Deepfake Detection - An Exploration of Dataset Size and Deepfake Mediums

Mark Glinberg, Charles McCartney, Andrew Gao, Jerome Perera
Georgia Institute of Technology
Github Repository

Abstract

This project aims to examine the effectiveness of different models using convolutional neural networks and visual transformers in detecting deepfake facial images. Using datasets with varying sizes, including 10,000 and 140,000 real and fake faces, this study compares the performance of these models with respect to metrics such as accuracy and F1 score. Furthermore, the project explores the impact of dataset size and depth on model performance and considers techniques such as data augmentation to improve results. Finally, this project delves deeper into deepfake detection by exploring a novel 3D CNN architecture to detect deepfake videos.

1. Introduction/Background/Motivation

Given recent advancements in the complexity and realism of generative deepfake technologies, the ability to detect artificially generated image, video, and audio artifacts has become an increasingly important research objective to prevent the consequences of misinformation and defamation in digital spaces. Modern deepfake techniques are capable of producing highly realistic fabricated images, videos, and audio recordings which present difficulties in distinguishing between real and fake media content. The primary objective of this project is to investigate the prevailing techniques and model types used to identify artificially generated facial images. Following this, an additional objective was established to experiment with these deep learning methods in the context of distinguishing real and fake videos and audio recordings. This was done in order to establish a deeper understanding of how various architectures and model frameworks influence the process of accurately identifying different types of artificially generated media.

Although existing approaches involving Convolutional Neural Networks, Transformers, and Generative Adversarial Networks have demonstrated promising results, further areas for research remain to investigate their performance across datasets with different characteristics as well as to ascertain which types of models are best suited for various objectives. This project seeks to address this gap by evaluating the Xception model, which is known for its high accuracy in deepfake detection, as well as an open-source deepfake Visual Transformer (ViT) model, which applies encoding and decoding techniques to produce highly accurate classification results. In particular, the Xception model was chosen as a preferred candidate model for this investigation given its ability to outperform similar CNN-based models on a variety of datasets. For example, Xception significantly outperforms Inception V3 on large image classification datasets including 350 million images and several thousands of classes [3]. Additionally, the open-source ViT model was selected such that the study could contrast the results of a model using an architecture distinct from primarily CNN-based models [5]. Thus, while many other approaches exist for this problem scope, the design of this study was effective in investigating the distinction between various types of architectures and identifying the most effective model characteristics in addressing the identification of artificially generated images.

Additionally, this study seeks to explore the relationship between the model characteristics that are conducive for success in detecting deepfake images as well as those which can be applied for deepfakes involving video and audio artifacts. To investigate the problem scope involving videos and audio recordings, a novel architecture was implemented involving a combination of a three-dimensional CNN as well as a standard neural network used to interpret video frames and Mel-Frequency Cepstral Coefficients (MFCCs).

An big problem faced by developers and stakeholders for deepfake detection algorithms is the fact that the quality and realism of deepfakes are constantly improving and can be expressed through an expansive range of content types. Notably, the results of a workshop consulting AI experts in 2020 indicated that deepfakes were identified as the most serious threat pertaining to artificial intelligence technologies by a panel comprised of individuals from academia, the private sector, and state agencies [2]. Among the foremost obstacles in mitigating the consequences of deepfakes, such artificially generated content is often proliferated through unregulated channels and is therefore not subject to the controls and policies which could otherwise be applied to reduce the spread of misinformation and defamatory materials. Thus, it follows that a successful investigation of the model characteristics conducive to detecting artificially generated content can serve to benefit a wide body of stakeholders comprised of individuals from both private and public sectors, as well as people in the online space who may yet be subject to the consequences of the proliferation of deepfake content.

For the experiments pertaining to artificially generated images, this study leveraged the Real vs. Fake Faces - 10k and Real vs. Fake Faces - 140k datasets sourced from Kaggle. Ensuring consistency in training the models, the dataset containing 10,000 images is a subset of the images included in the dataset containing 140,000 real and artifically generated facial images. This subset relationship between the datasets removes any potential confounds involving image composition and other characteristics of a given dataset that might present obstacles to an analysis of how dataset size influences the performance of the different models included in this study. Furthermore, all images included in each of these datasets were resized to 256x256 pixels, which were then reformatted depending on the structure of the models included in further experiments [8, 12].

For the experiments involving combined video and audio recordings, this study leveraged a subset of the Deepfake Detection Challenge Dataset (DDCD) from Kaggle. Specifically, this subset included 400 videos of the total dataset due to constraints involving time and computational resources. Each of these videos was uniformly processed and resized to 155x155 pixels to facilitate the training process for this model [1].

2. Approach

2.1. Data Preprocessing

2.1.1 Images

Before starting with data preprocessing, we first needed to extract the datasets and change the file structure within both the 10k and 140k datasets to match one another. We also needed to create a test set for the 10k dataset, since it included only a train and validation set, rather than a train, test, and validation set like the 140k dataset. We did this by splitting the 10k validation set in half, such that we had a test and validation set each with 1,500 images.

