Team 81: Detecting AI generated images

Anthony John Dsouza
7053485

Eva Morgan Toulouse Gavaller
7056847

1. Task and Motivation

1.1. Task statement & Motivation

With recent advancements in AI, generative models like the transformer, diffusion, SSMs make use of vast & diverse training data to generate text, images and audio indistinguishable from human generated data. These technologies have found use in machine translation, human machine interaction, image inpainting, text to video / image generation, etc. While many of these applications are beneficial to society, some bad actors use these tools for nefarious purposes, like spreading misinformation and revenge pornography. During this project we inted to explore methods to detect deepfakes and AI-generated media. To accomplish this, we intend to collect an AI-generated visual dataset, and annotate it.

We have 2 closely related ideas that we can look to implement

1. Idea 1 : Vision Language Model to detect AI generated images using reasoning
2. Idea 2 : Interpreting finetuned ViT hidden states

1.2. Related work

Most common approaches use a simple setup, by finetuning a CNN / Vision Transformer [9] with some simple augmentations on datasets like [8, 20, 24, 34]. The approach considered here is 'classification', where the detector tries to identify artifacts generated by the generators. The DeepFake Detection challenge by Meta and Kagggle [1] saw many submissions using the above approaches with the best performing submission [26] being an ensemble of many models with augmentations and more importantly face warping artifacts [19]. Other approaches used [11, 30, 31] for improved generalization.

In [10], the authors use motion magnification to amplify imperceptible movements in videos, making them observable. Initially developed to detect subtle vibrations in structures like high-rise buildings and bridges, this technique revealed that synthetic videos exhibit a higher flicker rate compared to real videos. To analyze these differences, the authors use InceptionV3 [28] as the backbone for extracting spatial features from video frames and an LSTM [13] to process temporal data, enabling effective discrimination between real and synthetic videos based on motion patterns.

In [33], authors reframe deepfake detection as a common sense reasoning task, aiming to address limitations of image classifiers that fail to identify subtle inconsistencies, such as unnatural skin tones or duplicate features like double eyebrows. Using a language model, the method reasons whether an image is real or fake, specifically highlighting inconsistencies in the temporal domain. This allows for a more nuanced analysis of visual content beyond standard detection techniques. For this, the authors use a BLIP [18] backbone, integrating a BERT [6] like text encoder and a Vision Transformer [9] based image encoder, with cross-modal attention mechanism for vision language grounding between the two modalities. The textual outputs generated by the encoders are then fed into the language model that reasons if the provided image is real or fake.

SurFake's [5] approach is based around concept that pixels in an image contain critical information about scene geometry and the acquisition process, which is often altered by deepfake generators, leaving detectable traces. By analyzing these pixel-level changes, SurFake aims to identify manipulations that traditional methods might overlook. The detection pipeline consists of three key steps- First, it performs face detection and extraction to isolate the relevant facial region. Next, a Global Surface Descriptor, based on UpRightNet [29], is used to capture the geometrical characteristics of the extracted face. Finally, these features are processed by a convolutional neural network (CNN), which analyzes the geometric data to determine whether the image has been manipulated, enabling robust deepfake detection.

[14] created the Social Media Image Detection Dataset, comprising 300,000 images that include real, AI-generated, and tampered images. This dataset emphasizes broad diversity and realism to ensure it reflects the complexities of images encountered on social media platforms, providing a robust resource for developing and evaluating detection mod-

[1] DeepFake Detection Challenge

els. For detection, the authors improve on the LISA [17] model by introducing two new tokens to its vocabulary: one for detection <DET> and one for segmentation <SEG>. The model takes an image and a prompt, such as "Can you identify if this image is real, fully synthetic, or tampered?" as input. It then generates a text description explaining its reasoning, while the last hidden layer contains the two additional tokens. The <DET> token identifies whether the image is manipulated, and if tampering is detected, the <SEG> token highlights the specific regions that have been altered, enabling precise localization of modifications.

1.3. Challenges

A significant portion of methods for deepfake detection systems were built before diffusion [12] was introduced-mainly due to the fact that training these models was computationally expensive. Soon, [7] showed that diffusion-based image generation models were better at image generation while being stable in comparison to their counterparts. [23], [21] build on this for improved generation, with methods developed to control the generation process like [25,32] and using fewer steps for denoising as in [21]. All these result in hyperrealistic image generation with few artifacts, making the process of detecting deepfakes highly difficult.

2. Goals

2.1. Challenges we aim to address

Our main goal is to detect AI generated images, with secondary goals being interpreting the hidden states of the ViT classifier using mechanistic interpretability / look at if logical reasoning can help in the detection of AI generated content.

2.2. Proposed Timeline

We have the following timeline for the project :

1. Week 1 & 2 (16.06 - 27.06) : dataset collection and labelling, writing the codebase and prototyping
2. Week 3 (30.06 - 06.07) : Model training, evaluations, writing reports.
3. Week 4 (07.07 - 12.07) : additional experiments, if needed

3. Methods

Since we have 2 ideas, we thought we could explain both of them here.

