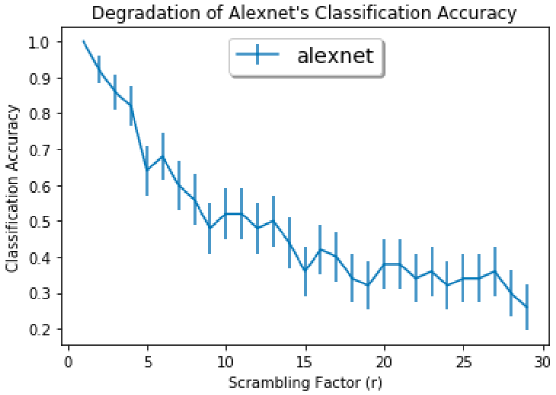


Figure 6: Alexnet's Classification Accuracy vs. Scrambling Factor.

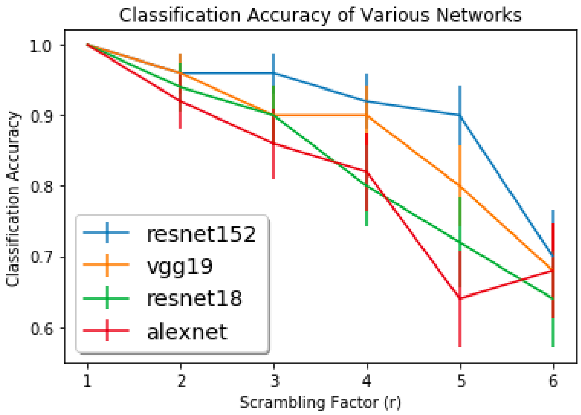


Figure 7: Classification Accuracy vs. Scrambling Factor for Various Network Architectures

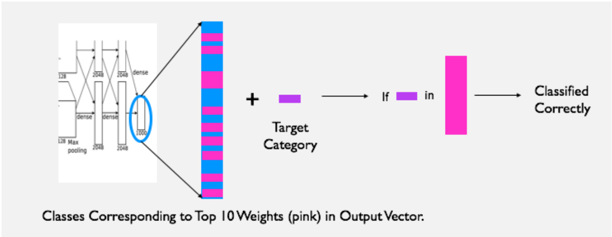

+
Target
Category
If
in
Classified
Correctly
Classes Corresponding to Top 10 Weights (pink) in Output Vector.

Figure 8: Top 10 metric to determine whether a DNN classifies an image correctly.

experiments, we were unsure whether the network was successful on the task due to poorly selected incorrect answer choices. We decided to introduce another method as a way to eliminate the bias from poorly selected answer choices. With this new criteria, a scrambled image is defined to be classified correctly if its class label is within the top 10 highest-weight classes in the DNNs output vector (see figure 8). We call this new metric the *Top 10 criteria*. Note that this is a harsher criteria than the multiple choice setting. We have evaluated Alexnet's performance on classifying scrambled images correctly over different scrambling factors with this metric (see figure 9).

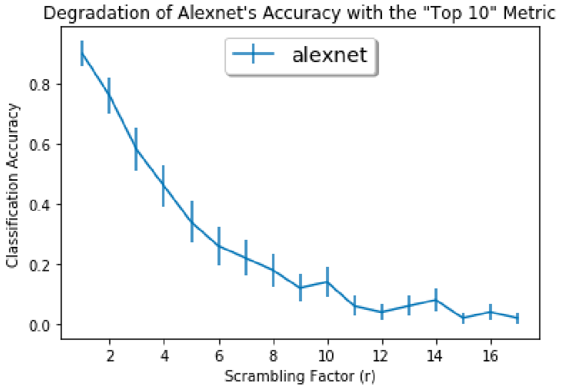


Figure 9: Alexnet's Classification Accuracy with Top 10 Criteria vs. Scrambling Factors

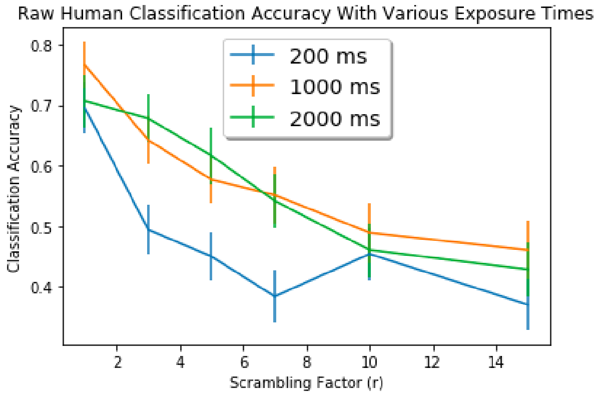


Figure 10: Raw Data: Human Classification Accuracy on Multiple Choice Questions vs. Scrambling Factors

Human Performance

We also assess human performance across various scrambling factors. We limit our assessment on human performance to scrambling factors $r = 1, 3, 5, 7, 10, 15$. In addition, we evaluate whether the performance at a given level of scrambling varies based on different exposure time. In the study, we had 86 human subjects total.

Raw Classification Accuracy for Various Exposure Times vs. Scrambling Factor We define that a human classifies an image correctly if they select the answer choice corresponding to the image from the same category as the scrambled stimulus image. When plotting the results, we observe that performance improves with longer exposure times. In general, humans perform worse when their is shorter; we can see that there is lower classification accuracy across the board for the 200 ms exposure time (see Figure 10).

Cleaned Classification Accuracy for Various Exposure Times vs. Scrambling Factor We observed that a few of the test subjects had very poor classification accuracy on all scrambling factors, including on images where $r = 1$, meaning that they were unscrambled. We removed all subjects that had worse than 50% accuracy on a random selected sample of half the $r = 1$ images on which they were evaluated. Out of 86 human subjects, there were 11 subjects that fell below this mark (see Figure 11.

