APS360 TEAM 53 PROJECT PROPOSAL

Shixuan (Shawn) Huang
Student# 1006964171
shixuan.huang@mail.utoronto.ca

Qingran Chen
Student# 1006724751
qingran.chen@mail.utoronto.ca

Jiahe Lu
Student# 1007213604
joeyjiahe.lu@mail.utoronto.ca

Yunxiang Zhang
Student# 10059850891
vincentyunxiang.zhang@mail.utoronto.ca

ABSTRACT

This project aims to address the pressing issue of deepfake detection within digital media, spotlighting the necessity for advanced verification methods against deceptive content, such as manipulated videos of well-known personalities. Utilizing deep learning techniques, the initiative develops a Convolutional Neural Network (CNN) model to differentiate authentic visuals from fabricated counterparts. The model is honed on a vast and varied image repository provided by Tushar Padhy, aligning with proven detection strategies from significant online challenges and AI security applications. This abstract presents a blueprint for a detection tool that is not only effective in identifying deepfakes but also adheres to ethical data usage and privacy standards. —-Total Pages: 7

1 INTRODUCTION

Deepfake technology, highlighted by the creation of convincing yet fake videos featuring celebrities like Taylor Swift, emphasizing the importance of improving security protocols for videos and images shared online. This project aims to focus on employing deep learning, an advanced computational approach, to distinguish between realistic and fabricated content. This technical approach is critical due to the potential of deepfakes to deceive individuals and disrupt sectors such as politics and entertainment. The objective of this project is to augment the safety and reliability of online environments, ensuring the integrity of visual content. Through the application of deep learning, the initiative aims to excel in identifying fabricated content, thereby contributing to a more secure internet landscape.

2 ILLUSTRATION

This project plans to utilize a CNN model to analyze visual imagery. As demonstrated in Figure 1, a human face image was first segmented into pixels and used as input for the model. The model then employs a kernel (e.g., a 3×3 matrix) and applies it to the input to extract convolved features. As these features are passed to deeper layers, the model can identify increasingly complex patterns. Ultimately, the features extracted from the CNN and pooling layers are flattened and transformed into learned features. The final output represents the identified features.

3 BACKGROUND RELATED WORK

3.1 KAGGLE DEEPFAKE DETECTION CHALLENGE:

AWS, Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee, and academics have come together to build the Deepfake Detection Challenge (DFDC) in 2020. (DFDC,

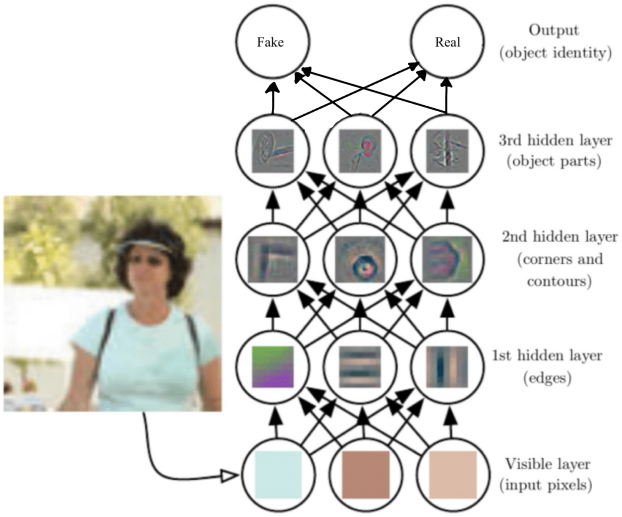

Figure 1: Illustration of CNN model. Image: (Mandal, 2024).

2020) This competition provides an example of mature criteria for evaluating the model, including the log loss function used for the final model evaluation. The competition also provides a large public database for reference.

3.2 Sentinel AI:

Sentinel is a leading AI-based protection platform that helps democratic governments, defence agencies, and enterprises stop the threat of deepfakes. It is used by leading organizations in Europe. As demonstrated on the website, this large commercial model also uses neural network classifiers, guiding the future development of the model (Sentinel, 2024).

3.3 Deepfake detection accuracy by human and machine:

This paper deeply explores the performance of human, machine, and crowd wisdom when facing DFDC. It collects human decisions through an experimental website, "Detect Fake." (Negar Kamali) The model generally outperforms individual participants, with an accuracy rate of 80 percent compared to participants' rates ranging from 66 to 69 percent. However, when participants accessed the model's predictions, their accuracy increased, indicating potential for human-AI collaboration in enhancing detection capabilities (Matthew Groh, 2021).

