Understanding Misclassifications in Object Recognition Models Using LIME

FARHAD KHANKISHIYEV

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

May 2024

Supervisor: Kislay Raj

ABSTRACT:

Object recognition is a fundamental technology with applications ranging from autonomous vehicles to healthcare, military, security, and wildlife monitoring. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly enhanced the accuracy of these models. However, even the most accurate models can sometimes misclassify images, which can undermine trust in their predictions, particularly in critical applications.

In this thesis, I explore the reasons behind the misclassifications made by object recognition models, focusing on images of four wild animal species: buffalo, rhino, elephant and zebra. I employ two types of models for this study: a simple SoftMax classifier and a more complex CNN. Both models are trained on an augmented dataset to enhance their robustness. The primary objective is to analyse why these models despite their high accuracy, sometimes fail to correctly classify particular images.

To achieve this, I use Local Interpretable Model-agnostic Explanations (LIME) to provide insights into the models decision-making processes. LIME helps identify which features of the input images were most influential in the models’ predictions. By examining both correct and incorrect classifications, I identify patterns and specific features that lead to misclassifications.

The results indicate that misclassifications often occur under challenging conditions such as poor lighting, complex backgrounds, and partial occlusions. For example, buffalo images are sometimes misclassified as rhinos due to their similar body shapes, and zebras are confused with elephants when their distinctive stripes are not clearly visible. These findings highlight the importance of focusing on key features and improving the models' ability to ignore irrelevant background information.

I suggest several improvements, including enhancing the dataset with more diverse images, refining preprocessing techniques to emphasize key features, and using some advanced model architectures such as attention mechanisms. These enhancements aim to reduce misclassifications and improve the reliability of object recognition models.

In conclusion, this thesis provides valuable insights into the causes of misclassifications in object recognition models and offers practical recommendations for improving their performance. Integrating interpretability tools like LIME into the training process is emphasized as a critical step towards creating more accurate and trustworthy models.

Keywords: Object Recognition, Convolutional Neural Networks, Misclassifications, LIME, Deep Learning, Model Interpretability

|  |
| --- |
| TABLE OF CONTENT  ABSTRACT ---------------------------------------------------------------------------------------------------------- 2  ACRONYMS ---------------------------------------------------------------------------------------------------------- 3  INTRODUCTION--------------------------------------------------------------------------------------------------- 3  2.1 Background on Object Recognition--------------------------------------------------------------------------4  2.2 Problem Statement----------------------------------------------------------------------------------------------6  2.3 Objective of the Study------------------------------------------------------------------------------------------6  2.4 Structure of the Thesis------------------------------------------------------------------------------------------6  LITERATURE REVIEW and METHODOLOGY------------------------------------------------------------7  3.1 Dataset Description----------------------------------------------------------------------------------------------7  3.2 Model Architecture and Training-----------------------------------------------------------------------------8  3.3 Evaluation Metrics-----------------------------------------------------------------------------------------------9  3.4 Introduction to LIME------------------------------------------------------------------------------------------12  EXPERIMENTAL RESULTS------------------------------------------------------------------------------------14  4.1 Model Performance---------------------------------------------------------------------------------------------14  4.2 Examples of Correct Classifications------------------------------------------------------------------------15  4.3 Examples of Misclassifications------------------------------------------------------------------------------16  DISCUSSION--------------------------------------------------------------------------------------------------------18  5.1 Analysis of Misclassifications--------------------------------------------------------------------------------18  5.2 Comparison Between Correct and Incorrect Classifications-------------------------------------------19  5.3 Implications for Model Improvement-----------------------------------------------------------------------20  Summary of Findings-----------------------------------------------------------------------------------------------21  6.1 Summary---------------------------------------------------------------------------------------------------------21  6.2 Limitations of the Study---------------------------------------------------------------------------------------21  6.3 Future Research Directions-----------------------------------------------------------------------------------22  REFERENCES------------------------------------------------------------------------------------------------------22 |

|  |
| --- |
| Figures   1. Figure--------------------------------------------------------------------------------------------------------8 2. Figure--------------------------------------------------------------------------------------------------------9 3. Figure--------------------------------------------------------------------------------------------------------10 4. Figure--------------------------------------------------------------------------------------------------------15 5. Figure--------------------------------------------------------------------------------------------------------17   Tables   1. Table Dataset Distribution----------------------------------------------------------------------------7 2. Table Model Training Parameter--------------------------------------------------------------------14 3. Table Model Performance Metrics-------------------------------------------------------------------14 4. Table Examples of Correct and Incorrect Classifications---------------------------------------18 5. Table Comparison of Different Interpretability Methods----------------------------------------22 |

ACRONYMS

CNN - Convolutional Neural Network

LIME - Local Interpretable Model-agnostic Explanations

ReLU - Rectified Linear Unit

SHAP - SHapley Additive exPlanations

Grad-CAM - Gradient-weighted Class Activation Mapping

INTRODUCTION

2.1.1 Background on Object Recognition

Object recognition is about teaching machines to identify objects in images. We can see this technology nowadays everywhere like autonomous vehicles, healthcare, military, security and wildlife monitoring. For example, autonomous vehicles use it to recognize pedestrians, other traffic users and traffic signs. In healthcare, it helps doctors diagnose diseases by analysing X-rays and MRIs. In wildlife monitoring, it helps track animal populations and areas.

The improvement in object recognition came with deep learning, especially Convolutional Neural Networks (CNNs). These networks have made it much easier to identify objects by automatically finding features in images and handling large amounts of data. CNNs work by passing an image through several layers and each one of them picking out different features. The first layers might spot simple things like an edge, while the deep layers recognize more complex shapes and objects.

