Detecting AI-generated Images

Yufan Yang

MSc Advanced Computer Science

Bailin Deng

School of Computer Science and Informatics, Cardiff University

02-08-2024

Table of Contents

Abstract

Today, the rapid development of generative AI technology, especially breakthroughs in cutting-edge areas such as generative adversarial networks (Gans) and diffusion models, is pushing generative AI to unprecedented heights. These advanced technologies make it easier and more efficient to generate high-quality, realistic images, and have been widely used in multiple industries and application scenarios, such as entertainment, advertising design, medical image analysis, and more. However, along with these technological advances have come new challenges, especially in how to accurately detect and identify false images. With the popularity of generative AI technology, the abuse of Deepfake technology is becoming increasingly serious, and the spread of false images and videos not only poses a threat to personal privacy and reputation, but also has a profound negative impact on the stability of society and the public's trust system.

Currently, while existing Deepfake detection methods show good results when dealing with images generated by generative adversarial networks (Gans), they are significantly less effective when dealing with images generated by diffusion models. This is because diffusion models can generate more detailed and diverse images than traditional GAN techniques, which makes existing detectors present significant generalization problems in the face of such more complex generated images. This limitation exposes the inadequacies of current detection techniques and demonstrates the urgent need to develop a new generation of detection methods to effectively address these emerging threats in the face of the challenges of diffusion model-generated images.

To solve this problem, an innovative detection method is proposed in this paper, which aims to improve the generalization ability of existing detection systems when processing images generated by diffusion models. Specifically, we adopt two main detection strategies: First, we use the diffusion data set to train the One-Class SVM model to play its advantages in the field of anomaly detection; Second, the diffusion data set is combined with the ImageNet data set, and the pre-trained ResNet50 model is learned and fine-tuned to adapt to the new image detection requirements. One-Class SVM was selected for its ability to detect abnormal images without large amounts of labeled data, while ResNet50 was fine-tuned to take advantage of its deep learning capabilities to significantly improve the recognition accuracy of complex generated images.

Through the research in this paper, we not only hope to provide a new perspective in the field of Deepfake detection, but also strive to promote the further development of detection methods for new generation

technologies such as diffusion models. We believe that with the continuous improvement and optimization of detection methods, we can effectively meet the challenges of AI-generated images, and provide stronger technical support for the safety and trust of society.

Acknowledgements

Throughout this journey of challenges and exploration, I have been fortunate to receive immense support and encouragement from many individuals. Their companionship and assistance have guided me through the vast expanse of academia, ultimately leading me to the completion of this master's thesis.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Professor Bailin Deng. Whether I was grappling with difficult research problems or uncertain about my direction, Professor Deng's exceptional expertise and patient guidance helped to dispel the fog, allowing me to move forward with confidence. His invaluable teachings will remain etched in my heart, serving as a precious asset on my future academic path.

I am profoundly grateful to my parents. Like towering mountains, they have provided me with unwavering love and support, forming the solid foundation upon which my life is built. During moments of setbacks and challenges, their silent guardianship and heartfelt expectations gave me the strength to regain confidence and persistently pursue my dreams.

I have also received tremendous support from many teachers and friends throughout this journey. I am thankful to my classmates, who have illuminated my academic path like stars in the night sky, making it less lonely. Together, we engaged in academic discussions, tackled challenges, and celebrated our successes. This friendship, like fine wine, grows richer with time and will be cherished forever.

I extend my sincere thanks to all the teachers and colleagues at the School of Computer Science. The fertile ground of knowledge and the support you provided have allowed my academic research to flourish like a thriving seedling.

The completion of this thesis reflects the wisdom and love of many people. I offer my deepest gratitude to everyone who has helped and supported me. May this gratitude flow like a gentle stream, forever remaining in your hearts.

1. Introduction

Today, due to the rapid development of generative AI technologies, especially advances in generative adversarial networks (Gans) and diffusion models, generative ai is being taken to a new level. These technologies make the generation of high-quality, realistic images easier than ever and are widely used in multiple fields. However, these advances have also created new challenges, especially when it comes to detecting false images. With the popularity of these technologies, the problem of abuse of Deepfake technology has become increasingly serious, and the spread of false images and videos poses a significant threat to social stability and public trust.

Existing Deepfake detection methods, while performing well against GAN-generated images, are significantly less effective when dealing with images generated by diffusion models. This is because the images generated by diffusion models surpass traditional GAN techniques in detail and variety, and existing detectors exhibit significant generalization problems in the face of this complexity. This limitation demonstrates the urgent need to develop new detection techniques that can effectively cope with images generated by diffusion models.

In order to solve this problem, this paper attempts to use the diffusion data set to train One-Class SVM and the diffusion data set mixed with the ImageNet data set to migrate the pre-trained model ResNet50 to detect the images generated by ai. One-Class SVM was chosen for its strong capabilities in the field of anomaly detection, while ResNet50 was fine-tuned to accommodate new detection requirements. The uniqueness of this approach is that it can significantly improve the detector's generalization ability in processing complex generated images without relying on large amounts of labelled data.

