Deepfake Detector

Cheng-You Lu Ji Won Chung Kelly Patel

BROWN

Motivation

Democratizing Tech

- Enable individuals to distinguish "real" content from fake content by automatically detecting deepfakes

Problematic Deepfakes

- Propagates false information
- Increases distrust in media
- Victimizes individuals
- Mass automatic creation

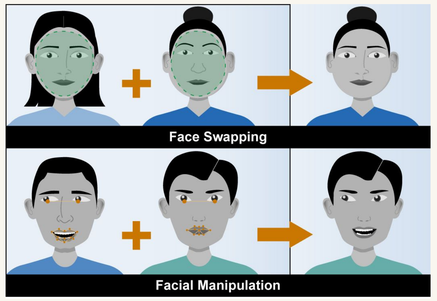


Figure 1: Visual Deepfake Generation
Img Src: GAO

The Problem

Binary Classification - Deepfakes vs Original

- Exploiting Visual Deepfake Artifacts with CNNs & ResNets
- Exploiting Temporal Information with ResNet50 + LSTM
- Image Processing to train models and increase accuracy
  - Size of video data
  - Face extraction technique
  - Number of frames from videos

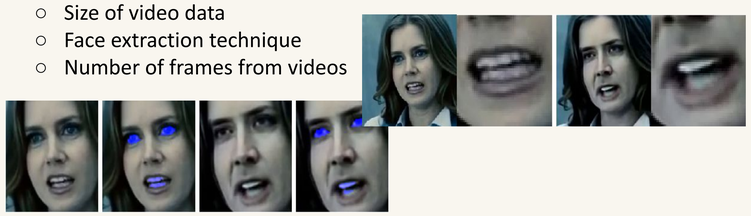


Figure 2: Visualization of Missing Geometry in DeepFakes (see teeth).
Img Src: Semantic Scholar

Goals

Main Goal

- Classify deepfakes from fakes by focusing on visual fakes

Intermediary Goals

- Maintain accuracy of at least 75%
- Beat ResNet50 with ResNet50+LSTM

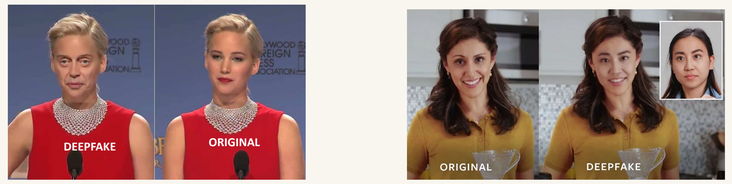


Figure 3: Examples of Deepfake vs Original
Img Src: Olhar Digital & Img Src: IEEE Spectrum

What is the insight/main technical solution?

Dataset
Deepfake Detection Challenge
Facebook AI on Kaggle
471.84 GB

→ **Processed Dataset**
Training, Validation, Testing

→ **Frame Extraction**
Extract 10 frames from Video with original resolution to maintain visual artifacts

→ **Image Processing**
Pad/Crop - 450 x 1800 for video size variance
Face Extraction - 700 x 700

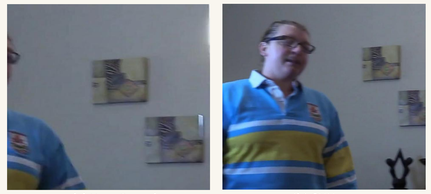
→ **Model Training**
ResNet50
and
ResNet50 +LSTM

Processed Dataset

- **Training**: 45,801 Videos (REAL: 6,377, FAKE: 39,424)
- **Validation**: 1,054 Videos (REAL: 515, FAKE: 539)
- **Testing** : 400 Videos (REAL: 200, FAKE: 200)

Image Processing / Face Extraction - Haar Cascade vs CNN

- **Haar Cascade**
  - 150x quicker, but sensitive to motion and brightness
- **CNN**
  - Slower but more stable



(a) Haar Cascade (b) CNN

Figure 4: Haar Cascade vs CNN Face Extraction
CNN shows better face extraction because Haar Cascade is sensitive to motion and brightness

Our Architecture

ResNet50 & ResNet50 + LSTM

- ResNet50
  - Mine spatial relationship
- Long Short-Term Memory (LSTM)
  - Capture long-term temporal information
- Global average pooling
  - Aggregate the global information
- Binary Cross-Entropy Loss Function
- Data Resampling
  - Mitigate the data imbalance issue