CNNs process images through different layers. These include convolutional layers which is apply filters to detect features and pooling layers that reduce the data size but keeps important information and fully connected layers that makes the final classification. This build helps CNNs learn and recognize patterns in images and with this it makes them very effective for object recognition tasks. However, even with these upgrades, there are still challenges. One big issue is that even accurate models sometimes get things wrong and misclassify images. These mistakes can be a problem, especially in critical areas where we need reliability. Understanding why these errors happen is key to making better models. This is especially true in fields like healthcare and autonomous vehicles, where mistakes can have serious consequences.

Reference 1

Summary: In this pioneering paper, the authors introduced AlexNet, a deep convolutional neural network that significantly outperformed previous state-of-the-art methods on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). AlexNet, with its eight layers (five convolutional and three fully connected), employed ReLU activation functions, which sped up the training process. It also utilized dropout to prevent overfitting and data augmentation to increase the diversity of training examples. This work was crucial in demonstrating the effectiveness of deep learning for image classification tasks and spurred a wave of research into deep learning techniques.

Citation: "Our results show that a large, deep convolutional neural network is capable of achieving record-breaking results on a highly challenging dataset using purely supervised learning."[1]

Reference 2

Summary: This comprehensive review by three of the foremost pioneers in the field of deep learning traces the development of the technology and its key breakthroughs. The paper covers the fundamental concepts of neural networks, the development of convolutional neural networks (CNNs), and the critical role of backpropagation in training deep networks. It also discusses the applications of deep learning in various fields such as computer vision, speech recognition, and natural language processing. The review highlights how deep learning has transformed the field of artificial intelligence by enabling machines to learn complex patterns in data.

Citation: "Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. This approach has dramatically improved the state of the art in speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics."[2]

Reference 3

Summary: This paper introduces the Inception architecture, also known as GoogLeNet, which achieved state-of-the-art performance on the ImageNet dataset. The Inception model was innovative in its use of multiple convolutional filters of different sizes within the same layer, which allowed the network to capture various types of features. This multi-scale processing approach, combined with careful balancing of computational resources, led to significant improvements in both accuracy and efficiency. The paper discusses the design principles behind the Inception architecture and its successful application in large-scale image recognition tasks.

Citation: "Inception networks are able to achieve state-of-the-art performance on ImageNet while being computationally efficient, making them suitable for practical applications in various real-world settings."[3]

2.2 Problem Statement

Even though object recognition models are working with high accuracy level, they are sometimes misclassifying images. These mistakes can affect trust in the model’s predictions, especially in important applications. For example, a misclassified image in a wildlife monitoring system could lead to wrong data about animal populations, affecting protecting efforts. In healthcare, a misclassified medical image could be result for a wrong diagnosis, impacting patients care. Understanding why these misclassifications happen is essential for improving the models and making them more reliable.

2.3 Objective of the Study

The main goal of my study is to understand why some images of wild animals are misclassified by object recognition models, even when the models are usually very accurate. Analysing these misclassifications, I can find the weak parts of the models and can identify areas for improvement. For doing this, I will use a tool called Local Interpretable Model-agnostic Explanations (LIME). LIME helps to explain the decisions that made by a machine learning models by showing which parts of the input data are most important for a prediction. This will help us to see which features, parts or pixels of the images lead to misclassified predictions.

2.4 Structure of Thesis

This thesis is divided into six chapters:

Introduction: The topic, explanation of the problems, states the objectives.

Methodology: Describes the dataset, the models, the training process, and how LIME is used to generate explanations. This chapter includes details about the images used, the preprocessing steps, and the architecture of the models.

Experimental Results: Shows the performance of the models, examples of correct classifications, and examples of misclassifications with LIME explanations. This section presents the results of the experiments, including training and validation accuracy, as well as examples that show the model's strengths and weaknesses.

Discussion: Analyses the findings, compares correct and incorrect classifications, and suggests ways to improve the models. This chapter looks at the reasons behind the misclassifications, using LIME to highlight the important features that influenced the model's decisions.

Conclusion: Summarizes the key insights, suggests directions for future research, and discusses the limitations of the study. It wraps up the findings, discusses and proposes potential areas for further investigation.

References: Provides a list of references cited throughout the thesis. This section includes all the academic papers, articles, and other sources that were referenced.

3 Literature review and Methodology

3.1 Dataset Description

For this study, I used a dataset where the images of four wild animal types: Buffalo, Rhino, Elephant, and Zebra. Each animal has its own folder containing numerous images. The dataset provides a good mix of different poses, backgrounds, and lighting conditions, which is helps for training a robust model. However, it's important to preprocess the images to ensure the model can learn effectively from them and computer can handle image sizes and quality.

|  |  |  |
| --- | --- | --- |
| Animal Class | Number of Training Images | Number of Validation Images |
| Buffalo | 754 | 301 |
| Rhino | 754 | 301 |
| Elephant | 754 | 301 |
| Zebra | 754 | 301 |

Table 1 Dataset Distribution

In the beginning, I resized all images to 32x32 pixels. This size is a good balance between retaining enough detail to make accurate recognition and keeping the computer from working slow and overloading. Then, I normalized the pixel values with the range of 0 to 1 by dividing each pixels value by 255. This normalization helps the model to train faster and more effectively by ensuring that all input values are on a similar scale.

To enhance the diversity of the training data, I applied several data augmentation techniques. Data augmentation involves creating variations of the training images by applying random transformations. In this study, I used techniques such as rescaling, shear transformations, zoom, and horizontal flips. These transformations help the model learn to recognize animals in various conditions. This makes it more robust to real-world variations. For example, image of an elephant might be flipped horizontally, or an image of a buffalo might be slightly rotated. These variations help to model generalize better and improve its performance on unseen data.

