Art Critique: Is It Real Or Fake?
Latent Feature Extraction from Convolutional Neural Network & Grad-CAM

Group 14
Wojciech Urban, Olga Stanczyk, Nastasija Maksimovic

Introduction

Recent advancements in artificial intelligence, particularly in art generation, have significantly blurred the lines between AI-generated and human-created artworks. Technologies like Latent Diffusion and Standard Diffusion models have propelled AI art to a level of sophistication where it's increasingly difficult to distinguish it from works made by human artists. This blurring raises substantial concerns for both the art community and society at large, where the authenticity and provenance of art are highly valued.

In response, the AI-ArtBench dataset was developed as part of a research initiative to create a comprehensive system for detecting AI-generated art. This dataset contains over 180,000 images, split evenly between human-created artworks from the ArtBench-10 dataset and those generated by Latent and Standard Diffusion models. The resolution of these images varies, with human-created and Latent Diffusion images at 256x256, and Standard Diffusion images at a higher resolution of 768x768.

The main goal of this study is to assess the effectiveness of using Convolutional Neural Networks (CNNs) for binary classification tasks that differentiate between AI-generated and human-created art. The approach involves extracting latent features from the CNN's final layer, reducing dimensionality with Principal Component Analysis (PCA), and performing clustering on these features to identify patterns that aid in correct and incorrect classifications.

Additionally, this report examines the application of Gradient-weighted Class Activation Mapping (Grad-CAM) to identify specific features within the images that influence classification decisions. Through these methods, this research aims to enhance our understanding of the distinct elements that separate AI-generated art from human-drawn pieces, potentially guiding the development of more sophisticated and accurate detection systems (Castellano and Vessio).

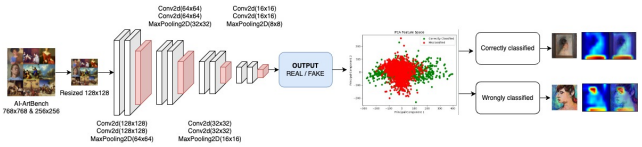

Figure 1: Architecture

Convolutional Neural Networks (CNNs)

We chose CNNs as the primary tool for binary classification of the images into AI-generated and human-drawn categories. CNNs excel in image recognition tasks due to their architecture, which efficiently captures the spatial hierarchies in data through convolutional layers that act as feature extractors. This capability makes them especially suited for differentiating fine details in artworks that distinguish AI-generated images from those created by humans. Furthermore, CNNs can manage varying image resolutions effectively, accommodating the different resolutions present in our dataset (Hamid et al.).

Feature Extraction from the Last Layer of CNN

Post-classification, we extract latent features from the last layer of the CNN. This layer is known to hold the most abstract and discriminative features of the input data, providing a rich representation that is crucial for the subsequent stages of our analysis. Extracting features from this layer simplifies the input data into a form that retains essential information required for further processing (Jogin et al.).

Dimensionality Reduction Using Principal Component Analysis (PCA)

To enhance the manageability and efficiency of our analysis, we apply PCA to the extracted features. PCA reduces the dimensionality of these features by transforming them into a smaller set of uncorrelated variables known as principal components. This reduction not only aids in alleviating computational burdens but also emphasizes the most significant features, facilitating clearer insights into data variances and aiding in more effective clustering (Ma and Yuan).

Clustering with K-means Algorithm

With the dimensionality reduced, we employ the K-means clustering algorithm to analyze patterns within the data, especially focusing on instances that were initially misclassified by the CNN. Clustering helps in identifying subtle groupings in the dataset that may indicate common features or discrepancies in the AI-generated versus human-drawn classification. We specifically opt for 10 clusters to parallel the 10 different art eras represented in the dataset, allowing us to investigate era-specific characteristics and their impact on classification accuracy (Boutsidis, Drineas, and Mahoney).

Gradient-weighted Class Activation Mapping (Grad-CAM)

To further our understanding of the CNN's decision-making process, we utilize Gradient-weighted Class Activation Mapping (Grad-CAM). This technique provides a visual representation of the areas within an image that influence the CNN's classification decision. Employing Grad-CAM aids in the interpretability of our model by highlighting which features within the image are pivotal for classification, and it proves instrumental in diagnosing and analyzing misclassified images.

Through the combined use of these methodologies, we aim to develop a robust system for distinguishing AI-generated art from human-drawn art, thereby contributing valuable insights into the ongoing development of detection and attribution mechanisms in the art domain (Selvaraju et al.).

Results

CNN

This study employs a CNN to classify images into two distinct categories: AI-generated and human-drawn. The chosen CNN architecture strategically utilizes its capabilities in feature extraction and hierarchical pattern recognition, enabling effective differentiation between the intricate details characteristic of AI-generated and human-drawn images.

The CNN architecture as visible in Figure 1 begins with an initial convolutional block comprising two layers, each employing 3x3 filters with 128 channels. Batch Normalization and Rectified Linear Unit (ReLU) activation follow each layer, facilitating rapid convergence and stability while capturing initial features. Subsequent convolutional blocks further refine feature extraction, with the second block featuring two layers of 64 channels, the third block with two layers of 32 channels, the fourth block with two layers of 16 channels, and the final block containing a single layer with 8 channels. Consistent application of Batch Normalization and ReLU activation ensures stability and efficiency across the architecture. Max Pooling layers follow each convolutional block, downsampling feature maps to reduce computational load and enhance pattern recognition at varying scales.

Following the convolutional blocks, the model transitions to fully connected layers. Pooled features are flattened and fed into a dense layer with 512 neurons, utilizing ReLU activation. The final classification layer comprises two output neurons, distinguishing between AI-generated and human-drawn images.

Over 50 epochs, training loss decreased significantly as visible in Figure 2, indicating effective learning, while validation loss exhibited fluctuations, suggesting challenges in model generalization. The model achieved an overall test accuracy of 91.49% and a test loss of 0.3549, with training accuracy reaching up to 99%, and validation accuracy peaking around 93%. These results underscore the CNN's efficacy in distinguishing between the two image categories through its layered architecture.

