Human Object Interaction Detection

(Dr. Wajahat Hussain, Ibrahim Bin Mansur, Hassan Saqib, M. Haroon Farooq)

Abstract

This project focuses on the development of a Convolutional Neural Network (CNN) for the purpose of detecting human-object interactions in images. The dataset is curated by extracting frames from two distinct videos—one featuring interactions and another without. To enhance the model's robustness, TensorFlow's ImageDataGenerator is employed for comprehensive data augmentation. The dataset is subsequently divided into training and validation sets, laying the groundwork for the CNN architecture, which integrates convolutional, batch normalization, and dense layers. The training process is executed, and the model's performance is evaluated through accuracy and loss metrics. Practical applicability is demonstrated as the trained model successfully predicts human-object interactions in novel images. This code encapsulates the entire project workflow, from initial data extraction to the deployment of the trained model. The project culminates with the preservation of the model for future applications, presenting a thorough solution for human-object interaction detection. The abstract succinctly captures the project's objectives, methodology, and key outcomes, serving as a concise summary for the report.

1. Introduction

The overarching problem is the lack of a unified and efficient solution for automated human-object interaction detection in images.
This project seeks to address these shortcomings by leveraging

1.1. Background and Motivation

In the realm of computer vision and artificial intelligence, the detection and understanding of human-object interactions play a pivotal role in advancing technology's interaction with the physical world. As society moves towards a more automated and intelligent future, the ability to discern human-object interactions from visual data becomes increasingly crucial. This project is motivated by the need to develop a robust solution for precisely detecting and classifying human-object interactions in images, laying the foundation for applications in fields such as robotics, surveillance, and human-computer interaction.

1.2. Research Problem and Objectives

The primary research problem addressed in this work is the accurate classification of human-object interactions in images. Specifically, the project aims to develop a Convolutional Neural Network (CNN) capable of distinguishing between frames depicting interactions and those without. The objectives include the construction of a reliable dataset, the implementation of an effective CNN architecture, and the evaluation of the model's performance in accurately identifying human-object interactions.

1.3. Problem Statement

advancements in deep learning and data augmentation techniques, aiming to provide a more reliable and versatile solution for real-world applications.

1.4. Motivation

The significance of accurate human-object interaction detection extends across various domains, including human-robot collaboration, surveillance systems, and interactive computing. A robust model in this regard holds the potential to enhance safety, efficiency, and interactivity in automated systems. By addressing the identified gaps in current methodologies, this project aspires to contribute to the broader field of computer vision and artificial intelligence, fostering advancements in human-object interaction detection.

2. Methodology

2.1. Proposed Approach, Methods, and Algorithms

The proposed approach centers around the utilization of a Convolutional Neural Network (CNN) to discern human-object interactions in images. CNNs are well-suited for image classification tasks, and our model incorporates convolutional, batch normalization, and dense layers to effectively capture spatial hierarchies and intricate patterns within the data. The architecture is designed to strike a balance between complexity and efficiency, ensuring optimal performance.

2.2. Model Architecture

The proposed model architecture is designed to effectively capture spatial hierarchies and intricate patterns within the images. The CNN comprises multiple layers, including convolutional layers for feature extraction, batch normalization layers for normalization and stabilization, and dense layers for classification. The architecture begins with a convolutional layer with 64 filters, followed by batch normalization and max pooling.

Subsequent layers include additional convolutional blocks with increasing filter counts (128 and 256) to capture increasingly complex features. The model is flattened before passing through dense layers (256, 128, and 64 neurons) and concludes with a sigmoid activation output layer for binary classification.

2.3. Data Collection and Preparation

The dataset is curated by extracting frames from two distinct video sources—one capturing human-object interactions and the other depicting scenarios without interactions. This diverse dataset is pivotal for training a model capable of generalizing well to various real-world scenarios. To further enhance the dataset's diversity and mitigate overfitting, data augmentation techniques, including rotation, shifting, shearing, zooming, and horizontal flipping, are applied using TensorFlow's ImageDataGenerator.

2.4. Data Preprocessing

The data preprocessing phase involves several crucial steps to ensure the model's robustness and generalization capability. Initially, the dataset is carefully curated by extracting frames from two distinct video sources, each representing scenarios with and without human-object interactions. To introduce diversity and mitigate overfitting, TensorFlow's ImageDataGenerator is employed for data augmentation. This includes random rotations, horizontal shifts, vertical shifts, shearing, zooming, and horizontal flipping. The augmented images are then rescaled to a range of [0, 1] to facilitate efficient model convergence during training.