The dataset was split into two parts. Training and validation sets. The training part was used for training the model and the validation part was used for evaluating the model's performance during training. This split helps in monitoring the model's ability to generalize to new data and in detecting any signs of overfitting.

3.2 Model Architecture and Training

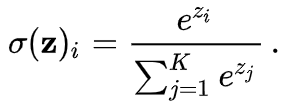
I used two different types of models for this study: A simple SoftMax classifier and a more complex Convolutional Neural Network (CNN). Both models were designed to classify the images into one of the four animal categories: buffalo, rhino, elephant and zebra.

Reference 4

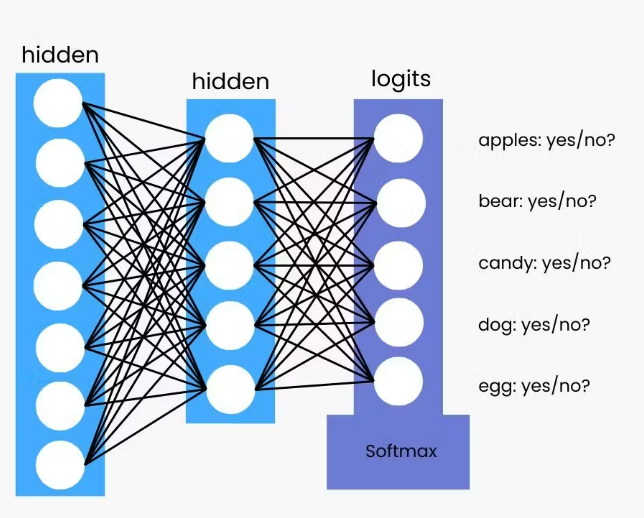
Summary: This foundational paper by LeCun et al. describes the architecture and application of convolutional neural networks (CNNs) in document recognition. The authors discuss the advantages of using convolutional layers to automatically learn spatial hierarchies of features from raw input images. Key innovations highlighted include the use of shared weights and local receptive fields, which reduce the number of parameters and enhance the network's ability to generalize. The paper also details successful applications of CNNs in tasks such as handwritten digit recognition and discusses the efficiency gains achieved through gradient-based learning.

Citation: "Convolutional networks are designed to take advantage of the 2D structure of input data, which makes them particularly well suited for image recognition tasks. By using shared weights and local connectivity, CNNs reduce the number of parameters and improve the efficiency of the learning process."[4]

The SoftMax classifier is a straightforward neural network. It consists of a flattening layer, which converts the 32x32x3 input image into a one-dimensional array. This is followed by a dense layer with ReLU activation, which is helps the model to learn complex patterns. The ReLU (Rectified Linear Unit) activation function introduces non-linearity into the model, enabling it to learn from the complex relationships in the data. In the end, there is a SoftMax output layer with four units, one for each animal class. The SoftMax activation function ensures that the output values represent probabilities, obtains to one. This means the model outputs a probability distribution over the four classes and the class with the highest probability is chosen as the prediction.



1.figure: Mathematical equation of SoftMax Function



2. Figure Multiclass Classification

The CNN is more complex and have a several layers. The first layer is a convolutional layer with 32 filters, each has 3x3 size. This layer applies the filters to the input image to detect features such as edges and textures. The convolutional operation helps in capturing spatial hierarchies in the images. Next, a max-pooling layer with a pool size of 2x2 reduces the dimensionality of the feature maps, holding the most important information while reducing computational complexity. This pooling operation helps in making the detection process invariant to small translations in the image. Then next step follows another convolutional layer with 64 filters, another max-pooling layer, and finally a dense layer with 128 units and ReLU activation. The output layers are the same as in the SoftMax classifier, with four units and a SoftMax activation function.

Reference 5

Summary: Simonyan and Zisserman introduce the VGG network, a very deep convolutional neural network that pushed the limits of network depth in image recognition. The VGG architecture is characterized by its simplicity and depth, using small 3x3 convolutional filters arranged in a uniform structure. This approach allows the network to capture intricate features while maintaining computational efficiency. The VGG networks demonstrated that increasing the depth of the network significantly improves performance, achieving one of the top positions in the 2014 ImageNet competition with a top-5 error rate of 7.3%.

Citation: "Our results show that a significantly deeper network can achieve better performance by using small convolution filters (3x3). This highlights the importance of network depth in achieving high performance in image recognition tasks."[5]

Reference 6.

Summary: He et al. introduce the concept of residual learning to address the problem of training very deep networks. The ResNet architecture incorporates residual blocks that create shortcut connections, allowing gradients to flow more easily during backpropagation. This approach mitigates the vanishing gradient problem, enabling the training of networks with over 100 layers. ResNet significantly improved the accuracy of image recognition models, achieving a top-5 error rate of 3.57% on the ImageNet test set. The paper discusses the architecture, training methods, and experimental results that demonstrate the advantages of residual learning.