Through the research in this paper, we aim to provide a new perspective on the field of Deepfake detection and promote the development of detection techniques for novel generation models.

2. Background information

The rapid development of generative AI technology, especially the breakthrough in the field of image generation, has attracted extensive attention from academia and industry. Generative adversarial networks (Gans) and Diffusion Models, as the current mainstream generative AI technologies, have shown strong capabilities and have made important progress in many application scenarios. For example, since GAN was first proposed, it has been widely used in many fields such as image generation, video generation, and image restoration [1]. Among them,

StyleGAN, proposed by Karras et al., further promoted the development of high-quality image generation technology through the antagonistic training of generator and discriminator, and was widely used in artistic creation, virtual image production and other fields, laying a foundation for the breakthrough of subsequent generation models [2].

Outside of academic research, commercial products are also beginning to widely adopt these generative AI technologies. For example, NVIDIA's GAN technology has been used in image generation tools such as GAN Paint Studio, which allows users to generate realistic scene images with simple brush manipulation [3]. At the same time, diffusion model, as an emerging technology in the field of generative AI, is gradually applied in content creation, advertising design and other industries due to its stability and diversity in generating high-resolution images.

However, with the popularity of these technologies, the abuse of AI-generated images, particularly Deepfake technology, has sparked widespread social concerns. Deepfake technology has been used to create fake videos, images, and spread disinformation in areas such as politics, finance, and social media, leading to serious social problems. Although there are some existing Deepfake detection methods, such as convolutional neural network (CNN) -based detectors and frequency domain analysis techniques, that perform well in certain scenarios, these methods are limited in their effectiveness against high quality images generated by diffusion models [4] [5].

a) Problems that need to be solved

The emergence of diffusion model has greatly improved the diversity and detail representation of generated images, which has brought unprecedented challenges to existing detection techniques. Traditional Deepfake detection methods mostly rely on detecting forgery traces in images, or on classification models based on known samples. However, images generated by diffusion models often have higher realism and diversity, which makes it difficult for traditional detection methods to identify subtle traces of counterfeiting. In addition, the generation process of diffusion model is complex and flexible, and the detector often presents significant generalization problems when facing different generation techniques [6]. These challenges indicate that existing detection methods are not effective in generating images in the face of diffusion models, and there is an urgent need to develop more advanced and adaptive detection techniques.

b) Literature evaluation and project uniqueness

Reviewing the existing research literature, although there have been many studies devoted to the detection of AI-generated images, most of the work has focused on GAN-generated images, and there is still little

attention to diffusion models. For example, Wang et al. 's research shows that traditional Deepfake detection methods perform well when dealing with GAN generated images, but the detection accuracy drops significantly when dealing with images generated by diffusion model [7]. In addition, other studies have pointed out the lack of generalization ability of existing detection techniques, especially when dealing with diverse backgrounds or complex images [8].

The uniqueness of this project is that it proposes a new method based on the One-Class SVM model and the pre-trained ResNet50 model to detect the images generated by the diffusion model. One-Class SVM was chosen for its excellent performance in anomaly detection, which can identify abnormal samples in unlabeled data and is particularly suitable for processing highly diverse generated images [9]. ResNet50, on the other hand, can fine-tune to new detection requirements through transfer learning, especially when it shows strong feature extraction capabilities for new generation models. By combining these two methods, this study is expected to improve the detector's generalization ability when processing images generated by diffusion models, especially in the case of not relying on large amounts of labeled data, and improve the detection accuracy.

By critically evaluating the existing literature, we found that there is currently a lack of effective detection methods for images generated by diffusion models. In response to this research gap, this project proposes a new detection strategy, hoping to promote the development of the field of Deepfake detection through innovative technical means, and provide technical support for dealing with more complex generative AI challenges in the future. We believe that this research will provide new insights for academia and industry and have a wide impact in practical applications."

3. Problem description

AI image generation is becoming increasingly realistic and diverse, while pushing the frontiers of image generation technology, it also brings unprecedented challenges, both in how to effectively detect false images generated by these advanced models. Current Deepfake detection methods have shown some effectiveness in dealing with images generated by Gans, but with the popularity of diffusion models, traditional detectors are powerless in the face of these new images. These new images are often far more detailed, realistic and diverse than ever before, making it difficult for existing methods to capture traces of forgery, resulting in a significant decline in detection accuracy.

a) Background information

In recent years, diffusion model has become an important field in generative AI research, which generates high-quality images by gradually adding and removing noise, and has been widely used in many fields such as image generation, restoration and art creation. These models are not only capable of producing extremely detailed images, but are also capable of generating diverse visual content in different contexts and backgrounds. The flexibility of this generation process and the high resolution of images make it difficult for traditional Deepfake detection methods, such as convolutional neural network (CNN) -based detectors and frequency domain analysis techniques, to maintain stable detection results when dealing with images generated by diffusion models. Especially when the images generated by these models are used in areas such as politics, journalism, or social media, the speed and reach of disinformation further add urgency to the problem.

