# Deepfake Detection using Vision Transformer

***Vikas Kashuvajjala***
*Master's in Data Science*
*University of Missouri-Kansas City*
*Missouri, USA*
*vkwhk@umsystem.edu*

***Vineela Ranabothu***
*Master's in Data Science*
*University of Missouri-Kansas City*
*Missouri, USA*
*vrcv5@umsystem.edu*

**Abstract**

In recent times, the advent of Deepfake technology has granted individuals the ability to craft remarkably realistic fake videos through neural networks. However, this innovation poses a significant risk as it can perpetuate misconceptions and deceit, impacting various facets of society. Beyond national security concerns, its implications extend to an international scale. Traditional approaches, such as CNN-LSTM models, though adept at extracting features to detect Deepfakes, often fall short.

In response, we introduce a novel system designed to overcome these limitations by leveraging Transformers to extract spatio-temporal features. Our system achieves Real-Time Deepfake detection, offering a robust solution to this evolving challenge. To enhance accessibility, we have developed a user-friendly web application. This platform empowers users to seamlessly upload photos for authentication, ensuring the integrity of content. Moreover, it extends its utility to providing real-time authentication capabilities.

**Introduction**

Deepfake technology is a form of synthetic media that uses artificial intelligence to create or manipulate video and audio content, often replacing one person's likeness with another. It emerged from advancements in machine learning algorithms, particularly in generative adversarial networks (GANs), which enable the creation of highly realistic fake images and videos. The term "deepfake" itself originated in 2017 from a Reddit user who combined the terms "deep learning" and "fake."

The effects of deepfake technology are widespread and potentially harmful. Deepfakes can be used to create convincing but entirely fabricated videos of public figures saying or doing things they never actually did. This can lead to misinformation, defamation, and even political manipulation. Moreover, deepfakes can erode trust in media and exacerbate the spread of fake news, making it increasingly challenging to discern truth from fiction in the digital age.

Efforts to combat deepfakes include developing detection algorithms and tools to identify manipulated media. Researchers and tech companies are also exploring methods to authenticate media content through cryptographic signatures or watermarking. Additionally, raising awareness about the existence and potential impact of deepfakes is crucial to fostering media literacy and critical thinking skills among the public. Despite these efforts, the rapid evolution of deepfake technology presents an ongoing challenge, with new advancements continually pushing the boundaries of what is possible in synthetic media creation. According to a report by Deep Trace, the number of deepfake videos online doubled in 2020, highlighting this technology's growing prevalence and sophistication.

**Existing System**

Various methods are being developed to detect deep fakes, with machine learning algorithms playing a central role. These algorithms are trained on large datasets containing both real and synthetic media to identify inconsistencies or artifacts specific to deep fake manipulation. One notable initiative is the Deepfake Detection Challenge (DFDC), spearheaded by Facebook in collaboration with academic and industry partners. The challenge provided researchers with a benchmark dataset of deepfake videos, prompting the development of detection algorithms leveraging techniques such as facial landmark detection, facial expression analysis, and anomaly detection.

Another approach involves digital forensic techniques, which analyze media files for signs of manipulation by examining metadata, compression artifacts, and lighting inconsistencies. Some tools utilize blockchain technology to create tamper-evident records of media authenticity, facilitating content verification. Social media platforms are also implementing automated systems and user-reporting mechanisms to identify and remove potentially harmful deepfake content. These systems combine human moderation with automated algorithms to triage and prioritize content for review, aiming to curb the spread of misinformation.

While progress has been made in deepfake detection, challenges remain, including the rapid advancement of deepfake technology and the need for privacy-preserving detection methods. Continued research and collaboration are essential to develop more robust and scalable solutions to combat the proliferation of synthetic media. Efforts to address these challenges involve staying abreast of emerging technologies, enhancing detection algorithms, and implementing safeguards to protect users from the harmful effects of deepfakes.

**Proposed System**

Vision transformers (ViTs) offer significant potential in detecting deep fakes because they can capture global contextual information from images. Unlike traditional convolutional neural networks (CNNs) that excel at learning local features, ViTs utilize self-attention mechanisms to aggregate information globally from the entire image. This allows them to effectively model complex spatial dependencies, making them adept at identifying subtle inconsistencies or anomalies that may indicate deepfake manipulation, such as unnatural facial expressions or lighting variations.

