Image Classification Under Attack

Contributors: Sai Bharath, Sai Harsha, Sairam

Introduction

The aim of this project is to analyze and understand the classification accuracy of Image Classifiers when given with images of increasing distortion. Through our exploration, we wanted to understand the model's performance at each stage of varying complexity in the image and identify the point at which model's performance drops significantly. Gaining this understanding could offer valuable insights that may help improve the robustness and reliability of image classification models in real-world scenarios, where images are often corrupted or distorted.

Approach

For our investigation, we utilized Microsoft/resnet-50 available on huggingface.co, Inference API subjected to a series of progressively corrupted elephant images. These corruptions involved overlaying a zebra-stripe-like pattern at various transparency levels keeping the subject image transparency same at all times.

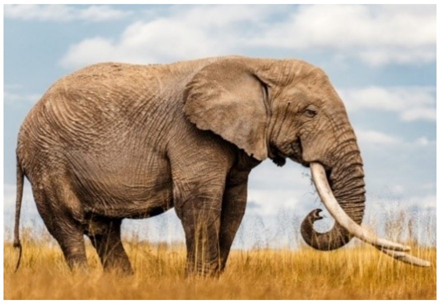


Figure 1 Image of an Elephant (subject): Source

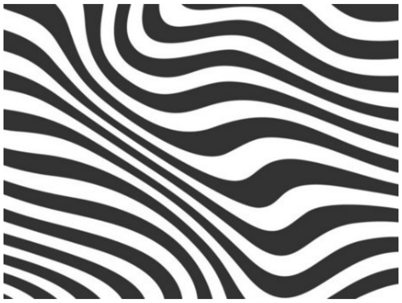
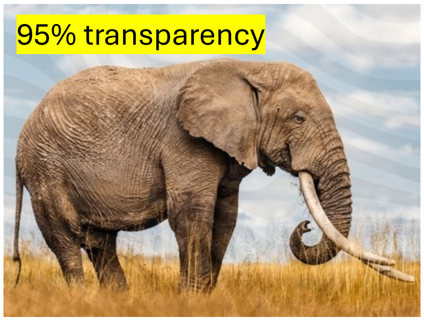
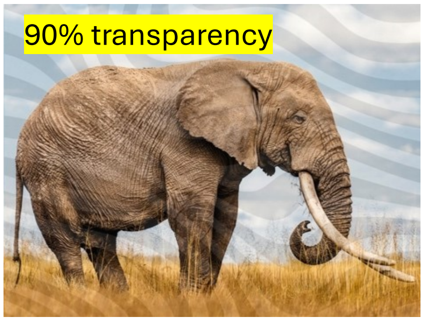
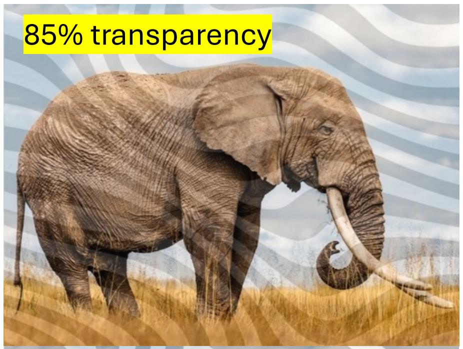


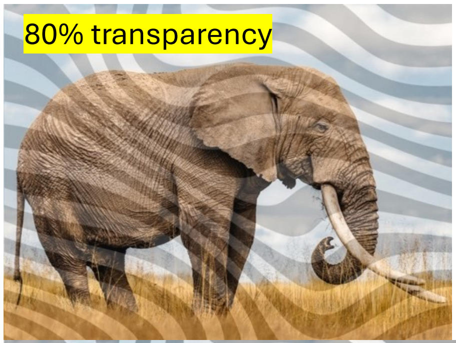
Figure 2 Noise (Zebra pattern) Source

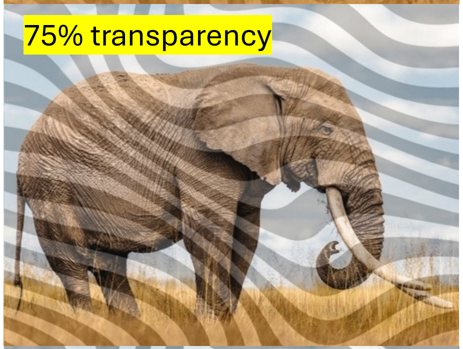
Following are the images that we generated using online-tool, to superimpose the pattern (noise) onto the elephant (primary image), by varying the transparency levels.