The main limitation of current detection techniques is that they lack the generalization ability to cope with the image generated by diffusion models. These detectors are usually trained on fixed feature sets or pre-defined image samples and are therefore prone to detection errors when faced with diffusion models with complex and flexible generation processes. In addition, because the images generated by these models often exhibit higher complexity in detail and realism, existing detection methods have difficulty capturing subtle traces of forgery in these images, resulting in a decrease in detection performance.

b) Expected research results and their benefits

The purpose of this study is to try the detection method using One-Class SVM and pre-trained ResNet50 model to try to find an effective method to deal with the false image generated by the diffusion model. By using the advantages of One-Class SVM single classification, it is possible to train using only data with only a diffusion dataset. In addition, leveraging ResNet50's powerful feature extraction capabilities by fine-tuning it to the specific task requirements of detecting AI-generated models can significantly improve the detector's generalization ability when processing complex generated images.

Expected successes include:

Develop a new detection algorithm: By comparing the two detection methods of One-Class SVM and ResNet50 model, find a model that can accurately identify AI-generated images, and can be used to detect false images generated by diffusion model.

Improved detection generalization: Improved detection methods improve the performance of detectors in different generation models and diverse image backgrounds, especially on previously unseen generation image types.

Provide practical application tools: To provide an efficient and reliable detection tool to combat the spread of disinformation, especially in the field of news and social media, to help maintain the authenticity of information and public trust.

The success of the research will bring significant benefits, not only fill the gap in the current detection technology in diffusion model image detection in theory, but also enhance the defense ability against false information in practical application. The technology developed in this research can effectively curb the spread of false images in sensitive areas, protect social stability, and enhance public trust in media content.

4. Method selection

In this study, which aims to detect images generated by AI, specifically for the Diffusion dataset, two main machine learning methods are used for comparative analysis: One-Class SVM (single-class support vector Machine) and ResNet50 transfer learning model represent the technical routes of unsupervised anomaly detection and supervised image classification respectively.

a) Source of data set

To train and evaluate the above models, this study used data from two main datasets:

ImageNet Dataset: ImageNe is a widely used large-scale image dataset containing millions of annotated high-resolution images covering thousands of object classes. In this study, we first selected and downloaded the data set of 2017 ImageNet Large-scale Visual Recognition Challenge (ILSVRC) from https://www.image-net.org. A total of 9928 images are downloaded to the local ImageNet dataset as normal samples for migration training of the ResNet50 model. These images represent natural images in the real world, providing the basis for ResNet50's "normal" distribution.

Diffusion data set: The Diffusion data set is a public image data set that contains the image generated by the diffusion model and the corresponding pair of real images. The images cover a variety of styles and topics, providing a rich sample for training and testing. The project will use the datasets package function to download 10,000

images of the Diffusion dataset from huggingface, an open-source website. The AI-generated image is labeled as type 1 and will form the base data for comparison with the real image in ImageNet in the detection task.

By combining these two datasets, we were able to capture the difference between the real image and the AI-generated image during training, allowing us to effectively train and evaluate the selected model.

b) One-Class SVM anomaly detection

One-Class SVM is an unsupervised machine learning algorithm. It does this by learning the feature distribution of a diffused dataset and then identifying samples of non-AI-generated pictures that differ significantly from that distribution. In this study, One-Class SVM was used to detect AI-generated images, which were treated as anomaly samples.

We used the pre-trained VGG19 model as a feature extractor to extract high-dimensional features from the ImageNet dataset and input them into a One-Class SVM for training. The VGG19 model has been pre-trained on the ImageNet dataset, so it has a strong image feature extraction capability to capture fine image structure and texture information, which is crucial for the detection of AI-generated images.

One-Class SVMS perform well in unsupervised learning tasks that provide only abnormal samples, especially when the data is asymmetrical (such as when the number of AI-generated images is not balanced with the number of real images). By utilizing the depth features extracted by VGG19, a One-Class SVM can effectively identify abnormal images that differ from the training set, thereby detecting the AI-generated images. This approach has been proven in several fields. For example, research by Khan and Madden (2014) shows that One-Class SVMS perform well in applications such as network anomaly detection and fraud detection [10].

c) ResNet50 migration learning

ResNet50 is a deep convolutional neural network commonly used for image classification tasks. Transfer learning is the process of fine-tuning an existing large-scale pre-trained model to adapt to new data sets and tasks. In this study, we use ResNet50's pre-trained model to binary classify the Diffusion data set and Imagenet data set by transfer learning, that is, distinguish the AI-generated image from the real image.

