On CLIP's Ability of Analyzing Fake Images at a Large Scale: Why are they fake?

JINBIN HUANG, Arizona State University, USA
CHEN CHEN, University of Maryland, College Park, USA
ADITI MISHRA, Arizona State University, USA
BUM CHUL KWON, IBM Research, USA
ZHICHENG LIU, University of Maryland, USA
CHRIS BRYAN, Arizona State University, USA

Generative AI now possesses the capability to generate highly realistic images that can deceive human viewers, raising concerns about misinformation and ethical quandaries in AI. This highlights the need for methodologies that can systematically analyze and summarize patterns in AI-generated, fake images. Though traditional learning-based techniques still need further improvements for the reliable detection of fake images, recent studies have shown that CLIP can provide favorable detection outcomes that can be generalized across diverse generative models. Building upon this, we introduce a novel interactive system that uses CLIP for summarization and analysis of patterns within AI-generated images on a large scale. Our method employs a backend pipeline to distill CLIP's complex embedding space into informative dimensions, identifying key image regions to evaluate authenticity. The outcomes generated from the computational pipeline are then fed to the interactive frontend, enabling users to explore clusters of images exhibiting specific patterns via a representation view and an image view, understand these patterns with a concept view, and pinpoint their origins within the images using an attention view. We also describe a practical use case to demonstrate how researchers can use our system to summarize and analyze the patterns of AI-generated images.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Visual analytics**; • **Computing methodologies** → **Image representations**; **Neural networks**.

Additional Key Words and Phrases: Generative AI, Multi-modal Machine Learning, Visual Analytics, Image forensics

1 INTRODUCTION

Generative AI has rapidly gained prominence, with many models, such as DALL·E 3 [15] and Midjourney [12], capable of creating highly realistic images that easily deceive human viewers. This advancement underscores the critical importance of effective detection and analysis of such realistic but fake images, as their misuse could lead to significant issues such as misinformation spread [21], copyright issues [18], increased difficulties in digital forensics tracking [7], and most importantly questionable AI ethnicity [16].

Despite various deep learning methods [4, 8, 10, 17, 20] proposed for identifying fake images, their performance often degrades with out-of-distribution samples. Recently, CLIP [14] has emerged as a promising tool, achieving high accuracy [13] across different generative models, showing strong generalizability. This effectiveness is due to CLIP's exposure to many images during training. Subsequent research [3] has further explored CLIP's role as a universal fake

Manuscript submitted to ACM

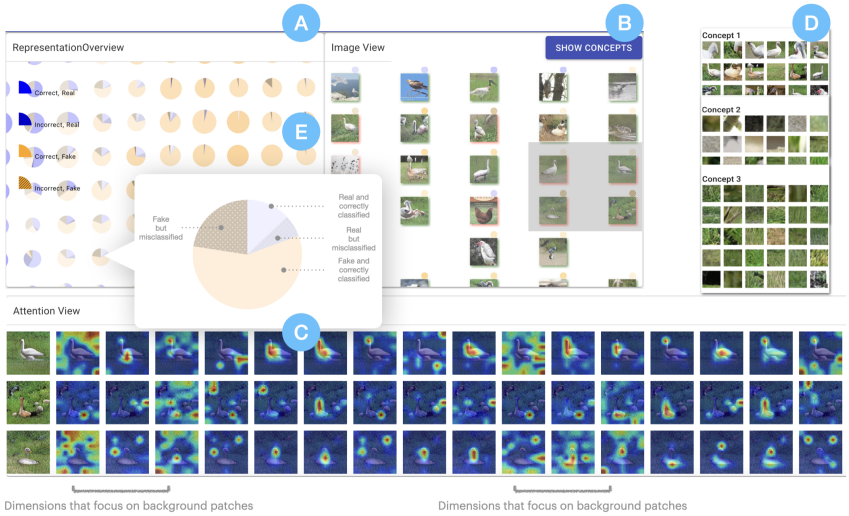

Fig. 1. The FaXplainer system employs the CLIP model to summarize and analyze fake images at a large scale.

image detector. This paves the way for developing systems that can analyze and summarize fake images at scale, aiding researchers and educating the public on discerning deceptive images online [9].

In this paper, we present a preliminary system, FaXplainer, that utilizes CLIP to summarize and analyze fake images at a large scale. To be specific, by extracting informative features from CLIP's intricate high-dimensional embedding space, FaXplainer identifies local pixel-level regions within images that are crucial for revealing fakeness and summarizes them into visual conceptual patterns. FaXplainer presents insights derived from the backend through an interactive interface (Fig. 1), where users can explore image clusters exhibiting certain fake patterns through an Representation Overview (Fig. 1A), review the image clusters in an Image View (Fig. 1B), understand these patterns with a concept view (Fig. 1D), and pinpoint their origins within the images using an attention view (Fig. 1C). In the following section, we present a use case to show how researchers can employ our system to efficiently summarize and analyze fake image patterns generated by a GAN-based model with the image data sourced from [20].

