Identical Origins: A Novel Approach to Verifying Text-to-Image Model Consistency

Guy Elovici
Ben Gurion Uneversity
guyelov@post.bgu.ac.il

Abstract

With the rapid advancement and proliferation of text-to-image models, accurately identifying the origins of AI-generated images presents a significant challenge that is essential for both digital forensics and the authenticity of digital content. In this paper, we introduce a novel approach to verifying whether the same text-to-image model generates two images. We employ Vision Transformers (ViTs) for feature extraction and integrate advanced machine learning models such as XGBoost, LightGBM, and CatBoost to discern and verify each pair of images. The cornerstone of our study is the creation of the Diffusion Verification Dataset (DVD), which encompasses a wide array of images from various text-to-image models. Our results demonstrate the method's effectiveness in accurately classifying images based on their source. This research fills a vital gap in understanding AI-generated imagery and lays a foundation for future advancements in digital content verification.

1. Introduction

With the rapid advancement of artificial intelligence, particularly in text-to-image generation, discerning the origins of digitally created images has become an intricate task. Verifying if the same source generates images is academically intriguing and vital for applications in digital forensics, copyright law, and content authenticity. Existing methods primarily focus on distinguishing AI-generated images from real images, with limited research on identifying the specific model or version responsible for generating these images. Our work presents a breakthrough in this area, proposing a method of classifying whether the same text-to-image model generated a pair of images.

This study explores the effectiveness of combining Vision Transformer (ViT) for feature extraction with various machine learning models to tackle the challenge of image source verification. Our approach is anchored in creating a substantial dataset, the Diffusion Verification Dataset (DVD), composed of images from multiple text-to-image models, fostering a robust basis for model training and evaluation. By analyzing pairs of images for shared source characteristics, we offer insights into the distinct visual styles of different AI models and their evolutionary versions. Applying machine learning models like XGBoost, LightGBM, and CatBoost, among others, to this verification task reveals their potential in accurately classifying images based on their source.

This paper is organized as follows: We begin with a detailed examination of related work in Section 2, offering context and highlighting the gaps our study aims to fill. Section 3 delves into our methods, detailing the Diffusion Verification Dataset (DVD) creation, the feature extraction process using Vision Transformer (ViT), and the specific machine learning models employed for image verification. In Section 4, we present our results, encompassing the dataset creation process, the performance of each machine learning model, and a sensitivity analysis of the verification models. Section 5 discusses the implications, challenges, and insights derived from our findings, critically evaluating our approach. We conclude in Section 6 with suggestions for future research directions, exploring potential expansions of our methodology and its applications.

2. Related Work

The identification of text2image models has become very challenging due to the variety of different generation models available to use [2, 5, 6]. Additionally, due to the availability of open source framework [8] that contains different versions of the same model, it becomes challenging to determine which version of a specific model generated a given image. Researchers have started studying how to identify if a particular image was generated by an AI model (text2image model) or if it is a real image [3]. The research of Jordan et.al [3] involves creating a CIFAKE dataset, mirroring CIFAR-10's ten classes using latent diffusion. The approach includes using Convolutional Neural Networks (CNNs) for image classification, achieving an accuracy of 92.98%. Additionally, the paper implements Explainable AI via Gradient Class Activation Mapping to identify features within the images crucial for classification, revealing that the model focuses on small visual imperfections rather than the main subject of the images. However, the model is specifically tailored to mirror the CIFAR-10 dataset [4], which may limit the model's generalize ability to other types of datasets or real-world images. Moreover, the model is only capable of determining whether an image was created by an AI; it cannot classify which type of model generated the given image. To the best of my knowledge, no existing methods can classify which model generates a given image or whether the same model generated a pair of images. In my proposed method, I was able to successfully classify if the same model generated a pair of images

3. Methods

In this study, we propose a novel method for creating an image verification model that classifies whether two images were generated by the same source. Our approach is structured into three primary steps: First, we focus on creating a diffusion verification dataset (DVD) using a large-scale prompt dataset [7], meticulously curated to encompass a diverse range of image pairs. The second step involves extracting and creating embedding features for each image pair. This process is crucial as it distills the essential characteristics of the images, enabling the model to make informed comparisons. Finally, the core of our methodology is training an image verification model, leveraging advanced machine learning techniques. This model is designed to analyze the extracted features and accurately determine the likelihood of the image pairs originating from the same source, thus addressing the challenge of image source verification with a high degree of precision.

