Distinguishing and Visulization of Animation Artworks between Human and Stable Diffusion

Yu Zhang, Mao Chenli
Department of Computer Science
Shanghaitech University
zhangyu9@shanghaitech.edu.cn

Abstract

We use Resnet to develop a classifier that can distinguish animation artworks between Human and Stable Diffusion and use this classifier to detect some frequency **features** of Stable diffusion generated artworks. We then use **fourier transform** to analyse artworks generated by Human and stable diffusion respectively. Finally, we found that diffusion generated model is more likely to generate **horizontal and vetical high-frequency signals** than human and that the high-frequency signals in human artworks are more well distributed in all directions.

1 Introduction

In recent years, AI generated contents are developing rapidly. The break through of latent diffusion[1] makes the animation artworks distinguishing even more difficult. Therefore, there's an urgent need to distinguish them and find the unique features that is only shared by diffusion models but not humans. There have been some researches on human perception of AIGC pictures[2] and assessments on the quality of AIGC pictures[3], but there isn't much study on the difference between works of human and diffusion.

In this work, we first build a dataset containing both animation artworks generated by human and diffusion animation model. And we use this dataset to train a resnet[4] based classifier. After that, we use back-propagation to visulize the feature that the classifier has leart and found that there're some frequency features in the animation artworks. Finally, we applied fourier transform to the dataset so that we can directly observe these features in the frequency domain, which is that diffusion generated model is more likely to generate **horizontal and vetical high-frequency signals** than human and that the high-frequency signals in human artworks are more well distributed in all directions.

2 Dataset preparation

Since we need to train a classifier, we need both animation artworks from human painters and diffusion models.

2.1 Human painter data

All the human painted artworks in our dataset comes from the website **danbooru**. Since most of the diffusion animation models excel in generating artworks that contains only one girl due to their training set mainly focusing on these works, I add the tag "1girl" to filter out these kind of works.

Preprint. Under review.

2.2 Diffusion generated data

For the diffusion generated data, we make use of multiple popular models nowadays, which are **Anything V3.0, Waifu 1.4, momoko-e**. By using multiple models, we can prevent the classifier from only learning information of painting styles to some extent such as the color features, line thickness, face charateristics, etc. These features are mostly model specified and may not be shared by all the diffusion animation models.

To improve the randomness and variety of the diffusion generated works, I use a random prompt generator to generate the prompt used for generating works. The generator first randomly choose several feature documents from the feature dir and then choose a random feature word from each choosen document. Finally, combine these feature words together to obtain a final prompt.

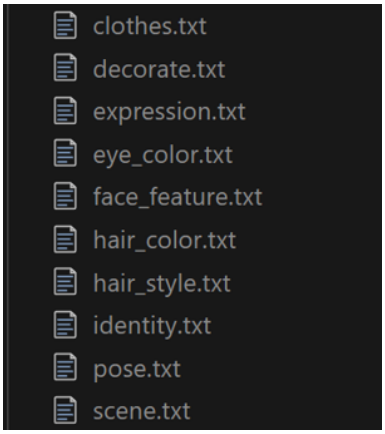


Figure 1: This is the feature dir

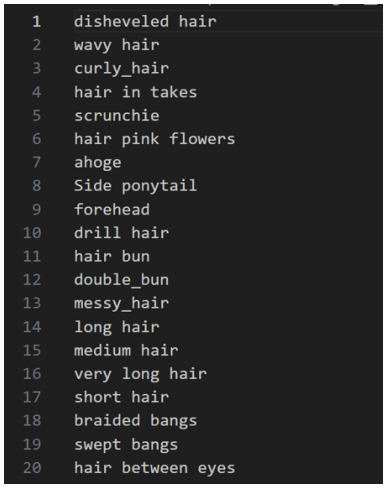


Figure 2: Featrue words from "hair_style.txt"

Besides these positive prompts, we also applied same negative prompts to all the works generated. As to the generated picture size, $\frac{1}{4}$ of the generated picture is squared and the rest are generated using Gaussian distribution to simulate the proportion of human dataset.

2.3 Dataset scale

In the end, we obtain a dataset with 6681 human works and 8819 diffusion works. We use 80% as training set, 15% as validation set and 5% as test set.

3 Classifier

3.1 Data Augmentation

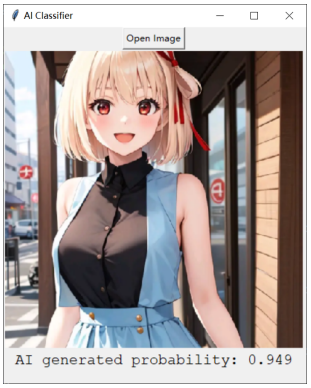
When the picture is fed into the classifier, its short edge will first be resized to 512 and maintain the scale ratio, and then random crop will be applied to crop a 512*512 region. We didn't directly resize the picture to 512*512 because this would break the scale ratio and may lose some information. We also did random flip, random rotation and normalization to the fed picture.

