Machine Learning Report: Lab 5

Archishman Ghosh
Department Of Computer Science
Amrita School Of Computing
Bangalore, India
bl.en.u4cse22103@bl.students.amrita.edu

Dhruv Vijay Kolhe
Department Of Computer Science
Amrita School Of Computing
Bangalore, India
bl.en.u4cse22111@bl.students.amrita.edu

Soumish Ghosh
Department Of Computer Science
Amrita School Of Computing
Bangalore, India
bl.en.u4cse22154@bl.students.amrita.edu

***Abstract*—In this paper, we explore various machine learning methodologies, including convolutional neural networks, generative adversarial networks, and hybrid models, which are used to differentiate between real and AI generated images. We also examine the key features and markers that are used to separate AI-generated images, such as texture analysis, frequency domain transformations, and inconsistencies in pixel-level statistics. Through a critical analysis of recent studies, we highlight the strengths and limitations of existing methods and identify emerging trends in this evolving field. Our findings underscore the necessity for strong and adaptive detection mechanisms to keep pace with the advancing capabilities of AI in image generation.**

***Index Terms*—Machine Learning, Classification, kNN, image classification**

I. Introduction

The widespread proliferation deep learning and AI technologies has resulted in widespread access to the ability of creating highly detailed and realistic images using generative models such as Stable Diffusion, Dall-E 2 and others. Such AI generated images have seen an extremely rapid rise in adoption across a range of industries such as entertainment, social media and advertising.

However, there is a growing concern in the potential of synthetic imagery to cause cause great disruptions in society through their use to spread misinformation, blackmail via deepfakes and to breach copyright laws. This has heightened the need of the creations of robust systems to detect AI generated images. Such systems can help in detecting misinformation and thus enable its prevention.

Given an image, our system is designed to classify it as either realistic or unrealistic, aiding in the detection of AI-generated content. By distinguishing between images that exhibit natural characteristics and those with features indicative of artificial creation, our approach helps identify potentially synthetic images. This classification provides a reliable method for discerning whether an image is likely to have been generated by AI.

II. Related Work

Chen et al.[2] talks recent advancements in generative models, such as GANs and diffusion models, which have enabled us to create realistic fake images, raising concerns about potential misuse and highlighting the need for effective detection methods. Traditional detection methods involves various techniques such as spatial and frequency domain methods. While data augmentation has improved generalization capabilities of detectors, challenges remain with diffusion models due to their unique attributes. Till date approaches, such as two-branch methods and pretrained models, have faced issues with robustness and generalization. Overall it proposes Single Simple Patch (SSP) network to enhance detection accuracy and generalization across various generators by focusing on extracting noise fingerprints from simple patches.

Corvi et al.[3] highlights advancements in synthetic media creation that has been done using Generative Adversarial Networks(GANs) and diffusion model(DMs). This portrays their ability to produce quality realistic images. Detection techniques which had been focused on GANs, are now switching to address DMs, which exhibit features like asymmetrical shadows and lighting inconsistencies due to lack of 3D modeling. The importance of data augmentation and training on diverse datasets is emphasized to improve robustness.

Moskowitz et al.[5] present an approach to detect AI generated image using the Contrastive Language–Image Pre-training (CLIP) model. This study leverages CLIP's ability to associate image embeddings with corresponding textual descriptions. The authors fine-tune CLIP by feeding it images from a diverse dataset containing both images generated from GANs as well as human generated images, each of which has been captioned to enhance the model's efficiency while learning. The authors claim that the fine-tuned CLIP model outperforms other approaches such as CNNDet and DIRE, in terms of accuracy, precision, recall, and F1 scores, particularly in detecting images from recent and sophisticated generative techniques.

Ojha el at. [6] explores various methods for manipulating images, such as DeepFakes demonstrating the advancement of image generation technologies. Traditional detection techniques rely on identifying alterations in image statistics, like compression artifacts and irregular reflections, while recent studies have utilized the frequency space to detect distinct artifacts in GAN-generated images. A significant challenge is the poor generalization of classifiers trained on specific generative models to other models. To address this, this study propose using frequency space for classification, effectively capturing artifacts in images from models like CycleGAN and StarGAN. The study also portrays using features from a CLIP-

ViT model for classification, which shows improved generalization capabilities and outperforms traditional classifiers on unseen generative models.