3.4 Deepfake Detection Techniques Using Deep Learning: A Survey:

This paper briefly reviews the deepfake creation and detection technology. It discussed various methods for detection using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid approaches. The paper also highlights datasets' importance in training, briefly introducing several reliable public datasets (Almars, 2021).

3.5 Deepfake Detection with Deep Learning:

CNN versus Transformers. This paper details the performance of eight deep learning architectures (four CNNs and four Transformers) and compares datasets, observing the strengths and uniqueness of the Deeper Forensics, DFDC, and FF++ 2020 datasets. Overall, CNN models did better in the same train-to-test dataset evaluations, and the Transformers models did better in cross-dataset evaluations. Throughout the eight models, the HRNet, XceptionNet, and EfficientNet B7 CNNs models performed very well consistently when trained and tested on the same dataset. In contrast, the XceptionNet, ViT, BeiT, and Swin Transformer performed better in cross-data validation (VrizlynnL.L., 2023).

4 Data Processing

In selecting an appropriate dataset for the deep learning model, the collection contributed by Tushar Padhy, comprising over 140,000 portrait images, was selected (Padhy, 2024). This dataset is divided into training, validation, and testing categories, with further segregation into real and fake images to ensure a balanced representation for in-depth analysis, maintaining a precise 1:1 ratio between the two. Each image is coloured and uniformly set to a resolution of 256x256 pixels, facilitating consistency in the data processing. Importantly, the dataset is characterized by its equitable gender distribution and encompasses a broad spectrum of ages and ethnic backgrounds.

The team will conduct a visual inspection with random sampling on the dataset across all segments—training, validation, and testing—to identify and remove duplicate images from individuals and across categories. This operation ensures the authenticity of each image file, to eliminate those identified as corrupted. Additionally, a review of the dataset will be conducted to uncover any latent biases towards particular demographic groups. This ensures that the model performs uniformly well across the wide spectrum of ages and ethnic backgrounds represented, maintaining the fairness and dependability of the deepfake detection initiative.

After a comprehensive review of the training, validation, and testing datasets. The team will elect to maintain the original segmentation of these datasets and plans to utilize the training set for the development of the model, the validation set for internal testing, and the test set for the ultimate evaluation of the model's performance. The immediate objective is to improve the model's accuracy in detecting deepfake images by incorporating more training images, subject to the limitations imposed by the project's timeline. However, the quantity of images integrated into the training process may be revised as the development of the model advances.

5 Architecture

The goal of this project is to develop a model capable of identifying deepfake images. Essentially, it functions as an image classification model. The model is composed of the following components:

5.1 Loading the Dataset:

The model needs to process a cleaned and well-categorized dataset, with each data labelled for subsequent output comparison. The dataset is organized into three distinct sections: Training data, to educate the model; Validation data, to ensure the training's effectiveness, detect any potential overfitting or underfitting, and for use in comparison against testing data during the evaluation phase; Testing data, intended for the final evaluation of the model's performance.

5.2 Training the model and providing results:

Once the data is loaded, the specific type of neural network will be established. A training function will then take this neural network and train the model using various hyperparameters. This training process is designed to yield two main outputs: error, which represents the rate of error encountered during training and validation phases, and loss, determined by the selected loss function to indicate the deviation of predictions from expected outcomes.

5.3 Plot the training curve:

After obtaining the results from training, the model will chart the training curve to investigate the presence of any overfitting or underfitting. To accomplish this, the team needs to store the training results in several files on the disk and read from them to plot the training curve.

5.4 Evaluate the training:

After completing the training, analysis, and fine-tuning processes, the team will proceed to assess the model's performance using the test dataset. This evaluation will yield the error rate and loss, which should then be benchmarked against the performance of the baseline model for comparison.

For this project, the team decided to choose CNN(convolutional neural network) as the neural network:

- The model is based on image processing and classification.
- CNN can effectively extract features from images and learn to recognize patterns, making them well-suited for object detection, image segmentation, and classification tasks (Tripathi, 2023).
- The convolutional layers of CNN can identify the simple features like texture and edges in the lower level and more complex features like objects in the higher level (Yamashita et al., 2018). This capability mirrors how humans visualize the picture.
- The same weight(filter) is applied across the whole image, thus allowing the network to search for the same features (Yamashita et al., 2018). This parameter sharing significantly reduces the parameters needed to compare with the fully connected networks. It makes CNN more efficient in image processing and less overfitting when processing high-dimensional image data.
- The choice of the loss function will be Binary Cross-Entropy(BCE). The BCE is used for binary classification tasks, where in the model, it will only decide if the image is either deepfaked or not deepfaked (normalized to 0/1). Therefore, BCE is an appropriate choice for the loss function.