Precision, recall, and F1-score metrics further validate the model's performance. For AI-generated images labeled as 'fake', precision stands at 0.92, recall at 0.96, and an F1-score of 0.94, reflecting strong predictive accuracy. For human-drawn images, classified as 'real', precision is 0.90, recall is 0.83, and the F1-score is 0.87, indicating robust performance in this category. Macro and weighted averages around 0.90-0.91 for precision, recall, and F1-score signify balanced performance across both image categories, with the weighted average reflecting consistent accuracy relative to sample proportions for each class.

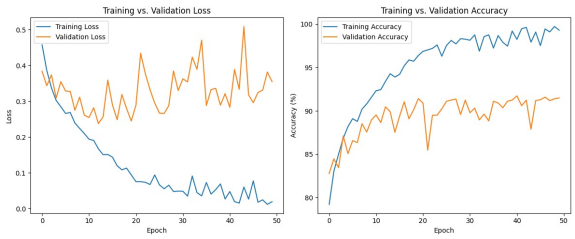

Figure 2: CNN results on training sets

```
Test Loss: 0.3549, Accuracy: 91.49%
Classification Report:
              precision    recall  f1-score   support

        fake       0.92      0.96      0.94     20000
        real       0.90      0.83      0.87     10000

    accuracy                           0.91     30000
   macro avg       0.91      0.89      0.90     30000
weighted avg       0.91      0.91      0.91     30000
```

Figure 3: CNN results on test set

Latent Features Extraction and PCA

Following the classification phase in our study, we proceed to extract latent features from the last layer of our CNN, specifically from the MaxPool2d layer. This particular layer captures the most abstract and discriminative features of the input data. By focusing on this layer, we effectively simplify the input data while retaining the essential information necessary for further analysis.

The latent features extracted are then subjected to PCA which is presented in Figure 4. This step is critical for enhancing the manageability and efficiency of our analytical process. PCA operates by reducing the dimensionality of these features, transforming them from a high-dimensional space—initially comprising 2,097,152 features—to a much more manageable two-dimensional space. This dimensionality reduction is achieved by isolating the principal components, which are a smaller set of uncorrelated variables that retain the most significant variances of the data.

By applying PCA, we not only alleviate computational burdens but also highlight the most pivotal features, thus providing clearer insights into the variations within the data. This streamlined dataset is then better suited for clustering and further analysis, enabling us to discern patterns and characteristics with greater clarity and precision.

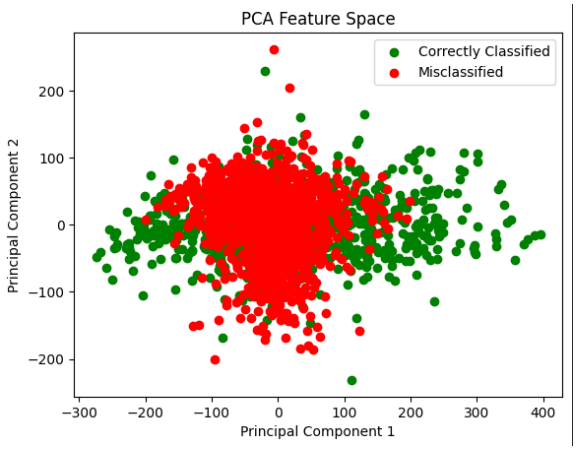

Figure 4: PCA Results

Clustering on misclassified features and GRAD-CAM

In our study, we employed a Convolutional Neural Network (CNN) to classify images representing ten distinct artistic eras: Art Nouveau, Baroque, Expressionism, Impressionism, Post-Impressionism, Realism, Renaissance, Romanticism, Surrealism, and Ukiyo-e. Subsequent to the classification, we conducted a thorough analysis using K-means clustering and Gradient-weighted Class Activation Mapping (Grad-CAM). As illustrated in Figure 4, the clusters formed by K-means exhibit well-defined circular shapes, indicating a clear delineation among the categorized eras.

The clustering results, alongside Grad-CAM visualizations, shed light on the underlying patterns that may influence CNN's decision-making process in classifying these images. Here is an overview of the findings (images provided in the appendix):

Cluster 0 (Figures 7, 8, 9) did not align with a specific artistic era, and no distinctive pattern emerged from the Grad-CAM visualizations for the misclassified images. In contrast, Cluster 1 (Figures 10, 11, 12), which primarily contained Baroque art, frequently showed the CNN focusing on the faces and bodies depicted in the artwork. This was indicated by the red areas in the Grad-CAM heatmaps, which signify significant influence on classification decisions.

Cluster 2 (Figures 13, 14, 15) predominantly featured Realism and demonstrated a pattern where CNN mainly focused on the lower sections of the paintings. This was evident from the red highlights in these areas on the heatmaps. Cluster 3 (Figure 16, 17, 18) consisted of Renaissance art and did not exhibit a distinctive pattern in the Grad-CAM analysis, suggesting a more dispersed focus by the CNN

across the images.

Cluster 4 (Figures 19, 20, 21), which did not predominantly represent a specific era, often included images with visible nature scenes. The CNN tended to focus on the background elements involving nature, as indicated by the red zones in the visualizations. Cluster 5 (Figure 22, 23, 24), dominated by Art Nouveau, observed a significant CNN activity in the bottom right of the paintings, where critical details often reside.

Clusters 6 (Figure 25, 26, 27) and 7 (Figure 28, 29, 30), filled with Post-Impressionist and Impressionist works respectively, did not show any clear pattern in the Grad-CAM results, indicating a varied influence across the images. Cluster 8 (Figures 31, 32, 33) contained primarily Expressionist art, with a consistent focus at the bottom of the paintings, suggesting key features in these areas heavily influence classification.

Lastly, Cluster 9 (Figures 34, 35, 36), which was dominated by Surrealism, did not reveal a specific pattern in the heatmap analysis, reflecting the eclectic nature of this art style.

Grad-CAM is crucial for understanding which parts of an image are most significant for a CNN classification decision. This technique generates a heatmap overlay on the image, where warmer colors such as red indicate regions that have a higher impact on the decision. The intensity of the color directly correlates with the influence that area had in guiding CNN's classification, making Grad-CAM a valuable tool for interpreting the model's behavior and highlighting potential areas for model improvement, especially in handling the diverse characteristics of different artistic styles. This visualization aids in identifying the focal points of the CNN, helping to refine and improve model accuracy in future iterations.

