Applying Explainable AI to AlexNet Models Performing Trash Classification

Samara Goltz Bryn Archambault

Abstract

With the complexity of models, we aim to use explainable AI to provide reasoning in model building. Comparisons between output images will be made to identify what the model is focusing on, and how different elements of the model impact this decision. With environmental issues becoming more crucial, we analyzed a trash classification model because of its environmental impacts. Properly sorting trash is a simple yet impactful thing everyone can do, so we wanted to assess this model. Through many ablation experiments with the addition of Grad-CAM, we found that the model with the best focus and intensity on objects was the final model He et al. chose. Removing CNN layers led to higher accuracy; however, our Grad-CAM images indicate the model looks at insignificant parts of the image. When dropout is added, normalization is removed, and ReLU activation is removed, the accuracy decreases and the model struggles to focus.

1. Introduction

Our project focuses on the 2020 paper, Trash Classification using Convolutional Neural Networks [1]. This is a continuation of the paper, Classification of Trash for Recyclability Status, written in 2016 by a group who recognized the growing waste problem [2]. They wanted to use CNN to sort trash in order to combat some of trash's negative environmental impacts. This puts this paper in an important frame of reference as topics like the environment and global warming have become more prevalent. It is estimated that 25% of items recycled are not actually recyclable or contaminated. This can cause the entire collection to be thrown in the trash, lower quality recyclables, or higher processing costs [3].

He's group improved Yang's AlexNet model and also experimented with ResNet50 and VGG architectures. AlexNet is normally 8 layers deep (5 CNN layers and 3 fully connected layers). He's group optimized the AlexNet architecture by removing 2 convolutional layers. The final model architecture is seen in Figure 1.

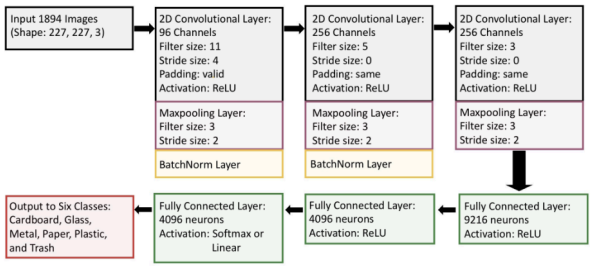

Figure 1. Baseline Model Architecture.

Yang and Thung compiled a dataset, TrashNet, which contains 2,527 images (See Figure 2) that are classified into 6 categories (cardboard, glass, metal, paper, plastic, and trash). The dataset can be found on Thung's GitHub, https://github.com/garythung/trashnet. They trained a support vector machines (SVM) model with scale-invariant feature transform features and a convolutional neural network

(CNN) model with AlexNet architecture. Due to suboptimal hyperparameters in the CNN model, they performed better in their SVM model (SVM accuracy: 63% compared to CNN accuracy: 22%) [2].

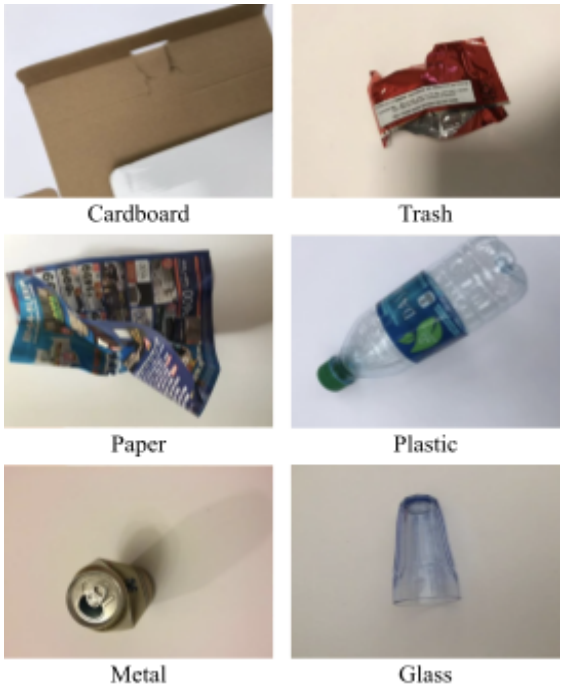

Figure 2. A sample image from each of the 6 classifications.

Our main interest in He's AlexNet model was to see how it performed through the lens of Explainable AI. Explainable AI (XAI) is a collection of methods and tools designed to make machine learning models more transparent and interpretable. Unlike traditional AI systems, which often operate as "black boxes" by providing predictions without clarity on how those conclusions were reached, XAI offers insights into the decision-making processes within these models. This transparency is achieved through techniques that analyze and visualize what the model is focusing on during its predictions, making it easier to understand and trust its outputs.