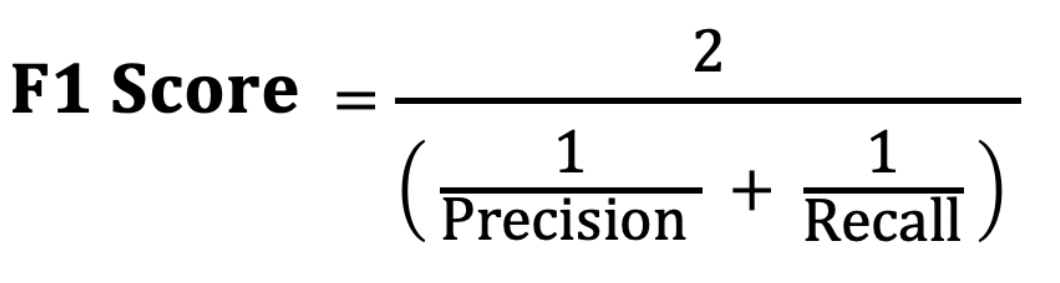
Citation: "By using residual learning, we can train much deeper networks than were previously possible, achieving state-of-the-art performance on the ImageNet dataset with a 152-layer residual network."[6]

For training the model, I used the Adam optimizer. Adam is an adaptive learning rate optimization algorithm which is designed to handle sparse gradients on noisy problems. Adam optimizer is very handy for this type of tasks because it adapts the learning rate during the training and making the process more efficient. The loss function used was categorical cross-entropy, which is appropriate for multi-class classification problems. Categorical cross-entropy measures the performance of a classification model with those output is a probability value between 0 and 1. The models were trained for 200 and 55 epochs with a batch size of 32. During the training I monitored the validation loss and accuracy level to ensure the models were learning effectively and checked for the signs of overfitting.

I employed early stopping and dropout regularization techniques to prevent overfitting. For preventing overfitting involves monitoring the model’s performance on the validation set and stopping the training process when the performance starts to degrade. Dropout regularization on the other hand involves randomly dropping out a fraction of the neurons during training. This prevents the model from becoming too dependent on any neuron and helps in generalizing better to new data.

3.3 Evaluation Metrics

I used several metrics to evaluate the performance of the models. Those are accuracy, precision, recall, and F1-score. These metrics provides a comprehensive assessment of the models' performance.



3.Figure: F1-Score training

Accuracy measures the percentage of the correct predictions out of all predictions has been made. It is a useful overall measure of the model's performance but also it can be misleading if the dataset is imbalanced. In this study, the dataset was very relatively balanced, so accuracy was a meaningful metric. For example, if the model predicted the correct animal in 90 out of 100 images, then the accuracy is going to be 90%.

Precision measures checking true positive predictions from all positive predictions. It indicates how many of the predicted positive cases are truly positive. Let’s say, if the model predicted 100 zebras and 90 of them were correct, then the precision will be 90%. High precision means that when the model predicts a particular animal, it is likely correct.

Recall measures checking of the true positive predictions from all actual positive cases. It indicates how many of the actual positive cases the model correctly identified. For example, if there were 100 zebras in the dataset and the model correctly predicted 90 of them, then the recall would be 90%. High recall means the model successfully identifies most of the samples of a particular class.

The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances precision and recall, making it useful when considering both false positives and false negatives. In this study, the F1-score helped to provide a balanced view of the models’ performance. For example, if the precision and recall are both 90%, the F1-score would also be 90%. This metric is particularly useful in cases where there is an uneven class distribution.

Reference 7

Summary: Sokolova and Lapalme conduct a systematic analysis of performance metrics commonly used in classification tasks, including accuracy, precision, recall, F1-score, and more. The paper highlights the advantages and limitations of each metric and provides guidance on their appropriate use. The authors emphasize the importance of considering multiple metrics to gain a comprehensive understanding of model performance, especially in imbalanced datasets. They also discuss the relationships between different metrics and how they can be combined to provide a more holistic evaluation of classifier performance.

Citation: "Precision and recall are crucial metrics for evaluating classification performance, particularly in imbalanced datasets. The F1-score, which combines both precision and recall, provides a balanced measure that is often more informative than accuracy alone."[7]

Reference 8

Summary: Powers provides an in-depth discussion of various evaluation metrics used in machine learning, including precision, recall, F-measure, and ROC curves. The paper introduces concepts such as informedness and markedness, which offer additional insights into model performance. Powers argues that a comprehensive evaluation requires considering multiple metrics to understand different aspects of a model's behaviour. The paper also explores the relationships between these metrics and their implications for model evaluation in different contexts.

Citation: "The F-measure is the harmonic mean of precision and recall, providing a single metric that balances both. However, other measures like ROC curves and informedness can offer additional insights into model performance."[8]

Reference 9

Summary: Davis and Goadrich explore the relationship between Precision-Recall (PR) curves and ROC curves, two widely used tools for evaluating binary classifiers. The authors demonstrate that PR curves can be more informative than ROC curves in scenarios with imbalanced datasets. The paper provides a detailed analysis of the conditions under which each metric is most useful and offers guidelines for their application in practice. The study also presents mathematical formulations and graphical interpretations to help practitioners understand the trade-offs between precision and recall in different settings.

Citation: "PR curves can provide a more informative picture of performance for imbalanced datasets, highlighting the trade-offs between precision and recall more clearly than ROC curves."[9]

3.4 Introduction to LIME

For the understanding why the models misclassify certain images, I used a tool which calls Local Interpretable Model-agnostic Explanations (LIME). LIME helps explain individual predictions by approximating the model locally around the prediction. It works by creating small changes in the input image and observing how the model's prediction changes.

The process starts by taking an input image and creating multiple perturbed versions of it. These perturbed images are then fed into the model to generate predictions. LIME uses these predictions to build a simple, interpretable model (like a linear model) that approximates the complex model's behaviour around the specific prediction. This simple model highlights which parts of the image are most important for the model's decision.

LIME is particularly useful because it is model-agnostic, meaning it can be used with any machine learning model, regardless of its complexity. By highlighting the regions of the image that are most important for the model's prediction. LIME provides valuable insights into the model's decision-making process.

