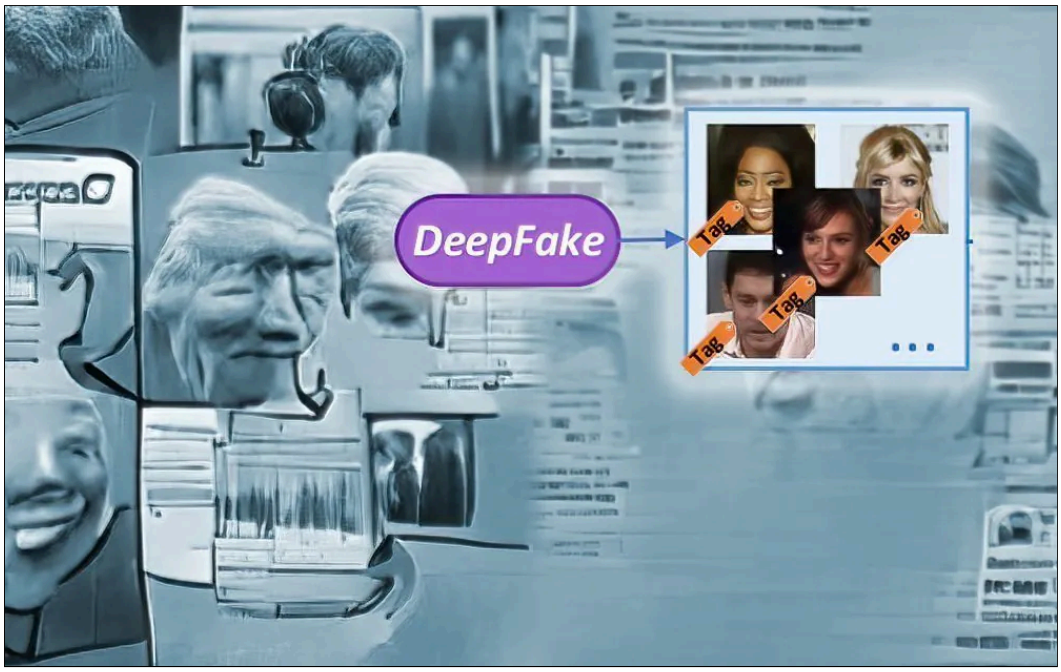
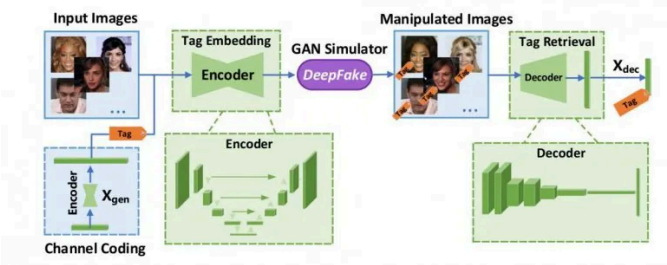
Identifying Deepfake Data Sources With AI-Based Tagging

Originally published at https://www.unite.ai/identifying-deepfake-data-sources-with-ai-based-tagging-faketagger/



A collaboration between researchers in China, Singapore and the US has produced a resilient system for 'tagging' face photos so robustly that the identifying markers are not destroyed during a deepfake training process, paving the way for IP claims that could put a dent in the ability of synthetic image generation systems to 'anonymize' illegitimately scraped source data.

The system, entitled *FakeTagger*, uses an encoder/decoder process to embed visually indiscernible ID information into images at a low enough level that the injected information will be interpreted as essential facial characteristic data, and therefore passed through abstraction processes intact, in the same way, for instance, as eye or mouth data.



An overview of the FakeTagger architecture. Source data is used to generate a 'redundant' facial characteristic, ignoring background elements which will be masked out through a typical deepfake workflow. The message is recoverable at the other end of the process, and identifiable through an apposite recognition algorithm. Source: http://xujuefei.com/felix_acmmm21_faketagger.pdf

The research comes from the School of Cyber Science and Engineering at Wuhan, the Key Laboratory of Aerospace Information Security and Trusted Computing at China's Ministry of Education, the Alibaba Group in the US, Northeastern University at Boston, and the Nanyang Technological University at Singapore.

Experimental results with FakeTagger indicate a re-identification rate of up to almost 95% across four common types of deepfake methodologies: identity swap (i.e. DeepFaceLab, FaceSwap); face reenactment; attribute editing; and total synthesis.

Shortcomings of Deepfake Detection

Though the last three years have brought a crop of new approaches to deepfake identification methodologies, all of these approaches key on remediable shortcomings of deepfake work-flows, such as eye-glint in under-trained models, and lack of blinking in earlier deepfakes with inadequately diverse face-sets. As new keys are identified, the free and open source software repositories have obviated them, either deliberately, or as a by-product of improvements in deepfake techniques.

The new paper observes that the most effective post-facto detection method produced from Facebook's most recent deepfake detection competition (DFDC) is limited to 70% accuracy, in terms of spotting deepfakes in the wild. The researchers ascribe this representative failure to poor generalization against new and innovative GAN and encoder/decoder deepfake systems, and to the often degraded quality of deepfake substitutions.