3.1. Diffusion Verification Dataset (DVD)

In the first stage of our methodology, we focus on creating a unique diffusion verification dataset (DVD), an essential foundation for our image verification model. To our knowledge, no pre-existing labeled dataset provides images alongside their corresponding sources, particularly for source verification. Therefore, we innovated by constructing our dataset independently rather than extracting images from the web. This approach was chosen to ensure that our dataset remains open-source, aligning with our commitment to public accessibility and transparency.

Our primary resource for image prompts was DiffusionDB [7], the first large-scale text-to-image prompt dataset. This database is notable for containing 14 million images generated by Stable Diffusion, created using various prompts and hyperparameters specified by real users. Drawing inspiration, we used five publicly available text-to-image models from Huggingface hub [8] to create our dataset. The specific models are Stable-diffusion-v1-4, Stable-diffusion-v1-5, Dallev3, Midjourney-v4 and Sdxl-turbo. For each model, we generated three distinct images for every prompt, ensuring diversity and comprehensiveness in the dataset. We selected 20,000 prompts for this purpose, resulting in a dataset that encompasses 300,000 images in total (as shown in Figure 1). Doing so ensured a rich and varied dataset comprising a broad spectrum of styles and themes. This vast and diverse dataset forms the backbone of our subsequent stages, providing a solid base for training and testing the image verification model with high accuracy and reliability.

Figure 1. Our large-scale image verification dataset

3.2. Converting Images Into Embedding

In the second stage of our methodology, we focused on creating embeddings from pairs of images derived from the DVD constructed in the previous stage. This process is vital for training the image verification model to discern whether two images share the same source. We meticulously generated pairs of images, ensuring a balanced representation of matching and non-matching sources. For each of the five text-to-image models used in our dataset, we created 60,000 pairs of images from the same model (intra-model pairs) and another 60,000 pairs from different models (inter-model pairs). This resulted in a comprehensive verification dataset comprising 600,000 pairs of images in total.

We utilized a state-of-the-art pre-trained backbone, the Vision Transformer (ViT) [9], to convert each image into a vectorized representation. The ViT is adept at capturing intricate visual details and patterns, transforming each image into a 768-dimensional embedding vector. This rich embedding captures the essential visual characteristics of each image, forming the foundation for the subsequent comparison.

This stage's critical aspect involves using a Siamese network approach [1], a widely recognized method in verification tasks. We subtracted the embedding vectors of each pair of images, which simplifies the task of comparing and contrasting their features. The result of this subtraction, along with the corresponding label (1 for pairs from the same source, 0 for pairs from different sources), forms the training dataset for our verification model.

3.3. Training a Verification Model

In the final stage, we trained a machine learning (ML) model using the embedding vectors generated in the previous stage. These embedding vectors, representative of each image's essential features, serve as the input features for our ML model, tailored to accomplish the verification task. The generated dataset of embedding vectors, each representing a pair of images and their associated source label, was randomly divided into training and testing sets. This split ensures that the models are trained on a diverse set of examples and validated on unseen data, thereby enhancing the generalizability and robustness of the trained models. We selected diverse ML models for this task, each known for its strengths in handling complex classification problems. The models include CatBoost, LightGBM, XGBoost, Random Forest, and Logistic Regression.

A critical component of this stage was hyperparameter tuning for each model. By meticulously adjusting and fine-tuning each model's parameters, we aimed to find the optimal configuration that achieved the best possible results. This process is fundamental in ML, as the correct set of hyperparameters can significantly enhance a model's performance, particularly in a task as nuanced as image verification. The outcome of this stage is a set of finely tuned models, each capable of effectively utilizing the embedding vectors to accurately classify whether a pair of images originates from the same source.

4. Results

4.1. Verification Dataset Creation

We used the prompts of the Diffusion-DB [7] dataset to create images from each of the five models we examined. To generate high-quality images, the prompt needs to be detailed and clear and play a key role in the quality of the generated image. Figure 2 presents the length distribution regarding the number of words in each prompt. Most prompts contain at least 15 words, which is well enough for text2image models. Furthermore, most of the text2image models are limited by the maximum length of the prompt as input, so very large prompts will be truncated. We can see from the plot that most of the prompts are not larger than 100 words.