In the data preprocessing for all of our models, we employed image augmentation. Image augmentation involves artificially expanding the size and diversity of a dataset by applying various transformations to the original images, such as rotations, shifts, flips, and changes in scale. This process helps us create a robust model by simulating a wider array of real-world scenarios that the model might encounter. This not only helps in reducing overfitting - where a model performs well on training data but poorly on new, unseen data — but also improves the model's overall accuracy and reliability in detecting deepfakes. We then resized the images to 299 x 299 for the Xception models, and 224 x 224 the ViT model.

Each model then ran predictions on the test set to measure its performance. The test and validation sets were made up of 1,500 images each for the 10k dataset and 20,000 images each for the 140k dataset. The train dataset had 7,000 images in the 10k dataset and 100k images in the 140k dataset. Each train, test, and validation set across both datasets had an even split of fake and real images.

2.1.2 Video and Audio

Video

Four frames were uniformly extracted across the total duration of each video to capture various motion states, resized to 224x224 pixels, and normalized.

Audio

The audio tracks were converted into 13-dimensional Mel Frequency Cepstrum (MFC), which is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The dimensions of an MFC are the Mel Frequency Cepstral Coefficients (MFCCs) [4]. Here, we used the Librosa Python library to extract MFCC's for each video, then used the mean of all MFCs in a video as input to the audio model.

2.2. CNN Model: Xception

The choice of Xception as the primary model for this task was influenced by several factors. Xception is a deep learning model based on an Inception architecture that modifies the typical Inception modules to use depthwise separable convolutions [3]. This architecture has shown excellent performance in tasks involving image classification and involves features that are particularly suited for processing the complex patterns found in images pertaining to artificial generation. The depthwise separable convolution layers allow for efficient model learning with fewer parameters than traditional convolutional layers, making it both computationally efficient and highly effective. Although models such as Inception with increased model size, depth and computational cost are traditionally thought to translate to performance gains in wide ranges of tasks, efficiency and lower parameter counts are often preferable for large data scenarios [11].

Our group also added additional layers to tailor the Xception model to more appropriately align with the nuances and characteristics of the deepfake facial image datasets. These changes included additional dense layers and adjustments in dropout rates to further refine the model's ability to discern subtle manipulations in images indicative of deepfakes.

The hypothesis was that Xception, with its architecture promoting feature independence and efficient use of model parameters through depthwise separable convolutions, would effectively capture the complex and subtle differences between real and manipulated images [3]. The expectation was that the robust feature extraction capabilities of Xception would result in high accuracy in classifying the images as real or fake, addressing the critical need for reliable automatic detection of deepfake content for images with diverse sets of characteristics.

The setup for the Xception model was more difficult than expected. Tensorflow errors in GPU detection necessitated clean reinstallations of CUDA and Python. Furthermore, it was necessary to monitor the training process to ensure that crashes did not occur which would nullify progress and thereby cause unnecessary repeated computations and labor.

2.3. Visual Transformer (ViT) Model

In contrast to the Xception model whose architecture is primarily comprised of convolutional neural network layers, the open-source Deepfake vs. Real Faces Detection ViT model was selected for this experiment in order to analyze the performance of an architecture which adopted techniques pertaining to vision transformers. Unlike CNN-based architectures like Xception, which demonstrate strong performance in capturing local features, visual transformers excel at capturing global features and relationships across the entire image [9]. The features used in the vision transformer were adapted from attention transformers that have often been used in the context of Natural Language Processing (NLP). In this process, images are divided into patches which are flattened into vectors and linearly embedded to produce a patch embedding matrix. Such transformers utilize positional encodings where each position is converted to a vector mapping, and each encoded object in the sequence is aggregated with the positional encoding [10]. The aggregated patch embeddings and positional encodings are fed into the transformer's encoder containing several layers of self-attention mechanisms. Thereafter, the output is fed into a classification head which outputs a prediction for whether the inputted image is classified as real or fake [6]. Although it is widely believed that convolutional neural network architectures have emerged as the preeminent method for computer vision tasks, this study posits that new and distinct architectures should be considered in contemporary investigations of deepfake detection approaches [3].

For the purposes of implementing this model, a preexisting code base from Kaggle was adapted to align with the characteristics of the Real vs. Fake Faces - 10k and 140k datasets, as well as the train, validation, and test dataset splits required to facilitate accurate results [5]. In particular, the dataset originally used to train this model contained different attributes, which necessitated an overhaul of the data augmentation process, instantiation and conversion of Pandas data frames, and the process of splitting the data to facilitate validation and post-training evaluation with the test set.

Given that this model is an open-source artifact found only on HuggingFace and Kaggle, there exist very few additional resources or documentation which describe this model's architecture in specific detail. To address this shortcoming, a variety of supplementary resources were referenced pertaining to visual transformers as well as their application to this particular use case in order to establish a more thorough understanding. Additionally, several obstacles were encountered during the training process with the 140k image dataset in which training computations alone took 11.5 hours to perform with GPU capabilities, alongside an additional 0.5 hours required to evaluate the results.

2.4. Audio and Video - Novel Architecture

The video sub-model consists of a 3D convolution layer with 32 filters, followed by batch normalization and max pooling. A second 3D convolution layer with 64 filters further processes the data, followed by batch normalization, max pooling, and global average pooling to condense the video features into a higher-level representation. This model also includes the computation of class weights to offset the imbalance in the DDCD dataset, ensuring that the model does not bias towards the more frequent class. By adjusting the model's focus towards the underrepresented 'REAL' class, the class weights help in improving the model's sensitivity towards 'REAL' videos, despite the predominance of 'FAKE' samples.