Yuan Rao et al.[8] presents a novel approach for image forgery detection. It uses convolutional neural network to learn hierarchical representations from input RGB colour images. The CNN is specific to image splicing and copy-move detector applications. The authors initialize the weights of the first layer of the CNN with high pass filter sets. These sets are used in finding the residual maps in spatial rich model. It acts as regulator to actively suppress the influence of image contents and capture the minute details which are used for forgery. A pre-trained model is used to get the detailed features from the test image data. This paper doesn't deal with other image types except RGB. It also has a high computational complexity and dependency on the quality of the pre-trained model.

v.Sasikala et al.[9] states that swarm intelligence is a rarely used technique in detecting fake and real image fingerprint classification researches. It is robust and accurate in tackling complex optimization problems. The fingerprint classification method used involves four key steps: image preprocessing, feature extraction, feature selection, and classification. For preprocessing, the method uses min-max normalization and median filtering. Multiple still attributes are extracted using Gabor filtering. The selection of optimal static features is achieved through the Artificial Bee Colony and Modified Artificial Bee Colony optimization algorithms. It chooses the best features relying on specific fitness values. The classification is done using a Fuzzy Feed Forward Neural Network. It differentiates between fake and real fingerprint images using a partial-supervised graph-based classification approach.

Wang et al.[10] explore the feasibility of creating a universal detector capable of distinguishing real images from those generated by a wide range of CNN-based models. The study utilized a dataset comprising fake images from 11 different generative models, including ProGAN, StyleGAN, and BigGAN, and discovered that classifiers trained on images from a single model (e.g., ProGAN) could surprisingly generalize well to images from unseen architectures and datasets. The research highlights that CNN-generated images exhibit common artifacts or "fingerprints," making them distinguishable from real images. The authors emphasize the importance of data augmentation and preprocessing techniques, such as JPEG compression and resizing, to enhance the model's generalization ability.

The detection of AI-generated images by existing methods performs poorly for images generated by increasingly sophisticated diffusion models.Wang et al.[11] highlight the failure of overfitting of a simple binary classifier trained on a dataset of real and diffusion generated images. The authors propose a new method using DIRE(Diffusion Reconstructed Error), which measures the error between an input image and its reconstruction counterpart by a pre-trained diffusion model. They leverage the idea that diffusion-generated images can be more accurately reconstructed by a pre-trained diffusion model compared to real images, which exhibit more complex characteristics and thus cannot be reconstructed as well. By calculating the reconstruction error between an input image and its counterpart produced by a pre-trained diffusion model, DIRE provides a distinguishing metric: lower errors indicate generated images, while higher errors suggest real images. The authors validate this approach using a newly created dataset, DiffusionForensics, which includes images from eight different diffusion models. Their experiments demonstrate that the DIRE method achieves superior performance in terms of detection accuracy and robustness compared to existing methods, highlighting its generalization capability to new, unseen diffusion models.

Zhou et al.[12] gives an outline of existing methods in image forgery detection, including noisy inconsistency, CFA pattern estimation, multi-task edge-enhanced FCN, and joint training with LSTM. These use many attributes such as colour filter array details, noise patterns, and edge inconsistencies to detect manipulated regions in image data.The proposed RBG-N network in this paper combines noise performance with RBG images using bilinear pooling. This gives a boost in the performance than the previous methods on standard datasets.

Zhu et at.[13] talks about the advancements in generative models in creating realistic images, which can be misused to spread wrong information, especially in sensitive areas like politics. Despite various detection methods, distinguishing real from fake images remains problematic with only a 61.3 percent accuracy. This case study introduces the GenImage dataset, containing over one million pairs of real and AI-generated images from GANs and diffusion models, enhancing training for detection methods. It proposes two evaluation tasks cross generator and degraded image classification to assess detector generalization and performance on images.

III. Methodology

A. Dataset

We use the CIFAKE dataset which contains real images gathered from Krizhevsky and Hinton's CIFAR-10 dataset for the real images[4] and images generated by Bird et al. [1].There are 100,000 images for training (50000 per class) and 20,000 for testing (10000 per class).

B. Feature Extraction

The Fast Fourier Transform transforms an image from the spatial or pixel representation to the frequency domain, where it is represented as a combination of sinusoidal components corresponding to different frequencies, enabling us to explore underlying patterns and structures not easily discernible in the spatial domain.

Low frequencies in the transformation correspond to gradual variations, for example gradual variation in brightness whereas high frequencies represent fine details such as edges and noise. The identification of such characteristics is crucial as AI generated images might display smoother textures and repeating patterns.