For each model, we generated three images from the same prompt to examine the difference in the image style generated by each model. Figure 4 illustrates an example of the generated images for a given prompt from the prompt dataset; each row in the figure represents a pair of images used for the verification task. As can be seen from the figure, the different models have different image styles. For example, the images generated by Stable-diffusion-v1-4 and Dallev3 are very different in terms of style, and as expected, images generated by the same model look very similar. In addition, we can see that images generated from models of different versions (Stable-diffusion-v1-4 and Stable-diffusion-v1-5) look very similar, but in the verification task, they should be classified as 0 (not the same source). This illustrates the challenge of this task due to the diversity of style of the images from the same prompt generated by different models.

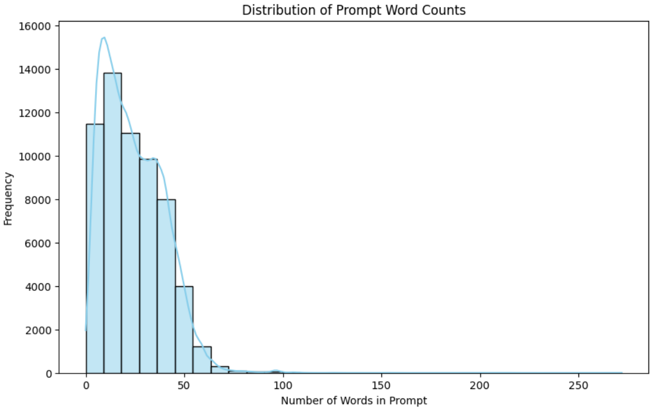

Figure 2. Visual representation of the prompts distribution in our dataset in terms of length of the prompt

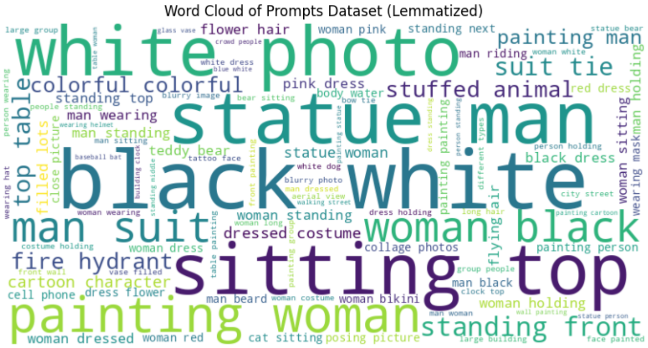

Figure 3. Word cloud visualization of the used prompts in our dataset

4.2. Verification Model Performance

Table 1 presents the experimental results from our image verification model, providing a comprehensive insight into the performance of various machine learning (ML) models. These results were evaluated based on key metrics: precision, recall, F1-score, and overall accuracy.

The XGBoost demonstrated a noteworthy performance, achieving a precision of 0.862877 and a recall of 0.816603. This high precision indicates its efficiency in correctly identifying pairs originating from the same source. It achieved an F1-score of 0.839103 and reported an accuracy of 0.842817, reflecting its overall effectiveness. In contrast, the LightGBM yielded a slightly lower precision of 0.861854 but surpassed in recall with a value of 0.836527