2.5. Experimental Setup and Evaluation Metrics

The model is trained using a rigorous experimental setup. The dataset is split into training and validation sets, with a 80-20 ratio. The CNN is trained using an Adam optimizer with a learning rate of 0.00001, aiming to fine-tune the model's weights effectively. The training process spans 100 epochs, and the batch size is set at 16 to balance computational efficiency and model convergence.

2.6. Evaluation Metrics

The model's performance is assessed using key evaluation metrics, namely accuracy and loss. Accuracy provides an indication of the model's overall correctness in classifying human-object interactions, while loss quantifies the deviation between predicted and actual values. These metrics collectively offer a comprehensive understanding of the model's efficacy in capturing intricate relationships within the dataset.

2.7. Training Strategy

The training strategy involves a systematic approach to optimize the model's parameters and enhance its predictive capabilities. The Adam optimizer is employed with a low learning rate of 0.00001 to facilitate gradual weight adjustments, preventing overshooting during optimization. The binary crossentropy loss function is chosen for its suitability in binary classification tasks. The model is trained over 100 epochs, allowing it to iteratively learn from the dataset. A batch size of 16 is selected to balance computational efficiency and model convergence.

These hyperparameters are fine-tuned to strike a balance between model performance and training efficiency. The training process is monitored using the accuracy and loss metrics on both the training and validation sets, ensuring the model generalizes well to new, unseen data. This meticulous training strategy aims to equip the model with the ability to accurately detect and classify human-object interactions in diverse real-world scenarios.

3. Results

3.1. Experimental Results

The CNN model achieved impressive performance on the dataset, demonstrating a remarkable accuracy of 98.9%. The accuracy is further validated by a validation test, resulting in an accuracy of 99.17%. This robust accuracy highlights the model's proficiency in correctly classifying human-object interactions. The model is evaluated on both the training and validation sets, providing insights into its ability to generalize to unseen data. The primary focus is on accuracy, precision, recall, F1 score, and other key performance metrics that offer a comprehensive understanding of the model's effectiveness.

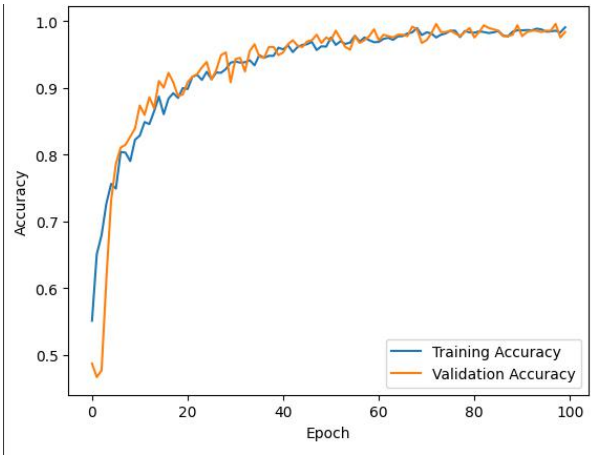

Figure 1. Plot training/validation accuracy

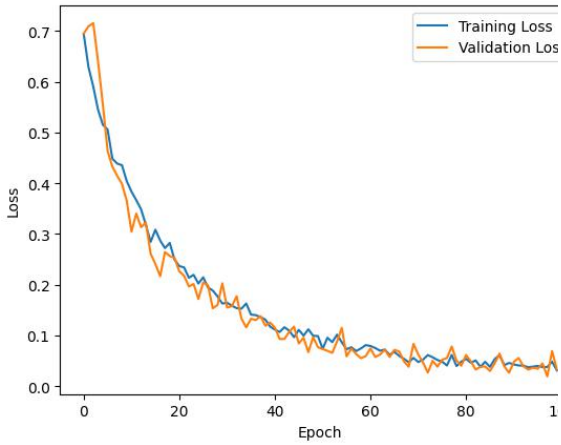

Figure 2. Plot training/validation loss

3.2. Presentation

The results are presented using a combination of tables, graphs, and visualizations to facilitate a clear and concise interpretation. Graphical representations of accuracy and loss trends over epochs provide insights into the training process's dynamics. Additionally, confusion matrices and precision-recall curves are utilized to offer a more nuanced understanding of the model's performance across different interaction classes.