We compute the magnitudes of each of the elements in the transformed image for our features, which reflect the strength of the various frequency components present in the image

C. Data Preparation For Regression

Let X denote the FFT output vector of length n, where $X = [X_0, X_1, X_2, \ldots, X_{n-1}]$. Here, X_0 represents the DC component. The feature vector $\mathbf{F}$ used for regression consists of the subsequent FFT coefficients, $\mathbf{F} = [X_1, X_2, \ldots, X_{n-1}]$.

D. Model Formulation

1) Linear Regression: The linear regression model is formulated as follows:

$$\hat{X}_0 = \beta_0 + \sum_{i=1}^{n-1} \beta_i X_i$$

Where:

- $\hat{X}_0$ is the predicted DC component.
- β_0 is the intercept term.
- β_i (for $i = 1, 2, \ldots, n-1$) are the regression coefficients corresponding to the FFT coefficients X_i.

2) K-Means Clustering: Here, we use the FFT coefficients as the feature-set. The K-Means algorithm is then applied to these feature vectors with the goal of grouping similar images together. The script iterates through a range of cluster numbers (K) and evaluates the performance of each clustering using metrics like the Davies-Bouldin Score, Silhouette Score, and Calinski-Harabasz Score.

E. Model Training

1) Linear Regression: To train the model, we utilize a dataset comprising multiple FFT outputs from various signals. The regression model is trained by minimizing the mean squared error (MSE) between the actual and predicted DC components. The performance of the model is assessed through standard metrics such as R-squared (R^2),root mean square error (RMSE), and mean squared error (MSE) enabling the evaluation of the model's predictive accuracy.

2) K-Means Clustering: We iterate over K from 2 to 20, fitting the model for each value and calculating various performance metrics such as Davies-Bouldin Score, Silhouette Score, Calinski-Harabasz Scores. The results are summarised in III. We then find the optimum value of K by using the algorithm described by Paez et. al. [7]

IV. Results And Analysis

A. Linear Regression

From II, we can see that the all the errors are extremely high. This is because the DC value represents the average value of the time-domain signal, while the fft contains the amplitude of the frequencies excluding the DC component. There is often no straightforward or linear correlation between these components.

	Actual DC Values	Predicted DC Values
0	104620.000000	115406.138900
1	98964.000000	118147.638900
2	137406.000000	121454.138900
3	109513.000000	108708.638900
4	98082.000000	105498.138900
5	110676.000000	121670.138900
6	83590.000000	115858.138900
7	105124.000000	128102.138900
8	98713.000000	119355.138900
9	153735.000000	115858.138900
10	91246.000000	111652.138900
11	109603.000000	136440.795150
12	131645.000000	118298.138900
13	110578.000000	118944.138900
14	105076.000000	110797.638900
15	96920.000000	121138.138900
16	112469.000000	108544.638900
17	100748.000000	115898.138900
18	148639.000000	124494.138900
19	150883.000000	119338.826400
20	101425.000000	109027.638900

TABLE I
COMPARISON OF ACTUAL AND PREDICTED DC VALUES.

R2-Score	RMSE	MSE
26635.333651	709440998.688633	0.070179

TABLE II
PERFORMANCE METRICS OF THE LINEAR REGRESSION MODEL.

B. K-Means Clustering

K	Davies-Bouldin Score	Silhouette Score	Calinski-Harabasz Score
2	1.836544	0.188282	5594.237559
3	2.155781	0.157203	4058.179983
4	1.988657	0.149200	3436.056640
5	1.918204	0.107182	3039.188757
6	2.182175	0.104831	2665.808842
7	2.124358	0.103585	2437.064482
8	2.085248	0.102616	2229.893722
9	2.111297	0.078787	2060.235376
10	2.316705	0.076146	1896.491410
11	2.335021	0.087901	1748.447594
12	2.283087	0.077078	1666.536401
13	2.286571	0.075017	1571.908952
14	2.243670	0.067880	1491.456174
15	2.352766	0.066318	1413.972123
16	2.410532	0.062754	1344.951792
17	2.430614	0.062341	1285.232579
18	2.460620	0.060935	1225.887605
19	2.458553	0.051818	1176.014958
20	2.458672	0.050920	1131.962862

TABLE III
CLUSTERING METRICS FOR DIFFERENT VALUES OF K.

1) Davies-Bouldin Score: The Davies-Bouldin Score (DBI) measures the average similarity ratio of each cluster with its most similar cluster. A lower DBI indicates better clustering, where clusters are more distinct from each other. From the table, the DBI value generally increases with K, starting from 1.836544 for $K = 2$ and rising to 2.458672 for $K = 20$. This trend suggests that as K increases, clusters become less distinct relative to each other, implying that too many clusters might create more overlap or less meaningful distinctions.

