Spot the Difference: Real vs. StyleGAN-generated Faces

Colden Bobowick, Spencer Dellenbaugh, Alexander Halpin

BROWN

CSCI 1430

Motivation

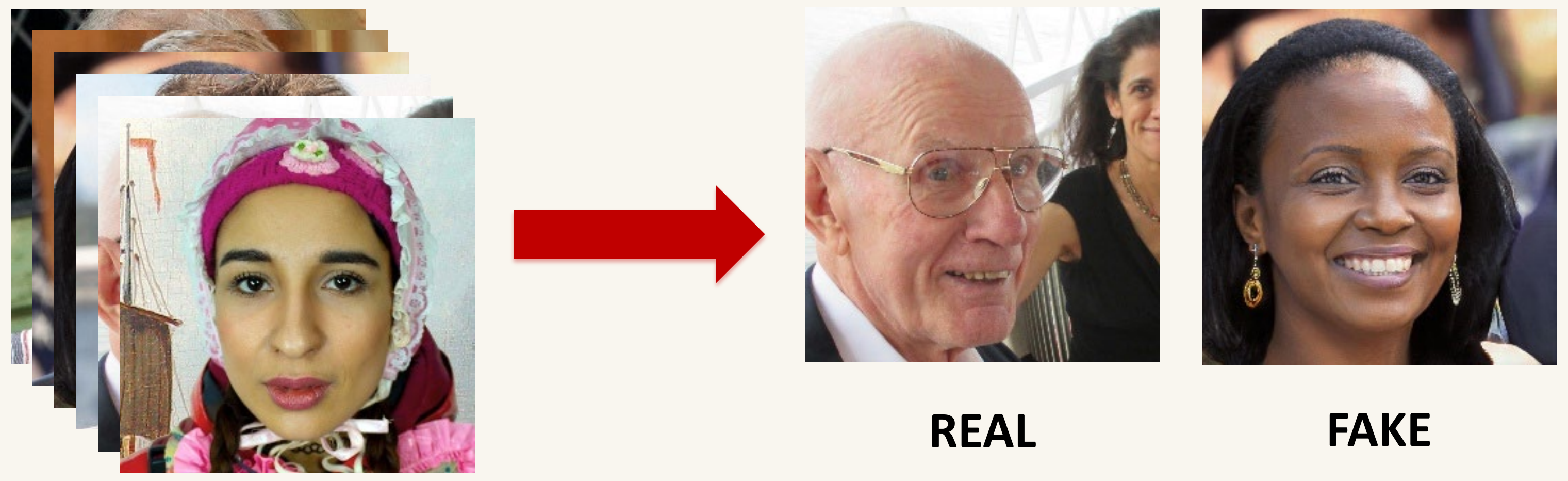
- Generative Adversarial Networks have created fake faces that have been used for misinformation campaigns, espionage, and other harmful uses.
- They have become so accurate that it can be difficult to determine if a face is real or GAN-generated.
- A convolutional neural network (CNN) can be used to help detect GAN-generated faces, which may be helpful to prevent widespread misinformation.

Problem

GAN-generated images are difficult to visually identify

Goal

Build a GAN-generated face detector that has >90% accuracy on labeling StyleGAN images