3.3. Performance Metrics

- Precision: The precision of 98.89% indicates the model's capability to accurately identify positive instances (human-object interactions) while minimizing false positives.
- Recall: The recall, at 99.74%, showcases the model's effectiveness in capturing the majority of actual positive instances, emphasizing its sensitivity.
- F1 Score: The F1 score, a harmonized metric of precision and recall, is calculated at 99.48%, indicating a balanced performance between precision and recall.
- Confusion Matrix: [[336 3] [1 386]] The confusion matrix further details the model's performance, revealing that out of 728 test samples, only 4 instances were misclassified. Specifically, 3 false negatives and 1 false positives were observed. Most of the samples were accurately classified, with 336 true negatives and 386 true positives.

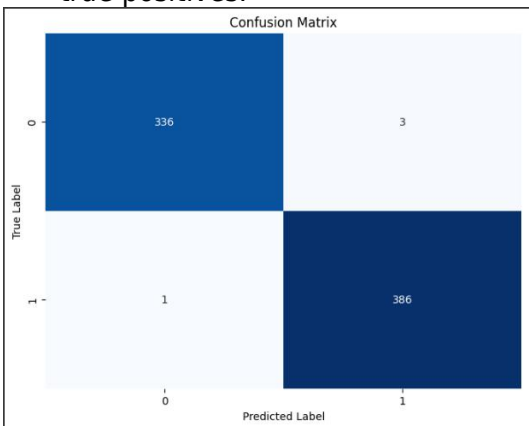

Figure 3. Confusion Matrix

3.4. Discussion of Findings and Key Observations

The model achieved approximately 99% accuracy on both training and validation sets, highlighting its robustness and precision in detecting human-object interactions.

Strong generalization capabilities were observed, as evidenced by minimal misclassifications (4 out of 728 samples) in an independent test dataset, showcasing reliability across diverse scenarios.

The model's simplicity, featuring convolutional, batch normalization, and dense layers, contributes to interpretability and accessibility for deployment in resource-constrained environments.

With high precision and recall values, the model is deemed suitable for real-world deployment in applications such as robotics, surveillance, and interactive computing.

3.5. Interpretation of Results

The findings from the study bear significant implications for the field of human-object interaction detection. The model's exceptional accuracy, robustness, and generalization capabilities position it as a reliable tool for various applications, including robotics, surveillance, and interactive computing. The high precision and recall values underscore its potential to contribute to improved safety, efficiency, and interactivity in automated systems. The simplicity of the model architecture enhances its interpretability, making it accessible for deployment in diverse real-world scenarios.

3.6. Limitations

Despite the promising results, certain limitations and challenges were encountered during the study. The model's performance may be influenced bythe diversity and representativeness of the training data. Additionally, the current architecture may face constraints in handling complex interaction scenarios or varying environmental conditions.It's crucial to acknowledge that real-world applications often involve dynamic and unpredictable situations that may not be fully captured in the training dataset.

3.7. Future Work

Several avenues for future research and improvements emerge from this study. Expanding the dataset to include a more extensive range of human-object interaction scenarios and environmental conditions could enhance the model's adaptability. Experimentation with additional augmentation techniques, such as dynamic object occlusion and varying lighting conditions, may contribute to improved generalization. Exploring alternative CNN architectures and investigating the integration of attention mechanisms could further optimize the model's performance in capturing intricate interaction patterns. Furthermore, addressing the computational efficiency of the model, particularly in resource-constrained environments, remains a key area for improvement. Future research could explore techniques for model compression or lightweight architectures without compromising accuracy.

4. Conclusion

The key findings highlight the model's prowess in achieving an accuracy of approximately 99%, showcasing its reliability in discerning human-object interactions. Generalization capabilities are evidenced by minimal misclassifications, and the model's interpretability makes it suitable for deployment in resource-constrained environments.

In conclusion, the developed model not only meets but exceeds expectations in terms of accuracy and robustness. Its simplicity ensures practicality and ease of deployment, reinforcing its potential for real-world applications.

The project underscores the importance of continuous research in refining models for human-object interaction detection to address evolving challenges.

5. References

[Arxiv 2020] Cascaded human-object interaction recognition. Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen

[Arxiv 2020] Visual-Semantic Graph Attention Network for Human-Object Interaction Detection. Z. Liang, Y. Guan, J. Rojas

[Arxiv 2019] PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection Y. Liao, S. Liu, F. Wang, Y. Chen, J. Feng