As mentioned in Data Preprocessing, or section 2.1.2, the Audio sub-model processes audio signals using MFCCs. This input is then fed into a dense neural network consisting of a sequence of layers with 256 and 128 neurons, each followed by a 0.5 dropout layer to reduce the possibility of overfitting negatively impacting the resultant performance of the model. These two sub-models were then combined with a dense layer and a sigmoid activation.

This dual-stream approach, especially the use of a three-dimensional CNN Which have shown promising accuracy metrics in classifying video data relative to two-dimensional CNNs, combined with targeted architectural choices and strategic class weighting, was anticipated to be highly effective in detecting deepfakes by leveraging both visual and auditory discrepancies that are typically harder to fake convincingly [7].

During this investigation, we encountered several critical obstacles. High-quality deepfake datasets such as the

DeepFake Detection Challenge Dataset (DFDC, different dataset than the DDCD from Kaggle) were difficult to access, often requiring access approval through form requests from which access was never acquired. Datasets such as the DFDC often include additional deepfake types and methods, more data, and are less skewed in the composition of real and fake videos. Google Colab Pro, which was used as our primary environment for this model, posed limitations with frequent crashes and timeouts due to the large data sizes and extensive operations required for extracting dataset files, extracting frames, and extracting MFCC data. Video data preprocessing, including frame extraction and MFCC feature computation, was overly time-consuming, with frame extraction for even 16 frames leading to a Colab timeout after 5 hours of runtime. This led to a reduction in the number of frames per video from the planned value of 32 to only 4. Furthermore, initial plans to employ an RNN LSTM network for audio processing were abandoned because of persistent RAM overflows in Google Colab, which eliminated the use of complex temporal MFCC sequences and forced the use of a basic neural network and the input consisting of the mean of the MFCCs.

3. Experiments and Results

Success in this project was quantitatively measured using several key performance metrics, including accuracy, precision, recall, and F1-score, across both facial image datasets (10k and 140k images). Additionally, the Receiver Operating Characteristic (ROC) curve, Confusion Matrix, and the Precision-Recall (P-R) curve were generated to evaluate the model's trade-offs between true positive rates and false positive rates and precision vs. recall, respectively.

3.1. Xception

3.1.1 Evaluation Metrics

The Xception model performed well across both datasets as you can see in tables 1, 2, with better metrics observed in the larger dataset as expected. While the model achieved high accuracy, there are areas of potential improvement, particularly in balancing recall rates between classes, as evidenced by the disparity in recall values between 'real' and 'fake' classifications.

	Precision	**Recall**	**F1 Score**	**Accuracy**
Fake	0.76	0.83	0.80	
Real	0.81	0.74	0.78	
Average	0.79	0.79	0.79	0.79

Table 1: Evaluation Metrics for Xception Model on 10k Dataset

	Precision	**Recall**	**F1 Score**	**Accuracy**
Fake	0.84	0.88	0.86	0.85
Real	0.87	0.83	0.85	0.85
Average	0.85	0.85	0.85	0.85

Table 2: Evaluation Metrics for Xception Model on 140k Dataset

3.1.2 Receiver Operating Characteristic Analysis

The Receiver Operating Characteristic (ROC) curve is a fundamental tool for diagnostic test evaluation. In this project, the ROC curves for the 10k and 140k images datasets demonstrate the model's capability to distinguish between 'real' and 'fake' classes under varying threshold settings.

The ROC curve for the 10k images dataset (Figure 1) shows an area under the curve (AUC) of 0.87, indicating a very good true positive rate (TPR) against the false positive rate (FPR). This suggests that the model, even with a relatively smaller dataset, is quite effective at distinguishing between the classes but could benefit from reduced false positives.

Xception 10k ROC Curve

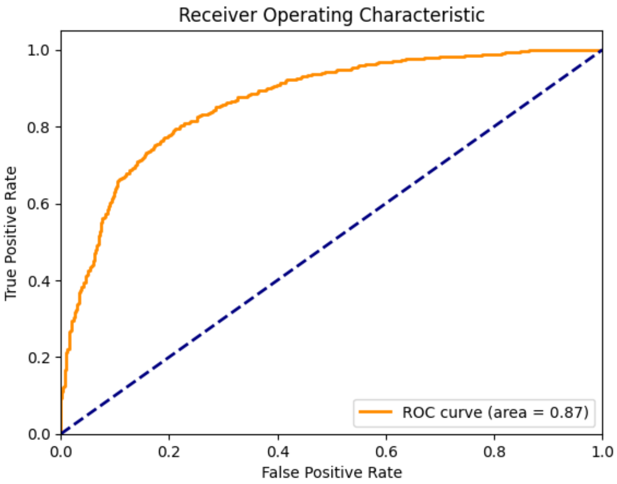

Figure 1: This figure illustrates the Receiver Operating Characteristic Graph for the Xception model on the Real vs. Fake Faces - 10k dataset.

For the larger dataset, the ROC curve (Figure 2) exhibits an AUC of 0.93, showing a significant improvement. This high AUC value reflects an excellent balance between sensitivity and specificity, implying that the model is highly capable of detecting 'real' images as real and 'fake' images as fake with a lower rate of incorrect classifications.