ResNet50 was chosen as the model architecture because of its deep feature extraction capability. By freezing the early layers of ResNet50 and fine-tuning only the late fully connected layers, we were able to

efficiently adapt to new image classification tasks. In addition, transfer learning techniques enable us to achieve high classification accuracy on limited data sets. He et al. (2016) pointed out that ResNet50, with its unique residual network structure, can effectively alleviate the problem of gradient disappearance in deep neural networks, thus improving the training effect of the model [11].

The advantage of transfer learning is the ability to take the knowledge gained on other tasks and apply it to the current task, thereby improving the performance and training efficiency of the model. ResNet50 performs well in image classification tasks due to its depth and structure. Through transfer learning, we can better adapt the Diffusion data set to recognize the AI-generated images. Zoph et al. (2018) also mentioned in their research that transfer learning can significantly improve the performance of deep learning models on new image datasets [12].

5. Description of One-Class SVM model structure

a) Feature mapping and kernel function

In this study, we intercept the output of the last convolutional layer of the VGG19 network and flatten it into a 4096-dimensional feature vector that preserves the core information of the image.

The core of One-Class SVM is to map the input high-dimensional features to a higher-dimensional or infinite feature space through a kernel function, in which the model can more easily separate normal and abnormal samples. In this study, we chose RBF kernel function. This kernel function maps the input data nonlinearly to a high-dimensional space by calculating the Euclidean distance between the eigenvectors. Gamma is an important parameter of the RBF kernel function, which determines the scale of the mapping. We experimentally adjust the value of gamma to ensure that normal samples are tightly enclosed in a high-dimensional space, while abnormal samples are separated outside the decision boundary.

b) Decision boundary and support vector

The main goal of the One-Class SVM model is to include normal samples within the boundary as much as possible, while excluding abnormal samples, by constructing a decision boundary in a high-dimensional space. This process relies on several core concepts:

The model learns to identify some key samples (i.e., support vectors) in the input features that are on or very close to the decision boundary. The choice of support vectors is critical because they determine the

location and shape of the decision boundary. These support vectors are often the most difficult samples to classify, and their presence ensures the robustness of the model. The One-Class SVM model forms a soft decision boundary by constructing a hyperplane containing all the support vectors. This boundary attempts to exclude all abnormal samples without misjudging normal samples as much as possible. The degree of "tightness" of the decision boundary is controlled by another key parameter of the model, nu, which defines the proportion of samples that are allowed to be exceptions. By adjusting nu, we can adjust the sensitivity of the model to abnormal samples.

c) Classification and prediction

After the model training is completed, the new input samples will enter the One-Class SVM model through feature extraction and kernel mapping, and the model will judge whether it is a normal sample or an abnormal sample according to the location of the sample in the high-dimensional space. For each new input sample, a One-Class SVM maps it into a higher-dimensional space and calculates the distance from the support vector. If the sample is within the decision boundary, it is classified as a normal sample. Otherwise, it is considered an abnormal sample. The result of the model output is a binary value indicating whether the sample belongs to the normal class (+1) or the abnormal class (-1).

d) Model structure diagram

The workflow of the One-Class SVM model structure is as follows:

1. Input image (feature extraction via VGG19) → high-dimensional feature vector.
2. The eigenvector (mapped to a higher-dimensional space via the RBF kernel function) → the support vector determines the decision boundary.
3. Decision boundary (defined by support vector) → new sample classification (based on its position relative to the decision boundary).

6. Description of ResNet50 model structure

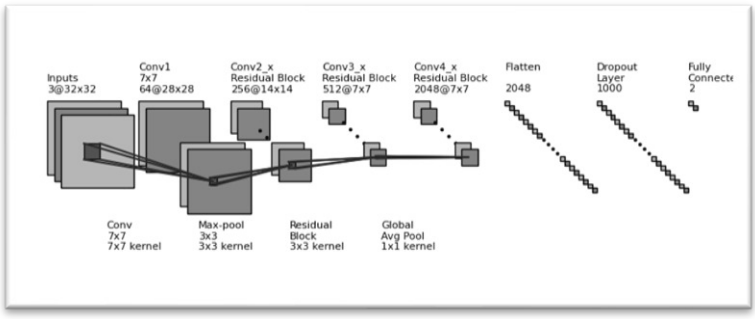
In this study, we adopted the ResNet50 model as the infrastructure for detecting AI-generated images. The ResNet50 model is widely used in image classification tasks mainly due to its unique "Residual Block" design, which effectively alleviates the problem of disappearing gradients as the network depth increases.