3.1. Vision Language Model to detect AI generated images

Modern AI-generated images often appear hyperrealistic, resembling works of art, yet they are not free from artifacts, with subtle logical inconsistencies that set them apart from real images. These artifacts, such as airbrushed skin, six fingers, weird illumination, or physically impossible body postures are easily noticeable to the human eye, allowing our logical reasoning to distinguish between real and synthetic images. However, vision-based models struggle to perform this kind of commonsense reasoning. To address this, we complement the vision encoder with a large language model (LLM) that can perform reasoning to determine whether an image or video is AI-generated, leveraging the strengths of both modalities for more accurate detection. So our approach is very similar to [33], with 2 key differences; the most important being the difference in data; variety of hyperrealistic images from more recent models and more logical explanations and different multimodal models like [1, 3, 4, 27]

3.2. ViT-based classifier

3.2.1 Stage 1: Training classifier

In this stage, a ViT model [9] will be fine-tuned on the collected dataset so that the model is able to learn features necessary for classifying between real and AI-generated images. The reason a vision transformer is chosen over a CNN is because the former learns robust representations from the data.

3.2.2 Stage 2: Interpretability - SAE and Attention rollout

The main challenge in interpretability is polysemanticity, the notion that neurons in a neural network can activate from multiple, unrelated inputs, likely due to a model using neurons efficiently, in order to leave other neurons available for important tasks [22]. Polysemanticity is attributed to superposition, where models represent more concepts than available neurons, leading to concepts being represented by various neurons at once [15]. This in turn makes interpreting specific model features difficult, which is why we use SAEs. SAEs create a matrix of sparse, more interpretable, encoded representations from LLM input [15]. Encoded representations can also be called features. These features may activate based on different inputs, and through these activations, we can better interpret what information these features activate on. Additionally, as SAEs are given the incentive to create sparse vectors as a result of imposing L1 loss, we get only a few non-zero values, which are theoretically the "true" features of the model [16].

The SAE will be trained by harvesting hidden representations from the model trained in 3.2.1 and help us understand what features the model looks at to classify between real and AI generated images.

Attention rollout [2] is another method to make later attention layers more interpretable.

4. Datasets

For this project, we intend to collect publicly available AI generated images and videos from social media sites like Facebook, Reddit, X, Instagram, etc. After collection, we plan to annotate them. This is because there is no publicly available dataset that is regularly updated to keep up with new generators like StableDiffusion, FLUX, etc. Our dataset needs to have the ground truths of the image / video and also a textual and logical description as to why it is labelled so.

5. Evaluation

5.1. ViT classifier finetuning

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (1)$$

For evaluating the ViT-based model, our metrics will be Accuracy and LogLoss / BCE Loss. We will use LogLoss because it penalizes classifications that are confidently wrong i.e. the model will be penalized when it assigns a low probability to a true event. Another factor we consider is that LogLoss is better suited for situations where confidence of prediction matters, as it looks at the probabilities of classes.

5.2. Vision-Language model

In this case, we consider the LogLoss as in 1 for classification and also inter-annotator agreement for the reasons generated by the LM backbone. The latter also provides us insight on what features the model reasons are important to observe.

Use of AI

We use Grok [2] by `x.ai` to improve the reading experience of teh reader. We use the following prompt

> Read the following text and suggest possible improvements for better reader understanding. You cannot add or remove any information, but only suggest rephrasing or spelling errorss.

We then may or may not incorporate the changes suggested by Grok in the final draft.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 2

[2] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers, May 2020. 2

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[5] Andrea Ciamarra, Roberto Caldelli, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Deepfake Detection by Exploiting Surface Anomalies: The SurFake Approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1024–1033, 2024. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

[2] Grok

pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 2

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset, Oct. 2020. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. 1, 2

[10] Jianwei Fei, Zhihua Xia, Peipeng Yu, and Fengjun Xiao. Exposing AI-generated videos with motion magnification. *Multimedia Tools and Applications*, 80(20):30789–30802, Aug. 2021. 1

[11] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*, Sept. 2019. 1

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 1

[14] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841, 2025. 1

[15] Adam Karvonen. An Intuitive Explanation of Sparse Autoencoders for Mechanistic Interpretability of LLMs. June 2024. 2

[16] Taras Kutsyk, Tommaso Mencattini, and Ciprian Florea. Do Sparse Autoencoders (SAEs) transfer across base and fine-tuned language models? Sept. 2024. 2

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model, May 2024. 2

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, Feb. 2022. 1

[19] Yuezun Li and Siwei Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts, May 2019. 1

[20] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, June 2020. 1

[21] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, Feb. 2023. 2

[22] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in. 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. IEEE Computer Society, June 2022. 2

[24] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces, Mar. 2018. 1

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[26] Selim Seferbekov. Deepfake Detection Challenge. https://kaggle.com/deepfake-detection-challenge. 1

[27] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. Paligemma 2: A family of versatile vlms for transfer, 2024. 2

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision, Dec. 2015. 1

[29] Wenqi Xian, Zhengqi Li, Matthew Fisher, Jonathan Eisenmann, Eli Shechtman, and Noah Snavely. UprightNet: Geometry-Aware Camera Orientation Estimation From Single Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9974–9983, 2019. 1

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 1

[31] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization, Apr. 2018. 1

[32] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models, Nov. 2023. 2

[33] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common Sense Reasoning for Deepfake Detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 399–415, Cham, 2025. Springer Nature Switzerland. 1, 2

[34] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking Semantic Artifacts for Generalized AI-generated Image Detection. *Advances in Neural Information Processing Systems*, 37:59570–59596, Dec. 2024. 1