Comparison of the Classification Accuracy of Humans and Machines We superimposed the AlexNet curve for classification accuracy with scrambling factor from $1 - 15$ on the graph of the cleaned human data. We see that machine classification accuracy is higher than human classification accuracy until $r = 15$.

Discussion

The above graph depicts a comparison between human performance and model performance at various scrambling factors.

The conclusions that we can draw from network performance include that networks can tolerate some degree of scrambling while maintaining a reasonably high classification accuracy. As expected, the network performance was near perfect on unscrambled images and declined in the range of scrambling factors $1 - 6$. After that, the network accuracy plateaued at a 40% accuracy rate. The network accuracy achieves random-guessing performance when $r = 29$. The network performs better than random-guessing on the four-choice classification task for scrambling factors

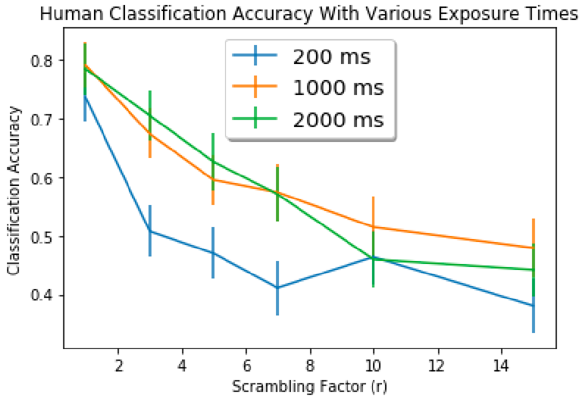


Figure 11: Clean Data: Human Classification Accuracy on Multiple Choice Questions vs. Scrambling Factors

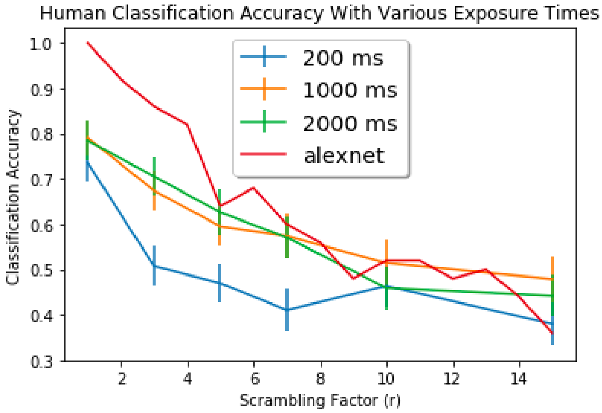


Figure 12: Comparison of Human and Machine Performance on Multiple Choice Questions Across Scrambling Factors

up to $r = 29$.

We expect global features and local features both impact a network's classification decision. Scrambling is a transformation of the image that distorts global shape but preserves local features (at least for scrambling factors less than 6). Since the network maintains classification accuracy higher than 70% on these images, we can conclude that local features play an important role in the network's classification decision. We observe very low machine performance for scrambling factors greater than 15, indicating that local textures after this point are distorted, as well.

Although the data suggests that DNNs outperform humans in this task of object classification on scrambled images, we are cautious in making such a sweeping claim because of the poor performance of the human test subjects on classifying unscrambled images. Since human performance on this baseline task is lower, it suggests that the human data may not be reliable. Possible explanations for this behavior include displaying ambiguous stimuli images in which the focal object of the image cannot be discerned, test subjects misunderstanding the task, or test subjects having a low incentive to complete the task correctly. Thus, we cannot claim that the data from the test subjects is representative of human performance overall.

The conclusions that we can draw from human performance are limited due to the noisiness of the data. Nevertheless, we observe that the classification accuracy for humans is higher with exposures of 1000 ms and 2000 ms than 200 ms. Humans have very high performance on *core object recognition*, which is the ability to rapidly ($<$ 200 ms viewing duration) discriminate a given visual object from all other possible visual objects without any object-specific or location-specific pre-cueing. Object recognition is solved in the brain via a cascade of reflexive, largely feedforward computations that culminate in a powerful neuronal representation in the inferior temporal cortex. If identifying objects in scrambled images was a core object recognition task, then we would expect no change in performance whether humans are exposed to the image for 200 ms or longer amounts of time such as 1000 ms or 2000 ms. However, an increase in performance with longer exposure indicates that this task requires more complex, or higher-level processing than feedforward object recognition.

That said, we do not expect human performance to increase linearly with exposure time. We observe a large jump in performance with exposure time increased from 200 ms to 1000ms, but quite similar performance with exposure times of 1000 ms and 2000 ms. This implies that there exists some exposure time after which human performance saturates and does not increase anymore.

Conclusion

Overall, the conclusions that we can draw from this experiment is that networks rely on both global and local features because we observe a degradation in performance with increased scrambling but still relatively high performance at scrambling factors between $1 - 6$.

Since DNNs perform better on this task than humans do, this may indicate that DNNs are more reliant on local textures in identifying objects.

Lastly, this task is more complex than *core object recognition* because we see an increase in classification performance with longer exposure time.

Acknowledgement

This work was completed at MIT in the class 9.60: Machine-Motivated Human Vision, taught by Pawan Sinha. We would like to thank Professor Sinha for providing guidance on the project. We would also like to thank our teaching assistant, Michael Lee, for assistance with the data collection and analysis.

References

[Bak+18] Nicholas Baker et al. "Deep convolutional networks do not classify based on global object shape". In: *PLOS Computational Biology* 14.12 (Dec. 2018), pp. 1–43. DOI: 10.1371/journal.pcbi.1006613. URL: https://doi.org/10.1371/journal.pcbi.1006613.