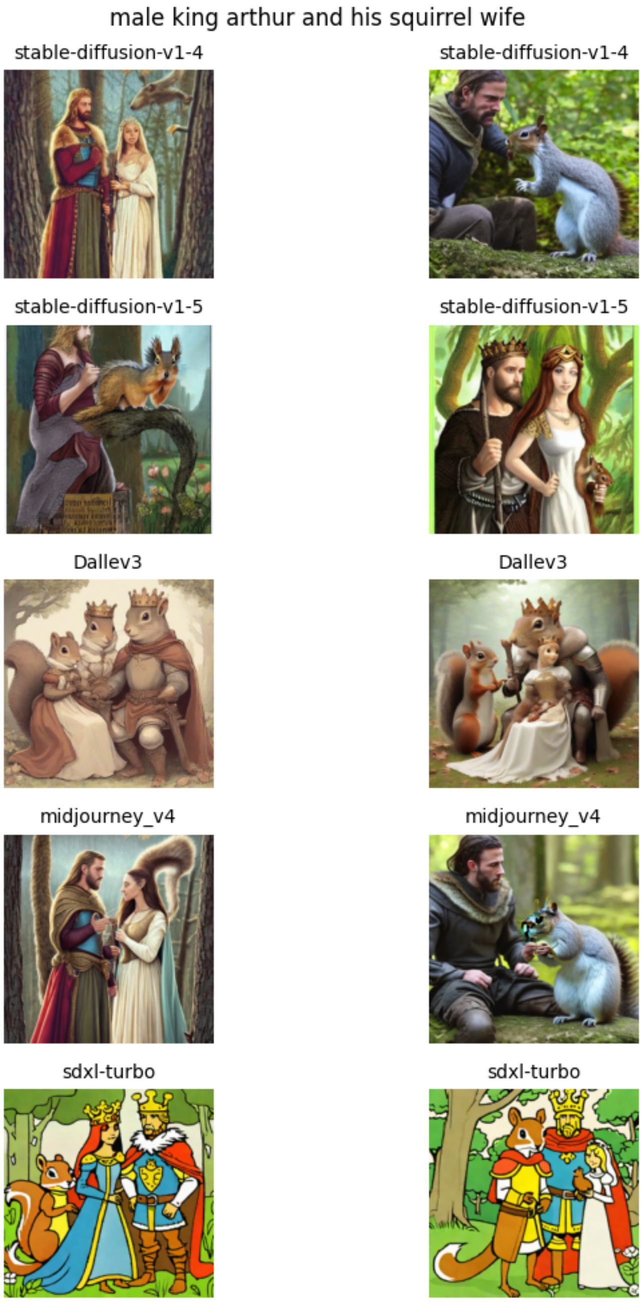

Figure 4. Visual representation of pair of generated images for a given prompt using five different text2image models

The model's F1-score stood at 0.849002, and it recorded the highest accuracy among the models at 0.850650, showcasing a balanced approach between precision and recall. The CatBoost also showed robust results, with a precision of 0.853565 and a recall of 0.843135. Its F1-score was 0.848318, and it attained an accuracy of 0.848667, indicating its strong performance in the verification task. On a different note, the Logistic Regression model under-performed relative to the others, with a precision of 0.628731 and a recall of 0.511273, leading to a lower F1-score of 0.563951 and an accuracy of 0.603167. This outcome suggests that Logistic Regression might be less suitable for this specific task. Finally, the Random Forest exhibited a precision of 0.855711 and a recall of 0.836560. It demonstrated its competence in the image verification domain with an F1-score of 0.846027 and an accuracy of 0.847167.

Overall, LightGBM emerged as the top performer in accuracy, closely followed by CatBoost and Random Forest. The XGBoost showed the highest precision, and all models, barring Logistic Regression, exhibited a strong balance between precision and recall, as indicated by their respective F1 scores. These findings demonstrate the ability to use simple ML models and a pre-trained backbone (without fine-tuning the backbone to the task) to accurately determine if a pair of generated images is of the same source.

Model	Precision	Recall	F1-score	Accuracy
XGBoost	**0.862877**	0.816603	0.839103	0.842817
LightGBM	0.861854	0.836527	**0.849002**	**0.850650**
CatBoost	0.853565	**0.843135**	0.848318	0.848667
Logistic Regression	0.628731	0.511273	0.563951	0.603167
Random Forest	0.855711	0.836560	0.846027	0.847167

Table 1. Performance of five different ML models on image verification task

4.3. Verification Model Sensitivity Analysis

To further examine how our model performed on the verification task, we measured how well each model succeeded in classifying pairs of images as the same source where they are actually from the same source and for pairs of images that are not from the same source. Figure 5 illustrates the confusion matrix of each model; the bottom row represents the model prediction where one is when the model classifies the images as the same source, and 0 is for not the same source. The values in the upper row are the labels of each pair of images, whereas the diagonal values are the values where the model was corrected for label one and label 0. We can see from the figure that all five models have a larger error ratio for predicting pairs of images as 1 (i.e., the same source) where the actual label is zero. This indicates the challenge of this problem since a pair of images from the same prompt but different sources can be visualized very similarly but are created from a different source.