2 EXTRACTING CRITICAL INFORMATION FOR FAKE IMAGE DETECTION

FaXplainer's backend is designed to identify pixel-level regions within images essential for CLIP to distinguish fake images from real ones, and reveal visual conceptual patterns across a certain image batch (specified by the user).

Vision-only Feature Disentanglement. FaXplainer starts with encoding a image into CLIP's embedding space (512-dimension in our case because we use the particular CLIP:ViT-b/32 architecture) to get its corresponding feature vector. The embedding, however, is not ideal due to the mixture of text and image information (as CLIP is trained with inputs of both formats), where text-relevant information is entangled with vision-relevant information in specific dimensions. Since we target only the visual features in fakeness detection, we introduce a specialized linear projection layer [11] in FaXplainer to obtain the embedding that only contains vision information, removing potential effects from its text counterparts. After this disentanglement, the input is reduced to a 256-dimension vision-centric feature vector.

Dimension Reduction and Classification. The 256-dimension visual-only feature is still high-dimensional for dimension-wise review and analysis. Additionally, high dimensionality leads to potential information coupling and

redundancy, making it difficult to identify unique information from each dimension. Thus, at this step, we aim at transforming the 256-dimension visual-only feature into a more compact embedding, where dimensionality is at a manageable low level and dimensions are as orthogonal to each other as possible. With this requirement in mind, we employ a two-layer fully-connected neural network: the first layer reduces the 256-dimension visual feature vector to a 16-dimension compact representation, with an orthogonal penalty term [1] applied to impose orthogonality between dimensions during training; the second layer produces logits which are used to calculate a binary cross-entropy loss for optimization. In short, FaXplainer uses the two layers to obtain a representation that has 16 (nearly) orthogonal dimensions that will be utilized for later analysis and to perform real or fake image classification, respectively.

Attention Map Generation. With the 16-dimension compact representation, we calculate each dimension's relevance to pixels in the original image, using gradient back-propagation techniques [2]. This step quantifies each pixel's contribution across the 16 dimensions towards CLIP's capability to distinguish fake from real images. By aggregating these pixel attributions dimension-wise, we pinpoint regions critical for identifying fake images.

Visual Concept Extraction. To uncover commonalities in images within a specific cell and identify fake patterns, FaXplainer computes common concepts across all cell images. Utilizing attention maps as masks, it segments important regions and obtains their 16-dimensional representations through three steps. FaXplainer then clusters these representations to identify concept clusters. Empirically, setting the displayed clusters to three captures critical foreground and background information across low, medium, and high resolutions. Using attention maps as masks reduces image segments for concept computation, removing the need for post-clustering filtering.

3 FAXPLAINER: AN INTERACTIVE TOOL FOR FAKE PATTERN SUMMARY AND ANALYSIS

FaXplainer presents an interactive summary and analysis of fake image patterns through an interface (Fig. 1) which consists of a representation view (Fig. 1A), an image view (Fig. 1B), an attention view (Fig. 1C), and a concept view (Fig. 1D).

Representation View. We implemented a customized representation view (Fig. 1A) to provide an overview of the fake image distributions against real ones. This view clusters images based on their Euclidean distance from each other in the 16-dimensional compact representation space during the dimension reduction phase described in Section 2, where images close to each other (usually they also look similar) are assigned to the same cell. Each cell in the representation view is rendered as a pie glyph (Fig. 1E), designed with the following rationale: the cell size indicates the number of images it contains; the cell color and shading denote the accuracy of CLIP's predictions where blue for real images (solid for correct predictions, shaded for incorrect ones) and orange for fake images (solid for correct predictions, shaded for incorrect ones); the cell saturation level represents the confidence of these predictions. Based on our observation, cells near the decision boundary often appear pale and multicolored, indicating a low prediction confidence and a mixture of real and fake images — these are the regions we think are most interesting to explore.

The benefits of the representation view are two-fold. On one hand, the user can easily compare and contrast similar images, facilitating the analysis process. On the other, such a design addresses the challenge of representing tens of thousands of images, which, if displayed as individual dots, would result in a cluttered and uninformative visualization. Our design enables easy navigation and provides a comprehensive overview of the dataset, highlighting regions near the decision boundary where fake images could potentially deceive the CLIP model. These areas are of particular interest as they allow us to examine images that might also mislead humans.

Image View. Upon opening an image cluster by clicking on a cell in the representation view, the corresponding images are displayed in a grid layout in the image view(Fig. 1B), arranged by similarity. This is achieved by mapping 2d

tSNE [19] coordinates to 2d grid coordinates using the IsoMatch method [5]. The images are framed in red or green, signifying CLIP's failures or successes in prediction. In this way, red-bordered fake images might be of particular interest to users since they may contain patterns that successfully deceive CLIP.