a) Hierarchical structure of ResNet50

The network structure of ResNet50 is divided into several phases, each consisting of a different number of residual blocks. Specifically, the initial part of the model consists of a 7x7 convolution layer, using 64 convolution cores, and convolution at a step of 2. This is followed by a 3x3 maximum pooling layer, which is responsible for extracting the initial low-level features of the image, such as edges and simple shapes. The core of the whole model is the residual block of 4 stages, each stage has a similar structure, but the number of convolution nuclei is gradually increased. The convolution kernel sizes inside the residual block are 1x1, 3x3, and 1x1, respectively. With skip connection, the input can bypass the middle convolution layer and be added directly to the output, resulting in a residual structure. Each residual block is designed to gradually improve the expressiveness of the model by capturing deeper features. After extracting advanced features through multiple residual blocks, ResNet50 uses a global averaging pooling layer to reduce the dimension of the feature map to reduce the risk of overfitting. Finally, the models are classified through a fully connected layer. Since our task is a binary classification problem, we adjust the output layer of the ResNet50 pre-trained model to a fully connected layer containing two neurons. To further enhance the generalization ability of the model, we added the Dropout layer before the fully connected layer to randomly drop the output of a subset of neurons during training, thus preventing the model from relying on features of specific neurons.

b) Visualization of the model

In order to more intuitively understand the structure of ResNet50, the network structure diagram of this model is as follows:



Network structure diagram of the model

This diagram shows the entire flow from the input image to the output category, including the type of each layer, the size changes in the input and output, and the specific design of the residual blocks. The design of residual blocks is particularly important because they allow

gradients to bypass multiple convolutional layers while backpropagating, directly updating the parameters of previous layers, which allows the network to go deeper while avoiding the problem of disappearing gradients.

c) Hierarchical analysis of the model

The hierarchical design of ResNet50 enables the model to extract image features at different levels. The original 7x7 convolution layer and maximum pooling layer are mainly used to extract image low-level features such as edges, corners, and simple shapes. These features are the basis for further advanced feature extraction. As the model moves deeper into the residual blocks, the receptive field of the convolution kernel gradually increases and begins to capture more complex structural information such as textures, patterns, and larger shapes. Feature extraction at this level mainly focuses on the details of the middle scale of the image, which is a bridge connecting low-level features and high-level features. The last few residual blocks and the global averaging pooling layer focus on extracting high-level semantic information from the image, such as object class and overall structure. In our AI-generated image detection task, these levels of features are critical to distinguish AI-generated images from real images.

By freezing the first few layers of ResNet50, we retain its low-level and intermediate feature extraction capabilities pre-trained on the ImageNet dataset, fine-tuning only the last few layers to suit our specific task needs. This transfer learning approach both retains the powerful capabilities of pre-trained models and enables them to handle new classification tasks through a small amount of data adjustment

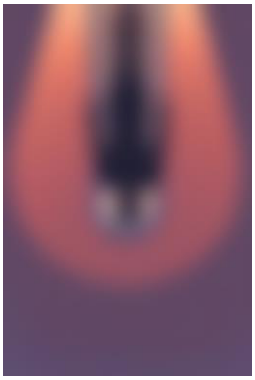
7. Implementation process

a) Data set preprocessing

For the ImageNet data set, the project first selects the data set of the 2017 ImageNet Large-scale Visual Recognition Challenge (ILSVRC) from the website https://www.image-net.org to download and save it locally. Since the original data set is a data set used for training the identification of everything, it contains label files for identifying target location and type, which is not needed in this project. This part of content is directly ignored, and only the pictures of the original data set are extracted and stored in the directory of type 0. When importing the data set, the content from this data set is labeled with type 0 by default.

For the diffusion dataset, the project uses the load_dataset function to load huggingface's diffusionDB dataset, which contains images that

are "not suitable for viewing in public places". In addition, this part of the image is shielded by adding blur effect during generation, and the effect is shown as the following figure.



Examples of images that need to be filtered

It is not practical to use this part of the blurred image and may affect the accuracy of the training results, so the filter function is used to filter the image. This part of the image is represented by the value of the image_nsfw attribute of 2 in the label of the dataset. Therefore, all images with an image_nsfw value of 2 can be filtered out. After image filtering, the data set needs to be further normalized. The project used torchvision's transforms module to accomplish this step, including scaling, cropping, tensor conversion, and normalization of images.First, we uniformly scale the image to a larger side length of 256 pixels, and then center crop to obtain a square image of 224x224 pixels.

```
class CustomTransform:
    FassonYang0616 *
    def __init__(self, target_size):
        self.target_size = target_size
        self.transform = transforms.Compose([
            transforms.RandomResizedCrop(target_size),
            transforms.RandomHorizontalFlip(),
            transforms.ColorJitter(brightness=0.4, contrast=0.4, saturation=0.4, hue=0.2),
            transforms.ToTensor(),
            transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
        ])

```

Preprocess part of the source code

This ensures that all images are the same size when input to the model. The image is then converted to a tensor and a normalization process is applied to normalize the pixel values of the image using predefined means and standard deviations. The images before and after processing are compared as follows:

