MSc Project - Reflective Essay

Project Title:	TruthLens: Deepfake detection using Xception Model
Student Name:	Samrudhi Bhosale
Student Number:	230854907
Supervisor Name:	Ammar Yasir Naich
Programme of Study:	Big Data Science

In today's world, artificial intelligence (AI) is rapidly advancing and becoming more available to everyone. However, these advancements bring with them significant ethical and societal challenges. This essay reflects on the development, strengths, and weaknesses of a deep fake detection project. It will also explore how this work can be expanded in the future and the broader effects of such technologies. By discussing the project's goals, scope, and methods, this essay highlights the importance of creating strong systems to prevent the misuse of AI, especially in media and communication.

The motivation for this deep fake detection project came from growing concerns about the misuse of AI-generated content. A recent example is the introduction of Grok 2.0 by Twitter, which allows users to create AI images without restrictions. While this technology is powerful, it also raises serious privacy concerns. For instance, the ability to create lifelike images of people, including public figures, without their consent, threatens personal privacy and can disrupt public discourse. This has already had real-world consequences, especially in politics, where AI-generated media has been used to spread false information and influence public opinion.

The impact of deep fake technology goes beyond individual incidents. It has the potential to erode trust in the media and weaken the foundation of factual information, which is essential for a healthy democracy. In some cases, deep fakes have been used to create fake news or explicit content without consent, violating personal privacy and causing significant harm. As this technology continues to improve, it poses an even greater risk to society.

While some social media platforms have introduced features that allow users to label AI-generated content, these measures are often not enough. Many people choose not to label their content, which only adds to the problem of misinformation. This growing issue of AI-generated media highlights the need for strict policies and regulations to protect privacy and ensure truthfulness. This is why I chose this topic for my dissertation. Developing effective tools to detect and stop the spread of deep fakes is more urgent than ever, as the potential for harm increases with each new technological advance.

Detecting deep fakes is a challenging task, especially as the methods for creating them become more advanced. AI models are getting better at mimicking human features and speech patterns, making it harder to tell the difference between real and AI-generated content. This challenge is made even more difficult by the large amounts of data available to train these models, allowing the creation of deep fakes that are almost indistinguishable from real media.

The ethical concerns surrounding deep fake technology are broad and complex. Developers, researchers, and policymakers have a significant responsibility to address these challenges. It's important to establish guidelines and regulations that prevent the misuse of AI while also encouraging its positive uses. Finding a balance between innovation and ethics is crucial to ensuring that deep fake technology benefits society rather than causing harm.

To address these issues, I developed a deep fake detection model using the CelebDF [1] dataset, a high-quality collection of data from YouTube videos. This dataset was chosen because it presents a significant challenge due to the realism of the deep fakes it contains. Unlike older datasets, the CelebDF[1] dataset was selected to push the limits of current detection capabilities. I used the Xception model, a well-trained architecture, to detect these subtle manipulations. Although the initial results were not as strong as expected, this project marks an important first step in developing real-time tools capable of identifying increasingly sophisticated AI-generated content.

The deepfake detection project highlighted several strengths that show its potential and importance in the field. Perhaps the most significant success was the detection algorithm where, despite the difficulties, the proposed approach demonstrated potential for detecting the difference between the original and manipulations. A key factor that contributed to this success was the Xception model, which is famous for its performance in image classification. It has a deep structure that enabled the model to detect minor discrepancies in the images, which are characteristic of deepfakes.

One of the strengths of this project was the choice of the CelebDF [1] dataset as the primary data source. This dataset is relatively realistic because of the high quality and the complexity of the deepfake videos, which challenged the existing detection approaches. By working on such a difficult dataset, the project was able to combat more complex kinds of deepfakes that are already emerging.

Despite these successes, the project faced several challenges and limitations that impacted its overall effectiveness. One of the primary weaknesses was related to the structure of the CelebDF [1] dataset itself. Because for each of the original videos in the dataset, three to four fake videos were created by replacing the original person with other candidates, this created redundancy. This redundancy likely confused the model, making it difficult to accurately distinguish between real and fake content. Having different videos that only differ a little from each other causes overfitting of the model, whereby the model is more inclined to the training data than to the new data. The CelebDF [1] dataset presented another challenge where participants wore accessories like spectacles. This issue became evident when participants moved their heads, causing spectacles to appear in one frame but not in others. The Xception model struggled to relate these inconsistencies between consecutive frames, leading to difficulties in accurately detecting deepfakes.

Time constraints also posed a significant limitation. With more time, the implementation of Generative Adversarial Networks (GANs) [2] could have been explored. GANs [2], which are often used to create deepfakes, can also be instrumental in their detection. By training a GAN-based model, it would have been possible to develop a more critical understanding of how deepfakes are generated, potentially leading to more effective detection methods.

Another challenge was the limited computational resources available for the project. The cloud storage space was capped at 20GB, which restricted the number of videos that could be used for training. As mentioned earlier, each video was of high quality: when frames were extracted for analysis, their size was rather large, which caused significant storage issues. Due to these factors, the model had to be trained on a smaller dataset than ideal, which likely affected its performance.