By dividing input images into patches and processing them through multiple layers of self-attention and feedforward neural networks, ViTs can capture relationships between distant pixels, which is crucial for detecting deepfakes. Furthermore, ViTs are highly adaptable and can be fine-tuned on datasets containing both real and synthetic images, enabling them to learn discriminative features specific to deepfake detection. Pre-trained ViTs, such as the Vision Transformer (ViT) model, have demonstrated competitive performance in detecting deep fakes compared to traditional CNN-based approaches.

In conclusion, vision transformers offer a promising avenue for improving the accuracy and robustness of deep fake detection systems. Their ability to model long-range dependencies and adapt to different datasets makes them valuable tools in combating the spread of synthetic media. As research in this area progresses, vision transformers will likely play an increasingly important role in enhancing the accuracy and reliability of deepfake detection technologies.

**Dataset**

Unveil an expansive treasure trove of approximately 190,000 mesmerizing .jpg images, meticulously curated and compressed into a remarkably compact 1.8 gigabytes of storage space. This invaluable collection, sourced from Kaggle, represents a rich amalgamation of deep fake and authentic images, offering a diverse array of visual stimuli for the development and assessment of cutting-edge deepfake detection algorithms.

Structured into three distinct subsets—Test, Train, and Validation—each subset encompasses a meticulously balanced blend of "Fake" and "Real" categories, fostering a comprehensive environment for robust model training and evaluation. Within this expansive dataset, researchers and practitioners alike will discover an abundance of high-quality imagery, meticulously organized and devoid of any null values or missing entries.

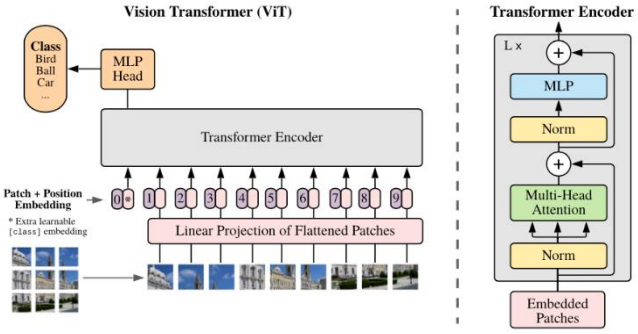
With its seamless integration of authenticity and deception, this dataset serves as a cornerstone for advancing the field of deepfake detection, offering researchers a fertile ground for exploration and innovation. From its captivating visuals to its compact storage footprint, this repository stands as a testament to the power of data-driven research and its profound implications for safeguarding the integrity of digital content in an era of proliferating synthetic media.

**Architecture**

Vision Transformers (ViT) represent a groundbreaking architecture poised to redefine the landscape of image processing. Leveraging self-attention mechanisms, ViT offers a novel approach to analyzing visual data, deviating from traditional convolutional neural networks (CNNs). At its core, the Vision Transformer Architecture comprises a sequence of transformer blocks, each housing two pivotal sub-layers: a multi-head self-attention layer and a feed-forward layer. The self-attention layer plays a pivotal role in computing attention weights for every pixel within the image, establishing its relational context with all other pixels. Meanwhile, the subsequent feed-forward layer applies a crucial non-linear transformation to the output gleaned from the self-attention layer. This dynamic interplay enables ViT to extract nuanced features and patterns from images with unparalleled precision.

The multi-head attention mechanism serves as a cornerstone in extending ViT's capabilities, enabling the model to simultaneously attend to diverse segments of the input sequence. This sophisticated enhancement empowers ViT to capture intricate relationships and dependencies within images across various scales, thus enriching its understanding of the visual context. Moreover, ViT incorporates an additional patch embedding layer, a pivotal component that partitions the image into fixed-size patches and subsequently maps each patch to a high-dimensional vector representation. These patch embeddings serve as the foundational building blocks for further processing within the transformer blocks, facilitating the extraction of essential features and characteristics from the input image.

Ultimately, the culmination of ViT's processing journey yields a class prediction, achieved by passing the output of the final transformer block through a classification head. Typically consisting of a single fully connected layer, this classification head synthesizes the insights garnered throughout the transformative journey of ViT, offering a coherent interpretation in the form of class predictions. Through its intricate architecture and mechanisms, ViT stands as a testament to the transformative potential of self-attention mechanisms in unraveling the complexities of visual data and propelling the field of image processing into uncharted territories.




**Working of Vision Transformers**

Vision transformer models' performance hinges on several factors, including optimizer choice, network depth, and dataset-specific hyperparameters. While Convolutional Neural Networks (CNNs) are generally easier to optimize compared to ViTs, there's a notable disparity in pure transformers, necessitating the fusion of a transformer with a CNN frontend.