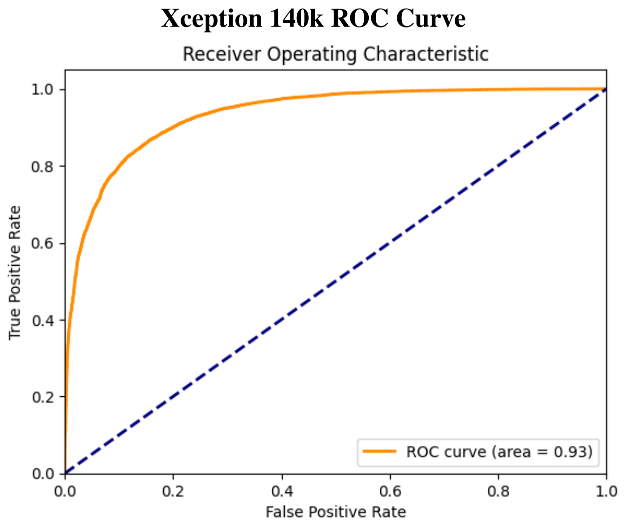
Xception 140k ROC Curve

Figure 2: This figure illustrates the the Receiver Operating Characteristic Graph for the Xception model on the Real vs. Fake Faces - 140k dataset.

3.1.3 Confusion Matrices Analysis

Confusion matrices provide a straightforward visualization of model performance, especially focusing on the types of errors made by the classifier.

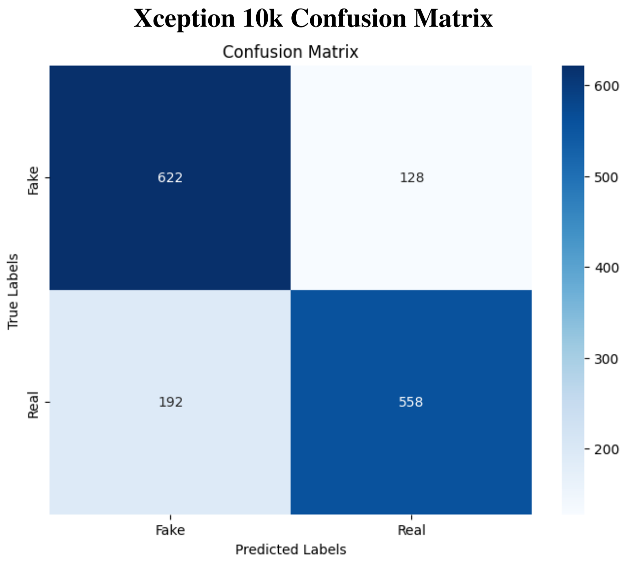
Xception 10k Confusion Matrix

Figure 3: This figure illustrates the Confusion Matrix for the Xception model on the Real vs. Fake Faces - 10k dataset.

The confusion matrix (Figure 3) reveals 622 true positives (TP) for ’fake’ and 558 true negatives (TN) for ’real’, against 128 false negatives (FN) and 192 false positives (FP). This is indicative of a propensity for the model to disproportionately identify images as falling under the fake categorization.

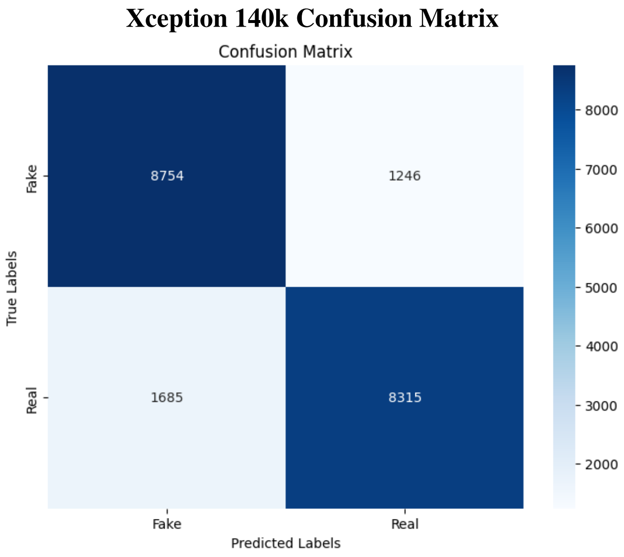
Xception 140k Confusion Matrix

Figure 4: This figure illustrates the Confusion Matrix for the Xception model on the Real vs. Fake Faces - 140k dataset.

Regarding the confusion matrix for the Xception model on the Real vs. Fake Faces - 140k dataset (Figure 4), with 8754 TP for ’fake’ and 8315 TN for ’real’, alongside 1246 FN and 1685 FP, the model shows substantial improvement in both detecting real images correctly and reducing false positives compared to the results obtained using the smaller dataset. The results are indicative of enhanced feature learning from a substantially larger dataset, leading to better generalization and task performance for a more diverse range of images.

3.1.4 Precision-Recall Curves Analysis

Precision-Recall (P-R) curves are critical in evaluating model performance when dealing with imbalanced datasets. They focus on the relationship between precision, denoting the accuracy of the positive predictions, and recall, denoting the ability to identify all positive instances.

