Detecting AI-Generated Images using ResNet

Ziyang Zeng and **Zhehu Yuan** and **Yifan Jin**
Dept. of Computer Science
New York University
251 Mercer Street, New York, NY
zz2960@nyu.edu, zy2262@nyu.edu, yj2063@nyu.edu

Abstract

Abstract goes here.

1 Introduction

Introduction goes here.

2 Related Work

Related work goes here.

3 Datasets

In order to train a classifier that can distinguish between real and generated images, we need a labeled dataset that consists of both real and generated images for supervised learning. Datasets of this kind is not common, and we'd like to experiment on the latest state-of-the-art image generation AI models, such as DALL·E 2 and Stable Diffusion. We started with real photos from existing public datasets as our raw datasets. Then, we use Diffusers-based image-to-image model to generate images from real photos. This way, we can get a labeled dataset that consists of both real and generated images.

3.1 Raw Datasets

The original real-world photos we experimented with are from public datasets: Indoor Scene Recognition Database and Weather Image Recognition Dataset. Additionally, we want to test how well our model performs on non-photo art work. A dataset that consists of pages of comic art (the Comic Books Images Dataset) is also used in our experiments.

3.1.1 Indoor Scene Recognition Database

Originally targeting at indoor scene recognition tasks, the Indoor Scene Recognition Database is a collection of 67 indoor categories (e.g. airport, living room, restaurant...) and at least 100 images per category. There are 15,620 images in the dataset in total. Figure **??** shows some examples of the dataset.

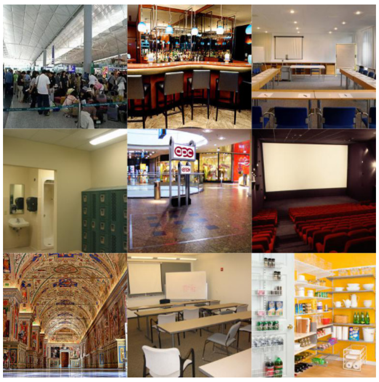

Figure 1: Indoor Scene Recognition Database Examples

This dataset is the biggest among the three datasets we used and serves as our main dataset. We use the AI-generated variations from these images and the images themselves to train our classifier.

3.1.2 Weather Image Recognition Dataset

The Weather Image Recognition Dataset contains labeled 6862 images of different types of weather, its catogories including dew, fog, frost, glaze, etc. This dataset is used in our experiment to evaluate our model's generalization ability and performance on other kinds of photos than indoors.

./assets/paper_weather_dataset.jpg

Figure 2: Weather Image Recognition Dataset Examples

3.1.3 Comic Books Images Dataset

3.2 Diffusers

3.2.1 CLIP guidance

3.2.2 Stable Diffusion

Some examples of the real-world pictures are:

Figure 3: airport

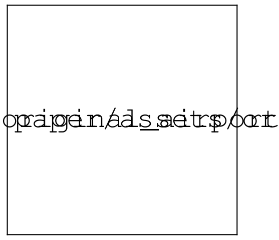

Figure 4: wine-Cellar

The corresponding AI-generated pictures are:

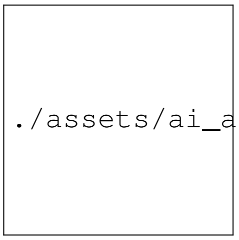

Figure 5: airport

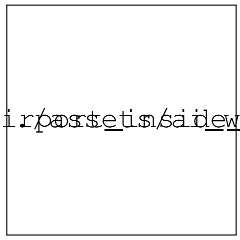

Figure 6: wine-Cellar

3.2.3 DALL·E 2

3.2.4 Image Variation Generation

4 Classifier

Classifier goes here.

5 Experiments

5.1 Experiment Settings

5.1.1 Training

The whole dataset includes 15620 real-world indoor pictures together with 15620 AI-generated pictures. Each real-world picture generates one AI pictures using the stable diffusion model.

Around 80% pictures for each type are used for training.

In terms of the model, we use the resnet34 intergrated inside the fastai.

5.1.2 Validation

We picked around 20% pictures of the above indoor pictures for each type as the validation dataset. The validation data is not used in the training process.

5.1.3 Testing

The testing dataset includes several types of pictures:

(1) 6738 real-world weather pictures together with 6738 AI-generated pictures. Each real-world picture generates one AI pictures using the stable diffusion model.

(2) Comic pictures

(3) Dalle dataset where AI pictures are generated by dalle model, not stable diffusion model.

5.2 Evaluation Metrics

To evaluate the performance of the model, we used accuracy, precision, recall and F1-score as evaluation metrics. Below are the formulas for calculating the accuracy, precision, recall and F1-score. Here we assume the AI-generated pictures as positive labels while original pictures as negative labels. Thus, TP is the number of AI pictures being correctly predicted as AI pictures, FP is the number of original pictures being wrongly predicted as AI pictures. Similarly, TN is the number of original pictures being correctly predicted as original pictures, FN is the number of AI pictures being wrongly predicted as original pictures.

$$Accuarcy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

6 Experiment Results

Experiment results go here.

7 Conclusion

Conclusion goes here.

8 Discussion and Future Work

Discussion and future work go here.

References

paper/training_confusion_matrix.jpg

Figure 7: Validation Confusion Matrix