Pre-treatment effect Post-treatment effect

This processing step can effectively eliminate the numerical offset of pixel values in the image, thus speeding up the training process of the model and improving the convergence speed. According to relevant studies, data preprocessing is crucial to improve the quality of image feature extraction (Simonyan & Zisserman, 2014) [13]. Considering a series of hardware constraints such as equipment computing power and storage conditions, the project only uses 10,000 data from each data set for training. After pre-processing and data set partitioning, the data set distribution of the two labels is as follows:

```
Dataset A label distribution: Counter({0: 9928})
Dataset B label distribution: Counter({1: 9906})
```

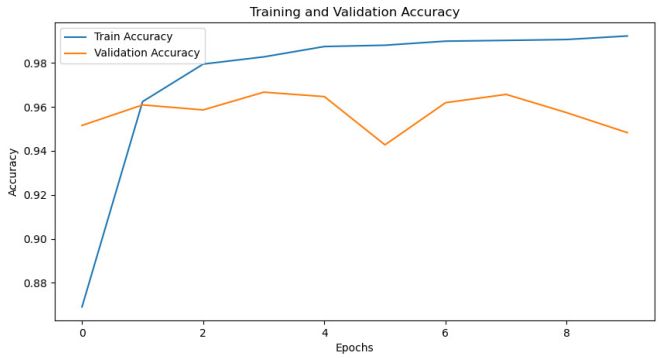
Data set label distribution (A is the real image, B is the fake image of the diffused data set)

b) One-Class SVM training

After the diffusion data set is preprocessed, the VGG19 model is invoked to extract the high-dimensional features of the image. VGG19 is a pre-trained deep convolutional neural network that removes the last fully connected classification layer and can extract feature vectors more conveniently. The extracted feature vectors were input into a One-Class SVM for training. During the training process, we used RBF (radial basis function) as the kernel function and set gamma='auto' parameter to ensure that the model can effectively adapt to the data distribution in high dimensional space. When the training is complete, the model is saved as one_class_svm_model_pytorch.joblib in the root directory of the project. In the test stage, the same preprocessing and feature extraction are performed on the test images in the Diffusion dataset, and then the features are input into the trained One-Class SVM model for prediction. The output of the model is 1, indicating that the image is "normal", that is, the real image; An output of -1 indicates that the image is "abnormal", that is, the AI generates the image.

c) ResNet50 migration training

Similar to the training part of the One-Class SVM, the program first loads and preprocesses images from the ImageNet and Diffusion datasets. After loading the ResNet50 pre-trained model, the early layers of the model are frozen and the last fully connected layer is replaced with a new classifier to accommodate the binary classification task, the purpose of freezing the early layers is to preserve the low-level features that ResNet50 has learned on large-scale datasets such as ImageNet, while only fine-tuning the later layers to accommodate the new data distribution. The fine-tuned ResNet50 model was trained using the pre-processed data set. The cross-entropy loss function and stochastic gradient descent optimizer were used to optimize the model parameters. The accuracy of the training process is shown in the figure below.

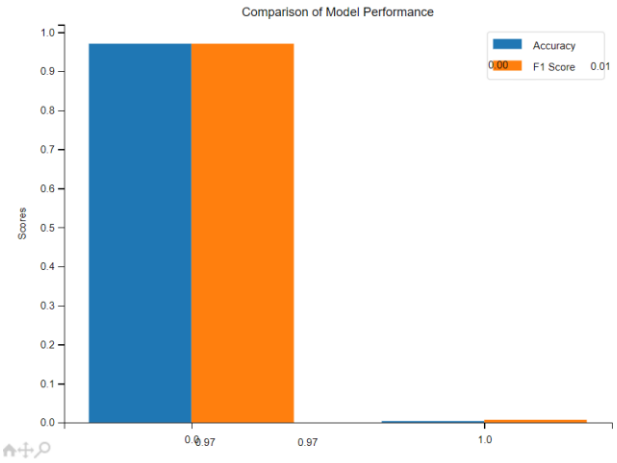


The accuracy of the training process

It can be seen that ResNet50 has a very good accuracy in the training process. To avoid overfitting, we also applied techniques such as Dropout and data enhancement. Literature has shown that these techniques play an important role in improving the generalization ability and stability of models (Srivastava et al., 2014) [14]. After the training is complete, the model is evaluated using the test set.

8. Evaluation of results

After completing the model training, we compared the performance of the two methods on the same test set in terms of accuracy and F1 scores. We first evaluated the test set using the ResNet50 model. The model's accuracy and F1 score are both close to 1.0, indicating that it performs very well on this task. On the other hand, the performance of the One-Class SVM model is extremely unsatisfactory, with accuracy and F1 scores close to zero, and hardly any test samples are correctly classified. To clearly show the performance of the two models, we plotted the results as a bar graph, as shown in the figure.