One of the primary benefits of XAI is the trust it fosters in AI systems, particularly in critical areas like finance, healthcare, and safety. For example, in healthcare, understanding why a model makes a specific diagnosis enhances its reliability for both practitioners and patients. In environmental applications, such as automated waste sorting, XAI helps stakeholders ensure the model is accurately classifying materials, which can significantly impact sustainability efforts. These insights make AI systems more accountable and reliable for practical implementation.

In summary, XAI provides a framework for understanding and refining machine learning models, enabling their responsible application across various domains. By making the internal mechanisms of AI systems more accessible, XAI promotes trust, accountability, and effectiveness, ensuring that these tools are not only powerful but also reliable and ethical [4].

2. Methodology

While there are many different models for explainable AI, we limited our options to those that are specialized in image classification. Models such as saliency maps and Grad-CAM provide a more visual and digestible representation of inputs that most influence the model's decision. In the end, we decided to go with Grad-CAM because of its more engaging appearance and its stronger compatibility with CNN models. While saliency maps can often include noise, Grad-CAM tends to provide a more clear image of influence [5]. Grad-CAM is

also class discriminative, allowing users to understand why the AlexNet model chose a certain class of trash over another for a particular image. We also make the decision to not explore Grad-CAM++. While Grad-CAM++ might have more refined weight calculations, it is more useful in the case where multiple classes are present in one input image. Since the TrashNet dataset consists of single images of trash, Grad-CAM++ would not be of much use.

There is use in applying Explainable AI to Trash classification for multiple reasons. First, since there is a growing concern over trash contamination or the overall influx of pollution, there is a coinciding increase in companies producing ways to combat these issues. One way is through smart garbage or trash-sorting apps. These models analyze photos of trash and determine where the trash should be placed (trash, recycling, compost, etc). With these machines offering a sound resolution to trash contamination, it is important to ensure that these models are effective and can identify key details of glass, plastic, etc. Explainable AI assists in this improvement by showcasing where the model wants to decide. If the model incorrectly classifies objects, we can see where in the photos the model is focusing on and make tweaks as necessary. Similarly, if the model has a high accuracy of classification, we can analyze what aspects of different objects are most important to the model.

While He provides similar trash classification using VGG and ResNet, we are focusing on a vertical comparison with AlexNet. While we hoped to make additional horizontal comparisons between the three models, the time and space complexity of running VGG and ResNet were vastly inefficient, with expected runtimes of 16 and 28 hours respectively. Perhaps this will be a point of further study or included in a proposed Milestone 2.

In terms of vertical comparison, we chose to look at the effect of multiple factors within He's fine-tuned AlexNet model. He noted that removing dropout improved the model's accuracy. We wanted to see whether adding the dropout back in would affect the model's focus with different images. Furthermore, we also experimented with the impact of Normalization and ReLU on the model's performance. Lastly, He noted that the AlexNet model's five CNN layers were unnecessary and removed two unimportant layers. We decided to test this further by removing two more CNN layers, leaving only one. While we also intended to explore the impact of max pooling on Grad-CAM, the process required far too much memory, seeing as the max pooling operation retains only the maximum activations within each pooling window, potentially discarding critical gradient information and requiring additional computational overhead for backpropagation. We selected the first five images from each class in the TrashNet dataset for time, space, and consistency. Upon applying these minor tweaks, we looked at the heatmaps of selected images and the impact of the model's accuracy and loss. We explored whether there was a relationship between variations in accuracy and loss and the corresponding heatmaps.

3. Results

3.1 Baseline Model

Our experiments yielded a series of insights into the role of specific architectural components in our trash classification model. The baseline model achieved an accuracy of 0.8709 and a loss of 0.3401. Grad-CAM visualizations for this model demonstrated well-defined attention heatmaps, showing the model's ability to focus on relevant regions of the input images for classification.