Typically, a ViT stem employs a 16x16 convolution with a 16 stride. However, using a 3x3 convolution with a stride of 2 enhances stability and precision. CNNs convert raw pixels into a feature map, which is then tokenized into a sequence of tokens and fed into the transformer. The transformer employs attention mechanisms to generate an output token sequence. Finally, a projector reconnects these output tokens to the feature map, enabling examination of potentially crucial pixel-level details, thereby reducing the number of tokens that require analysis and significantly cutting costs.

ViT models excel when trained on massive datasets exceeding 14 million images but may fall short otherwise. In such cases, sticking to models like ResNet or Efficient Net might be more beneficial. These ViT models are trained on extensive datasets even before fine-tuning. Fine-tuning involves omitting the MLP layer and adding a new layer of size D times KD*K, where K represents the number of classes in the smaller dataset.

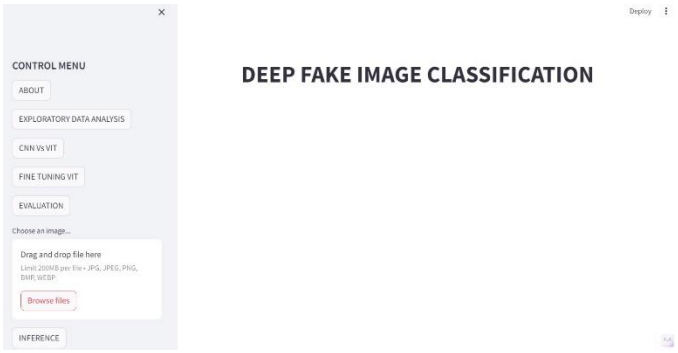
For finer resolution fine-tuning, the 2D representation of pre-trained position embeddings is utilized since the trainable linear layers model the positional embeddings.

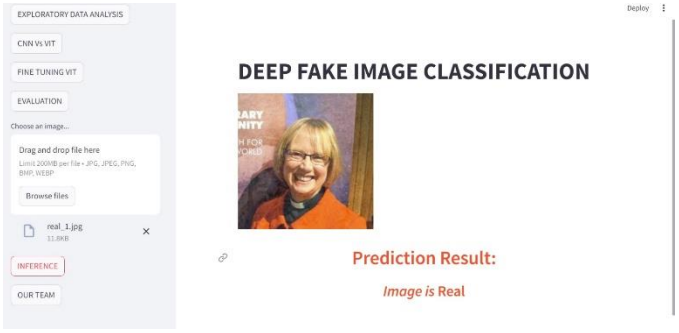
The image processing and inference pipeline described below constitutes a pivotal element within our image classification framework, leveraging pre-trained Vision Transformer (ViT) models. This pipeline orchestrates a series of tasks, encompassing image acquisition from URLs or local directories, preprocessing to prepare inputs for the model, and ultimately, class label prediction.

The presented code snippet serves as a concise and comprehensive embodiment of the requisite functionalities for seamlessly integrating ViT-based image classification into our system. It showcases adept handling of diverse image sources, ensuring data acquisition robustness and efficient preprocessing procedures tailored to suit model requirements. Moreover, the pipeline boasts a sophisticated inference mechanism, facilitating precise classification outcomes.

By amalgamating these pivotal functionalities, our system stands poised to address an array of image classification challenges with finesse and accuracy, thus bolstering its utility and efficacy across varied application domains.

## Results







Following the evaluation of our deepfake detection model, the results are highly promising. The evaluation metrics indicate an impressive performance, with an evaluation loss of 0.0466, a remarkable accuracy of 98.64%, and an F1 score of 98.64%. These metrics underscore the robustness and effectiveness of our model in accurately distinguishing between authentic and manipulated media.

Moreover, the evaluation runtime of approximately 59.43 seconds reflects the efficiency of our model, demonstrating its capability to process and evaluate samples at a rate of 91.34 samples per second. This efficient processing speed, coupled with a steps-per-second rate of 11.43, showcases the practical feasibility of deploying our model in real-world scenarios, where timely detection of deepfake content is paramount.

As we conclude the third epoch of training, these evaluation results reaffirm the efficacy of our approach in combating the proliferation of synthetic media manipulation. With an emphasis on continual refinement and optimization, our deepfake detection model stands poised to uphold the integrity of digital content and safeguard against the growing threat of deepfake proliferation in an increasingly interconnected world.