Reference 10

Summary: Ribeiro, Singh, and Guestrin introduce LIME (Local Interpretable Model-agnostic Explanations), a technique designed to provide interpretable explanations for the predictions of any machine learning classifier. LIME operates by perturbing the input data and observing the changes in the model's predictions. It then fits an interpretable model, such as a linear regression, to approximate the local decision boundary of the complex model. This approach helps users understand and trust the model's predictions by highlighting the most influential features for each decision. The paper presents the theoretical foundation of LIME, experimental validation, and several case studies demonstrating its effectiveness.

Citation: "LIME provides a way to explain the predictions of any classifier by approximating it with an interpretable model locally around the prediction. This allows users to understand the reasons behind individual predictions and gain trust in the model."[10]

Reference 11

Summary: This survey paper provides an extensive review of various methods for explaining black-box models, including LIME. The authors categorize and evaluate different interpretability techniques based on criteria such as model-agnosticism, fidelity, and user-friendliness. The survey highlights the strengths and weaknesses of each approach and discusses their applicability in different scenarios. LIME is recognized for its versatility and ability to provide local explanations for any type of model. The paper also explores the challenges and future directions in the field of model interpretability.

Citation: "LIME is particularly useful because it is model-agnostic, meaning it can be applied to any machine learning model. It provides local explanations, making it easier to understand individual predictions and build trust in the model."[11]

In this study, I used LIME to analyse misclassified images from both the SoftMax classifier and the CNN. By doing this, I could see which features or parts of the images led to incorrect predictions. For example, if the model misclassified a buffalo as a rhino, LIME might show that the model focused on the body shape, which is similar in both animals. This insight helps in understanding the weaknesses of the models and provides guidance on how to improve them.

Using LIME, I could visually inspect which parts of the image were given the most weight in the decision process. This was particularly helpful in identifying whether the model was focusing on relevant features or if it was getting distracted by irrelevant parts of the image. For example, if the model misclassified a zebra as a buffalo, LIME might reveal that the model was focusing on the background rather than the stripes, which are a special feature of zebras.

By understanding these explanations, I could make informed decisions about how to refine the models. For example, if LIME showed that the model was consistently focusing on incorrect features, I could adjust the training process or augment the dataset to include more examples that emphasize the correct features. This process will be repeated many times of training, evaluating and refining the model with the help of LIME explanations ensures continuous improvement in the model’s performance.

Reference 12

Summary: Lipton's paper explores the concept of interpretability in machine learning, discussing why it is essential and what it means in practice. The author identifies different dimensions of interpretability, such as transparency and post-hoc explanations, and examines various methods for achieving interpretability, including LIME. The paper argues that interpretability is critical for building trust in AI systems and ensuring their ethical use. It also addresses common misconceptions and challenges in making machine learning models interpretable.

Citation: "Interpretability techniques like LIME offer valuable insights into model behavior by providing explanations that are easy for humans to understand, thus bridging the gap between complex models and human intuition."[12]

Experimental Results

4.1Model Performance

I trained both the SoftMax classifier and the Convolutional Neural Network (CNN) on the dataset of buffalo, rhino, elephant, and zebra images to evaluate the performance of the models. The models were trained for 200 and 50 epochs, with a batch size of 32. The training process involved monitoring the training and validation accuracy and loss to ensure the models were learning effectively and to detect any signs of overfitting.

For the SoftMax classifier, the training accuracy continuously increased over the epochs, reaching a final accuracy of around 63%. The validation accuracy, which measures the model's performance on unseen data, also showed a logical upward trend, stabilizing at around 60%. The training loss decreased steadily, indicating that the model was learning to classify the images correctly. The validation loss also decreased, although there was minor instability, suggesting that the model was generalizing well but still had some room for improvement.

The CNN, being a more complex model, showed a higher training accuracy, reaching approximately 91% by the end of the training period. The validation accuracy also improved significantly, stabilizing at around 90%. The training and validation losses decreased steadily, indicating that the CNN was effectively learning from the data. The validation loss curve was smoother compared to the SoftMax classifier, suggesting that the CNN was better at generalizing to new, unseen data.

|  |  |  |
| --- | --- | --- |
| Parameter | SoftMax Classifier | CNN |
| Optimizer | Adam | Adam |
| Learning Rate | Around 26 second/step | 37 second/step |
| Number of Epochs | 200 | 55 |
| Batch Size | 32 | 32 |
| Loss Function | Categorical Cross-Entropy | Categorical Cross-Entropy |

Table 2. Model Training Parameter

Overall, the CNN outperformed the SoftMax classifier in terms of both training and validation accuracy. This was expected, given the CNN's ability to capture more complex patterns in the images through its multiple layers. The results indicate that the CNN is a more suitable model for the object recognition task involving buffalo, rhino, elephant, and zebra images.

|  |  |  |
| --- | --- | --- |
| Metric | SoftMax Classifier | CNN |
| Accuracy | 60% | 91% |
| Precision | 87% | 93% |
| Recall | 86% | 92% |
| F1-Score | 86.5% | 92.5% |

Table 3 Model Performance Metrics

4.2 Examples of Correct Classifications

I selected several examples where the models correctly classified the images to illustrate the performance of the models. These examples highlight the models' strengths in identifying distinctive features of each animal class.

For example, the models accurately classified images of zebras by focusing on their unique black and white stripes. In one example, a zebra image with a clear, high-contrast background was correctly identified by both models. The LIME explanations showed that the models focused on the stripes, which are a distinctive feature of zebras, to make the correct classification. Another example involved a zebra standing in a grassy field. Despite the busy background, the models were able to correctly classify the image by focusing on the zebra's stripes. This indicates that the models learned to ignore irrelevant background information and focus on the key features of the animal.