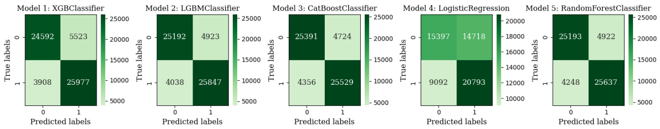

Figure 5. A confusion matrix of each of the five models on the verification task

In addition, we examined how well each model can accurately classify a pair of images for each type of text2image model. This examination can emphasize which text-to-image models are more challenging to classify (given a pair of these images) and which models can be classified more easily. Figure 6 presents the accuracy score for each of the ML models on each type of text2image model. We can see from the plot that all the models were very successful on images generated by the Dallev3 text2image model; the Logistic regression model, which got meager results, succeeded the most on images generated by this model. The main reason for this is that the style of images generated by the Dallev3 model is unique and very different from the other models (as seen from Figure 4). Additionally, we can see that for images generated from stable-diffusion1.4/5, the models got approximately the same accuracy score; this indicates that even though these models generated very similar images, our models can still distinguish between pairs of images of these models.

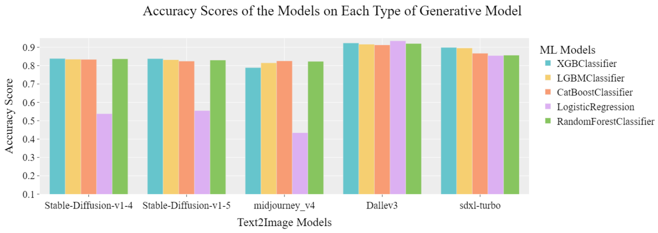

Figure 6. A bar plot of each model accuracy score for each type of text2image model

5. Discussion

This study introduces a novel approach to verifying the consistency of text-to-image models, a significant step forward in understanding and identifying the sources of AI-generated images. The methodology employed in this research demonstrates high accuracy in distinguishing images generated by different models and highlights the complexities and challenges involved in such tasks. The utilization of various machine learning models in combination with the Vision Transformer (ViT) for feature extraction underlines the potential of current AI techniques in addressing intricate problems in image verification.

One of the key findings of this research is the discernible differences in image styles produced by different text-to-image models, even when prompted with identical inputs. Our verification models have effectively captured this distinctiveness in style and characteristics, indicating a path toward more advanced image source identification methods. However, it is also observed that images generated from different versions of the same model (e.g., Stable-diffusion-v1-4 and v1-5) pose a higher challenge for accurate classification, suggesting the need for further refinements in the model.

Another important aspect of this study is the construction of the Diffusion Verification Dataset (DVD), which plays a crucial role in training and evaluating the models. The creation of this dataset from scratch, using a variety of prompts and models, sets a precedent for future research in this domain, particularly in terms of dataset creation and curation methodologies.

6. Future Work

Given the findings and the current limitations of our study, several areas can be explored in future research. Firstly, expanding the dataset to include more text-to-image models and different versions of existing models would be beneficial. This expansion would enhance the robustness of the verification models and help understand the nuanced differences between various model generations.

Experimenting with different feature extraction techniques and machine learning models could improve verification accuracy. Exploration of more sophisticated deep learning architectures, such as different variants of transformers or even custom neural network designs, could provide new insights into image source verification.

Additionally, adapting the methodology for real-world applications, such as detecting deepfakes or verifying the authenticity of digital media, represents a significant direction for future work. This adaptation would require the model to handle a broader range of image qualities and conditions, presenting an opportunity to test and enhance the model's practical applicability.

Finally, an intriguing direction for future research is the development of a classifier capable of not only verifying whether images are from the same source but also identifying which specific text-to-image model generated a given image. This advancement would entail a deeper analysis and classification of unique stylistic and structural elements inherent to

7. Acknowledgments

I want to thank Michael Fire for the guidance and advice during this research project and for providing the required tools to complete this task. In addition, I used ChatGPT for the programming part of the research, specifically for creating the DVD and for creating functions to plot the different figures presented in the GitHub repository.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[3] Jordan J Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 2024.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[5] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[7] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[9] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.