RD Vision Takehome Assessment

Vuong Nguyen

1 Design an AI-generated Fake Image Detector

1.1 Idea Explanation

- To generalize the detector over various generative models like GANs or Diffusion models, I think rather than directly training a binary classifier from the spatial domain, we need to extract a universal artifact/fingerprint across the generative models.
- During generative process, from random noise, it is difficult for generative models to approximate a distribution with high entropy, i.e. large number of image categories.
- Due to entropy discrepancy between poor texture regions and rich texture regions in an AI-gen image, these two regions could behave differently, leaving artifacts.
- We can leverage this artifacts for detecting fake images. Artifacts can be represented by inter-pixel correlation fingerprint between rich and poor regions. For real images, their inter-pixel correlations of different regions behave consistently, unlike fake images.
- Fake images behave abnormally in high frequency [2, 1]. High frequency or noise of fake images exhibits large gaps from real images. Thus, rather than classifying on spatial domain, I transform it to some high frequency domain and pass this processed output to a classifier.

1.2 Model baseline

Model baseline is shown in Figure 1. An image is first broken into patches. Then, for each patch, texture diversity is computed. 64 patches with richest texture diversity are used to reconstructed rich texture image, while 64 patches with poorest texture diversity are used to reconstructed poor texture image. These two reconstructed images are then passed through a set of high pass filters to extract noise patterns, then passed to a convolutional block. Fingerprint is then measured by the residual of the outputs of the learnable convolution block. Finally, fingerprint is fed into a classifier to classifier fake or real image.

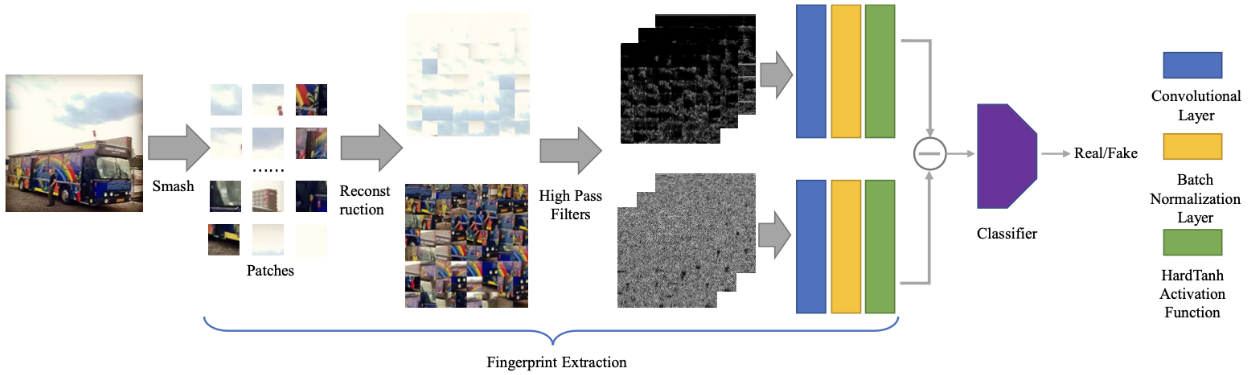

Figure 1: Model Baseline with three steps: 1) Extract rich texture and poor texture regions, 2) Get artifact (fingerprint) from these two parts, and 3) Classification.

1.3 Discussion

Strengths: High generalizability across generative models, since the fingerprint feature is based on the inherent weakness of distribution approximation during generative process.

2 How to extend to videos?

The approach can be extended to video easily by classifying one or many random frames extracted from the video.

Similar to images in the real world, when a video is uploaded to a media platform (like YouTube or Twitter) it may get re-encoded by a proprietary codec. Encoding can be another adversarial attack which comes from manipulation of encoding parameters.

In this case, I guess we can leverage compression artifacts for detection

3 In-painting Detection

The step that breaking the original image into patches is beneficial for in-painting detection where only part of an image/video is fake. We can extract artifacts among patches using inter-pixel correlation, then use the artifacts to detect anomaly occurring in some in-painted patch.

References

[1] Frank, Joel, et al. "Leveraging frequency analysis for deep fake image recognition." In *ICML*, 2020.

[2] Durall, Ricard, Margret Keuper, and Janis Keuper. "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions." In *CVPR*, 2020.