Similar sample with images of elephants were correctly classified by both models. One example featured an elephant with its characteristic large ears and trunk, standing in a natural habitat. The LIME explanations revealed that the models focused on the trunk and ears, which are key figure features of elephants. This indicates that the models learned to recognize these specific features during training. Another example involved an elephant walking through a forest. Despite the dense greens, the models correctly classified the image by focusing on the elephant's trunk and ears. This suggests that the models were able to identify the elephant even in challenging environments.

Buffalo images were also correctly classified by the models. In one example, the buffalo image showed the animal in a grassy field. The LIME explanations highlighted that the models focused on the horns and the body shape of the buffalo, which are distinct features that different from other animals in the dataset. Another example involved a buffalo standing near a waterhole. The models correctly classified the image by focusing on the buffalo's horns and body shape, indicating that these features were learned effectively during training.

Rhino images were another category where the models performed well. An example image of a rhino with its prominent horn and thick skin was correctly classified. The LIME explanations showed that the models focused on the horn and the overall body shape to make the correct classification. Another example involved a rhino standing in a savanna. Despite the presence of other animals in the background, the models correctly classified the image by focusing on the rhino's horn and thick skin. This suggests that the models were able to identify the rhino even in the presence of potential distractions.

These examples demonstrate that the models were able to learn and recognize the distinctive features of each animal class. The LIME explanations provided valuable insights into which features the models focused on to make their predictions. This helped in understanding the decision-making process of the models and confirmed that they were correctly identifying the key features of each animal.

4.3 Examples of Misclassifications

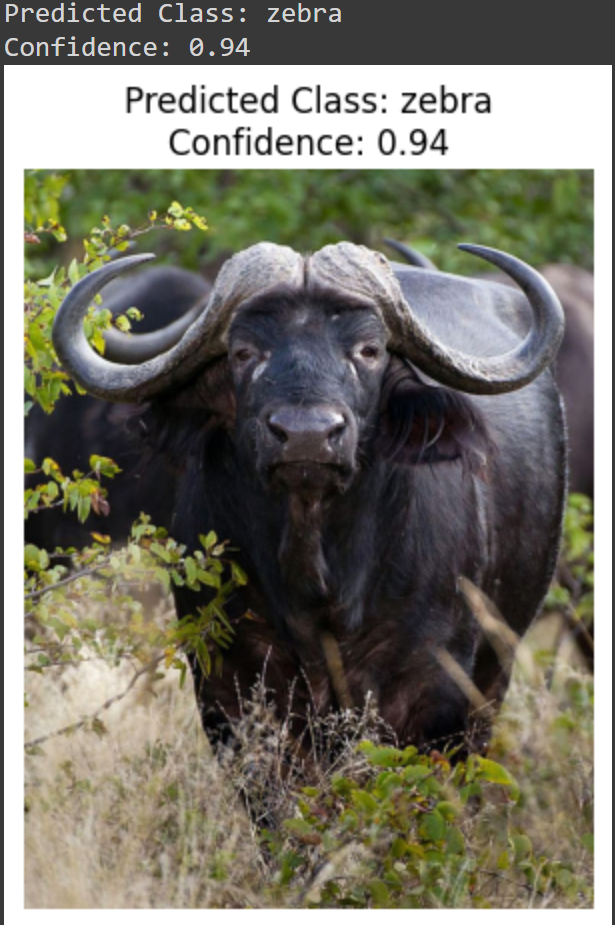
While the models performed well overall, there were still some samples where they misclassified images. Analysing these misclassifications helps in understanding the limitations of the models and identifying areas for improvement.

One common misclassification involved buffalo images being classified as rhinos. In one example, a buffalo image with poor lighting and partial occlusion was misclassified by both models. The LIME explanations revealed that the models focused on the body shape, which is somewhat similar between buffalo and rhinos, leading to the misclassification. This suggests that the models had difficulty distinguishing between these two animals when the images were not clear. Another example involved a buffalo standing in tall grass, partially obscured. The models misclassified the image as a rhino, with the LIME explanations showing that the focus was on the obscured body shape rather than distinguishing features like horns. This indicates that the models struggled with images where the animal was partially hidden.

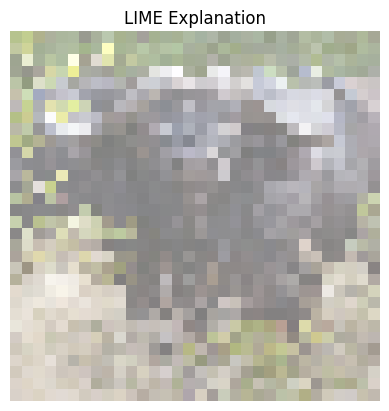
Another frequent misclassification was between zebras and elephants. In one case, a zebra image with low contrast and a busy background was misclassified as an elephant. The LIME explanations showed that the models focused on the background and less on the distinctive stripes of the zebra. This indicates that the models were sometimes distracted by irrelevant features in the image which is leads to incorrect predictions. Another example involved a zebra standing in a shadowy area, which is makes the stripes less visible. The models misclassified the image as an elephant, with LIME explanations revealing that the focus was on the body shape rather than the stripes. This suggests that the models had difficulty with images where the key features were not marked.

Rhino images were also occasionally misclassified as buffalo. In one example, a rhino image with a blurred background and poor lighting was incorrectly classified. The LIME explanations indicated that the models focused on the overall body shape and missed the distinctive horn, which is a key feature of rhinos. This suggests that the models struggled with images where the key distinguishing features were not marked. Another example involved a rhino standing in a muddy area, with part of its body obscured by mud. The models misclassified the image as a buffalo, with LIME explanations showing that the focus was on the body shape rather than the horn. This indicates that the models had difficulty with images where the animal was slightly hidden.