In addition to these successes, the project demonstrated robustness in the field of face feature extraction, which is one of the key components of deepfake detection. While prior models were able to achieve good results on conventional datasets such as FaceForensics++ [3] and DFDC [4] but they failed to give good results on CelebDF[1] dataset, this project showed that my model was uniquely effective in identifying facial

structures. This was done using the GradCAM [5] library which helped to explain the decisions made by the model and indeed the model was focusing on the facial areas which were most likely to be manipulated.

The theoretical concepts of deep learning and computer vision were instrumental in this project and the practical solutions derived from them. Firstly, my literature review [6] suggested that VGG16 model performed well in previous datasets as stated in survey papers. Theoretically, VGG16 appeared to be the most suitable model for use in this particular case since it was known to provide high classification accuracy. However, when I tried using VGG16 on the CelebDF [1] dataset, the accuracy was very low. This gap between the theoretical perspective and the real results made me try to use the Xception model, which has a deeper structure and better feature extraction to identify deepfakes.

This shift highlights the challenge of translating theoretical models into real-world applications. Although VGG16 was successful within the experiments with FaceForensics++ [3] that is a traditional dataset, it was not able to accommodate the more intricate and diverse data available in CelebDF [1]. The reason why the Xception model which is computationally expensive was better suited to work on this high-quality dataset is because of the complexity that comes with it. This experience proves that there should be no rigidity in the choice of models to apply when implementing theoretical frameworks as the nature of the tasks may require certain approaches.

The second theoretical approach I had in mind was merging the MTCNN [7] for face detection as I thought it would give even higher accuracy. Nonetheless, the practical application was not implemented due to the Out of Memory errors in CUDA and, therefore, I had to select a more lightweight approach. This experience is a good example of how one has to take into account not only the theoretical aspects of the work, but also the real capabilities of the hardware that is to be used.

To overcome some of these challenges, I utilized Optuna for hyperparameter tuning. This technique helped optimize the model's parameters, leading to a noticeable improvement in accuracy. The use of Optuna was a practical application of theoretical knowledge about the importance of hyperparameter tuning in machine learning, illustrating how fine-tuning can sometimes compensate for initial shortcomings in model performance.

Last one was an attempt to connect Xception with LSTM, believing that LSTM would be able to detect temporal dependencies in sequences of video frames and Xception would work on spatial ones. In theory, such a combination should have enhanced the detection of deepfakes since temporal and spatial information were being used. Nevertheless, the practical need to bring it down to a size that would make this method usable once more meant a drop in accuracy. This outcome underlined the need for enough data in training of complex models and challenges of applying theoretically sophisticated models under constraint of resources.

Deepfake detection is an area of law that is still relatively unexplored and has many implications in terms of privacy, IP, and regulation. Deepfake technology by its very nature involves the interference with the existing media and in most cases produces content that infringes on the right to privacy of an individual. For instance, unauthorized use of a person's likeness to create fake videos or images can lead to severe legal consequences, including defamation, harassment, and identity theft. These legal factors are also applicable to the deepfake detection project and the project should ensure that the technology should not infringe any laws and the rights of people.

Another important aspect is Intellectual property. In most deepfakes, the original content is not the owner's creation but rather copied from other sources, thus raising issues of ownership when an AI creates new content and whether the creator of the content is legally responsible for any infringements. To mitigate these issues, the project deals with deepfakes in the form of detection of the generated content rather than its creation, and utilizes the data set that does not infringe on the copyright laws. Also, the project is in line with the current regulatory activities that are being undertaken to fight fake news and to promote the use of artificial intelligence in a responsible manner.

It was noted, that the ethical issues connected with deepfake technology are very significant. Despite the need for deepfake detection tools in preventing malicious use of the AI generated content, one has to wonder how these tools could be used. For example, the danger of using detection technology could be in its ability to be employed as a tool to discredit lawful content or to limit freedom of speech. This shows the question of how to increase security without compromising on privacy and free speech. My project acknowledges these ethical dilemmas by adhering to strict guidelines that ensure the technology is applied only in ways that protect individuals' rights and promote transparency.

However, the effects of deepfakes on the society cannot be underestimated either. Deepfakes are a threat to society and can reduce the credibility of media, spread fake news and pose a risk to people and companies. This project wants to counter these negative impacts by creating a tool that will be able to distinguish between deep fake content and authentic content. Ethical standards were prioritized throughout the project by guaranteeing that all the conducted research and development was done in an ethical manner while attempting to cause least harm.

Another limitation that can be associated with deepfake detection is sustainability, which refers to the ability to continue the AI projects' development and the effects that they have on the environment. Deep learning models, like the ones used in this project, often require substantial computational resources, leading to high energy consumption. This is important from the environmental perspective as the amount of energy needed to train the large scale AI is not insignificant and contributes to carbon footprints.

To address these concerns, the project incorporated several strategies aimed at reducing its environmental footprint. For instance, the resources in cloud computing were properly utilized in such a way that energy was conserved and data and computational power were properly utilized. While the Xception model and dlib library are computationally heavy, careful planning and the use of hyperparameter tuning with tools like Optuna helped to streamline the process, reducing unnecessary computations and conserving energy.

