# Vulnerabilities in Food Image Classification

**Tomasz Siłkowski** ts407106@students.mimuw.edu.pl
*University of Warsaw*

## Abstract

This study assesses the resilience of the 'nateraw/food' Visual Transformer, a food classification model, against common data manipulation attacks. Employing methods like LIME and Attention Rollout for insight, the research finds that the model withstands most transformations, some extreme photographic effects and methods of overlaying key non-food features can significantly alter the predictions. These results highlight the model's robustness, with implications for understanding the vulnerabilities of advanced computer vision systems.

## 1. Introduction

Deep learning models, particularly in computer vision, have seen remarkable advancements in recent years. As these methods become more popular and widespread, it has become crucial to be able to understand their process and explain their results. This study examines the robustness of Visual Transformer [4], a modern solution, against various data manipulation attacks.

Our study centers on the 'nateraw/food' [5] Visual Transformer, a highly popular model, available on HuggingFace. We chose this model due to the quality of training and overall execution. The task is properly defined, and appropriate architecture was selected with a robust pretraining regime on ImageNet-21k [3]. The crowdsourced nature of the dataset is a promising sign of a model invulnerable to deviations from typical images.

This paper aims to present an evaluation of user-accessible image data manipulations and their impact on the performance of this model, with the objective of uncovering any inherent vulnerabilities. We investigate not only the model's susceptibility to such attacks but also seek to understand the underlying reasons for any observed vulnerabilities. Insights gained from this robust model are likely to be applicable to similar models, offering broader implications for the field of computer vision.

## 2. Targeted model

**Dataset** In this study, we utilize food101 dataset [2] which consists of 101000 images spanning 101 categories of food representative of cuisines from across the world. The images are crowdsourced providing a realistic representation of composition styles, backgrounds, lighting quality and overall photo quality. This variety is essential for training a robust model as well as testing its biases.

**Architecture** The model assessed in our study is 'nateraw/food', an instance of the Visual Transformer [7]. Initially pretrained on the ImageNet-21k dataset [3], it was sub-

sequently fine-tuned on the Food101 dataset. In validation tests, this model achieved an accuracy of 89%. To measure its performance on outside data, we independently compiled a targeted dataset comprising 303 images, with 3 images per category, primarily sourced from Wikipedia and well-produced articles. This evaluation delivered promising results, with the model attaining a 92% accuracy rate with high confidence scores for correct labels. This focused assessment indicated that the model is resilient to out-of-distribution attacks.

## 3. Methods

**LIME** Local Interpretable Model-agnostic Explanations [6] provides model behavior insights by perturbing input data and observing changes in the model's output, offering a view into the model's decision-making process for specific inputs. This versatility involves dividing the image into regions independently of the division into patches made by the transformer model. However, the mismatch between LIME's regions and the transformer's patches pollutes the explanations with overlap. To avoid that, a higher density of regions is needed, which drives already high computational complexity.

**Attention Rollout** Attention Rollout [1] is a method from a family of Transformer-specific methods using attention matrices to give insights about model functioning. It reveals where the model's attention is focused and what it deemed to be key in a given inference. A significant advantage of Attention Rollout is its alignment with the model's processing scale, as it utilizes the attention matrices generated as a byproduct of inference. However, it's important to note that this method does not account for gradient information, which influences the model's outputs. Since its calculations overlap with the inference process, this method is computationally inexpensive.

## 4. Data manipulation attack

To provide a baseline for all attacks, a testing set was created. It consists of 1010 samples, extracted from the validation set with an equal distribution of 10 images per category. Our evaluation involved running inference on both original, unaltered images and their transformed counterparts. We then compared the scores for the correct label in each case, saving the change induced by the given transformation. This approach allowed us to focus on the transformation's impact on the model's prediction, without being influenced by the change of the top prediction. Additionally, it also allowed us to measure the positive effect of transformation on the accuracy of inference, which then could be interpreted as an inherent level of noise of the evaluated method.