The performance of the two models on test set

The chart visually shows the accuracy and F1 scores of the ResNet50 and One-Class SVM models on the test set. As can be seen from the figure, the accuracy and F1 score of the ResNet50 model are both 0.97, while the accuracy and F1 score of the One-Class SVM model are almost 0. The ResNet50 model successfully distinguished most samples in the test set, showing high classification ability. On the contrary, the One-Class SVM model cannot distinguish the samples in the test set effectively, and the performance is very unsatisfactory.

By comparing the performance of the two models, we can draw the following conclusions:

i) Advantages of ResNet50 model: As a pre-trained deep convolutional neural network, ResNet50 model can effectively extract advanced features of images and accurately classify them through the final fully connected layer. In our experiments, the ResNet50 model demonstrated its strong performance in this task, correctly classifying almost all test samples.

ii) Limitations of the One-Class SVM model: The performance of the One-Class SVM model suggests that this unsupervised learning approach based on anomaly detection may not be suitable for such classification tasks. Although we used the advanced features extracted by the VGG19 model, the One-Class SVM model was not effective in classifying the test set. This may be because One-Class SVMS are better suited to handle exception detection problems than multi-class classification tasks.

iii) Implications for model applicability: The experimental results suggest that for complex image classification tasks, deep learning-based convolutional neural network models (such as ResNet50) may be more suitable for direct applications, while methods like One-Class SVM may require more specific scenarios or further tuning to work.

9. Error analysis and improvement direction

In this section, we will analyse in depth the error performance of ResNet50 and One-Class SVM models in detecting AI-generated image tasks. Although the ResNet50 model performs well in accuracy and F1 scores, there are still some potential sources of error that deserve our attention and analysis.

a) Error analysis of ResNet50 model

The ResNet50 model demonstrated up to 97% accuracy and F1 scores in our test sets. This result shows that the model performs well in recognizing AI-generated images and can effectively distinguish AI-generated images from real images. However, even at this high accuracy rate, there are still some possible sources of error:

i) Misjudgment of boundary samples: The boundary between the AI-generated image and the real image may sometimes be very fuzzy, especially when the AI-generated image has been highly realistic processing, or the real image has undergone a lot of post-processing, the visual features of the two may become very similar. For these samples with blurred boundaries, the model may not be able to make accurate classification, leading to misjudgement.

ii) Distribution difference between training set and test set: If there is a difference in data distribution between training set and test set, the generalization ability of the model may be affected. This can cause the model to perform poorly in certain scenarios, for example, if the training set contains a lot of picture data in the style of painting or animation, but the test set does not, it may cause the model to misjudge.

iii) Overfitting risk: Although ResNet50 performs well on training sets, there is a risk that the model may be overfitting due to limited training data. In this case, the model may perform poorly when encountering new or unfamiliar AI-generated images.

iv) Novelty of AI-generated images: AI generation technology continues to advance, and newly generated images may use different techniques or methods than those in the training set. If the images generated by these new techniques differ significantly from the training data, the model may not be able to recognize them effectively.

b) Error analysis of One-Class SVM model

The One-Class SVM model performs extremely poorly in detection tasks, with almost zero accuracy and F1 scores. This suggests that the model almost completely fails at classifying AI-generated images. The

following are potential sources of error that lead to poor performance of the One-Class SVM model:

i) **Model suitability problem:** One-Class SVM is an unsupervised learning method for anomaly detection, usually used to identify abnormal samples that are significantly different from the normal class. In this project, a One-Class SVM may not be suitable for distinguishing between two classes (AI-generated images and real images), especially if the feature differences between the two are not obvious.

ii) **Limitations of feature extraction:** Although we extracted 4096 dimensional advanced features using VGG19, these features may not be sufficient to effectively distinguish AI-generated images from real images. The VGG19 model may have lost some key details when extracting these features, resulting in a One-Class SVM that cannot effectively learn to distinguish boundaries.

iii) **Deviations between data distribution and model assumptions:** One-Class SVM assumes that the data distribution in the feature space is single (i.e., single class), which is fundamentally different from our binary classification task. If the feature distribution of the AI-generated image does not deviate significantly from the feature distribution of the real image, it will be difficult for a One-Class SVM to detect anomalies.

iv) **Sensitivity of hyperparameter Settings:** One-Class SVM is very sensitive to hyperparameters such as kernel type, gamma, and nu. Even if these parameters perform well during training, they may adversely affect generalization performance when tested.

v) In summary, although the ResNet50 model has excellent performance, it still has the risk of error under the difference of boundary samples and data distribution. However, the poor performance of the One-Class SVM model in this task may be mainly due to the contradiction between the model method and the task applicability. In future research, stronger feature extraction network or multi-layer feature fusion technology can be used to improve the feature representation ability of the model. Second, consider replacing a One-Class SVM with a model that is better suited for multi-class classification tasks, such as a support vector classifier (SVC) or other supervised learning methods. In addition, training set is also a very important starting point. In the future, it is necessary to expand the diversity of training set, reduce the distribution difference between training and test data, and improve the generalization ability of the model. Finally, it is also necessary to explore novel generation methods, update the training data of the model in time for the continuous development of AI generation

technology, and maintain the ability to identify new generation technologies.