In the latter case, this can be caused by low-quality work on the part of deepfakers, or compression artifacts when videos are uploaded to sharing platforms that seek to limit bandwidth costs, and re-encode videos at drastically lower bit rates than the submissions. Ironically, not only does this image degradation *not* interfere with the apparent authenticity of a deepfake, but it can actually enhance the illusion, since the deepfake video is subsumed into a common, low-quality visual idiom that's perceived as authentic.

Survivable Tagging as an Aid to Model Inversion

Identifying source data from machine learning output is a relatively new and growing field, and one that makes possible a new era of IP-based litigation, as governments' current permissive screen-scraping regulations (designed not to stifle national research pre-eminence in the face of a global AI 'arms race') evolve into stricter legislation as the sector becomes commercialized.

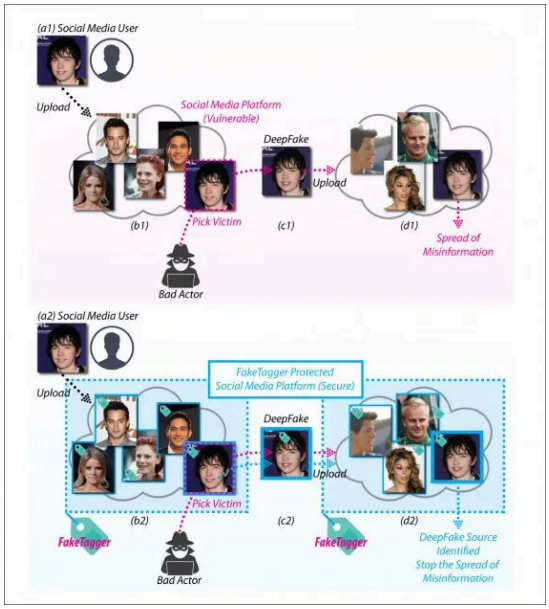
Model Inversion deals with the mapping and identification of source data from the output generated by synthesis systems in a number of domains, including Natural Language Generation (NLG) and image synthesis. Model inversion is particularly effective at re-identifying faces that were either blurred, pixelated, or else have made their way through the abstraction process of a Generative Adversarial Network or encoder/decoder transformation system such as DeepFaceLab.

Adding targeted tagging to new or existing facial imagery is a potential new aide to model inversion techniques, with watermarking an emergent field.

Post-Facto Tagging

FakeTagger is intended as a post-processing approach. For instance, when a user uploads a photo to a social network (which usually involves some kind of optimization process, and rarely a direct and unadulterated transfer of the original image), the algorithm would process the image to apply supposedly indelible characteristics to the face.

Alternately, the algorithm could be applied across historical image collections, as has happened a number of times over the last twenty years, as large stock photo and commercial image collection sites have sought methods to identify content that has been re-used without permission.



FakeTagger seeks to embed recoverable ID characteristics from various deepfake processes.

Development and Testing

The researchers tested FakeTagger against a number of deepfake software applications across the aforementioned four approaches, including the most widely-used repository, DeepFaceLab; Stanford's Face2Face, which can transfer facial expressions across images and identities; and STGAN, which can edit facial attributes.

Testing was done with CelebA-HQ, a popular scraped public repository containing 30,000 face images of celebrities at various resolutions up to 1024 x 1024 pixels.

As a baseline, the researchers initially tested conventional image watermarking techniques, to see if the imposed tags would survive the training processes of deepfake workflows, but the methods failed across all four approaches.

FakeTagger's embedded data was injected at the encoder stage into the face-set imagery using an architecture based on the U-Net convolutional network for biomedical Image segmentation, released in 2015. Subsequently, the decoder section of the framework is trained to find the embedded information.

The process was trialed in a GAN simulator that leveraged the aforementioned FOSS applications/algorithms, in a black box setting with no discrete or special access to the work-flows of each system. Random signals were attached to the celebrity images, and logged as related data to each image.

In a black box setting, FakeTagger was able to achieve an accuracy exceeding 88.95% over the four applications' approaches. In a parallel white-box scenario, accuracy increased to nearly 100%. However, since this suggests future iterations of deepfake software that incorporates FakeTagger directly, it's an unlikely scenario in the near future.

Counting the Cost

The researchers note that the most challenging scenario for FakeTagger is complete image synthesis, such as CLIP-based abstract generation, since input training data is subject to the very deepest levels of abstraction in such a case. However, this does not apply to the deepfake work-flows that have dominated headlines over the last several years, as these are dependent on faithful reproduction of ID-defining facial characteristics.

The paper also notes that adversarial attackers could conceivably attempt to add perturbations, such as artificial noise and grain, in order to foil such a tagging system, though this would be likely to have a detrimental effect on the authenticity of deepfake output.

Further, they note that FakeTagger needs to add redundant data to imagery in order to ensure the survival of the tags that it embeds, and that this could have a notable computational cost at scale.

The authors conclude by noting that FakeTagger may have potential for provenance tracking in other domains, such as adversarial rain attacks and other types of image-based attacks, such as adversarial exposure, haze, blur, vignetting and color-jittering.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More