The Precision-Recall curve for this dataset (Figure 5) shows a relatively impressive balance with an average precision (AP) of 0.86. The curve begins with high precision but undergoes a decrease as recall increases, which might be typical in scenarios where achieving high recall slightly compromises precision.

Xception 10k Precision Recall Curve

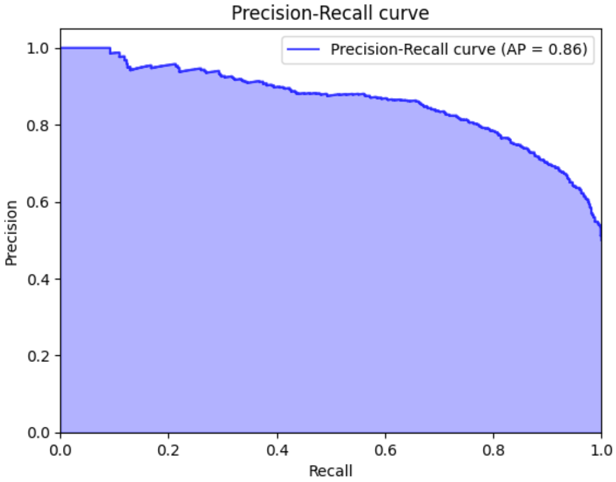

Figure 5: This figure illustrates the Precision Recall curve for the Xception model on the Real vs. Fake Faces - 10k dataset.

The Precision-Recall curve for the larger dataset (Figure 6) begins with very high precision and sustains this value along the recall axis, ending with an AP of 0.93. This indicates that the model is exceptionally effective at maintaining high accuracy in its positive predictions across various thresholds, benefiting significantly from the larger training set.

Xception 140k Precision Recall Curve

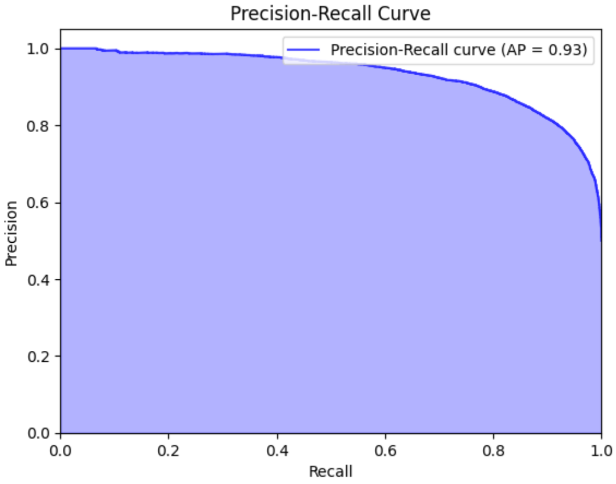

Figure 6: This figure illustrates the Precision Recall curve for the Xception model on the Real vs. Fake Faces - 140k dataset.

3.2. Visual Transfomer (ViT) Model

3.2.1 Evaluation Metrics

The visual transformer model performed similarly to the Xception model on the 10K dataset. As you can see in (Table 3), the 78% accuracy is only a 1% difference when compared to the Xception model accuracy on the 10K dataset.

	Precision	**Recall**	**F1 Score**	**Accuracy**
Fake	0.72	0.92	0.81	
Real	0.89	0.65	0.78	
Average	0.81	0.78	0.78	0.78

Table 3: Evaluation Metrics for ViT Model on 10k Dataset

While the above is a quality result, the ViT model's evaluation metrics on the 140K dataset (Table 4) indicate an extremely impressive accuracy of 97%. Given that this model was evaluated using an unseen test dataset which was split from the training / validation sets in the pre-training phase, this cannot be due to overfitting and is a reliable indication of the strength of this visual transformer model's performance for this task.

	Precision	**Recall**	**F1 Score**	**Accuracy**
Fake	0.97	0.98	0.97	
Real	0.98	0.97	0.97	
Average	0.97	0.97	0.97	0.97

Table 4: Evaluation Metrics for ViT Model on 140k Dataset

3.2.2 Receiver Operating Characteristic Analysis

The ROC curve for the visual transformer model on the 10k image dataset (Figure 7) demonstrates relatively impressive results, with an area under the curve (AUC) of 0.90. This AUC value indicates a strong true positive rate against the false positive rate, thereby suggesting the model is slightly more effective at distinguishing between classes than the Xception model on the same dataset.

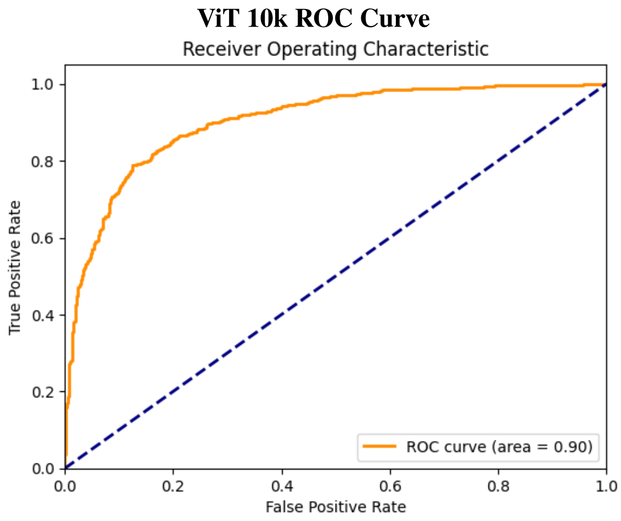
ViT 10k ROC Curve

Figure 7: This figure illustrates the Receiver Operating Characteristic curve for the visual transformer model on the Real vs. Fake Faces - 10k dataset.