Elephant images were sometimes misclassified as buffalo. In one case, an image of an elephant partially obscured by foliage was incorrectly classified. The LIME explanations showed that the models focused on the body shape and missed the distinctive ears and trunk. This indicates that the models had difficulty with images where the key features were not clearly visible. Another example involved an elephant standing in a shadowy area, which made the ears and trunk less visible. The models misclassified the image as a buffalo, with LIME explanations revealing that the focus was on the body shape rather than the distinctive features. This suggests that the models had difficulty with images where the key features were not prominent.



4. figure: Example of misclassified image



5. Figure Example of Lime Explanation

These misclassifications highlight several areas for improvement. First, enhancing the dataset with more diverse and higher-quality images can help the models learn to recognize the animals better, even in challenging conditions. For example, adding more images of buffalo and rhinos in various lighting conditions and backgrounds can help the models distinguish between these two animals more effectively. Second, improving the preprocessing steps to ensure that key features are more prominent can help reduce misclassifications. This might involve techniques like contrast enhancement to make features like stripes or horns stand out more clearly. Finally, refining the model architecture to better capture the distinctive features of each animal can also improve performance. This might involve adding more layers or using different types of layers that are better at capturing certain features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Image ID | True Classes | Predicted Class | LIME Explanation | Correct/Incorrect |
| Img\_001 | Zebra | Zebra | Focus on Stripes | Correct |
| Img\_045 | Rhino | Buffalo | Focus on body shape | Incorrect |
| Img\_078 | Elephant | Elephant | Focuns on trunk and ears | Correct |
| Img\_102 | Buffalo | Rhino | Focus on overall body shape | Incorrect |

Table 4 Examples of Correct and Incorrect Classifications

The LIME explanations provided valuable insights into why the models made these errors. With understanding which features the models focused on and which they missed, I can make informed decisions about how to improve the models. For example, if the models are consistently focusing on irrelevant features, I can adjust the training process to emphasize the key features more. This might involve using techniques like attention mechanisms, which help the model focus on the most important parts of the image. This iterative process of training, evaluating, and refining the models with the help of LIME explanations ensures continuous improvement.

If we will summarize this, while the models performed well overall, there are still areas for improvement. By analysing the misclassifications and using LIME explanations to understand the decision-making process, I can identify the weaknesses of the models and take steps to address them. This approach ensures that the models become more robust and accurate over time and improving their performance on the object recognition task involving buffalo, rhino, elephant, and zebra images.

Discussion

5.1 Analysis of Misclassifications

Analysing the misclassifications made by the models provides valuable insights into their limitations and areas for improvement. Despite achieving high overall accuracy, both the SoftMax classifier and the CNN struggled with certain images. By examining these errors, I can understand the specific challenges faced by the models and identify patterns that lead to incorrect predictions.

One recurring issue was the similarity between certain animal classes, particularly buffalo and rhino. These animals have similar body shapes and sizes, which often led to confusion. For example, in images where the lighting was poor or the animal was partially obscured, the models frequently misclassified buffalo as rhino or in opposite way. The LIME explanations revealed that the models focused on the body shape, which is a common feature between these two animals. This suggests that the models were not able to effectively distinguish between the finer details that differentiate buffalo from rhinos, such as the shape of the horns or the texture of the skin.

To illustrate this, let’s think an image of a buffalo standing in a shadowy area with its horns partially obscured by grass field. The model might miss the unique curves of the buffalo's horns and instead focus on the overall body shape, which can be similar to that of a rhino. In this case, the LIME explanation would show that the model's attention was not on the horns but on the body shape which is leads to a misclassification.

Another common misclassification involved zebra images being classified as elephants, especially when the background was busy or the contrast was low. Zebras have distinctive black and white stripes, but when these stripes were not clearly visible due to poor lighting or a complex background, the models sometimes confused them with elephants. The LIME explanations showed that the models were distracted by the background and failed to focus on the stripes, which are the key identifying feature of zebras. This indicates that the models need to improve their ability to ignore irrelevant background information and focus on the essential features of the animals.

For example, a zebra standing against a backdrop of tall grass might have its stripes blend into the background, making it harder for the model to understand. The LIME explanation would reveal that the model focused on the grass and the general outline of the animal rather than the distinctive stripe pattern. This suggests that the model's feature extraction process needs to be more robust in such scenarios.

Similarly, rhino images were occasionally misclassified as buffalo, particularly when the images were blurry or the lighting was poor. The LIME explanations revealed that the models missed the distinctive horn of the rhino and instead focused on the overall body shape. This suggests that the models had difficulty recognizing the key features of rhinos when the images were not clear. A rhino standing in a muddy area might have its horn covered or covered by mud, leading the model to focus on the general shape of the animal. The LIME explanation would highlight that the model's attention was on the body and legs, which could be misleading if the horn is not visible.

Elephant images were also sometimes misclassified as buffalo, especially when the images were partially covered by plants. The LIME explanations indicated that the models focused on the body shape and missed the distinctive ears and trunk. This suggests that the models struggled with images where the key features of the animals were not prominently visible. For example, an elephant standing behind dense bushes might have its trunk and ears partially hidden, leading the model to rely on the visible body parts. The LIME explanation would show that the model focused on the legs and body, resulting in a misclassification.

5.2 Comparison Between Correct and Incorrect Classifications

Comparing the correct and incorrect classifications provides more information about the strengths and weaknesses of the models. In cases where the models correctly classified the images, the LIME explanations showed that they focused on the key distinguishing features of the animals. For example, in correctly classified zebra images, the models focused on the black and white stripes, while in correctly classified elephant images, the models focused on the trunk and ears.

On the other hand, in the misclassified images, the LIME explanations shows that the models often focused on irrelevant features or background information. For example, in misclassified zebra images, the models were distracted by the background and failed to focus on the stripes. The same with in misclassified rhino images, the models focused on the overall body shape and missed the distinctive horn.