Looking forward, the sustainability of the detection techniques developed in this project will depend on their adaptability and efficiency. As deepfake technology continues to advance, so must the detection methods. This requires ongoing research and development to ensure that the models remain effective without requiring excessive computational resources. The sustainability consideration in the project is an understanding of the social responsibility of the AI researchers and developers towards the environment.

Given more time, there are a few things I would have liked to do to improve the reliability of the deepfake detection model. One prominent open-ended area was the application of Generative Adversarial Networks (GANs) to their full potential. GANs are particularly relevant in this context because they are the very technology often used to create deepfakes. My hypothesis was that if GANs are capable of creating realistic fake data, they can also be used for the identification of such data due to their ability to discern

typical patterns and outliers characteristic of deepfake images.. Unfortunately, due to time constraints, I could not fully develop and integrate a GAN- based detection model into the project. This approach could have potentially provided deeper insights and improved the model's performance by leveraging the same techniques used to create the deepfakes.

Additionally, I had planned to conduct more extensive data collection and augmentation. The CelebDF dataset was chosen for its high quality and challenging deepfakes, but it could have been useful to have a more diverse data set, either by adding other inputs or by generating new data through GANs to improve the training and testing of the model. This additional dataset could have helped in extending the generalization of the model across different deepfake types that can be seen in the real world.

Furthermore, extending the application of deepfake detection technology to other domains of use apart from videos is another promising area. For instance, furthering the research on its ability to identify audio deepfakes or the text-based manipulations by AI can help in combating the increasing problem of misinformation in both audio and text media. Such cross-domain could result into the synthesis of multi-modal detection systems that are capable of detecting deep fake in text, audio and video.

The complexity of deepfake technology and its societal implications call for interdisciplinary collaboration. In future work, partnerships with experts in law, ethics, and social sciences could be instrumental in shaping the development and deployment of deepfake detection tools. Policy-makers and legal experts could provide valuable insights into the regulatory frameworks needed to ensure that these tools are used responsibly and effectively, while social scientists could help assess the societal impact and guide the ethical deployment of the technology.

Collaboration with AI researchers and developers from other fields, such as natural language processing or audio signal processing, could also lead to innovations in detecting deepfakes across various media types. By bringing together diverse perspectives and expertise, it would be possible to create a more holistic approach to combating the spread of AI-generated misinformation.

Scaling the deepfake detection solution for broader applications is another critical consideration for future work. Creating plugins for the web browsers, integrating into the social networks or the content management systems, real-time detection systems, could help to immediately identify the deepfakes and potentially halt the sharing of fake-news at the source. Developing lightweight, scalable models that can be deployed on various devices, including smartphones and low-power servers, would be essential for achieving this goal.

Deploying the technology at scale presents both challenges and opportunities. One challenge is ensuring that the detection system remains efficient and accurate as the scale of data increases. This may require optimization techniques, such as model pruning or distillation, to reduce the computational load without sacrificing performance. Another challenge is the need for continuous updates to the detection model to keep pace with the evolving nature of deepfake technologies. This could be done by incorporating features for automated re-training of the system using the latest deepfake data in the market.

In this essay, I have reflected on the development and outcomes of the deepfake detection project, exploring its strengths, weaknesses, and the complex relationship between theoretical models and practical applications. The project highlighted the importance of integrating advanced AI models, like Xception, and the need for continuous adaptation to tackle the evolving challenges posed by deepfake

technologies. Despite the obstacles faced, such as computational limitations and dataset constraints, the project made significant strides in developing a robust detection tool.

Through this project, I experienced considerable personal and professional growth. I honed my technical skills in deep learning, computer vision, and model optimization while also gaining a deeper understanding of the ethical and societal implications of AI. This experience has reinforced the importance of responsible AI development and has prepared me for future work in this dynamic field.

References –

[1] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, ”Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” 2020 IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3207-3216, doi: 10.1109/CVPR42600.2020.00327.

[2] Y. Lin, Y. Qu, Y. Li and Z. Nie, "Exploring Generalization Capability for Video Forgery and Detection based on Generative Adversarial Network," *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2020, pp. 1575-1580, doi: 10.1109/CSCI51800.2020.00291. keywords: {Training;Computational modeling;Mean square error methods;Harmonic analysis;Generative adversarial networks;Feature extraction;Forgery;Deepfake;Deepfake Detection;Generative Adversarial Network;Face Swap;Detection Generalization}

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Chris- tian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In ICCV, 2019.

[4] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854, 2019.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[6] A. Das, K. S. A. Viji, and L. Sebastian, "A Survey on Deepfake Video Detection Techniques Using Deep Learning," Second International Conference on Next Generation Intelligent Systems (ICNGIS), 2022.

[7] J. Xiang and G. Zhu, "Joint Face Detection and Facial Expression Recognition with MTCNN," 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 2017, pp. 424-427, doi: 10.1109/ICISCE.2017.95. keywords: {Face detection;Face;Face recognition;Training;Feature extraction;Neural networks;convolutional network;face detection;facial expression recognition;MTCNN},