# Opening up the blackbox: Explaining AI decision-making through object recognition

Jiro Mizuno
Northeastern University
mizuno.j@northeastern.edu

## Abstract

In this paper, we explore the use of Explainability models for Computer Vision applications for both education and the general public. We used LIME [1] and GRAD-CAM [2], applied to a CNN model used to classify between dogs and cats. We also utilized an accessible GUI to make the process of prediction and explanation more simplified and streamlined for better accessibility. We show that the explainability models can both improve the transparency of a model's decision-making and sometimes encourage the users to be more cautious due to the model's understanding of the subject being questionable. Significantly, it successfully highlighted weak points in the underlying classification systems that would not be obvious from accuracy and loss metrics on validation data sets.

## 1. Introduction

Computer Vision models have recently received a lot of attention in various fields from engineering, security, medicine and even education. According to some estimates by experts, the industry of educational technology is still rapidly growing as technology improves, currently raising the total worth of the industry to approximately $8 billion [3]. With adjacent technologies such as virtual reality and alternate realities also rapidly becoming further publicly accessible, computer vision is also growing as a critical field in education. While the main focus of the field remains on pattern recognition, another critical field under consideration is how to provide the public with acceptable reasons for them to trust the provided computer vision model.

Beyond the expanding business horizons of computer vision, this current age of publicly available AI that can seemingly magically give any answers to any student wishing to ask, is a critical one for current students. Now more than ever, the utility of AIs such as ChatGPT have been recognized by students and the notion of studying as we did before is becoming obsolete.

However, despite such advancements, the central issue remains that for the vast majority of people, there is no clue to exactly how such a powerful tool makes its decisions. Without any clues there can be no trust between students and the results the AI returns. How can we be sure the AI can be trusted? Thus, to establish trust, we must make models explain themselves, so that mistakes can also be addressed as soon as possible.

Explainability models seek to reveal what dense neural networks concentrate upon to make their final decisions [4]. We seek to use explainability models to see if we can improve trust in Computer Vision models for the general public, and, even if trust does not increase, we can use explainability models to expose how our classification models can make unnoticed mistakes. If we cannot understand how a model decided upon its final prediction, we cannot fully trust it, and thus explainability technologies help bridge the gap of understanding between machines and users.

## 2. Related Work

Computer vision has not been widely utilized in the current state of the field of education, but its applications can be predicted to have a large impact sooner than later. Evidence by researchers suggest that many students are comparatively better as visual learners than rote text-based learners, and as such, integrating greater visual components into an educational curriculum can be expected to have a greater impact on their studies than purely formula and textual based explanations [5].

With further reports of students embracing large language models such as ChatGPT as part of their educational experience, the critical importance of AI in the daily lives of students is only growing stronger [6]. However, in many cases students are also growing increasingly concerned with the usage of AI in their academic lives. Thus, to utilize both the incredible potential of AI in education, but also temper its influence in students, it is increasingly critical for a way to increase both user trust and skepticism to increase the utility of AI overall for students and educators.

Although there are some researches indicating the importance of AI explainability as part of case studies involving existing platforms such as RiPPLE and FUMA, the field is comparatively under researched, and thus its full potential remains to be seen [7].

Finally, Local Interpretable Model-agnostic Explanations (LIME) was developed in a paper by Ribeiro et al [1]. It is a system that can highlight which parts of any input are most critically impacting the decision-making process of any classification network. This technology is not only limited to conventional neural networks and images, but also with different models and classification types such as Random Forest and text. Gradient-weighted Class Activation Mapping (GRAD-CAM) was introduced in the paper by Selvaraju et al [2]. It is a technique that highlights the image regions that contribute most to the model's prediction for a particular class. In comparison to LIME, it is only meant to be used for image classifications.

## 3. Methods

### 3.1 Model Architecture

Our dataset was pulled from Kaggle [8], and consists of 12,500 images of cats and dogs each for a total of 25,000 total, but due to the both data storage and time constraints,

we opted to only utilize the first 1,500 of cats and dogs each for a total of 3,000 training and testing images.

The image dataset itself was shuffled and randomized, but also stratified to make sure both training and testing pools had equal amounts of cats and dogs to reduce the risk of over and under representation during training and validation. Our training and testing split was 80-20. The images were also of varying sizes, so to preprocess the images, we resized and normalized them. Additionally, we applied a random horizontal flip and rotation in the range of (-10, 10) degrees, to help guide the network to learn features that are rotation and translation invariant.