Attention View. For a deeper analysis, users can select a rectangular area within the image view to examine specific images in the attention view (Fig. 1C), where it reveals the important regions within those images for CLIP's prediction, computed in the backend (the attention map generation phase in Section 2). These attention maps enable users to identify specific features that are most likely to confuse the CLIP model, offering a granular understanding of deceptive patterns at the ground level.

Concept View. The image view includes a toggle button in the top right corner, enabling users to access a concept view (Fig. 1D). This view aggregates visual patterns within a cell, providing insights into prevalent deceptive patterns at a glance. The generation of the concepts is described in the visual concept extraction phase in Section 2.

4 USE CASE

This section demonstrates how the system could assist a user, Ryan, in summarizing and analyzing fake bird images generated by the ProGAN model [6]. Upon initiating the load for the bird-class dataset, a representation overview updates, displaying processed image cells, in which most blue cells are positioned to the left and orange cells to the right, delineating a clear boundary marked by cells of less saturated, mixed colors (Fig.1A). Intrigued by the decision boundary, Ryan starts to examine the cells more closely. One cell (Fig.1E), in particular, draws his attention. Ryanclicks on it, which leads to a collection of fake images featuring goose-like birds amidst grassy backgrounds, possibly depicting geese or cranes near ponds, in the image view. The grass texture appears remarkably real, initially convincing Ryan of their authenticity. However, upon closer inspection, Ryan notices that the grass pattern's abnormally low resolution, making it appear unnatural; though, from a distance, the images' authenticity remains ambiguous. Ryan also notices that this images are misclassified as real by CLIP.

Ryan proceeds to open the concept view (Fig.1D) to identify prevalent patterns among these images, unsurprisingly finding grassy concepts. He selects three images for further analysis, dragging a box around them to engage the attention view (Fig.1C). Here, Ryan observes that while there are ten dimensions focused on the goose's body, there are also six dimensions highlighting the background's grassy patches in selected images (marked in Fig.1), indicating the CLIP model also pay a decent amount of attention to their realistic backgrounds despite the foreground subjects being evidently artificial. Further utilizing the system, Ryan identifies additional deceptive patterns by inspecting images contained in cells along the decision boundaries, making notes on particularly convincing ones: 1) bird in the sky, 2) bird in water, 3) bird in yellow grass, 4) bird in green grass, and 5) bird in a tree with crisscrossing branches in the background, as depicted in Fig. 2. This marks the end of an exploration run of FaXplainer by Ryan.

Fig. 2. Fake patterns that Ryan found with FaXplainer

REFERENCES

[1] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. 2016. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093* (2016).

[2] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 397–406.

[3] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2023. Raising the Bar of AI-generated Image Detection with CLIP. *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2312.00195

[4] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510* (2018).

[5] Ohad Fried, Stephen DiVerdi, Maciej Halber, Elena Sizikova, and Adam Finkelstein. 2015. IsoMatch: Creating informative grid layouts. In *Computer graphics forum*, Vol. 34. Wiley Online Library, 155–166.

[6] Hongchang Gao, Jian Pei, and Heng Huang. 2019. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1308–1316.

[7] Ankita Guleria, Kewal Krishan, Vishal Sharma, and Tanuj Kanchan. 2023. ChatGPT: Forensic, Legal, and Ethical Issues. *Medicine, Science and the Law* (2023). https://doi.org/10.1177/00258024231191829

[8] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5039–5049.

[9] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2024. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems* 36 (2024).

[10] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. 2018. Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 384–389.

[11] Joanna Materzyńska, Antonio Torralba, and David Bau. 2022. Disentangling Visual and Written Concepts in CLIP. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16389–16398. https://doi.org/10.1109/CVPR52688.2022.01592

[12] Midjourney. 2024. Midjourney. https://www.midjourney.com/home.

[13] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards Universal Fake Image Detectors That Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 24480–24489. https://doi.org/10.1109/CVPR52729.2023.02345

[14] OpenAI. 2024. CLIP: Connecting text and images. https://openai.com/research/clip.

[15] OpenAI. 2024. DALL·E 3. https://openai.com/dall-e-3.

[16] Sebastian Porsdam Mann, Brian D Earp, Sven Nyholm, John Danaher, Nikolaj Møller, Hilary Bowman-Smart, Joshua Hatherley, Julian Koplin, Monika Plozza, Daniel Rodger, et al. 2023. Generative AI entails a credit–blame asymmetry. *Nature Machine Intelligence* (2023), 1–4. https://doi.org/10.1038/s42256-023-00653-1

[17] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

[18] Matthew Sag. 2023. Copyright safety for generative AI. *Forthcoming in the Houston Law Review* 61, 2 (2023). https://doi.org/10.2139/ssrn.4438593

[19] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[20] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8695–8704.

[21] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating Misinformation in the Era of Generative AI Models. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9291–9298. https://doi.org/10.1145/3581783.3612704