COMP576 Final Project Proposal: Distinguish AI Generated Images

Chuxuan Yin(cy48)

Nov 2022

1 Introduction

In the last mouths, AI painting suddenly began to attract people's attention. With the implementation of the stable diffusion framework, AI painting finally showed its great potential on large scale commercial usages. People who know nothing about painting could also easily use this model to generate high quality images in just one minute. While AI painting looked really awesome, it also raised some critques, one of these is that those generated images are very homogenized, so they are unable to substitue the hand-writting pictures. Although this is a very controversial topic, we could try to think this problem in a different way: do AI generated images really have some features in common? While it might be a hard task for human eyes to distinguish these slight differences, AI itself could be a good classifier. We could apply these traditional image classification models on the two different kind of images, and figure out whether the AI generated images really have "something in common".

Figure 1: Can we really distinguish which image is generated by AI?

2 Proposed Solution and Contribution

Although this is a rather new problem, we could still apply some mature methods in image classification. Nowadays, there are already many great pre-trained image classification models for us to choose, such as VCG-16[1], ResNet50[2] and AlexNet[3]. We

could train these models on our dataset. But these traditional models may perform bad on this problem since all these images may have same local features. We should instead use some fine-grained image classification models such as visual attention model[4], or modify those regular models to make them fit this problem. We may also need to understand the stable diffusion model[5], try to find a way to extract the common features for the images it generated.

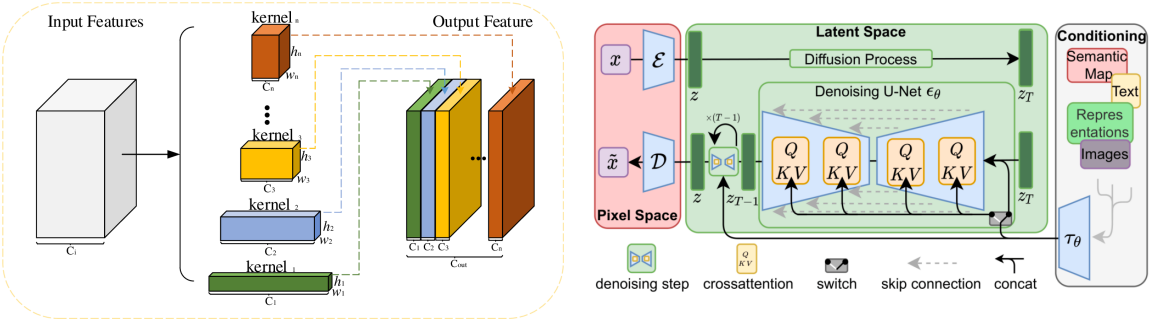

Figure 2: fine grained model and stable diffusion model

3 Goal

The goal of this project is to find whether we could distinguish AI generated images by current image classification technique. In order to do this, we need to try our best to build a classification model and figure out its ability. If the accuracy of prediction is greater then 60%, then we can declare that our network could distinguish AI generated images to some extent. Otherwise we should say even nerual network could not identify these AI generated images and we should admit they are really good images like hand-drawings.

Figure 3: Good or Bad?

4 Method

First we need to prepare our dataset. The key issue is that we should try to make the two genre look like pretty close at a whole, so that our classification model could mainly foucus on the tiny differences between AI generated and hand-drawings. So we will collect data based on tags: we collect images on three different kind of tags and for each kind of tag we will give a sufficient description to restrict them in a specific scenario. We will prepere training set and testing set for each tags and test the models' performance seperately. We will also mix them together to make a dataset, testing generalizability of the model. Considering our limited computation resources, we will preliminarily collect 300 images on each tag, 200 for training and 100 for testing.

Figure 4: Ganyu, head only

Figure 5: Castle, fantasy

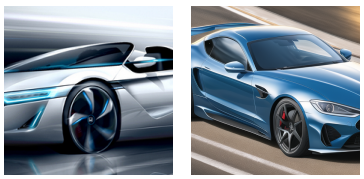

Figure 6: Sports car, realistic

Then we resize them to 200 $\times$ 200 images and put them into our classification models. We will try three different kind of models:

- Fine-tuning of pre-trained image classification model, such as VCG-16
- Fine-grained image classification models, such as vision attetion
- Self-developed classification model, based on CNN

As descriped above, we will train each model on each dataset with a specific tag, and then train all the models on a generalized dataset. Since we only care about the accuracy of prediction, the performance contrast of each model is not our interest.

5 Feasibility

The most challenging part of our plan is collecting data and training our model on limited computation resources. Due to these two reasons, we choose to use 200 images in our training set and 100 images in our testing set for each tag. Laion dataset, Pixiv and Google search provide tons of great images that we can use, while NovelAI and Wombo are good enough for generating useful AI images. With a dataset of this size, the computation resources provided by our own laptop, RiceNots and Google Colab is enough for us to use. If we found our time and computation resources is sufficient, we may consider to enlarge our dataset in further.

6 Project Execution Plan

This project will be done on one person. There are several steps we need to follow:

- Design a feasible plan and write a proposal for this project
- Collect data from website, resize them to 200 $\times$ 200 images
- Survey on image classification models and stable diffusion models
- Implement classification models we need with tensorflow or pytorch
- Conduct experiment on each model for our dataset, conclude results
- Formulate the final report, prepare for the presentation

7 Potential Impact

In this project we review the AI generated images at the view of AI. It really help us understand more about AI painting, let us carefully review the difference and connection between AI painting and human painting, which may affect how we use AI painting in the future. This project also focuses on the cutting-edge generative models and classification models. It is kind of like "generative algorithm" vs "classification algorithm", and the result may help us figure out the limitation of each model, enhancing our understanding of them.

References

[1] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs.CV]*, Sept 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, May 2017.

[4] Volodymyr Mnih, Nicolas Heess, Alex Graves and KorayKavukcuoglu. Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems*, June 2014.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *IEEE Conference on Computer Vision and Pattern Recognition*, Dec 2021.