For our model, we used a conventional neural network that consisted of five fully convolutional layers formed from a 2d convolutional layer, a ReLU activation layer, a 2x2 max pooling layer. The fully convolutional layers were then followed by a dropout layer with a rate of 20% and finally a sigmoid layer to determine the final classification. The dropout layer was used to help the model be less overfitted and reduce reliance on training data. For each 2d convolutional layer we doubled the filter size starting from the initial 32x32, we progressed to 64x64, 128x128, 256x256, and finally 512x512. They were all accompanied by a 3x3 kernel. With other layers considered this resulted in a network with a total of 8,123,201 parameters.

During training, our CNN employed binary cross entropy as our primary loss metric, utilized Adam as our optimizer with a learning rate of 1e-4, grouped images in a batch size of 32, and finally ran training for a total 75 epochs.

## 3.2 Cat Dog GUI

For our small, but accessible GUI to showcase explainable AI, we mostly utilized the PyQt6 package. The application itself only consists of four components: a radio button group to determine whether to use LIME or GRAD-CAM as the primary explainable technology, a button to upload an image of a cat or dog to be predicted and explained, an image showcasing the results of explainability technology, and finally a label that prints the final prediction of the model and its confidence level expressed as a percentage.


Figure 0. Explainability AI app in action with LIME

The CNN model is not trained as part of the GUI but separately in a dedicated Jupyter Notebook file, where it is trained and stored into the project directories. The GUI then loads up the model to use in prediction and explainability explanations.

## 3.3 LIME explainability

LIME explainable AI was first pioneered by Ribeiro et al, 2016 [1]. The primary focus of the LIME explainability technology is that it is model-agnostic, able to be used on any sort predictive model ranging from CNNs to Random Forest without the need to examine the weights of the model itself. This allows the technology to be used in diverse situations, but it cannot explain the predictions of individual nodes or groups within models.

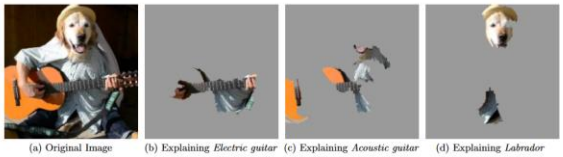


Fig 1. LIME explaining contributions to each class [1]

LIME for images works by generating a collection of slightly altered input images relative to the target image, feeding them through the model to gather results, and then seeing which alterations cause the most significant changes in predicted class. Thus, LIME can determine which parts of the base image have the largest weight on the output class. It then creates a mask on the image corresponding to those parts, which can be used to draw boundary lines or simply select those parts in isolation.

## 3.4 GRAD-CAM explainability

GRAD-CAM explainable AI was developed by Selvaraju et al [2]. It is an explainability technique specifically designed for convolutional neural networks (CNNs). In comparison to the previous LIME example, GRAD-CAM is limited to CNN-based models.

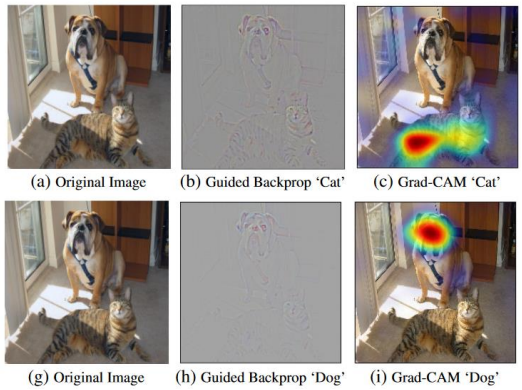


Fig 2. GRAD-CAM tested on both a dog and a cat [2]

GRAD-CAM works by extracting a model's final convolutional layer, and subsequently computing the gradient of the top predicted class for the input with respect to the activations of the last convolutional layer.

The gradients are then used to determine which areas of the image are most important to the model's predictions based on magnitude over all channels. These magnitudes are then

converted visually into a heatmap where important areas are labelled “hot” painted dark red while less critical areas are gradually painted in “cooler” colors.

# 4. Results

## 4.1 CNN training and validation

The most successful training session for our CNN model yielded a final training accuracy of 0.9187, a training loss of 0.2018, a validation accuracy of 0.8583 and finally a validation loss 0.3532.