6 Baseline Model

The initial architecture of the project employs a Convolutional Neural Network (CNN) designed to process and analyze image data efficiently. This baseline model includes several key components, each chosen for its specific contribution to the model's performance:

- Convolutional Layer: To identify critical features such as textures and edges within the images. The team intends to use 32 filters of size 3x3, a setup that strikes a balance between capturing detailed features and managing computational resources.
- Activation Function: The ReLu (Rectified Linear Unit) function is selected due to its effectiveness in adding non-linearity, which is crucial for learning complex patterns. ReLu is favoured for its simplicity and ability to mitigate the vanishing gradient problem.
- Pooling Layer:A max pooling layer with a 2x2 size follows the convolutional layers to reduce the spatial dimensions of the feature maps. This step is vital for decreasing computational load while retaining the most significant features.
- Fully Connected Layer: This layer integrates the learned features for the final classification. The size of this layer will be adjusted based on the dataset's complexity and the number of output classes to prevent overfitting.

The baseline model aimes to be simple but eligible as a reference point to compare with the complex model. The developed advanced model is expected to outperform the baseline model (lower error rate and less loss) to justify the model's improvements in performance. This baseline model's architecture is not only geared towards understanding complex visual inputs but also set up to be a comparative standard for evaluating subsequent models. As the model development progresses, alternative models will be compared against this baseline by varying several aspects:

- Adjusting Hyperparameters: The hyperparameters of the baseline model will be set to a default value, where the team will tune the advanced model with different learning rates, batch size, number of layers, the number of filters in convolutional layers, and the size of the fully connected layer. These adjustments can significantly impact the model's learning ability from the dataset. However, both models will be trained with the same number of epochs to be evaluated and visualize how well the advanced model can outperform the baseline model.

- Changing Functions:Experiment with different activation functions like Leaky ReLu or sigmoid to assess their impact on model performance.
- Model Architecture Modifications: Introducing variations in the architecture, such as adding dropout layers to reduce overfitting or incorporating batch normalization to improve training stability.

The team intends to employ a suite of performance indicators, including accuracy, precision, recall, and the F1 score, for the assessment of these models. Additionally, the computational efficiency and generalization capability of the models will be evaluated by analyzing their performance on a validation set. This strategy aims at enhancing the model's effectiveness and efficiency, ensuring the development of a robust and optimized solution.

7 Ethical Considerations

As the team strives to create an AI model to counteract deepfakes, it is crucial to emphasize the ethical gathering and use of data, ensuring that it respects individuals' rights. The project will be guided by the following principles:

Consent Acquisition:

It is essential to obtain explicit consent from individuals before utilizing their images, videos, or audio recordings for AI training purposes. Utilizing personal data without prior authorization is unethical and can lead to dissatisfaction. Commitment to using data exclusively from individuals who have provided explicit consent will be upheld.

Privacy Protection:

Safeguarding the privacy of individuals whose data is utilized is crucial. Efforts will include anonymizing data whenever feasible and ensuring that access is restricted to authorized personnel only. Adherence to privacy regulations such as the GDPR is mandatory, yet the commitment extends beyond legal compliance to a moral obligation to protect personal information.

Ensuring Equity and Mitigating Bias:

The risk of incorporating biased data, which might not accurately represent diverse populations, thereby introducing unfairness into AI behaviour, is acknowledged. To counteract this, a commitment to utilizing a broad spectrum of data is made, aiming for AI models that are equitable and effective across diverse groups.

Transparency and Accountability:

A commitment to transparency regarding the methods of data collection, utilization, and safeguarding is established. Additionally, there is an acknowledgment of responsibility for the decisions made in the development and application of AI technologies. Openness to acknowledging errors and rectifying them is essential for maintaining public trust.

By adhering to these ethical guidelines, the objective is to not only enhance the AI models' capability in detecting and distinguishing deepfakes but also to ensure the approach respects and upholds the dignity and rights of all individuals. The team is dedicated to advancing this project with a commitment to ethical integrity, aiming to contribute positively to the digital environment.