### 4.1 Photographic Effects

The first family of distortions we examine is a range of common photographic effects. They are particularly relevant to our study as they are readily accessible to an average user and result in low information loss. Their accessibility and intuitive nature make them a prime choice for initial attempts by users to deceive the model.
In this category, we explore the most popular effects with distinct impacts on the im-

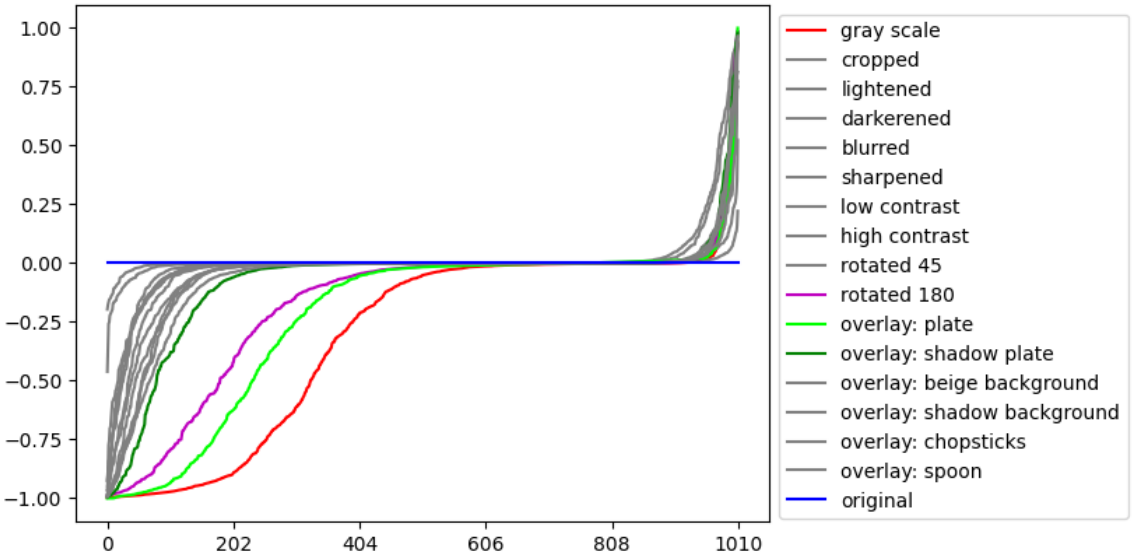


Figure 1: Plot of change of score for the correct label after manipulating the image. Note difference between overlay – lime – and its 'shadow' – dark green.

age quality and information content: blurring and sharpening, brightening and darkening, change in contrast, rotation, conversion to grayscale, cropping

### 4.2 Overlaying

In another approach, we consider overlaying distinct features onto different images. This technique involves a manual review of data instances where the model predicts incorrectly with LIME and attention rollout. The goal is to identify non-food features that consistently appear across many images and could potentially confuse the model. These features include background setting, utensils present or even the framing of the photo.
Our experimentation revealed that larger overlays tend to obscure more of the original image, which can significantly impair the model's inference accuracy. To control for the impact of blocking, tests with additional overlays, called 'shadows' were introduced. These 'shadows' are overlays of shape identical to the original but filled with black, allowing us to isolate the effect of blocking information given overlay has.

## 5. Conclusion

Our analysis shows that the model is generally resilient to data manipulation. As illustrated in Figure 1, out of 16 transformations (10 photographic and 6 overlays), only 3 impacted the model's predictions significantly: 180-degree rotation, an overlay of the plate and the application of grayscale filter. These techniques vary in their destructiveness of information, however, it's important to note that plate overlay impacted the model significantly more than its 'shadow'. This success leads us to conclude that strategically selected non-food features when overlaid on unrelated images, can effectively alter model predictions. As for