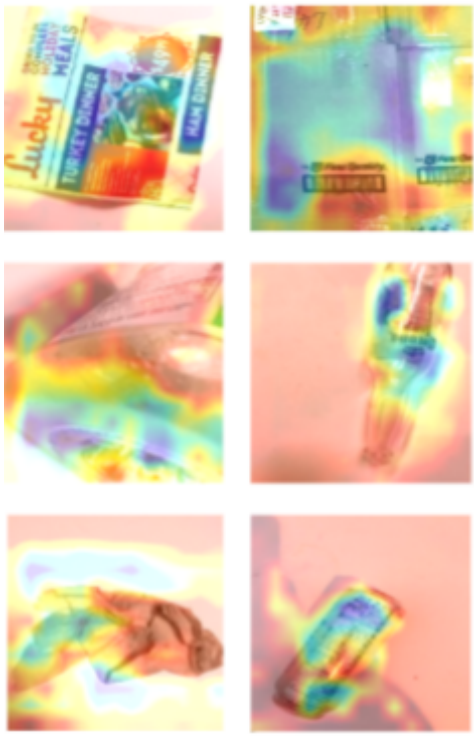

Figure 3. Grad-CAM for the baseline model.

3.2 Adding Dropout

Introducing dropout led to a marked decline in performance, with accuracy dropping to 0.6147 and loss increasing to 0.9988. Grad-CAM visualizations showed significantly less focused heatmaps. This can be attributed to the inherent randomness of dropout, which disables a subset of neurons during training to prevent overfitting. While dropout can help generalize the model in cases of overfitting, it appears to have hindered our model's ability to consistently extract meaningful patterns, likely due to image neurons being highly correlated. The model learns better when it has information from the surrounding neurons to identify more complex relationships. Additionally, the TrashNet dataset is relatively small. With fewer data points, the model struggled to recover the lost information caused by deactivating neurons, leading to poorer results and less cohesive attention.

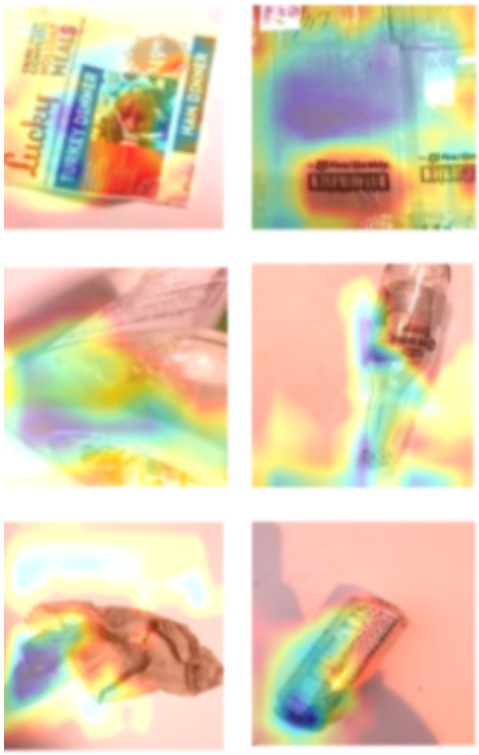

Figure 4. Grad-CAM results when dropout is returned.

3.3 Removing Normalization

Removing normalization resulted in a slight decline in accuracy to 0.8161 and an increase in loss to 0.4870. Although these metrics were not drastically affected, Grad-CAM visualizations revealed sparse and less coherent attention heatmaps. Normalization ensures that input data and intermediate activations remain on a consistent scale, which stabilizes gradient updates and prevents exploding or vanishing gradients. Without normalization, the model likely struggled to converge effectively, as the distribution of activations became unbalanced. While the performance metrics appear relatively stable, the sparse heatmaps suggest that removing normalization hindered the model's ability to form clear and focused attention, resulting in weaker interpretability.

Figure 5. Grad-CAM results when normalization is removed.

3.4 Removing ReLU

Eliminating the ReLU activation function had the most pronounced effect, with accuracy dropping sharply to 0.6021 and loss increasing to 1.1777. Grad-CAM visualizations for this configuration revealed relatively sparse attention. ReLU introduces nonlinearity to the model, allowing it to learn complex, non-linear feature representations critical for effective classification. Without ReLU, the model essentially functions as a linear classifier, limiting its capacity to differentiate between intricate patterns in the data. This significantly impaired both the model's ability to classify trash categories accurately and its capacity to focus on meaningful regions of the input images.