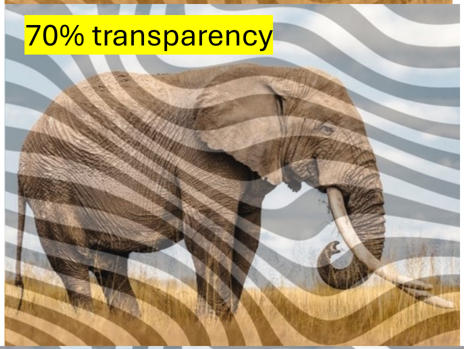


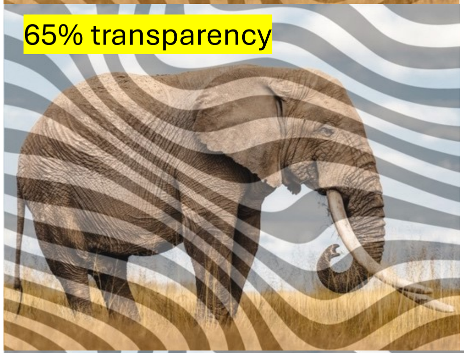


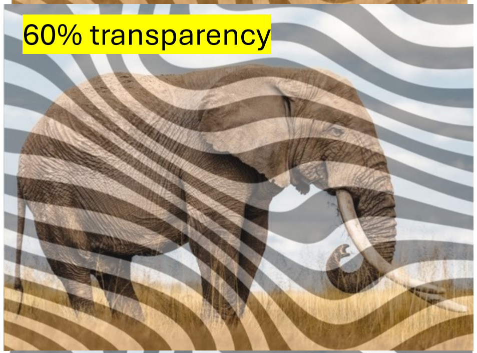

85% transparency

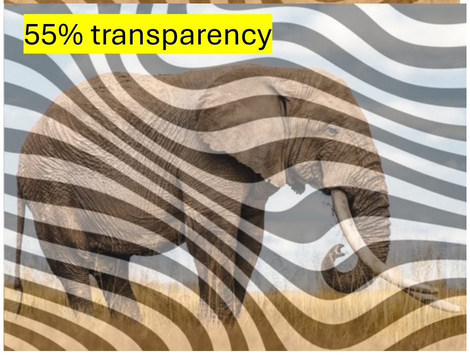

80% transparency

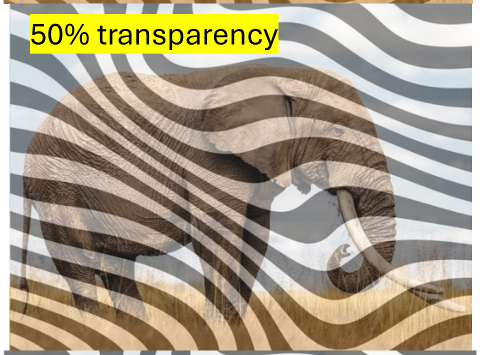

75% transparency

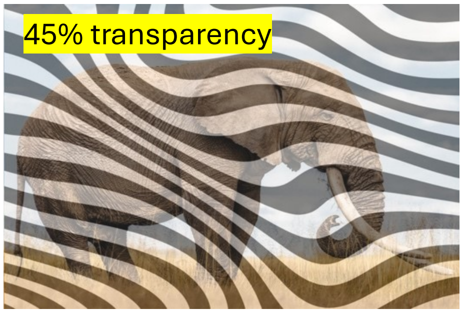

70% transparency

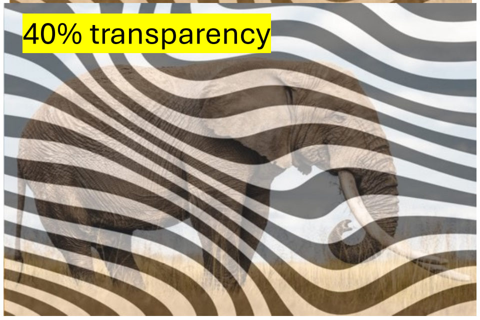

65% transparency


60% transparency


55% transparency


50% transparency


45% transparency


40% transparency

The motive to use this approach by superimposing two images and by using ‘different level of transparency’ as corruption mechanism is

- On a given spectrum, at the transparency level of 0% or 100% of the ‘noise’(zebra pattern) image, the model can accurately classify the object. In between these two transparency levels, neither of the image (subject and noise) is clearer to view in a single image. We want to test how would model interpret the super imposed image in between this spectrum
- Transparent images at different levels are easier to generate