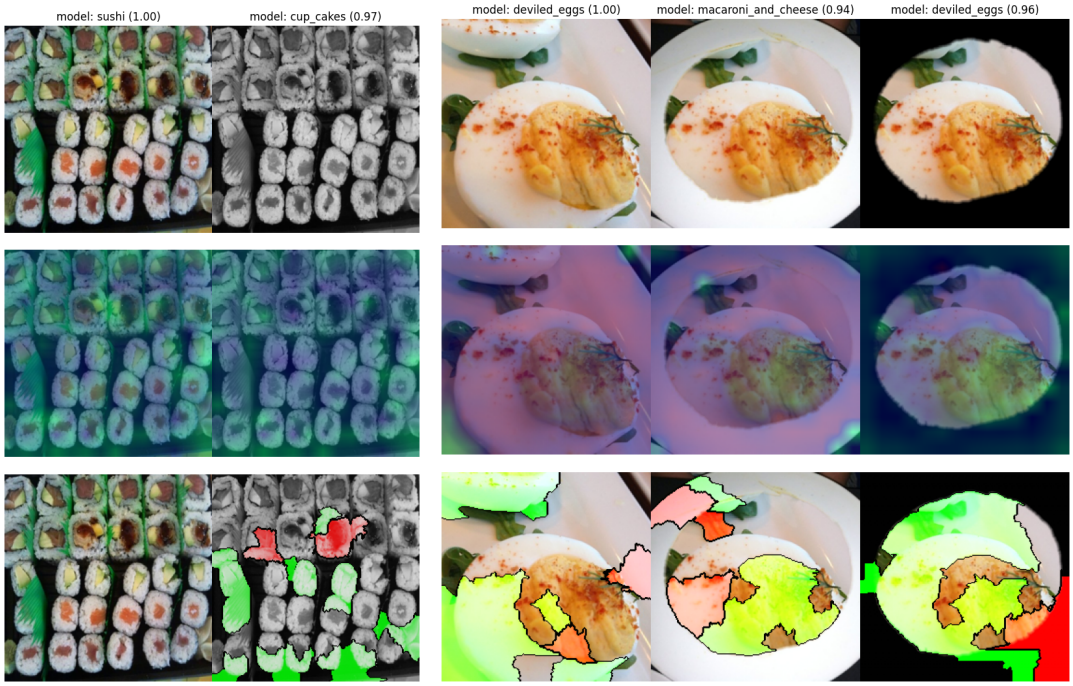


Figure 2: Representative examples of grey filter and plate overlays working with Attention Rollout and LIME explanations. scale of Attention Rollout: 1 (Green) – 0.5 (Blue) – 0 (Red); LIME: weights above 0.5 are shown, positive as green, negative as red; (Left) Applying grey filter changes model prediction from sushi to cupcakes with very high confidence. (Right) Overlaying a plate over the image causes the model to predict macaroni and cheese instead of deviled eggs despite no adjustment for 'shadow' overlay.

photographic distortions, the impact of extreme measures, such as half-full rotation and grey filter, is consistent with loss of information in the image. This outcome demonstrates that only drastic photographic transformations to the image have a significant effect on the model's reasoning.

## Appendix A.

In this appendix, we show additional plots.
Firstly, we have an evaluation plot for all transformations in Figure 3.
We also provide additional visualisations of explanations for prediction before after applying all tested transformations: Figure 4.
Some more examples of explanation visualisations are show in Figure 5.

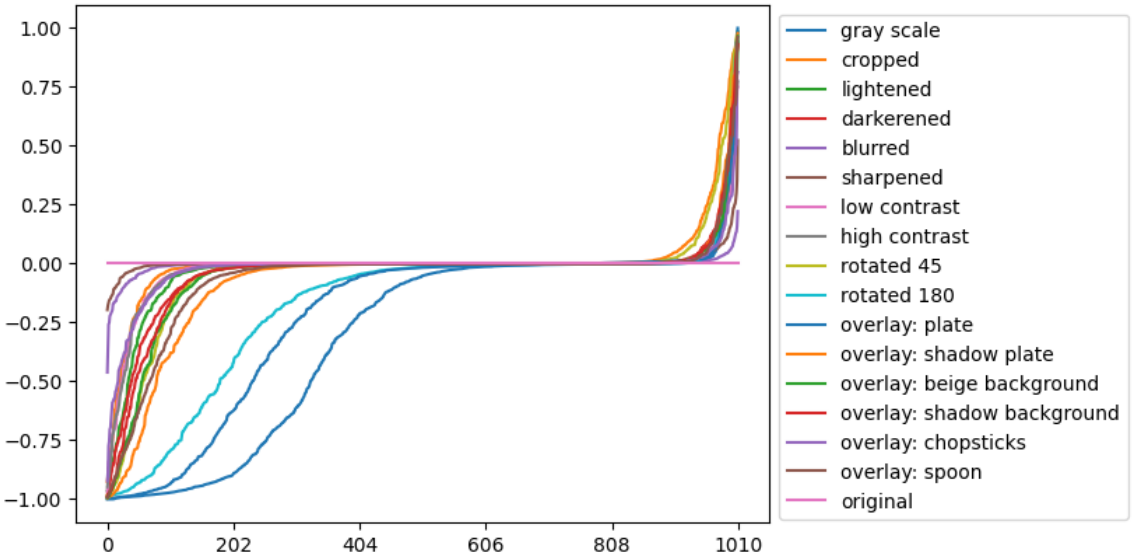


Figure 3: Plot of change of score for the correct label after manipulating the image.