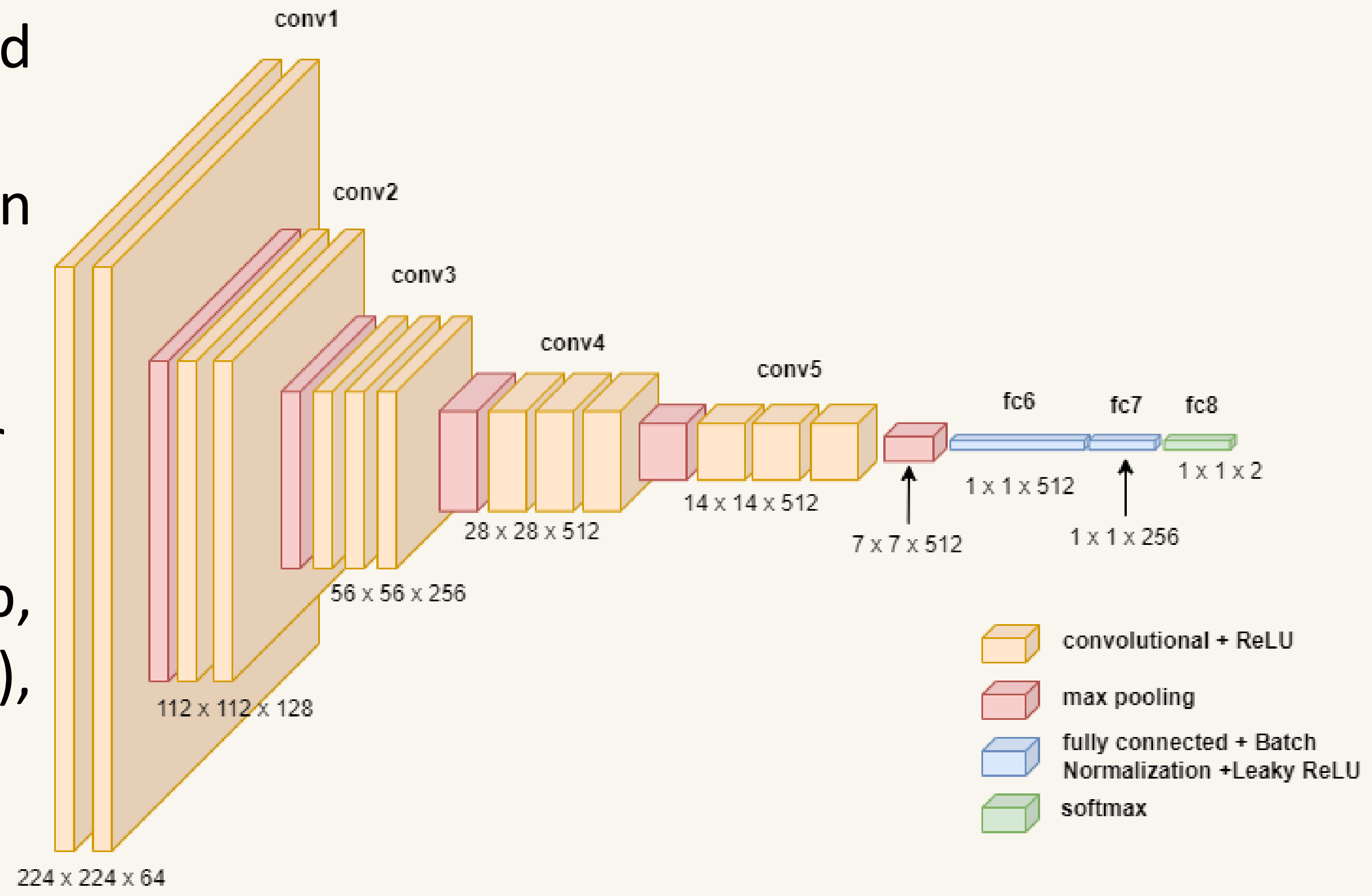


Design Process

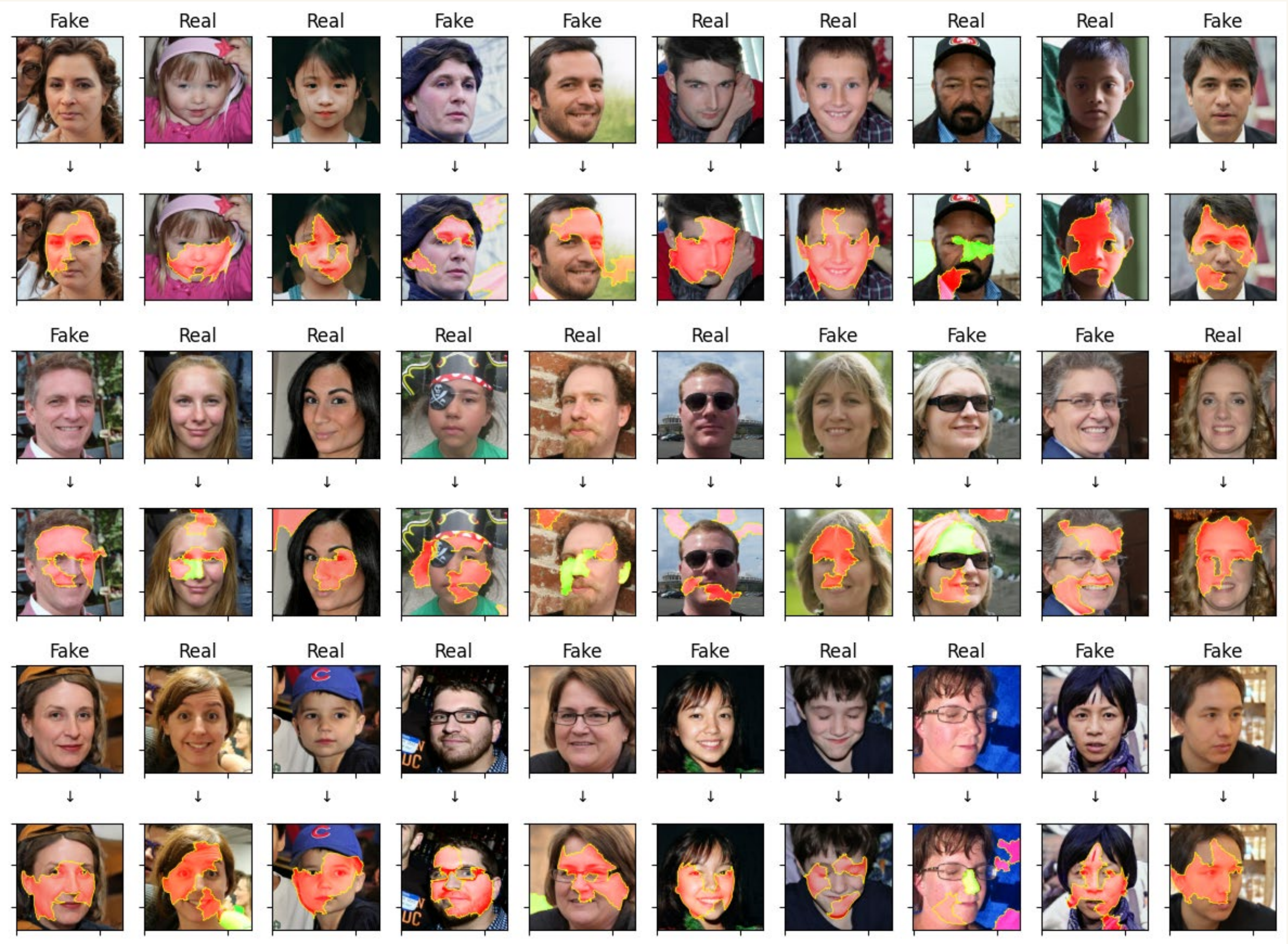
Variable	Initial Attempt	Initial Accuracy	Final Attempt	Final Accuracy
Batch Normalization	0 Layers	84.95%	2 Layers	87.20%
Dropout Layers	1 Layer of 0.5	90.45%	0 Layers	91.60%
Number of Dense Layers	3 Layers	82.75%	2 Layers	86.40%
Leaky ReLU alpha value	0.1	90.45%	0.5	91.00%
Number of Neurons	26 (24+2)	87.45%	770 (512+256+2)	91.90%
Augment. (Bright., Cont.)	0.1, 0.1	91.40%	(0.3,0.2)	93.40%
Trainable vgg16	Trainable	55.34%	Untrainable	92.92%