Clustering on correctly classified features and GRAD-CAM

In our analysis of image classification using a CNN, we employed Grad-CAM to gain insights into the features within art images that most influenced classification decisions. This report presents a detailed examination of patterns in correctly classified images across different artistic eras, demonstrating CNN's ability to discern distinguishing features.

For Cluster 0 (Figures 37, 38, 39), which did not correspond to any specific artistic era, Grad-CAM visualizations did not reveal a consistent pattern, suggesting a general proficiency of the CNN in dealing with a variety of artistic expressions without era-specific characteristics (images provided in the appendix).

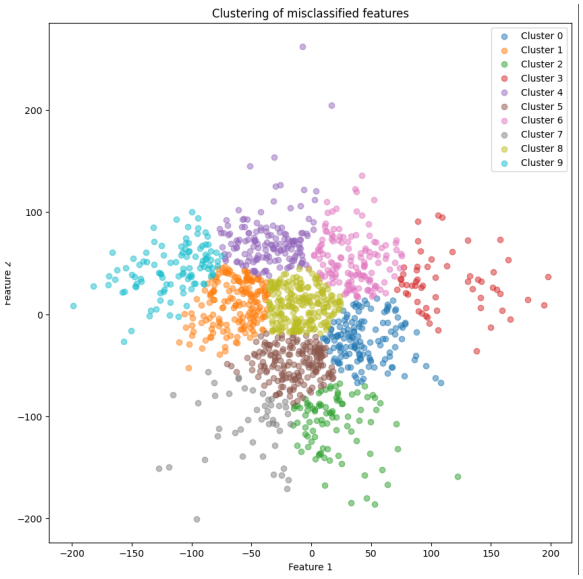

Figure 5: Clustering Misclassified Features

Cluster 1 (Figures 40, 41, 42), predominantly comprising Baroque art, demonstrated a significant focus on the lower sections of the paintings. The heatmaps consistently showed red highlights in these areas, indicating these regions played a crucial role in the accurate classification of these artworks.

Similarly, Cluster 2 (Figures 43, 44, 45), which mostly included artworks from the Realism era, exhibited a pattern where CNN effectively utilized information from the lower parts of the paintings. This suggests that characteristic elements located in these sections are key for recognizing the Realism style.

The Renaissance artworks in Cluster 3 (Figures 46, 47, 48) also followed this trend, with CNN primarily leveraging features from the bottom sections of the paintings for successful classification, although the pattern was less pronounced compared to Baroque and Realism.

In Cluster 4 (Figures 49, 50, 51), which included a mix of various eras, a combined focus on both the bottom and middle sections of the paintings was observed. This pattern suggests that for a diverse set of images, the CNN adapts by considering multiple key regions to optimize classification accuracy.

Art Nouveau artworks in Cluster 5 (Figures 52, 53, 54) similarly displayed a focus on the bottom and middle sections. The recurrent emphasis on these parts across different clusters indicates a generalizable strategy by CNN to pinpoint era-defining features in similar locations.

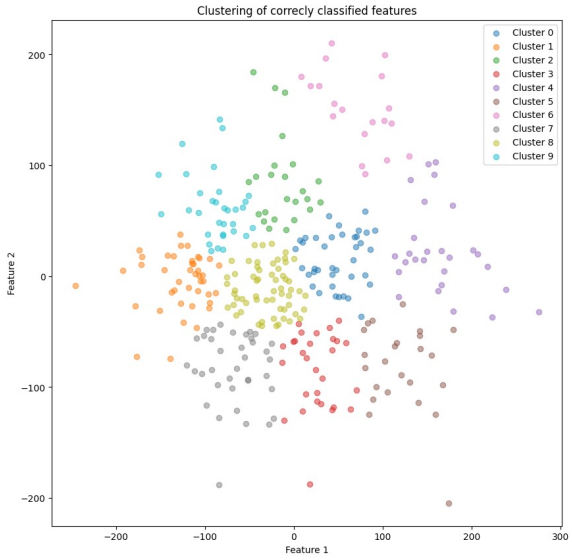

Figure 6: Clustering correctly classified Features

Clusters 6 through 9 (Figures from 55 to 66), containing Post-Impressionist, Impressionist, Expressionist, and Surrealist works respectively, all displayed a consistent pattern where the CNN focused on the lower and, to a lesser extent, middle sections of the paintings. This consistency across various artistic styles reinforces the effectiveness of CNN's strategy in focusing on these areas to achieve accurate classification.

This pattern analysis through Grad-CAM has proven crucial for understanding CNN's decision-making process in image classification. The focus on the lower parts of the paintings, as revealed by the heatmaps, played a pivotal role in successfully classifying images according to their corresponding artistic eras. These findings not only validate CNN's effectiveness but also highlight potential areas for further refinement and model enhancement, especially in terms of capturing and utilizing features from diverse and eclectic art styles.

Conclusions

We deployed a CNN to effectively classify images as either AI-generated or human-drawn, utilizing advanced techniques such as Grad-CAM for deeper analysis. The CNN's architecture, comprising multiple convolutional and pooling layers followed by dense layers, proved adept at distinguishing between the nuanced features characteristic of both image types.

Throughout the training process, as evidenced by significant reductions in training loss and a high test accuracy of 91.49%, the CNN demonstrated its robustness in feature recognition and pattern extraction across diverse datasets. The architecture's effectiveness is further exemplified by the precision, recall, and F1-score metrics, which showed strong performance, particularly in identifying AI-generated images with high confidence.

The application of Grad-CAM provided crucial insights into the CNN's decision-making process. For correctly classified images, the CNN consistently focused on specific regions—primarily the lower parts of the images—across various artistic eras, indicating a successful strategy in recognizing era-defining features. This was particularly notable in clusters where the CNN's focused areas aligned well with the key characteristics of the classified era, thus confirming the model's effectiveness.

Moreover, the study extended beyond classification to explore the latent features of the images using Principal Component Analysis (PCA). This step not only reduced the dimensionality of the data but also highlighted the most significant features for further analysis. The resulting two-dimensional space facilitated an efficient examination of the underlying patterns and variations within the data, enhancing our understanding of the CNN's classification logic.