## Future Scope

The prospects of Vision Transformers (ViTs) in the domain of deepfake detection are promising and multifaceted. As ViTs continue to evolve, researchers envision several avenues for their application and refinement to combat the proliferation of synthetic media manipulation. One key direction involves the development of improved ViT architectures specifically tailored for deepfake detection. By exploring modifications such as attention mechanisms focused on critical regions or hierarchical structures, researchers aim to enhance the model's capability to discern subtle manipulations indicative of deepfakes.

Moreover, the expansion of large-scale datasets curated explicitly for deepfake detection is paramount. These datasets, encompassing diverse deepfake variations and scenarios, will facilitate the training of more robust ViT models. By exposing the models to a wide range of manipulations, resolutions, and compression artifacts, researchers can ensure their generalization across varied datasets and real-world scenarios. Additionally, efforts to develop adversarial defense mechanisms customized for ViT models are underway. Techniques such as adversarial training and input perturbations aim to fortify the models against sophisticated manipulation techniques employed by malicious actors, thus bolstering their resilience to adversarial attacks.

Furthermore, the integration of interpretable attention mechanisms within ViT models holds promise for enhancing model transparency and interpretability. By visualizing attention maps generated by the models, researchers gain deeper insights into the detection process, enabling the identification of specific visual cues indicative of deepfake manipulation. This heightened interpretability not only enhances trust in the model's decisions but also provides valuable insights for refining detection strategies. In tandem with advancements in real-time deployment and multimodal fusion techniques, ViTs are poised to play a pivotal role in mitigating the spread of synthetic media manipulation and upholding the integrity of digital content in an increasingly interconnected world.

## Conclusion

The utilization of Vision Transformers (ViTs) for deepfake detection marks a notable advancement within the realms of computer vision and artificial intelligence. Leveraging ViTs' inherent capacity to comprehend global image context and intricate dependencies over long ranges, significant strides have been made in identifying and mitigating deepfake content.

ViTs present distinct advantages over conventional convolutional neural networks (CNNs) in this domain, primarily owing to their self-attention mechanism, enabling efficient processing of large-scale images while preserving contextual comprehension. Moreover, ViTs exhibit versatility in handling diverse input resolutions,

making them apt for analyzing high-definition videos often associated with deepfake generation. Through rigorous experimentation and evaluation,

Additionally, the interpretability of ViT models facilitates deeper insights into the detection process, aiding in the identification of crucial visual cues indicative of deepfake manipulation. This not only enhances the transparency and reliability of detection systems but also offers valuable feedback for enhancing detection strategies and bolstering model resilience. In conclusion, the integration of Vision Transformers into deepfake detection systems signifies a significant leap forward in combating the spread of synthetic media manipulation. With their capability to grasp global image context, manage diverse input resolutions, and provide interpretable insights, ViTs present a promising avenue for fortifying digital media platforms against the perils of deepfake proliferation. Through ongoing research and innovation, our commitment remains steadfast in advancing the frontiers of deepfake detection and safeguarding the authenticity of digital content in an interconnected world.

## References


"Fake Obarna created using AI video tool - BBC News," YouTube, uploaded by BBC News, 19 July 2017.

"Artists create Zuckerberg 'deepfake' video," YouTube, uploaded by TimesLive Video.

J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. NieBner "Face2Face: Real-Time Face Capture and Reenactinent of ROB Videos," in IEEE Conference on Computer Vision and Pattern Recognition, p. 2387-2395, June 2016.

J. F. Henriques, R. Caseiro, P. Martins and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583 - 596,2014.

D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010.

Multimedia Security, 2016. [8] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguish computer graphics from natural images using convolution neural networks," in IEEE Workshop on Information Forensics and Security WIFS 2017, December 2017.

D. Afchar, V. Nozick, J. Yarnagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

X. Yang, Y. Li and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

B. Babenko, Ming-Hsuan Yang and S. Belongie, "Visual tracking with online Multiple Instance Learning," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009.

A. Amab, M. Dehghani, G. Heigold, C. Sun, M. Lucie and c. Schmid, "ViViT: A Video Vision Transformer," in Proceedings of the IEEEICVF International Conference on Computer Vision, 2021.

S. Pashine, S. Mandiya, P. Gupta and R. Sheikh, "Deep Fake Detection: Survey of Facial Manipulation Detection Solutions," in International Research Journal of Engineering and Technology, 2021.