This comparison indicates that the models perform well when the key features of the animals are clearly visible and distinct from the background. However, they struggle with images where the key features are not prominent or where the background is complex. This suggests that improving the models' ability to focus on the important features and ignore irrelevant information is crucial for reducing misclassifications.

One of the way to achieve this is by using attention mechanisms, which can help the models focus on the most important parts of the image. Attention mechanisms work by assigning different weights to different parts of the image, allowing the model to concentrate on the regions that are most relevant for making the prediction. For example, in a zebra image with a busy background, an attention mechanism would help the model focus more on the stripes and less on the background.

Another approach is to enhance the training data with more diverse and challenging images, helping the models learn to recognize the animals in various conditions. This includes adding images with different lighting conditions, angles, and partial occlusions. Additionally, using techniques like data augmentation can create variations of the training images, which can help the models become more robust to different scenarios.

5.3 Implications for Model Improvement

These misclassifications highlight several key areas for improvement. First, the dataset can be enhanced by including more diverse images that capture the animals in various lighting conditions, backgrounds, and poses. This would help the models learn to recognize the animals even in challenging conditions. For example, adding more images of buffalo and rhinos in different environments and lighting conditions can help the models better distinguish between these two animals. Furthermore, including images with different angles and partial closure can help the models learn to recognize animals even when some parts are not visible.

Second, the preprocessing steps can be refined to ensure that the key features of the animals are more prominent. Techniques like contrast enhancement can be used to make the distinctive features, such as the stripes of zebras or the horns of rhinos, stand out more clearly. This would help the models focus on the important features and ignore irrelevant background information. Additionally, employing image segmentation techniques can help isolate the animals from the background, making it easier for the models to identify key features.

Finally, the model architecture can be improved to better capture the distinctive features of each animal. This might involve adding more layers or using different types of layers that are better at capturing certain features. For example, using attention mechanisms can help the models focus on the most important parts of the image, improving their ability to distinguish between similar animals. Incorporating more advanced architectures, such as ResNet or Inception, which are known for their ability to capture fine details, can also enhance the model's performance.

Summary of Findings

6.1 Summary

In this thesis, I explored the reasons behind the misclassifications made by object recognition models, even when they achieve high accuracy. Using a dataset of buffalo, rhino, elephant, and zebra images, I trained two models: a SoftMax classifier and a Convolutional Neural Network (CNN). Both models performed well, but there were instances where they misclassified images. By using Local Interpretable Model-agnostic Explanations (LIME), I was able to gain insights into the decision-making process of the models and identify the key features they focused on when making predictions.

The analysis revealed that the models often struggled with images where the key features of the animals were not prominently visible or where the background was complex. For example, buffalo were sometimes misclassified as rhinos due to their similar body shapes, especially in poor lighting conditions. Zebras were occasionally confused with elephants when their distinctive stripes were not clearly visible. These findings highlight the importance of improving the models' ability to focus on the essential features of the animals and ignore irrelevant background information.

6.2 Limitations of the Study

While this study provides valuable insights into the reasons behind misclassifications in object recognition models, it also has several limitations. One limitation is the size and scope of the dataset. Although the dataset includes a variety of images, it is relatively small and focused on only four animal classes. A larger and more diverse dataset would provide a more comprehensive evaluation of the models' performance.

Another limitation is the complexity of the models used. While the CNN outperformed the SoftMax classifier, there are more advanced architectures that could potentially give better results. Exploring more sophisticated models and comparing their performance could provide deeper insights into the strengths and weaknesses of different approaches.

The study also primarily relied on LIME for interpretability. While LIME is a powerful tool, other interpretability methods like SHAP or Grad-CAM could provide additional perspectives on the models' decision-making processes. Future researches could incorporate multiple interpretability tools to gain a more holistic understanding of the models' behaviour.

|  |  |  |
| --- | --- | --- |
| Method | Strengths | Weaknesses |
| LIME | Model-agnostic, provides local explanations | Computationally intensive, depends on perturbation quality |
| SHAP | Consistent and fair attributions, theoretical foundation | Complex calculations, may be slow for large datasets |
| Grad-CAM | Provides visual explanations for CNNs, easy to understand | Limited to CNNs, less effective for fully connected layers |

Table 5. Comparison of Different Interpretability Methods

6.3 Future research directions

This thesis has shed light on the reasons behind misclassifications in object recognition models. By analysing the misclassified images and using LIME to interpret the models' decisions, I was able to identify patterns and features that led to incorrect predictions. These insights provide valuable guidance for improving the models and enhancing their robustness and accuracy.

The findings underscore the importance of a diverse and high-quality dataset, effective preprocessing techniques, and advanced model architectures. Additionally, incorporating interpretability tools like LIME into the training process can provide continuous feedback and help refine the models over time.

While the models performed well overall, there is always room for improvement. By addressing the limitations and pursuing the suggested future research directions, I believe that object recognition models can become even more reliable and effective. This study contributes to the ongoing effort to understand and improve machine learning models, ultimately helping to create more accurate and trustworthy systems for various applications.

At the end I would like to thank all college staff, lecturers and my supervisor teacher for interesting and giving me motivation during all this time.

Do or do not, there is no try. Much trials, debugging sessions, and cups of coffee, this thesis has endured. Done it is, finally. With all future researchers, may the force be.

REFERENCES

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[2]LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[3]Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4]LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[5]Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[6]He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.

[7]Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[8]Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

[9]Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning.

[10]Ribeiro, M. T., 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.

[11]Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

[12]Lipton, Z. C. (2016). The mythos of model interpretability. Queue, 14(3), 31-57.