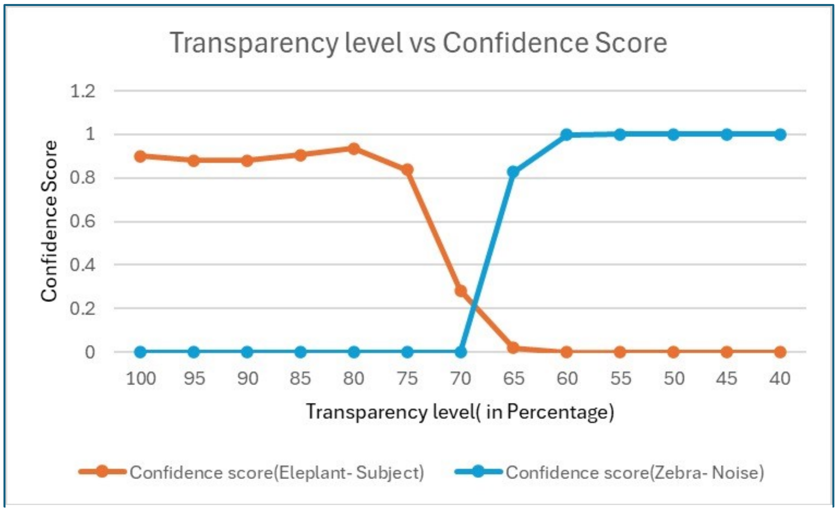
Analysis

When we fed these images to the Microsoft/resnet-50 classifier, it was able to correctly identify the animal in the image until the pattern became more pronounced. The model performed efficiently, with the confidence score for the class "Elephant" ranging between 83% and 93% for transparency levels between 75% and 100%. However, there was a significant drop in confidence to 28% when the transparency level reached 70%. From that point, the model struggled to classify the image correctly, with the confidence score decreasing at each further reduction in transparency, ultimately reaching 0% at a transparency level of 60%.

At the same time, at a transparency level of 65%, the model began predicting "Zebra" as one of the classes, with a confidence score of 82%. This score continued to increase, reaching 100% at a transparency level of 55%. When the transparency level dropped to 50% and below, the model no longer listed "Elephant" in its predicted classes, consistently predicting "Zebra" with a 100% confidence score.

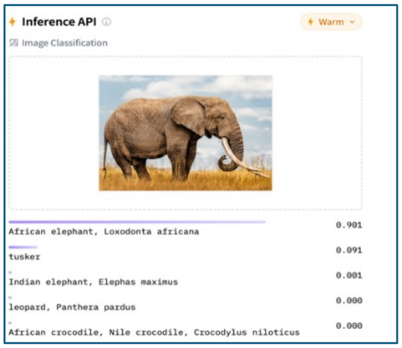
The results section shows the images provided by the model as response, starting from a transparency level of 100 down to 40.

The following plot gives a visual representation of how the confidence score of each class keeps changing at varying transparency levels.

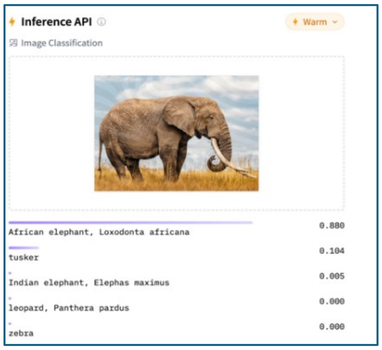


Graph: *Confidence score of elephant vs transparency level of noise image*

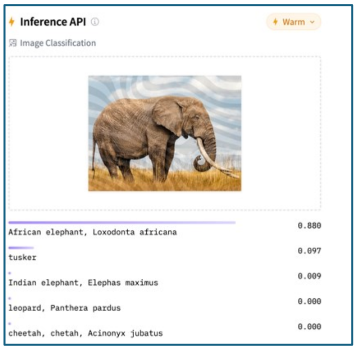
Results:



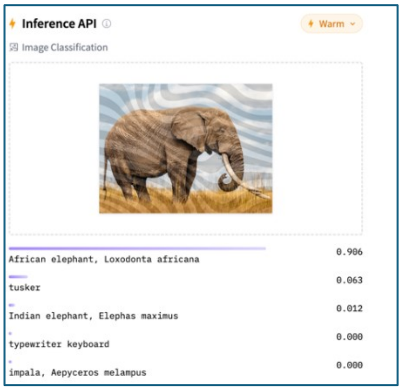
```
African elephant, Loxodonta africana                 0.901
tusker                                               0.091
Indian elephant, Elephas maximus                     0.001
leopard, Panthera pardus                             0.000
African crocodile, Nile crocodile, Crocodylus niloticus  0.000
```