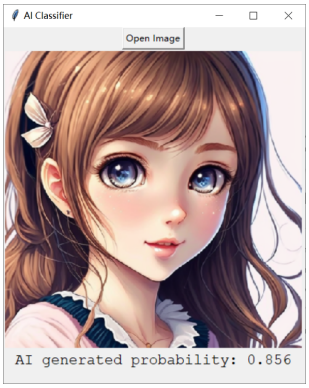
3.2 Model structure

We use a very simple model to build the classifier. We use Resnet101 as the base model. The model will take a 512*512 picture as input and output a vector containing 2 numbers, which indicates the possibility of the picture being human works and the opposite after they are transformed using a softmax function.

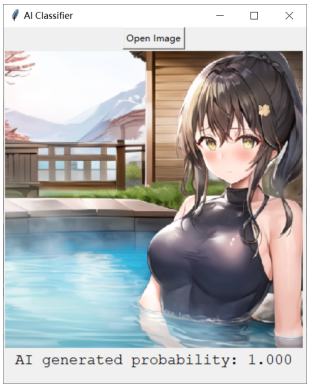
3.3 Model result

The model reaches 99.0% accuracy in the test set. As for the pictures beyond our dataset(other diffusion animation models), our model can also reach a great performance, which means that our model has the ability to detect the shared features of diffusion generated works regardless of the painting style. Here are some examples of the works that are generated from other diffusion animation models with different styles.





And for the same style artworks. The model can also distinguish the works from human and diffusion. Here's the example of momoko drawn picture(a famous painter) and a picture generated from the momoko-e diffusion model.



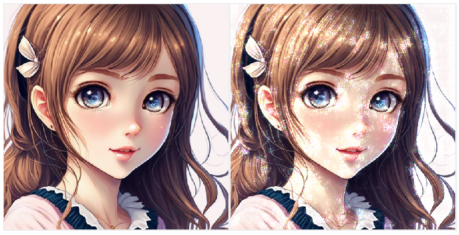
4 visulization

4.1 Gradient visulization

Since the our model can tell whether the picture comes from human or diffusion, if we treat the input picture as the parameter and get its gradient, we can acknowledge which part of the picture the classifier treats as a feature and uses to classify the picture. However, if we directly treats the gradient as depth and directly use heat map to represent it, the result wouldn't be clear since the distribution of gradients are very disperse. Therefore, I use the following equation to visualize the gradient.

$$X' = X + log(\frac{|\nabla X| + \beta}{\beta})$$

In the equation above, X' represents the visualized graph, X represents the original picture that was fed to the network, β represents the threshold(gradients below this threshold almost doesn't affect the X'), ∇X represents the gradients.
The point of adding a threshold β and a log function is that this can let us only focus on a portion of significant gradients so that the distribution are more concentrated and easier to be visualized.
Here're some examples of the gradient visulization.($\beta = 2e - 2$)



As you can see, we already can obtain a lot information from this visulization, for example, the intersection of lines and the intersection of light and shadow are the place where diffusion can't handle very well. However, making the graph lighter isn't very clear for a light color picture. Therefore, we use an opposite approach by substituting the add with minus.

$$X' = X + log(\frac{|\nabla X| - \beta}{\beta})$$

And in some diffusion pictures we found some unusual high-frequency patterns like the following.

4.1.1 Fourier visulization

Since there may be some high-frequency features in the diffusion generated works, it's very natural to apply fourier transform to analyse it. Therefore, we first apply transform to all the photos in our dataset, and get the average frequency-domain graph of both human drawn and diffusion generated picture.4.1.1

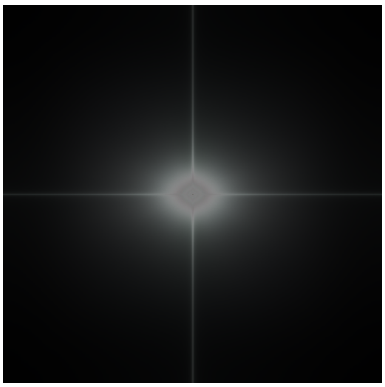


Figure 3: Diffusion average freq

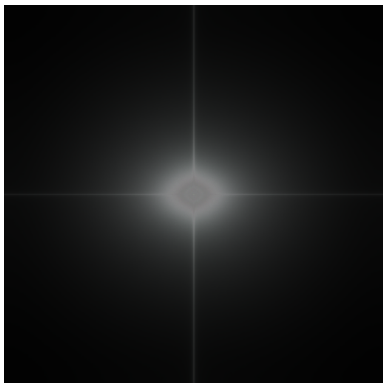


Figure 4: Human average freq

It's not easy to directly see the difference between them, therefore, we did some subtraction to them.