10. Reflect and learn

Reviewing the whole research process, the successful application of ResNet50 model makes me deeply understand the working principle of convolutional neural network (CNN), especially its performance in image classification tasks. We learned how to effectively leverage pre-trained models to improve their performance by fine-tuning task-specific layers. In the process, we learned how to adjust the learning rate, batch size, and hyperparameters such as optimizers, which are important knowledge for optimizing the performance of deep learning models.

During the whole project, I explored the process of extracting features for one-class svm model with VGG19 and applying these features to One-Class SVM for image classification. This led me to deepen my understanding of feature extraction, especially in feature engineering using pre-trained networks. Despite the poor performance of the One-Class SVM model on this task, the process was a valuable learning experience that made me realize the limitations of unsupervised learning in specific tasks and the challenges it can have when faced with complex classification tasks. In addition, data preprocessing is also a key part of the machine learning workflow. In this study, we learn how to effectively perform image normalization, crop and other pre-processing operations in order to adapt to the input requirements of deep learning models. In addition, we also apply the various model evaluation indicators we have learned, such as accuracy and F1 score, and learn how to comprehensively evaluate the performance of the model through these indicators, so as to obtain more accurate model performance information. Finally, we get an in-depth understanding of the model's performance in different scenarios through error analysis. By analyzing the sources of error of ResNet50 and One-Class SVM in detecting AI-generated image tasks, we learned how to identify weaknesses in the model and think about ways to improve. This skill is important for future model tuning and further research. During the whole research process, I realized the rigor of scientific research, especially in every step of data processing, model training and evaluation, care should be taken to ensure the accuracy of results. At the same time, I also learned and understood that innovation is the core of research, especially in the face of emerging AI-generated image detection tasks, the need to constantly explore and try new methods and techniques to combat deep falsities.

One of the major contributions of this study is the exploration of AI-generated image detection. Through the detailed analysis of ResNet50 and One-Class SVM model, we provide a valuable reference and foundation for future research. These results can help other researchers better understand the performance of different models in AI-generated image detection tasks, and provide an important basis for further improvement and development of more efficient detection methods.

In the future, we will consider the integration of different types of models and integrated learning. Consider the fusion of multiple models, combining the advantages of deep learning and traditional machine learning, and improve the detection accuracy and robustness of the model through ensemble learning technology. Using the advantages of different models in different data classification and combining their own advantages, fusion processing may be a good direction for improvement. In addition, AI-generated image technology will continue to progress, and counter-counterfeiting may also face more severe challenges. Therefore, research and development of AI-generated image detection technology is still an important direction. The detection system based on such research needs to help people avoid fake news rumors and other negative affairs brought by AI-generated content to a certain extent.

11. References

1) Goodfellow, I. et al. (2014) 'Generative Adversarial Nets', in Advances in Neural Information Processing Systems (NeurIPS).
2) Karras, T. et al. (2019) 'A style-based generator architecture for generative adversarial networks', in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
3) Ho, J. et al. (2020) 'Denoising Diffusion Probabilistic Models', in Advances in Neural Information Processing Systems (NeurIPS).
4) Dhariwal, P. & Nichol, A. (2021) 'Diffusion Models Beat GANs on Image Synthesis', in Advances in Neural Information Processing Systems (NeurIPS).
5) Kingma, D.P. & Welling, M. (2014) 'Auto-Encoding Variational Bayes', in Proceedings of the International Conference on Learning Representations (ICLR).
6) Chesney, R. & Citron, D. (2019) 'Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics', Foreign Affairs.
7) Wang, S. et al. (2020) 'Exposing DeepFake Videos by Detecting Face Warping Artifacts', in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

8) Schölkopf, B. et al. (2001) ‘Estimating the support of a high-dimensional distribution’, Neural Computation, 13(7), pp. 1443-1471.
9) He, K. et al. (2016) ‘Deep Residual Learning for Image Recognition’, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.
10) Khan, L. & Madden, M.G. (2014) ‘One-class classification: taxonomy of study and review of techniques’, Journal of Computer Science, 10(3), pp. 145-162.
11) He, K., Zhang, X., Ren, S. & Sun, J. (2016) ‘Deep residual learning for image recognition’, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.
12) Zoph, B., Vasudevan, V., Shlens, J. & Le, Q.V. (2018) ‘Learning transferable architectures for scalable image recognition’, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697-8710.
13) Simonyan, K. & Zisserman, A. (2014) ‘Very deep convolutional networks for large-scale image recognition’, arXiv preprint arXiv:1409.1556.
14) Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014) ‘Dropout: A simple way to prevent neural networks from overfitting’, The Journal of Machine Learning Research, 15(1), pp. 1929-1958.