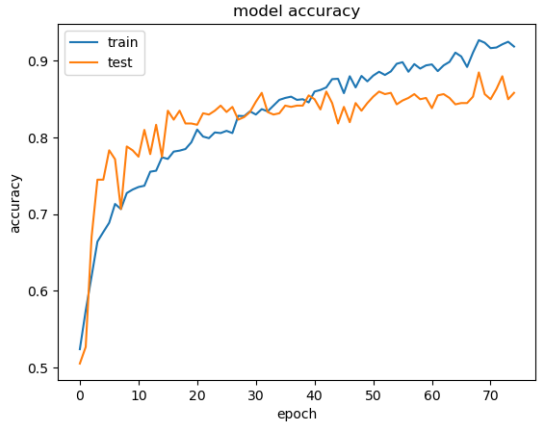


Fig 3. Model accuracy over time for training and testing

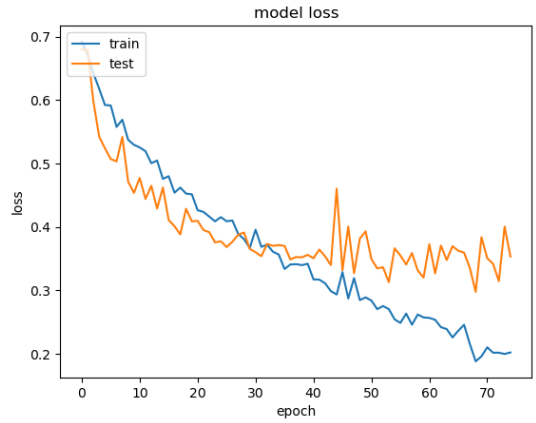


Fig 4. Model loss over time for training and testing

As seen in figures 3 and 4, while the training accuracy and losses continued to decrease for the entirety of the training session, both the validation accuracies and losses practically plateaued around epoch 40.

This suggests a degree of overfitting is occurring in the model for unclear reasons. Most likely, the model maybe overengineered with the number of parameters it handles, or it did not have a high enough dropout rate to destabilize the hold the training data has on the model. Whatever the case, with a validation accuracy of around 86% we determined the model acceptable for use in the GUI and saved the model.

## 4.2 Impressions from observing AI explanations

As this project is not based on strict numerical analyses and goals, the following section may lack hard data weight, but we believe the conclusions are valuable nonetheless as a window into the inside of CNN models.

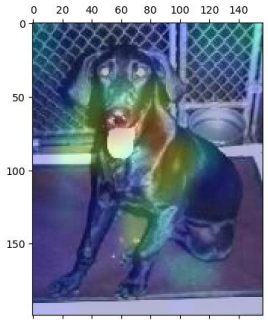

Fig 5.

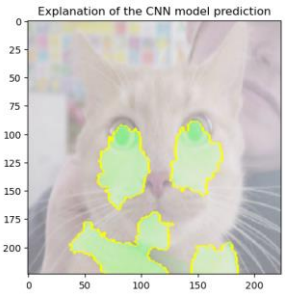


Fig 6.

Overall, the accuracy of the predictions resulting from our current model are correct and result in comprehensible explanations for both explainability technologies. For example, as seen in figures 5 and 6, both GRAD-CAM and LIME resulted in correct predictions where the most significant factors were focused upon the heads or areas close to the heads of the cats and dogs.

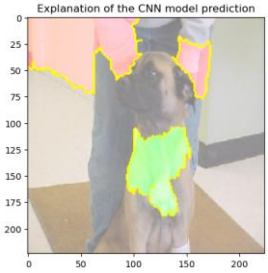


Fig 7.

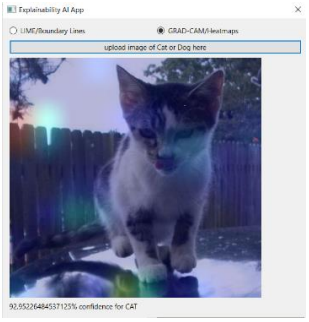


Fig 8.

However, for a significant section of the images, while the predictions themselves are correct, the explanations provided by LIME and GRAD-CAM that led to such decisions are puzzling. For example, in figure 7, while there is some focus on the dog’s torso by the model, the other half of the model’s focus is squarely on the human’s legs behind and the linoleum floor next to them. In figure 8, most of the focus is not on the cat, but on the fence and the tree beyond it instead.

These confusing results suggest that for a significant minority of the training data, the CNN model was trained to focus upon the environments the animals live in instead of the actual animals themselves. As the actual predictions themselves are correct, it is debatable whether this sort of misguided fixation by the model is acceptable. Despite the confusing focus on aspects of images that have no bearing to the actual subject, which may partially be the cause of validation losses, the predictive results of such images remain correct.

In another perspective, the fact that we can ask questions and weigh the pros and cons of trusting this model is only possible, because we can identify what sections of an image the model is focusing upon. Thus, the usefulness of

explainability technologies such as LIME and GRAD-CAM are proven.

For predictions that are incorrect, while it can be assumed that the explainability technologies will allow us to understand what sorts of mistakes were made, in a surprising amount of cases we simply don't know still.