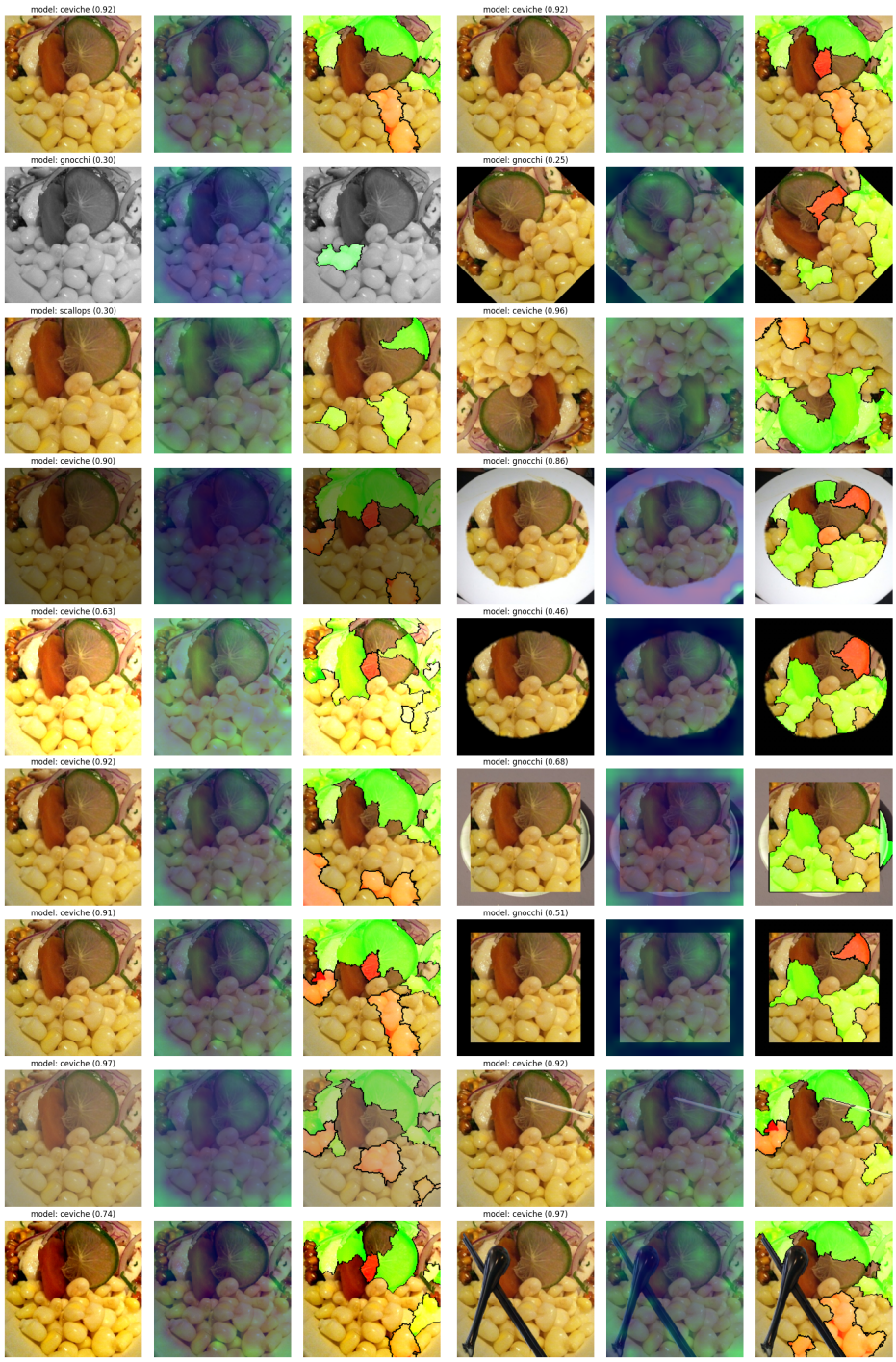


Figure 4: All transformations applied to the same image with explanations from Attention Rollout and LIME; scale of Attention Rollout: 1 (Green) – 0.5 (Blue) – 0 (Red); LIME: weights above 0.5 are shown, positive as green, negative as red;