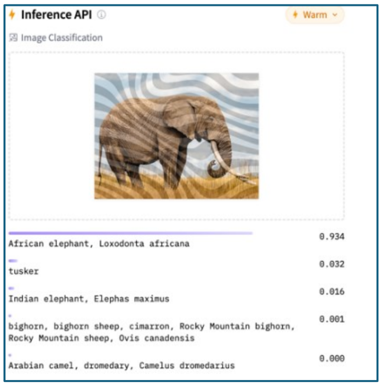
```
African elephant, Loxodonta africana                 0.880
tusker                                               0.104
Indian elephant, Elephas maximus                     0.005
leopard, Panthera pardus                             0.000
zebra                                                0.000
```



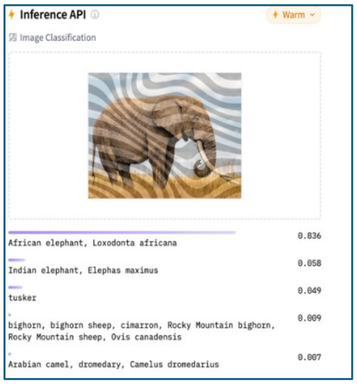
```
African elephant, Loxodonta africana                 0.880
tusker                                               0.097
Indian elephant, Elephas maximus                     0.009
leopard, Panthera pardus                             0.000
cheetah, chetah, Acinonyx jubatus                    0.000
```



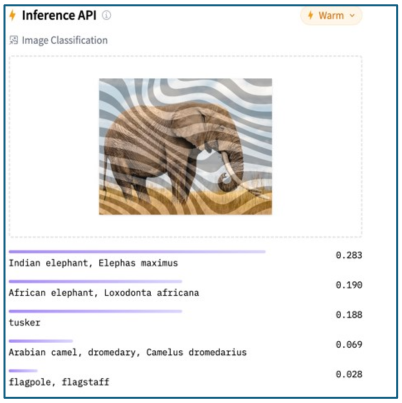
```
African elephant, Loxodonta africana                 0.906
tusker                                               0.063
Indian elephant, Elephas maximus                     0.012
typewriter keyboard                                  0.000
impala, Aepyceros melampus                           0.000
```



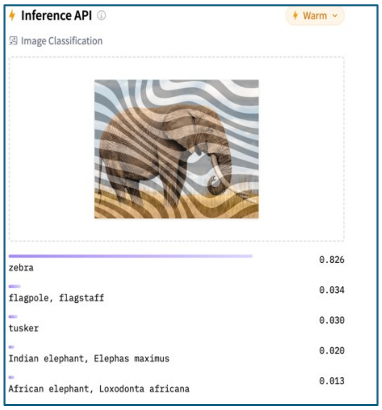
```
African elephant, Loxodonta africana                 0.934
tusker                                               0.032
Indian elephant, Elephas maximus                     0.016
bighorn, bighorn sheep, cimarron, Rocky Mountain bighorn,  0.001
Rocky Mountain sheep, Ovis canadensis
Arabian camel, dromedary, Camelus dromedarius        0.000
```



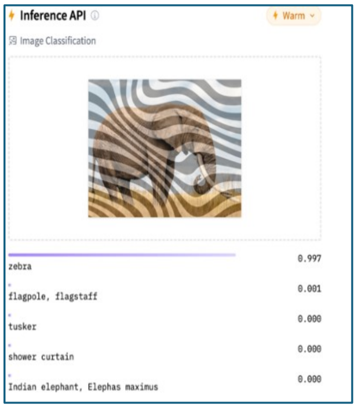
```
African elephant, Loxodonta africana                 0.836
Indian elephant, Elephas maximus                     0.058
tusker                                               0.049
bighorn, bighorn sheep, cimarron, Rocky Mountain bighorn,  0.009
Rocky Mountain sheep, Ovis canadensis
Arabian camel, dromedary, Camelus dromedarius        0.007
```