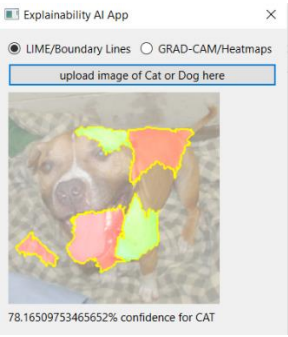


Fig 9.


Fig 10.

For example, in figure 9, the model believes the dog is a cat with 78% confidence, and it appears such confidence is based on the jaws and torso of the dog. Unfortunately, we do not have a clear answer for why focusing upon the distinctive body parts of a dog results in the model thinking that it is a cat.

Fortunately, not all incorrect cases are incomprehensible, and instead in other cases such as figure 10, the cause of model's mistakes are clear. The primary focus of the model is on the floor surrounding the cat. Therefore, with the tendency of dogs to be photographed on indoor, linoleum floor, it is reasonable to think that the model became too hyper-focused on how dogs tend to be inside photographed on clearly distinctive floors.

Finally, it should be acknowledged that while the final model's predictions remain the same even when using different explainability technologies, their explanations of how the model arrived can differ greatly. As mentioned earlier, LIME cannot observe the inside workings of the model and must extrapolate from repeated partial image testing. In contrast, GRAD-CAM is directly fed from the last convolutional layer of the model. Therefore, the two technologies can and frequently end up with different explanations and perspectives.

Fig 11.

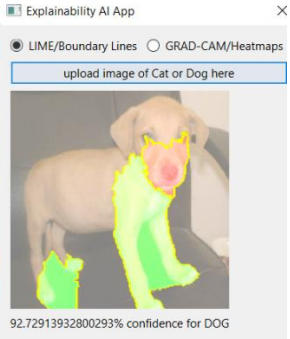


Fig 12.

For example, in figure 11 and 12, GRAD-CAM believes the model correctly identified this dog due to its surroundings. On the other hand, LIME believes that the most significant factor was the head and front legs of this dog. This showcases the utility in advancing different explainability technologies to gradually figure out why the model behaves the way it does. More perspectives may allow us to better determine the inner mechanics of the model.

## 5. Conclusions, Discussions and Future Work

From our continued testing of the CNN model, it is clear now that despite the general success in the model's decision-making, there is a still a great need for caution. Both LIME and GRAD-CAM's various explanations show that the model has a troubling tendency to focus on environments instead of the subject to make decisions.

Despite these failures, the fact that we can address them shows the immediate benefits of explainable AI. We no longer need to trust a model's predictions purely based on validation accuracy, but can view the inner workings of the model to either trust or distrust its decision-making as we see fit. This valuable insight compels us to improve our models and stay cautious of their decisions so that they may serve us better.

We also believe this informed caution can bring benefits to the field of education as students can be taught to both understand how previously incomprehensible machines make decisions, and how they make mistakes. It is critical that these students who are faced with a mountain of powerful but mysterious AIs to not be afraid to embrace the future, but remain cautious of their supposed powers.

Finally, even as the GUI is relatively barebones, we believe it accomplished its goal of making explainability technology far more accessible for the average person. By streamlining out all directory file manipulation and Jupyter Notebook's sprawl of code to simply selecting one of the explainability technologies and uploading a file, we massively simplified the process of getting an explanation. We hope this helps others use the exciting new technologies of today more effectively. Moreover, with a such a simple design, it maybe useful to use in a classroom without the need for much preparation, but it still may make as much of an impact in student understanding of these technologies. Although, explainability technology remains quite recent, the great potential to further technological understanding throughout society is exciting.

## 6. References


1. Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

2. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.. *ICCV* (p./pp. 618-626), : IEEE Computer Society. ISBN: 978-1-5386-1032-9.

3. Soni, H. (2022, December 5). *Significance of Computer Vision in the EdTech Industry: Challenges and Uses*. eLearning Industry. https://elearningindustry.com/significance-of-computer-vision-in-the-edtech-industry-challenges-and-uses

4. *What is Explainable AI (XAI)? | IBM*. (n.d.). https://www.ibm.com/topics/explainable-ai

5. Raiyn, J. (2016). The Role of Visual Learning in Improving Students' High-Order Thinking Skills. *Journal of Education and Practice*, *7*(24), 115-121.

6. Farhi, F., Jeljeli, R., Aburezeq, I., Dweikat, F. F., Al-shami, S. A., & Slamene, R. (2023). Analyzing the students' views, concerns, and perceived ethics about chat GPT usage. *Computers and Education: Artificial Intelligence*, 100180.

7. Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074.

8. Will Cukierski. (2013). Dogs vs. Cats. Kaggle. https://kaggle.com/competitions/dogs-vs-cats

9. Team, K. (n.d.). *Keras documentation: Grad-CAM class activation visualization*. https://keras.io/examples/vision/grad_cam/