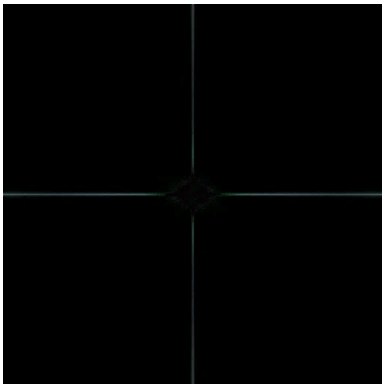


Figure 5: Diffusion - Human(brightness+50%)

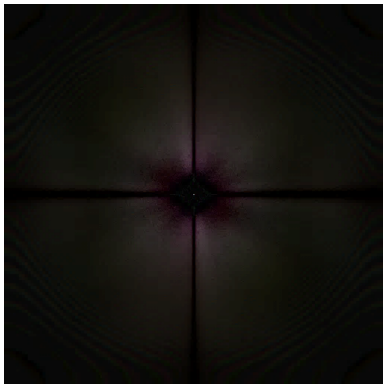


Figure 6: Human - Diffusion(brightness+50%)

As we can see from the frequency-domain subtraction graph, diffusion generated animation works has more horizontal and vetical high-frequency signals are more concentrated in the x,y axis. In the contrary, the high-frequency signals in human artworks are more well distributed in all directions. 4.1.1

To avoid that this is only the result of the bias caused by distribution difference in our whole dataset between human data and diffusion data. We also randomly choose a small batch of pictures (100 pictures) from both sets, and get the mean-frequency graph and do the same subtraction just like in the whole dataset for 10 times. And we obtain the same result for all the graph.

Here're four of the ten graph.4.1.1

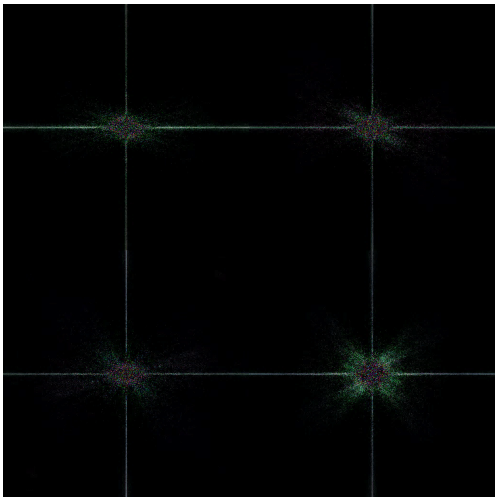


Figure 7: Diffusion - Human(brightness+50%) (small batch)

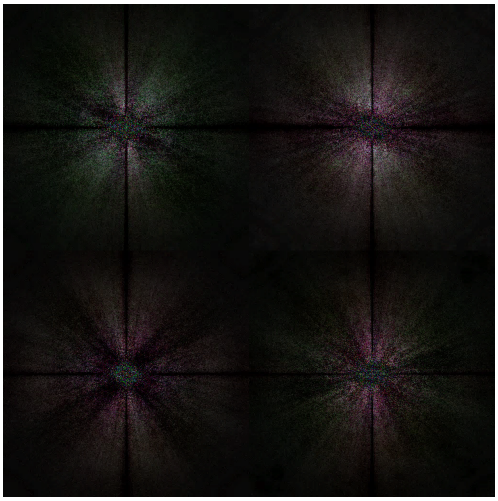


Figure 8: Human - Diffusion(brightness+50%) (small batch)

5 Conclusion

Though the diffusion model has achieved a great success in creating beautiful animation pictures and can mix the false with genuine in most people's eyes, there're still some ways to distinguish them. Using resnet based classifier can reach 99.0% accuracy. And the freq-domain graph also shows some frequency features in diffusion generated artworks. Be more specific, diffusion generated model is more likely to generate horizontal and vetical high-frequency signals than human and that the high-frequency signals in human artworks are more well distributed in all directions in the contrary.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P.,& Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

[2] Lu, Z., Huang, D., Bai, L., Liu, X., Qu, J., & Ouyang, W. (2023). Seeing is not always believing: A Quantitative Study on Human Perception of AI-Generated Images. arXiv preprint arXiv:2304.13023.

[3] Li, C., (2023). "AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment", arXiv:2306.04717.

[4] 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).