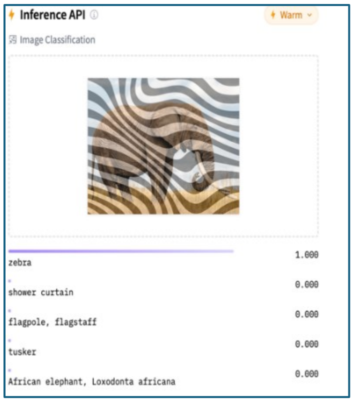
```
Indian elephant, Elephas maximus                     0.283
African elephant, Loxodonta africana                 0.190
tusker                                               0.188
Arabian camel, dromedary, Camelus dromedarius        0.069
flagpole, flagstaff                                  0.028
```



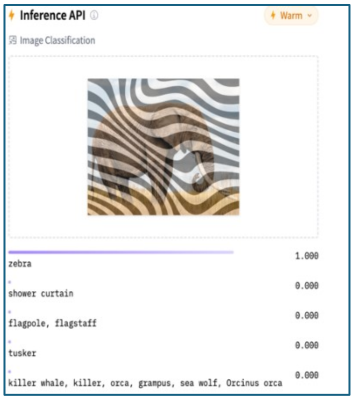
```
zebra                                                0.826
flagpole, flagstaff                                  0.034
tusker                                               0.030
Indian elephant, Elephas maximus                     0.020
African elephant, Loxodonta africana                 0.013
```



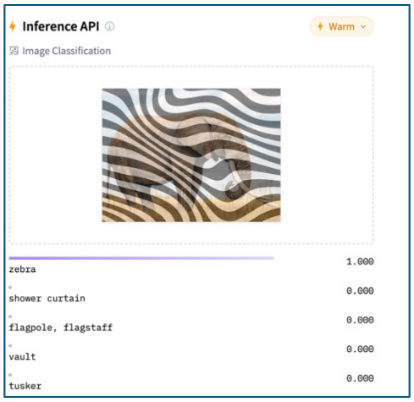
```
zebra                                                0.997
flagpole, flagstaff                                  0.001
tusker                                               0.000
shower curtain                                       0.000
Indian elephant, Elephas maximus                     0.000
```



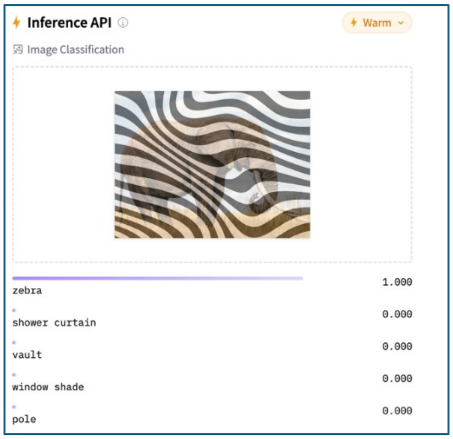
```
zebra                                   1.000
shower curtain                          0.000
flagpole, flagstaff                     0.000
tusker                                  0.000
African elephant, Loxodonta africana    0.000
```



```
zebra                                                       1.000
shower curtain                                              0.000
flagpole, flagstaff                                         0.000
tusker                                                      0.000
killer whale, killer, orca, grampus, sea wolf, Orcinus orca 0.000
```



```
zebra                 1.000
shower curtain        0.000
flagpole, flagstaff   0.000
vault                 0.000
tusker                0.000
```



```
zebra            1.000
shower curtain   0.000
vault            0.000
window shade     0.000
pole             0.000
```

These results show that, Resnet50 performs efficiently until the dominant and distinctive features (large body, tusks, trunk of the elephant) of the subject in the image are visible. At transparency level 70%, the zebra-like pattern becomes more visible, and its features (such as stripes) start to confuse the model. The elephant's distinctive features become less prominent, which explains the sharp drop in the confidence score for the "Elephant" class to 28%.

As the transparency continues to decrease (below 70%), the elephant's features are gradually masked by the zebra pattern. The model's ability to distinguish the elephant is compromised, which is why the confidence drops further until it reaches 0% at 60% transparency.

Summary

We believe that this behaviour of this model is caused due to high sensitive to textures and repetitive patterns. The model associates the zebra-like stripes with the "Zebra" class. This texture bias can cause the model to prioritize the zebra-like pattern over other features, leading to misclassification when such patterns become dominant in the image. The model predicted the class as "Zebra" with 100% confidence, even though there wasn't any zebra in the image. However, for the "Elephant" class, which was present in the image, the highest confidence score was only 93%. This experiment highlights a weakness in the model's robustness to image distortions.