The ROC curve for the ViT model on the 140k image dataset (Figure 8 illustrates an AUC of 1.00. This indicates a near-perfect TPR and FPR of almost 0% across all threshold settings. This can be explained by the possibility that the ROC curve is simply not granular enough to capture the misclassifications that were known to have occurred in the confusion matrices provided.

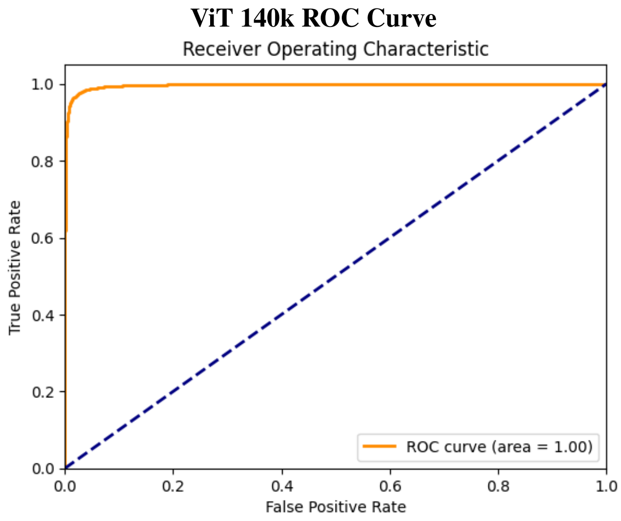
ViT 140k ROC Curve

Figure 8: This figure illustrates the Receiver Operating Characteristic curve for the visual transformer model on the Real vs. Fake Faces - 140k dataset.

3.2.3 Confusion Matrices

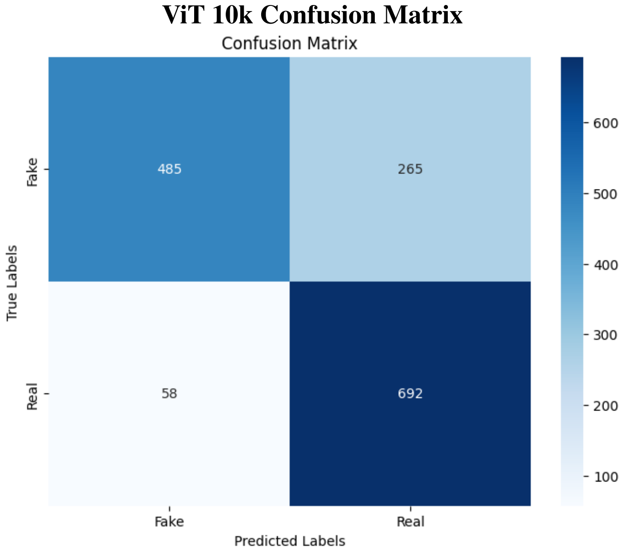
ViT 10k Confusion Matrix

Figure 9: This figure illustrates the Confusion Matrix for the Visual Transformer model on the Real vs. Fake Faces - 10k dataset.

The confusion matrix for the ViT model on the 10K dataset (Figure 9) reveals an interesting occurrence - there is an unusual rate of false negatives, evidenced by the 265 occurrences. This is reflected in the classification report, with a 0.76 precision.

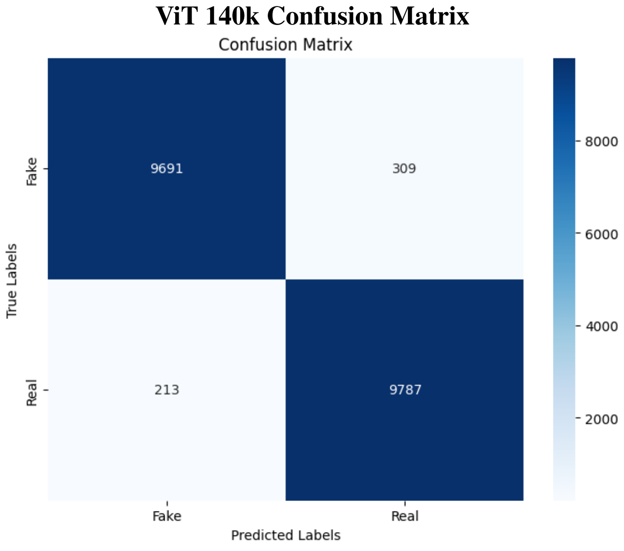
ViT 140k Confusion Matrix

Figure 10: This figure illustrates the Confusion Matrix for the visual transformer model on the Real vs. Fake Faces - 140k dataset.

However, similar to accuracy, we see a massive improvement when tested and trained on the 140K dataset (Figure

10) - with a lower rate of false negatives and a precision of 0.84. Clearly, the larger amount of training data greatly benefits the ViT.

3.2.4 Precision-Recall Analysis

The precision-recall curve for the ViT model for the 10K dataset (Figure 11) starts off strong but tapers off over time. Still, the overall average precision of 0.90 indicates that even for a smaller dataset, the ViT model is still very precise in its predictions.