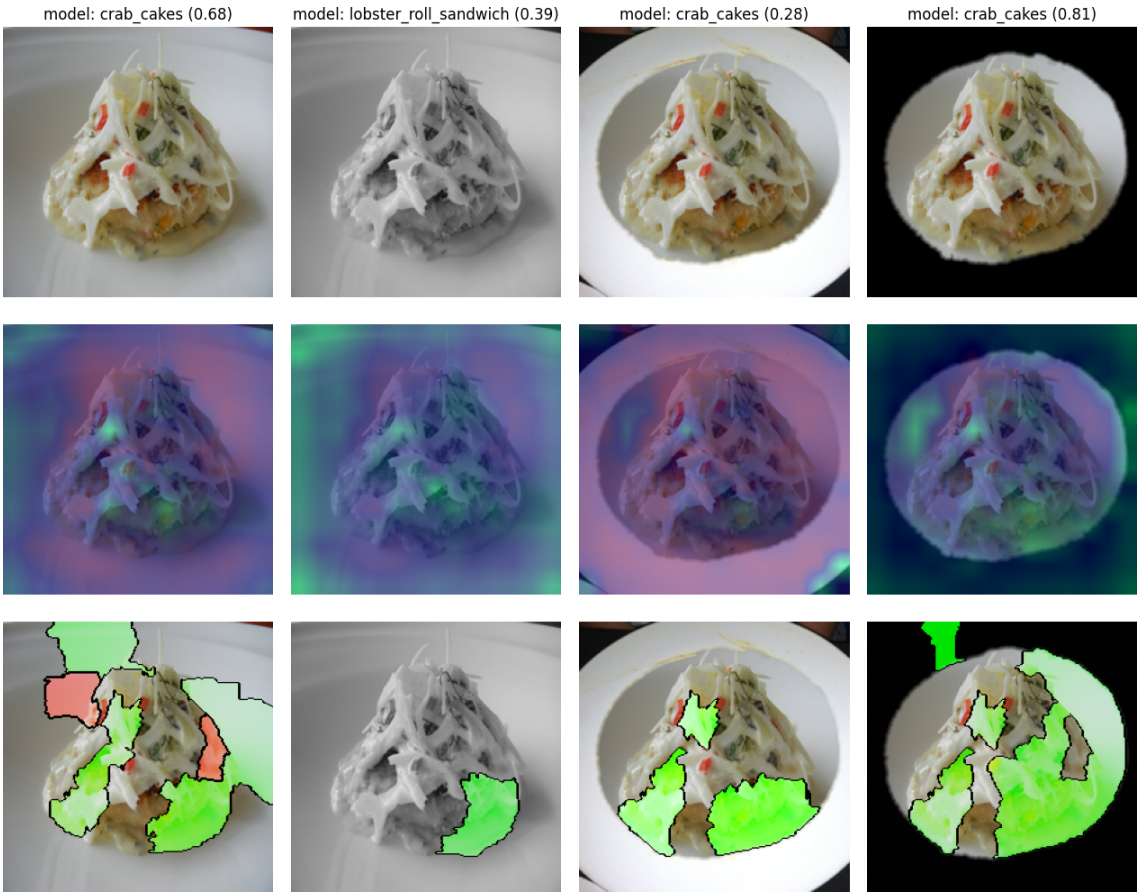


Figure 5: More successful examples; scale of Attention Rollout: 1 (Green) – 0.5 (Blue) – 0 (Red); LIME: weights above 0.5 are shown, positive as green, negative as red;

## References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] Nate Raw. Food Classifier. `https://huggingface.co/nateraw/food`, 2023. [Online; accessed November-2023].

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[7] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.