In conclusion, this study underscores the CNN's capability to not only distinguish between AI-generated and human-drawn images with high accuracy but also to provide interpretable insights through advanced visualization techniques like Grad-CAM. These findings suggest that the CNN, with its structured layering and strategic feature analysis, is a potent tool in the field of image classification. The study also highlights areas for future improvements, particularly in enhancing model generalization to reduce fluctuations in validation loss. Through ongoing refinement and leveraging insights from techniques such as PCA and Grad-CAM, there is potential for further enhancing the model's accuracy and applicability to broader image datasets.

References

Boutsidis, C.; Drineas, P.; and Mahoney, M. W. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Castellano, G., and Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. 33(19):12263–12282.

Hamid, Y.; Elyassami, S.; Gulzar, Y.; Balasaraswathi, V. R.;

Habuza, T.; and Wani, S. An improvised CNN model for fake image detection. 15(1):5–15.

Jogin, M.; Mohana; Madhulika, M. S.; Divya, G. D.; Meghana, R. K.; and Apoorva, S. Feature extraction using convolution neural networks (CNN) and deep learning. In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2319–2323.

Ma, J., and Yuan, Y. Dimension reduction of image deep feature using PCA. 63:102578.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. 618–626.

Appendix

Misclassified Images

Cluster 0 Misclassified