2) Silhouette Score: The Silhouette Score measures how similar each point is to its own cluster compared to other clusters. It ranges from -1 to 1, with a higher score indicating better cluster separation and cohesion. The table shows that the Silhouette Score generally decreases with increasing K, starting from 0.188282 for $K = 2$ and dropping to 0.050920 for $K = 20$. This suggests that as the number of clusters increases, the clusters may become less cohesive or more poorly separated, leading to lower scores.

3) Calinski-Harabasz Score: The Calinski-Harabasz Score (CH Score) evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. A higher score indicates a better-defined clustering structure. The CH Score decreases as K increases, starting from 5594.237559 for $K = 2$ and decreasing to 1131.962862 for $K = 20$. This decreasing trend suggests that increasing K leads to clusters that are less well-defined, with higher within-cluster dispersion relative to between-cluster dispersion.

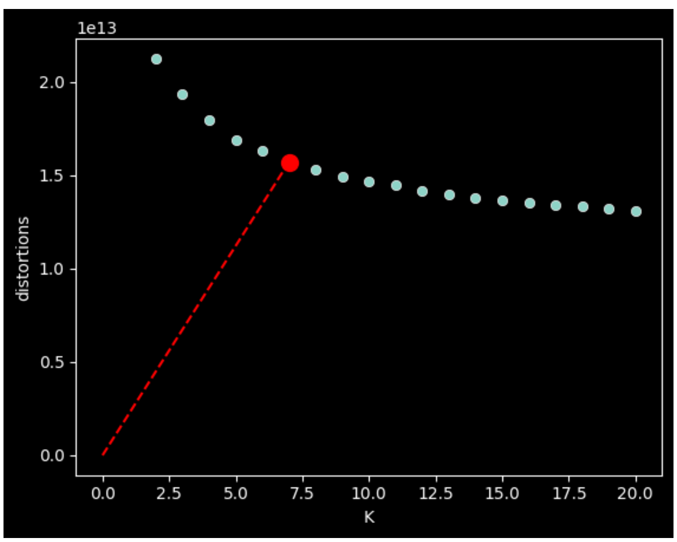

Fig. 1. Elbow Graph

From the 1, we see that the value of K which produces the most optimum clustering is 7

References

[1] Jordan J. Bird and Ahmad Lotfi. *CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images*. 2023. arXiv: 2303.14126 [cs.CV]. URL: https://arxiv.org/abs/2303.14126.

[2] Jiaxuan Chen, Jieteng Yao, and Li Niu. "A single simple patch is all you need for ai-generated image detection". In: *arXiv preprint arXiv:2402.01123* (2024).

[3] Riccardo Corvi et al. "On the detection of synthetic images generated by diffusion models". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

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

[5] A. G. Moskowitz, T. Gaona, and J. Peterson. *Detecting AI-Generated Images via CLIP*. 2024. arXiv: 2404.08788 [cs.CV]. URL: https://arxiv.org/abs/2404.08788.

[6] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. "Towards universal fake image detectors that generalize across generative models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 24480–24489.

[7] Barath Á. Páez. *Kneedle: An Algorithm for Detecting "Knee" Points in a Curve*. Technical Report TR-2011-04. ICSI, University of California, Berkeley, 2011. URL: https://www1.icsi.berkeley.edu/~barath/papers/kneedle-simplex11.pdf.

[8] Yuan Rao and Jiangqun Ni. "A deep learning approach to detection of splicing and copy-move forgeries in images". In: *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2016, pp. 1–6. DOI: 10.1109/WIFS.2016.7823911.

[9] V. Sasikala and V. LakshmiPrabha. "A comparative study on the swarm intelligence based feature selection approaches for fake and real fingerprint classification". In: *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*. 2015, pp. 1–8. DOI: 10.1109/ICSNS.2015.7292421.

[10] Sheng-Yu Wang et al. "CNN-Generated Images Are Surprisingly Easy to Spot... for Now". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8692–8701. DOI: 10.1109/CVPR42600.2020.00872.

[11] Zhendong Wang et al. "DIRE for Diffusion-Generated Image Detection". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 22388–22398. DOI: 10.1109/ICCV51070.2023.02051.

[12] Peng Zhou et al. "Learning Rich Features for Image Manipulation Detection". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1053–1061. DOI: 10.1109/CVPR.2018.00116.

[13] Mingjian Zhu et al. "Genimage: A million-scale benchmark for detecting ai-generated image". In: *Advances in Neural Information Processing Systems* 36 (2024).