	ResNet50	ResNet50 + LSTM
Frames	5	5
Trainable Parameters	23,566,848	90,897,537
Training Time	1 day	3 days

Table 1: ResNet50 & ResNet50 + LSTM training Parameters

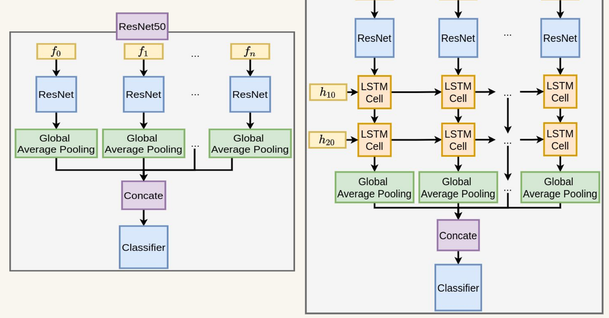


Figure 5: ResNet50 Architecture (*left*) ResNet50 + LSTM Architecture (*right*)

Results

Analysis

- ResNet50 vs ResNet50 + LSTM
  - ResNet50 + LSTM can outperform ResNet50 due to the LSTM module
- Pad/Cropped Data vs Face Extraction Data
  - Pad/Cropped Data outperforms Face Extraction Data because deepfake artifacts may exist in non-face region
- Global average pooling
  - Global average pooling is affected by padding size during inference

Model	Train-Process	Test-Process	Frames	Val	Test
ResNet50	Crop & Pad	Pad 1920 x 1920	5	89.08	83.99
ResNet50 + LSTM	Crop & Pad	Pad 1920 x 1920	5	95.73	88.00
ResNet50 + LSTM	Crop & Pad	Pad 2500 x 2500	5	95.16	86.50
ResNet50	Face Extraction	Pad 1920 x 1920	10	84.16	68.00
ResNet50 + LSTM	Face Extraction	Pad 1920 x 1920	10	85.95	78.50
ResNet50 + LSTM	Face Extraction	Pad 2500 x 2500	10	75.42	63.99

Table 2: ResNet50 & ResNet50 + LSTM Accuracy Performance

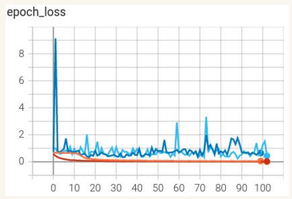


Figure 6: Loss for Training and Validation
Loss decreases exponentially in both training and validation as expected

More Results

(a) Fake Mislabeled as Real Video 1 (c) Real Video Mislabeled as Fake Video 1
(b) Fake Mislabeled as Real Video 2 (d) Real Video Mislabeled as Fake Video 1

Figure 6: Fake and Real Videos Mislabeled

Misclassified Videos

- Analysis on small sample of misclassified images without face extraction
- Images misclassified as Deepfake but actually Real
  - Contain more of the subject's body - Figure 6(a)(b)
- Images misclassified as Real but actually Deepfake
  - Subjects not facing camera squarely Figure 6(c)(d)

Conclusion

- Observed trends were based on a small sample, so it is possible that other factors like color, shadows, gaze-information, etc. played a significant role in misclassification.

References

[1]Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2814–2822, 2020. 1
[2] Atmik Ajoy, Chethan U Mahindrakar, Dhanya Gowrish, and Vinay A. Deepfake detection using a frame based approach involving cnn. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pages 1329–1333, 2021. 1
[3] Ben Pflaum Jikuo Lu Russ Howes Menglin Wang Cristian Canton Ferrer Brian Dolhansky, Joanna Bitton. The deepfake detection challenge dataset, 2020. 1, 5
[4] Ilke Demir and Umur Aybars Ciftci. Where do deep fakes look? synthetic face detection via gaze tracking. In ACM Symposium on Eye Tracking Research and Applications, ETRA '21 Full Papers, New York, NY, USA, 2021. Association for Computing Machinery. 1
[5] Jia Deng, Wenjun Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 1
[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. 1
[7] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term ¨ memory. In Neural Computation, 1997. 1
[8] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media Society, 6(1):205630512090340, January 2020. 1, 4
[9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, pages I–I, 2001. 1, 2

Acknowledgements

Thank you Monica Roy, our assigned TA, the general CSCI 1430 TAs, Professor Sridhar, and Google Cloud Platform for making this final project possible.