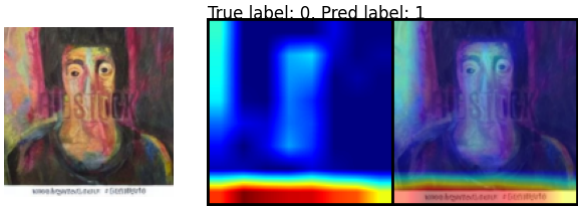

Figure 7: GRAD-CAM cluster 0 misclassified 0

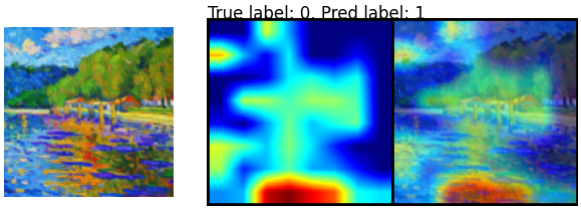

Figure 8: GRAD-CAM cluster 0 misclassified 1

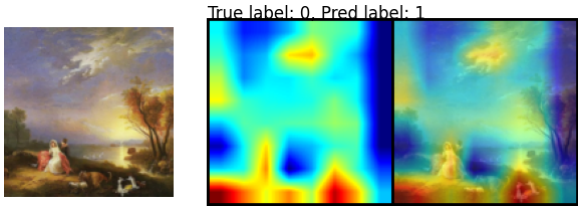

Figure 9: GRAD-CAM cluster 0 misclassified 2

Cluster 1 Misclassified

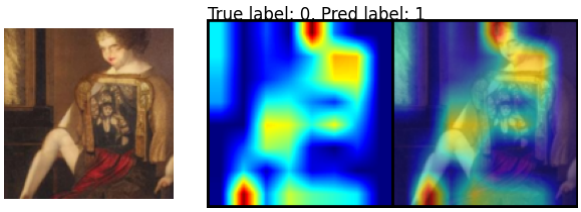

Figure 10: GRAD-CAM cluster 1 misclassified 0

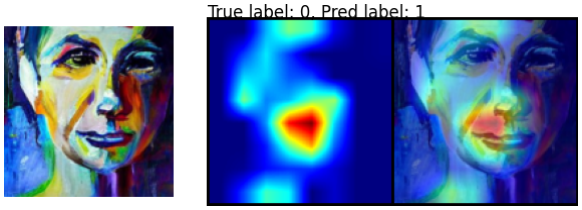

Figure 11: GRAD-CAM cluster 1 misclassified 1

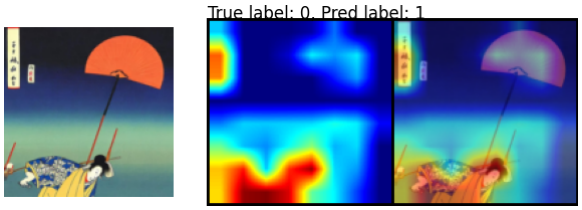

Figure 12: GRAD-CAM cluster 1 misclassified 2

Cluster 2 Misclassified

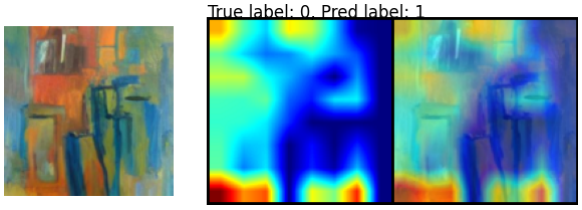

Figure 13: GRAD-CAM cluster 2 misclassified 0

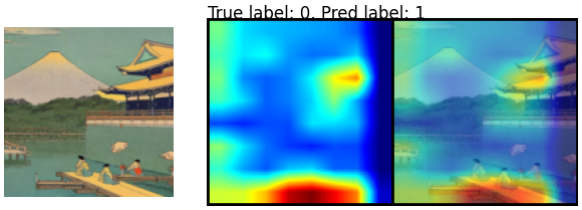

Figure 14: GRAD-CAM cluster 2 misclassified 1

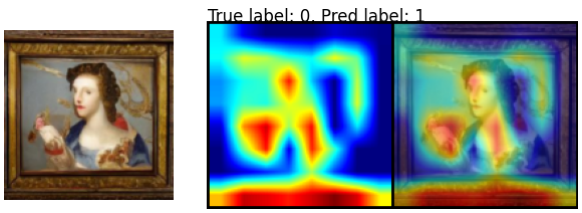

Figure 15: GRAD-CAM cluster 2 misclassified 2

Cluster 3 Misclassified

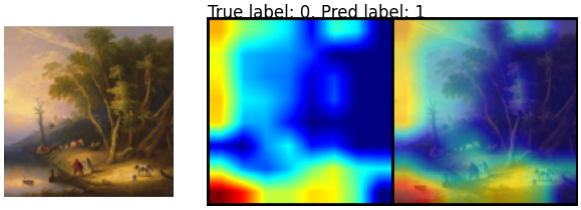

Figure 16: GRAD-CAM cluster 3 misclassified 0

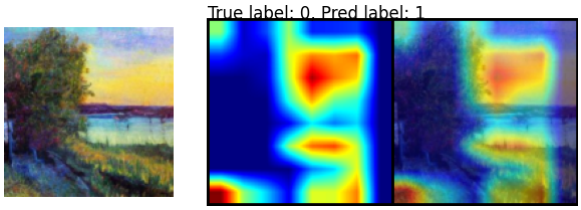

Figure 17: GRAD-CAM cluster 3 misclassified 1

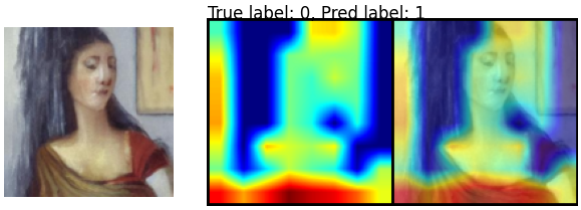

Figure 18: GRAD-CAM cluster 3 misclassified 2

Cluster 4 Misclassified

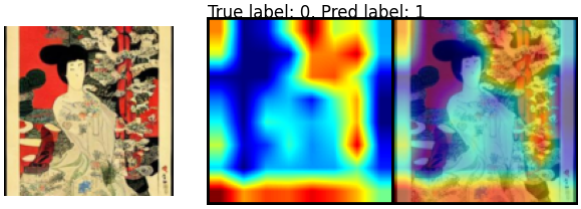

Figure 19: GRAD-CAM cluster 4 misclassified 0

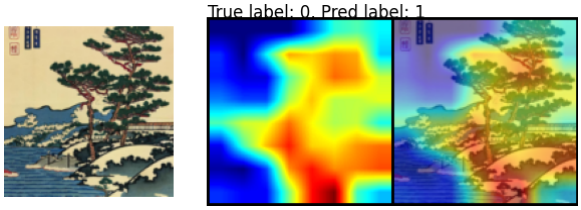

Figure 20: GRAD-CAM cluster 4 misclassified 1

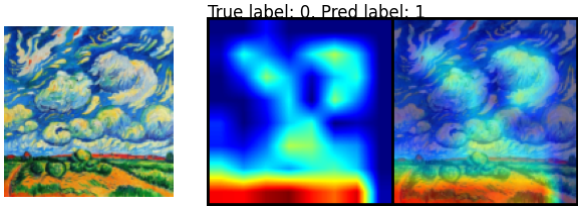

Figure 21: GRAD-CAM cluster 4 misclassified 2

Cluster 5 Misclassified

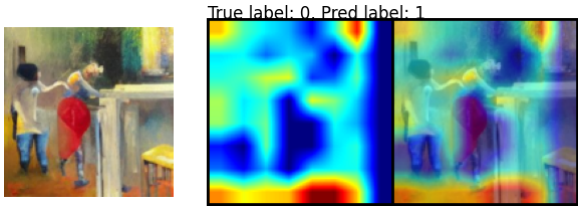

Figure 22: GRAD-CAM cluster 5 misclassified 0