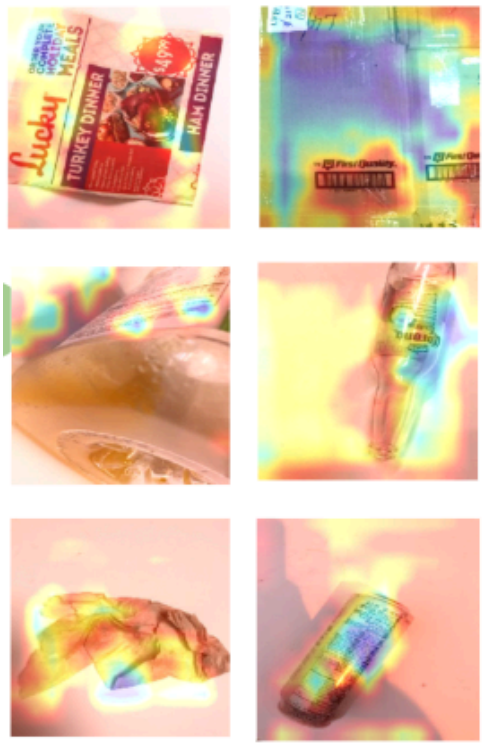

Figure 6. Grad-CAM results when ReLU is removed.

3.5 Removing Two CNN Layers

Further experimentation involved reducing the number of CNN layers from three to one. Surprisingly, this adjustment resulted in an accuracy increase to 0.8920 and a loss reduction to 0.2812. However, Grad-CAM visualizations revealed an almost complete lack of focus or attention. The improved metrics might be explained by the reduction in model complexity, which could mitigate overfitting on our relatively small dataset. However, the single-layer architecture lacks the depth required to capture hierarchical features that are essential for robust classification. The absence of meaningful heatmaps suggests that the model resorted to simpler heuristics or overfit to superficial patterns, rather than learning generalized representations of the input.

Figure 7. Grad-CAM results when two CNN layers are removed.

3.6 Observations

These results underscore the importance of each architectural component in developing a high-performing and interpretable trash classification model. Dropout, while beneficial for regularization, requires careful tuning to balance preventing overfitting with preserving the model's ability to identify consistent patterns. Normalization is crucial for stabilizing activations and ensuring effective gradient flow, which contributes to both performance and explainability. ReLU is indispensable for introducing the nonlinearity necessary for learning complex features, and its absence significantly diminishes both accuracy and interpretability. Finally, the depth of the network plays a critical role in capturing hierarchical and abstract features; reducing the number of CNN layers may improve performance metrics in some cases but at the cost of attention and explainability, as evidenced by Grad-CAM visualizations. Each of these architectural elements, when combined, contributes to the overall performance and interpretability of the model, demonstrating the delicate balance required to build an effective trash classification system

4. Conclusion

Overall, we found that the initial AlexNet model implemented by He was not only the most effective when looking at strictly accuracy and loss, but also when applied to Grad-CAM. Compared to the other variations of AlexNet that we imposed, we found that He's model showed the most competency when classifying images. The model looked at the most

logical areas of an object to make its decision and did not fall victim to diffused attention, or many areas of the image instead of focusing on specific, relevant regions, as well as sparse attention, when the Grad-CAM fails to focus on any features.

5. Challenges and Further Research

Many of the challenges we faced during this research resulted from a shortage of both time and space. Our initial research proposal included a horizontal comparison of Grad-CAM on AlexNet, ResNet, and VGG. However, testing on ResNet and VGG was exponentially longer than AlexNet and more labor-intensive than our computers could manage. Similar reasons also prevented us from exploring how Max Pooling affects the Grad-CAM.

For Milestone Two, we would like to continue further with a more extensive vertical comparison on AlexNet, preferably with Max Pooling, additional CNN layers, etc. Furthermore, we would like to carry out horizontal comparisons to see whether there is a vast difference between Grad-CAM implementation on ResNet and VGG. We would also like to implement other forms of explainable AI such as saliency maps. Doing so could provide a deeper understanding and insight into how CNN models carry out image classification.

6. Resources

[1] He, Yujie, et al. *Trash Classification Using Convolutional Neural Networks*, cs230.stanford.edu/projects_spring_2020/reports/38847029.pdf.

[2] Yang, Mindy, and Gary Thung. Classification of Trash for Recyclability Status, cs229.stanford.edu/proj2016/report/ThungYang-ClassificationOfTrashForRecyclabilityStatus-report.pdf.

[3] Planet, Heal The. “#25 Recycling Consciously.” *HEAL THE PLANET*, 5 Mar. 2024, healtheplanet.com/100-ways-to-heal-the-planet/recycling-consciously. Accessed 11 Dec. 2024.

[4] Neha Vishwakarma. “A Guide to Grad-CAM in Deep Learning.” https://www.analyticsvidhya.com/blog/2023/12/grad-cam-in-deep-learning/

[5] Molnar, Christoph. “10.2 Pixel Attribution (Saliency Maps) | Interpretable Machine Learning.” Github.io, July 31, 2024. https://christophm.github.io/interpretable-ml-book/pixel-attribution.html#grad-cam.