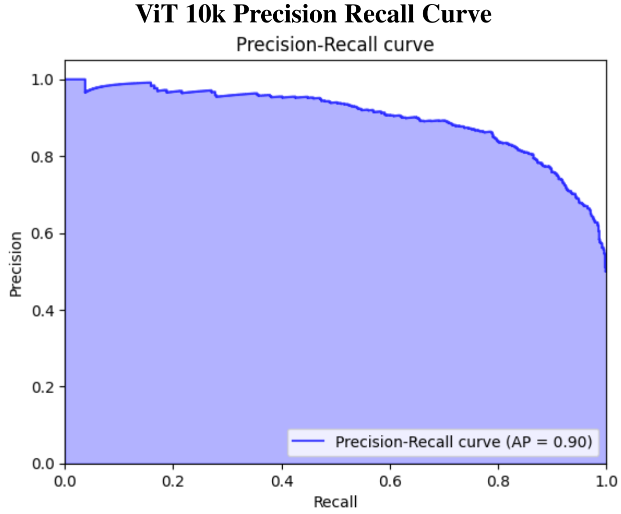

Figure 11: This figure illustrates the Precision-Recall curve for the visual transformer model on the Real vs. Fake Faces - 10k dataset.

Looking at the 140k dataset (Figure 12), however, it is apparent that the ViT model is almost perfectly precise with an average precision of 1.00 (the highest score possible). This shows that the larger dataset had a positive effect on the precision of the model even with the precision starting off from a high point.

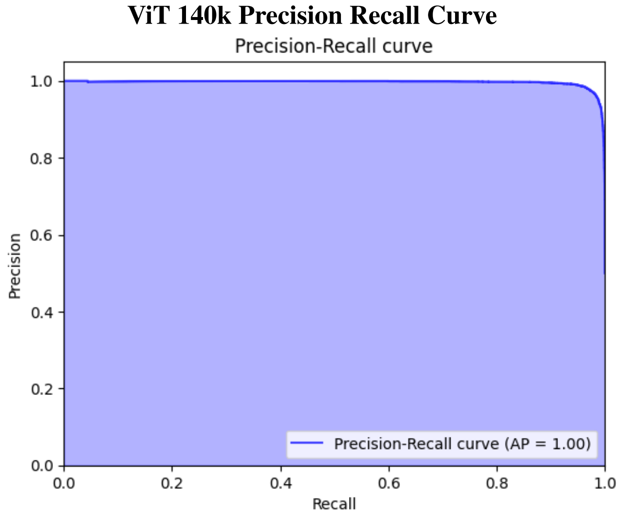

Figure 12: This figure illustrates the Precision-Recall curve for the visual transformer model on the Real vs. Fake Faces - 140k dataset.

3.3. Novel Architecture - Audio and Visual

	Precision	Recall	F1 Score	Accuracy
Fake	0.86	0.90	0.88	
Real	0.22	0.17	0.19	
Average	0.76	0.77	0.77	0.79

Table 5: Evaluation Metrics for ViT Model on 140k Dataset

The novel architecture for video and audio deepfake detection achieved a test accuracy of 78.75%. While this indicates a relatively high overall performance, a more detailed analysis of the metrics reveals potential underlying issues. For this, it is necessary to investigate the Confusion Matrix (Figure 13) and Classification Report (Table 5). For example, the Confusion Matrix highlighted a considerable number of false positives and false negatives, revealing a tendency to incorrectly classify real videos as fake.

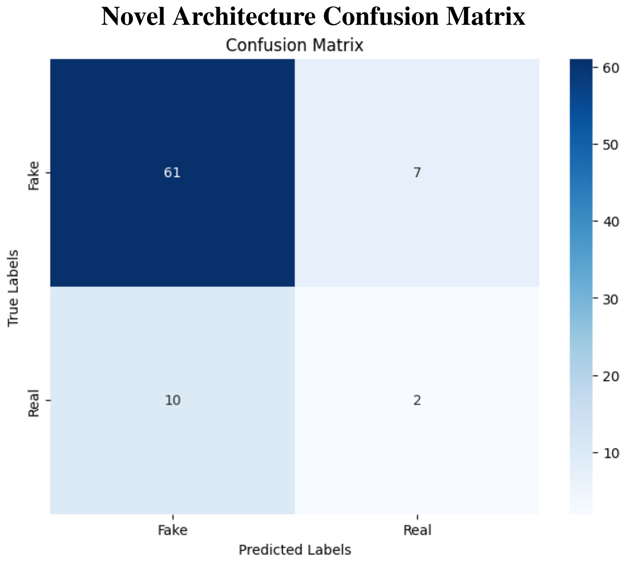
Novel Architecture Confusion Matrix

Figure 13: This figure illustrates the Confusion Matrix for the Novel Architecture Model ran on the Audio-Video dataset

At 0.22, the precision for real videos indicates a low accuracy in predicting real videos when the model labels a video as real. Furthermore, the recall of 0.17 for real videos demonstrates the model's inability to identify all actual real instances correctly. The F1-Score for real videos was notably low at 0.19, indicating poor performance for this class. Despite the implementation of class weights, the model favored the majority (Fake) class and suffered from a lack of availability of data items corresponding to this class necessary for learning the features to facilitate accurate classification. This suggests that while class weights helped to some extent, they were insufficient to fully correct the imbalance impact.