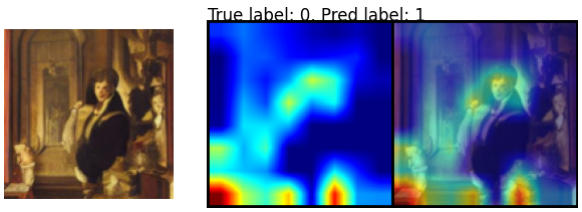

Figure 23: GRAD-CAM cluster 5 misclassified 1

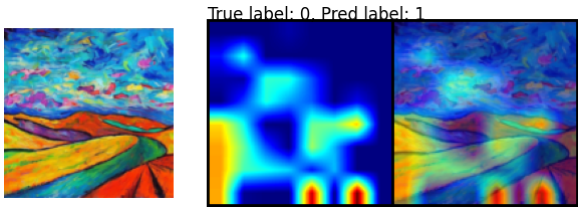

Figure 24: GRAD-CAM cluster 5 misclassified 2

Cluster 6 Misclassified

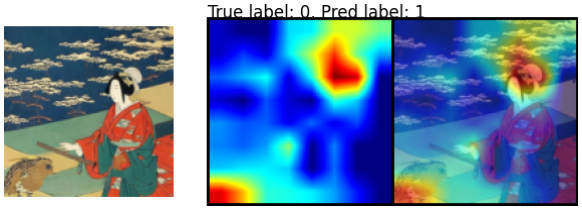

Figure 25: GRAD-CAM cluster 6 misclassified 0

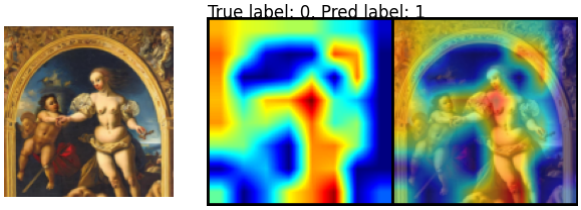

Figure 26: GRAD-CAM cluster 6 misclassified 1

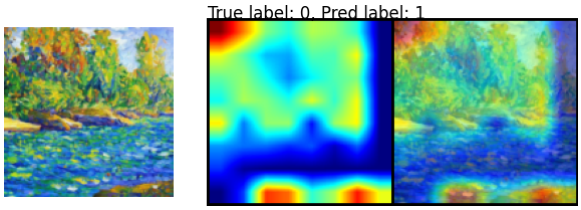

Figure 27: GRAD-CAM cluster 6 misclassified 2

Cluster 7 Misclassified

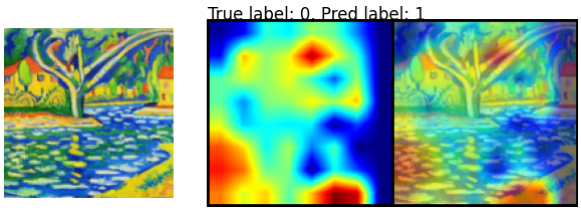

Figure 28: GRAD-CAM cluster 7 misclassified 0

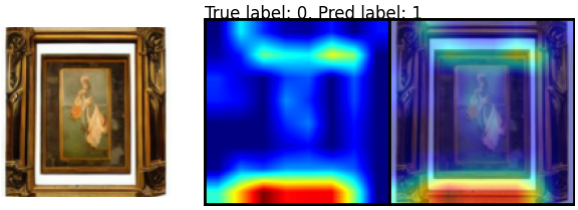

Figure 29: GRAD-CAM cluster 7 misclassified 1

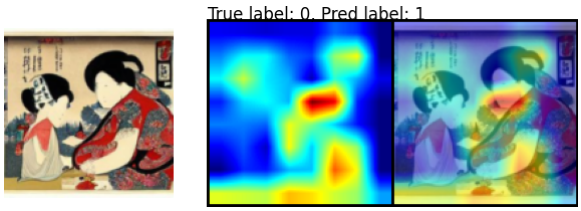

Figure 30: GRAD-CAM cluster 7 misclassified 2

Cluster 8 Misclassified

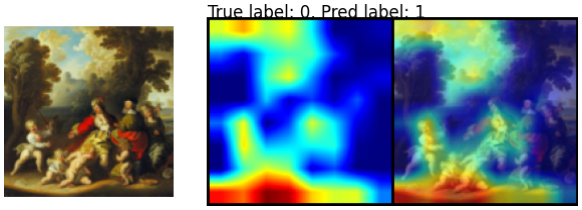

Figure 31: GRAD-CAM cluster 8 misclassified 0

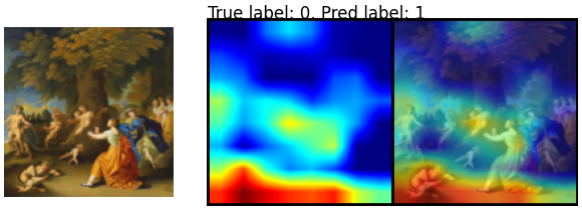

Figure 32: GRAD-CAM cluster 8 misclassified 1

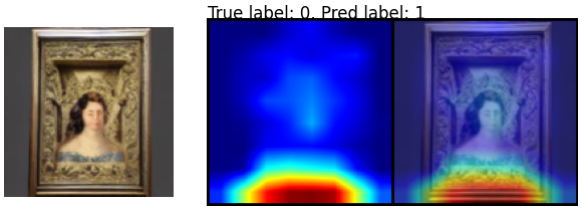

Figure 33: GRAD-CAM cluster 8 misclassified 2

Cluster 9 Misclassified

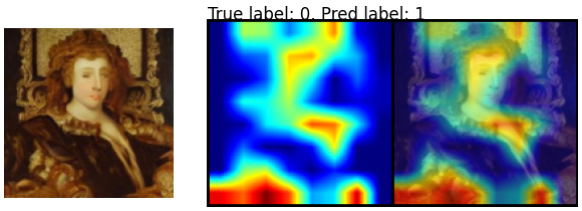

Figure 34: GRAD-CAM cluster 9 misclassified 0

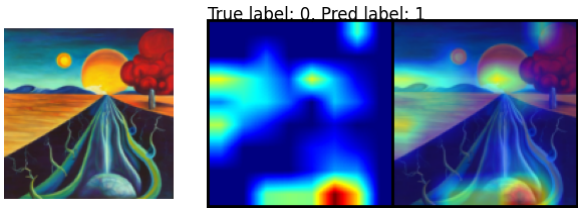

Figure 35: GRAD-CAM cluster 9 misclassified 1

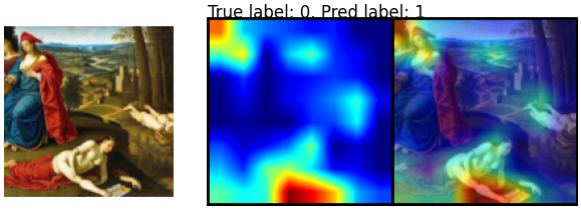

Figure 36: GRAD-CAM cluster 9 misclassified 2

Correctly Classified Images

Cluster 0 Correctly Classified

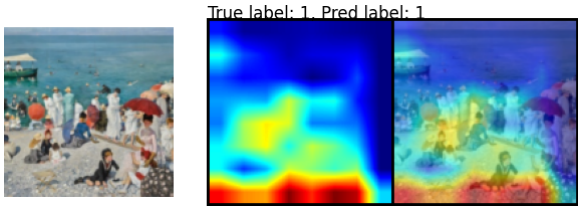

Figure 37: GRAD-CAM cluster 0 correctly classified 0