8 Project Plan

The internal dealine and milestone for the Project has been settled.See Table 1

Table 1: Task Milestone

Task	People	Internal Deadline
Project Proposal		
Data collection	Vincent/Isabel	5/Feb
Data cleaning	Shawn/Joey	7/Feb
Related-Research	Isabel/Joey	7/Feb
Proposal documentation	All	9/Feb
Progress Report		
CNN model construction	Vincent/Shawn	5/Mar
Baseline model construction	Isabel/Joey	7/Mar
Primary Model set up	Vincent/Shawn	7/Mar
Progress report documentation	All	9/Mar
Project Presentation		
Report Draft	Shawn/Isabel	1/Apr
PowerPoint Setup	Vincent/Joey	3/Apr
Recorded Presentation	All	3/Apr
Final Report		
Final Model Testing	Vincent/Shawn	5/Apr
Data collection	All	6/Apr

8.1 WAY OF WORKING TOGETHER

- Team collaboration on coding will be conducted through Google Colab
- Members must immediately inform the project manager of any issues or delays, along with the work completed so far
- Weekly meetings will be held for open communication and assignment of coding tasks.
- To prevent code conflicts, new code blocks should be clearly labelled with their purpose and the author's name.
- Code should be well-commented for clarity, ensuring easy maintenance and readability.
- Use Colab's version control to monitor changes; always save a version before significant updates.
- Team members should proactively share knowledge and provide constructive feedback in the group chat.
- For conflicts, schedule an urgent meeting for resolution; unresolved issues should be taken to the TA by the project manager or a designated person.
- Meeting Schedule: Weekly Zoom or Microsoft Teams meetings, with extra meetings arranged as needed.
- Communication: Team interaction through WeChat, expecting responses within 24 hours for urgent matters.

9 Risk Register

Drop-out of member

- Task re-assignment: If a member drops out, their workload and responsibilities must be immediately reassessed, and tasks should be reassigned within the team to ensure the smooth progress of the project.
- Work transfer: Ensure the leaving members have documented and uploaded their work so members can take over their part.

Model-training over time

- Change modelling strategy: change current model training strategies like changing the batch size or other hyperparameters
- Dataset Processing: Check the quality and size of the dataset and try to reduce the dataset's size or optimize the data preprocessing pipeline to speed up training.
- Internal deadline: set-up internal deadline to prevent emergency issues

Work not finished within the internal deadline

- Task reassignment: Evaluate unfinished work and reassign tasks based on priority to ensure that important tasks are completed
- Priority management: Communicate with relevant team members, reassess project goals and work breakout, negotiate with them to adjust deadlines or reschedule work to ensure the quality and timely completion of work.

10 Link to GitHub/Colab Notebook

https://colab.research.google.com/drive/1FIdtQSsjdUm4Gcgm96ySg1RoQFDVHSYs?usp=sharing
Link to Dataset: https://www.kaggle.com/datasets/tusharpadhy/deepfake-dataset

References

Abdulqader M. Almars. Deepfakes detection techniques using deep learning a survey, May 2021. URL https://www.scirp.org/journal/paperinformation?paperid=109149.

DFDC. Deepfake detection challenge, 2020. URL https://www.kaggle.com/c/deepfake-detection-challenge/overview.

Manav Mandal. Introduction to convolutional neural networks (cnn), Feb 2024. URL https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/.

Chaz Firestone Rosalind Picard Matthew Groh, Zip Epstein. Deepfake detection by human crowds, machines, and machine-informed crowds, December 2021. URL https://doi.org/10.1073/pnas.2110013119.

Jessica Hullman Matt Groh Negar Kamali, Angelos Chatzimparmpas. Deepfake, can you spot them? URL https://detectfakes.kellogg.northwestern.edu.

Tushar Padhy. Deepfake-dataset (140k + dataset real or fake), Jan 2024. URL https://www.kaggle.com/datasets/tusharpadhy/deepfake-dataset.

Sentinel. Defending against deepfakes and information warfare, 2024. URL https://thesentinel.ai/.

Mohit Tripathi. Image processing using cnn: A beginners guide, May 2023. URL https://www.analyticsvidhya.com/blog/2021/06/image-processing-using-cnn-a-beginners-guide/.

VrizlynnL.L. Deepfake detection with deep learning: Convolutional neural networks versus transformers, April 2023. URL `https://arxiv.org/pdf/2304.03698.pdf`.

Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: An overview and application in radiology - insights into imaging, Jun 2018. URL `https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9`.