Overall, the model demonstrated competency in identifying fake videos, as indicated by the outputted metrics corresponding to this class. However, the model did not adequately address the detection of real videos, resulting in a high rate of false positives. This aspect of the model's performance indicates that the design and data preprocessing methods require significant improvement.

3.4. Discussion

Comparing the performance of the Xception model, the Visual Transformer (ViT) model, and the novel architecture combining 3D CNN and standard neural network for video and audio analysis provides valuable insights into their effectiveness in detecting deepfakes for various types of media. Although this study's investigations primarily revolved around artificially generated facial images, it is important to highlight the application of these techniques and architectures with respect to video and audio content.

The Xception model, with its CNN-based architecture, demonstrated strong performance in capturing local features, achieving high accuracy, precision, recall, and F1-score particularly for the smaller 10k images dataset. In contrast, the ViT model exhibited a longer training time but obtained a significantly higher accuracy (97%) on the 140k images dataset. Despite the extremely impressive results of the visual transformer model on the larger dataset, it can be concluded that the model did not overfit given that it was evaluated with an unseen test set. As discussed previously, visual transformers are generally known to excel at capturing global features as well as learning relationships across full images. This evidence indicates that, while CNN-focused architectures like Xception may be preferable for smaller datasets due to faster training and high accuracy, vision transformers could be more suitable for larger datasets where higher accuracy is desired despite longer training times.

The ViT model's performance was further highlighted when compared to the novel architecture involving a combination of 3D CNN and standard neural network for video and audio analysis. While the ViT model achieved high accuracy, precision, recall, and F1-score, the novel architecture showed a lower precision and recall for real videos, indicating challenges in accurately identifying real videos. Aside from this, the results from the experimentation with the novel architecture are not reliable given the lack of abundant data items corresponding to real videos and audio.

These results suggest that the choice of architecture depends on the dataset size and the balance between training time and desired accuracy. For smaller datasets where faster training is preferred, CNN-based architectures like Xception may be more suitable. Despite this, such conclusions may not be accurate given that previous research has indicated that Xception is a preferred model for very large image datasets [3]. However, for larger datasets where higher accuracy is paramount, architectures like visual transformers that excel at capturing global features may be more effective. Future research could focus on optimizing novel and/or hybrid architecture approaches for better performance on real video detection, possibly by incorporating more advanced features or refining the training process.

Student Name	Contributed Aspects	Details
Mark Glinberg	Debugging and execution, GitHub, Paper	Executed and Debugged Xception and ViT Runtimes, Created maintained Github, Proofread and Edited Paper, along with significant contributions to Introduction and Results.
Andrew Gao	Xception Models, AV Model, Approach and Results	Implemented and Ran both Xception Models on both 10K and 140K datasets. Further responsible for the full implementation of the Video and Audio Model and debugged and executed the model. Wrote signifiacant portions of Approaches and Results sections.
Charles McCartney	ViT Code Implementation, Intro and Approach	Formatted the Real vs. Fake Faces - 10k dataset and implemented ViT code for 10k and 140k datasets. Further contributions involve conducting analysis of results, drafting the introduction, ViT approach, discussion, and results sections, and refining project scope.
Jerome Perera	Debugging, Paper, Github	Assisted in debugging 10k and 140k Xception Runtimes, Debugged data preprocessing, Extracted and formatted images in the paper and made additional general formatting improvements, added sizeable portions to Experiments and Results section.

Table 6: Contributions of team members.

References

[1] benpflaum, Brian G, djdj, Irina Kofman, JE Tester, JLElliott, Joshua Metherd, Julia Elliott, Mozaic, Phil Culliton, Sohier Dane, and Woo Kim. Deepfake detection challenge, 2019. 2

[2] Sergi D Bray, Shane D Johnson, and Bennett Kleinberg. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1):tyad011, 06 2023. 1

[3] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 1, 2, 3, 9

[4] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah Wan Ahmad, and Arpit Bansal. Feature extraction using mfcc. *Signal Image Processing : An International Journal*, 4:101–108, 08 2013. 2

[5] Dmytro Iakubovskyi. Deepfake vs real faces detection vit, 2024. 1, 3

[6] Hafsa Ilyas, Aun Irtaza, Ali Javed, and Khalid Mahmood Malik. Deepfakes examiner: An end-to-end deep learning model for deepfakes videos detection. In *2022 16th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6, 2022. 3

[7] S. M. Ariful Islam, Mst Sunjida Sultana, Uttam Kumar Roy, and Jubayer Al Mahmud. A review on video classification with methods, findings, performance, challenges, limitations and future work. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, 2021. 3

[8] Sachchit Kunichetty. Real vs fake faces - 10k, 2024. 2

[9] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23296–23308. Curran Associates, Inc., 2021. 3

[10] Misaj Sharafudeen, Andrew Onesimu, and Vinod Chandra. Leveraging vision attention transformers for detection of artificially synthesized dermoscopic lesion deepfakes using derm-cgan. *Diagnostics*, 13:825, 02 2023. 3

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 2

[12] xhlulu. 140k real and fake faces, 2020. 2