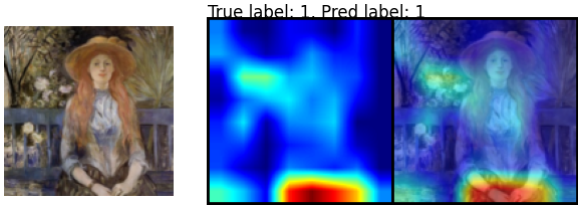

Figure 38: GRAD-CAM cluster 0 correctly classified 1

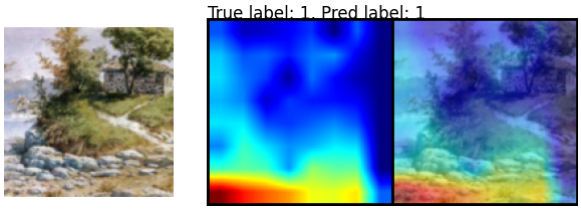

Figure 39: GRAD-CAM cluster 0 correctly classified 2

Cluster 1 Correctly Classified

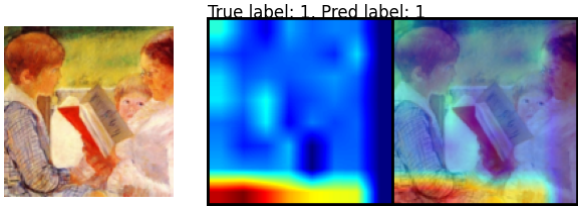

Figure 40: GRAD-CAM cluster 1 correctly classified 0

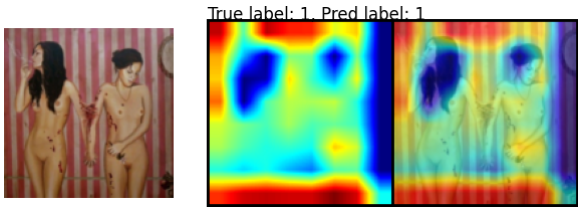

Figure 41: GRAD-CAM cluster 1 correctly classified 1

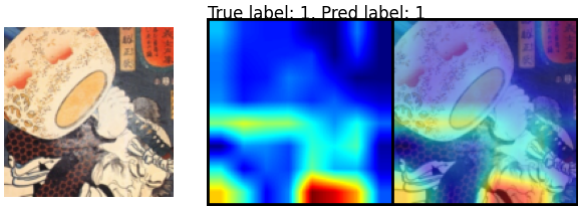

Figure 42: GRAD-CAM cluster 1 correctly classified 2

Cluster 2 Correctly Classified

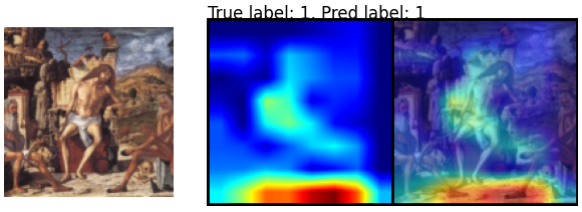

Figure 43: GRAD-CAM cluster 2 correctly classified 0

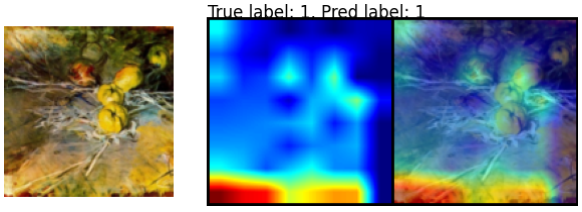

Figure 44: GRAD-CAM cluster 2 correctly classified 1

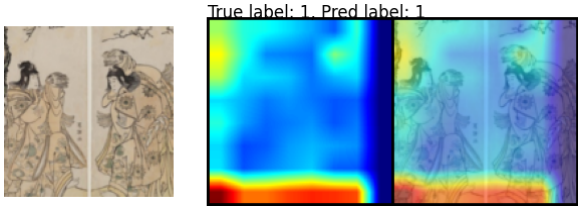

Figure 45: GRAD-CAM cluster 2 correctly classified 2

Cluster 3 Correctly Classified

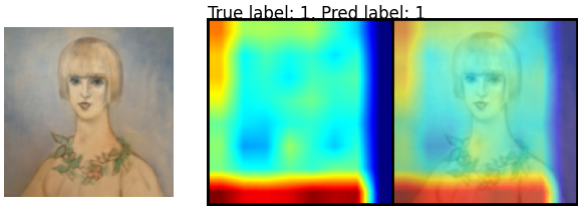

Figure 46: GRAD-CAM cluster 3 correctly classified 0

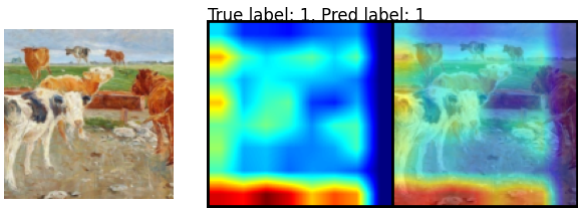

Figure 47: GRAD-CAM cluster 3 correctly classified 1

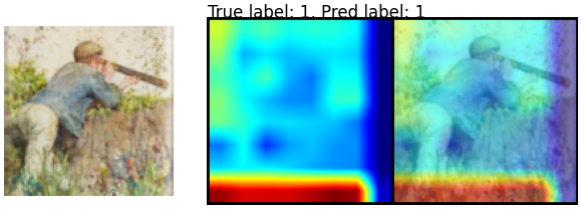

Figure 48: GRAD-CAM cluster 3 correctly classified 2

Cluster 4 Correctly Classified

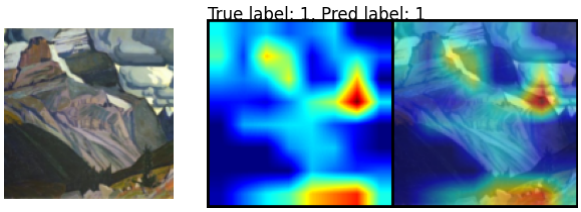

Figure 49: GRAD-CAM cluster 4 correctly classified 0

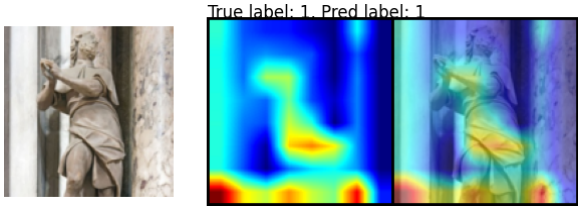

Figure 50: GRAD-CAM cluster 4 correctly classified 1

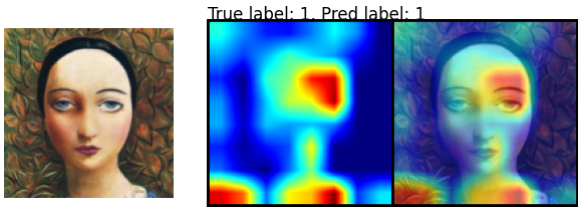

Figure 51: GRAD-CAM cluster 4 correctly classified 2

Cluster 5 Correctly Classified

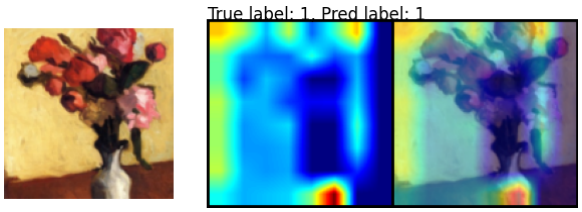

