Team 81: Deepfake Detection

Anthony John Dsouza
7053485

Eva Morgan Toulouse Gavaller
Matriculation 2

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.

- would like to explore methods to detect deepfakes / ai generated media - collect ai generated visual dataset, annotate it for

– train VIT based model to detect deepfakes

– train a VL model [PaliGemma, phi, etc] to reason with features extracted by image encoder and why provided image / video is real or fake

– what features do classifiers / detectors focus on? [look at attention rollout, SAE features, any more techniques?]

1.2. Motivation

...

1.3. Related work

Most common approaches use a simple setup, by finetuning a CNN / Vision Transformer [6] with some simple augmentations on datasets like [5, 15, 18, 27]. The approach considered here is 'classification', where the detector tries to identify artifacts generated by the generators. The DeepFake Detection challenge by Meta and Kaggle [1] saw many submissions using the above approaches with the best performing submission [20] being an ensamble of many models with augmentations and more importantly face warping artifacts [14]. Other approaches used [8, 23, 24] for improved generalization.

In [7], 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 [21] as the backbone for extracting spatial features from video frames and an LSTM [10] to process temporal data, enabling effective discrimination between real and synthetic videos based on motion patterns.

In [26], 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 [13] backbone, integrating a BERT [3] like text encoder and a Vision Transformer [6] 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 [2] 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 [22], 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.

[11] 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

[1] DeepFake Detection Challenge

images encountered on social media platforms, providing a robust resource for developing and evaluating detection models. for detection, the authors develop on LISA [12] 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.

1.4. Challenges

A significant portion of methods for DeepFake detection systems were built before diffusion [9] was a thing- mainly because of the fact that training these models was computationally expensive. Soon, [4] showed that diffusion based image generation models were better at image generation while being stable in comparison to their counterparts. [17], [16] build on this for improved generation, with methods developed to control the generation process like [19, 25] and using fewer steps for denoising as in [16]. 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

...

2.2. Proposed Timeline

...

3. Methods

3.1. ViT based classifier

3.1.1 Stage 1: Training classifier

In this stage, a ViT model [6] 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 CNNs is because the former learns robust representations from the data.

3.1.2 Stage 2: Interpretability - SAE and Attention rollout

Sparse autoencoders are small models that are trained on hidden representations for a model with an aim to sparsify entangled features in the hidden representations of large models. This makes models more interpretable, essentially helping in deriving human understandable explanations for behaviours of the model. The SAE will be trained by harvesting hidden representations from the model and help us understand what features the model looks at to classify between real and AI generated images.

Attention rollout [1] is another method to make later attention layers more interpretable.

4. Datasets

For this project, we intend to collect publicly available AI generated images / 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.

5. Evaluation

5.1. ViT based model

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

For evaluating the ViT based model, our metrics will be Accuracy and LogLoss. The reason for the latter is that LogLoss 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 also better suited for situations where confidence of prediction matters, as it looks at 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.

References

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

[2] 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

[3] 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

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

[11] 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

[12] 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

[13] 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

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

[15] 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

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

[17] 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

[18] 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

[19] 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

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

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

[22] 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

[23] 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

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

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

[26] 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

[27] 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