Figure 52: GRAD-CAM cluster 5 correctly classified 0

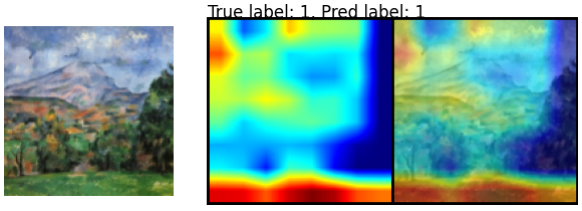

Figure 53: GRAD-CAM cluster 5 correctly classified 1

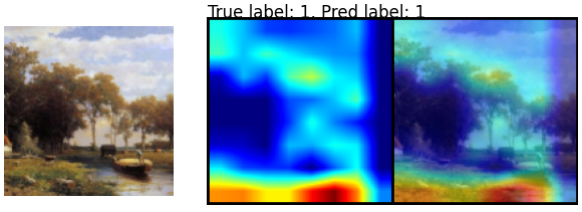

Figure 54: GRAD-CAM cluster 5 correctly classified 2

Cluster 6 Correctly Classified

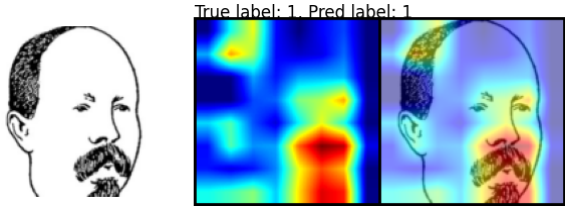

Figure 55: GRAD-CAM cluster 6 correctly classified 0

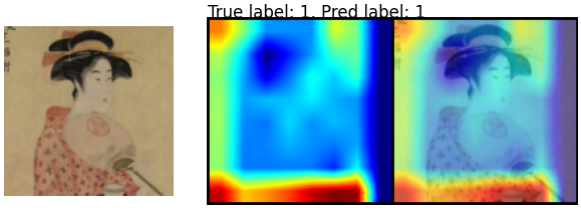

Figure 56: GRAD-CAM cluster 6 correctly classified 1

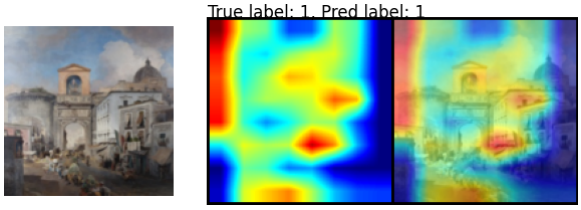

Figure 57: GRAD-CAM cluster 6 correctly classified 2

Cluster 7 Correctly Classified

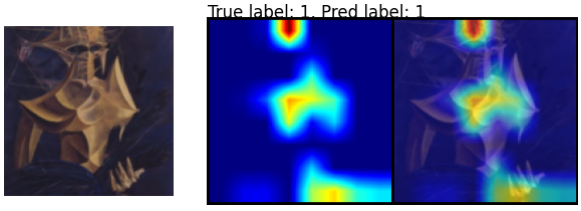

Figure 58: GRAD-CAM cluster 7 correctly classified 0

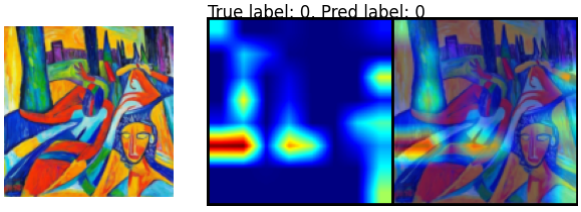

Figure 59: GRAD-CAM cluster 7 correctly classified 1

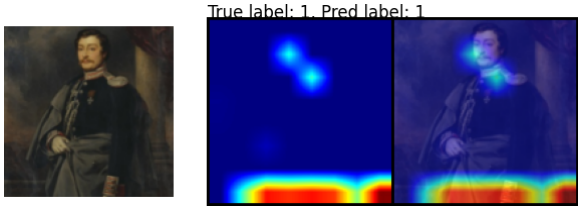

Figure 60: GRAD-CAM cluster 7 correctly classified 2

Cluster 8 Correctly Classified

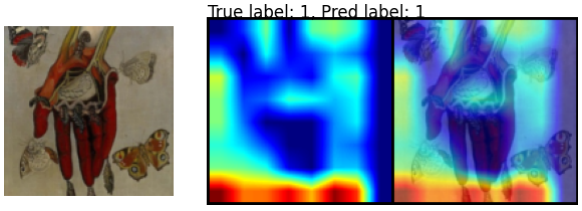

Figure 61: GRAD-CAM cluster 8 correctly classified 0

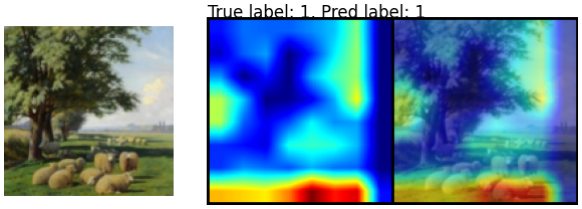

Figure 62: GRAD-CAM cluster 8 correctly classified 1

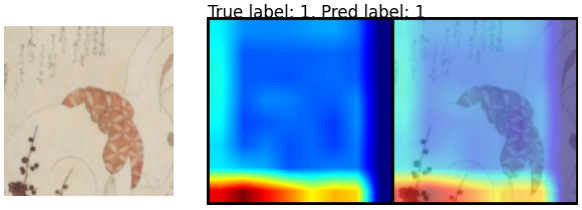

Figure 63: GRAD-CAM cluster 8 correctly classified 2

Cluster 9 Correctly Classified

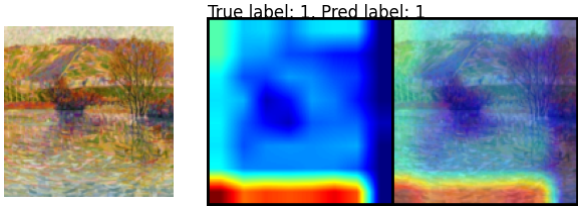

Figure 64: GRAD-CAM cluster 9 correctly classified 0

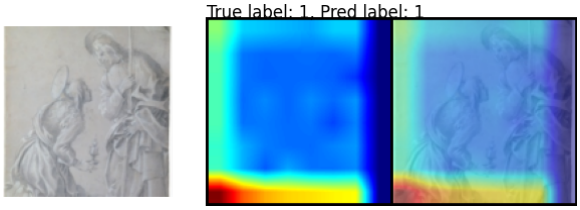

Figure 65: GRAD-CAM cluster 9 correctly classified 1

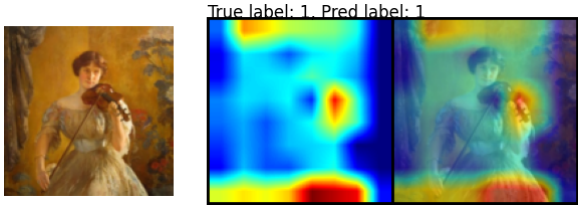

Figure 66: GRAD-CAM cluster 9 correctly classified 2