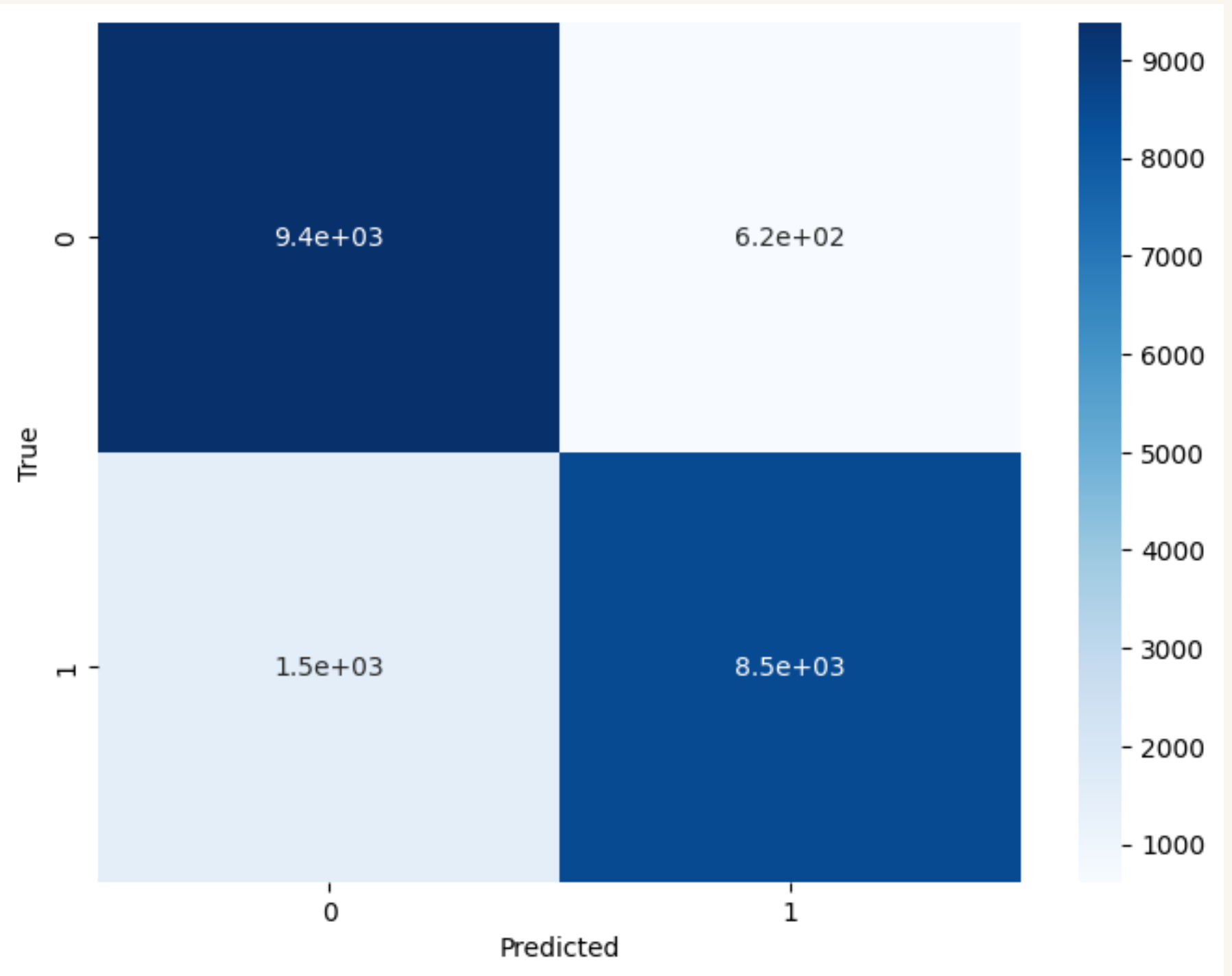
Proposed Solution

- Proposed solution is based off the pretrained vgg16 model
- Specialized feature classifier that has been fine-tuned for face detection
- Optimal modifications:
  - Batch normalization after each dense layer
  - Leaky ReLU activation function with α=0.5
  - Augment block with random horizontal flip, zoom (0.1), rotation (0.1), brightness (0.3), and contrast (0.2)
  - Adam optimizer with learning rate 0.001



LIME Visualization & Confusion Matrix

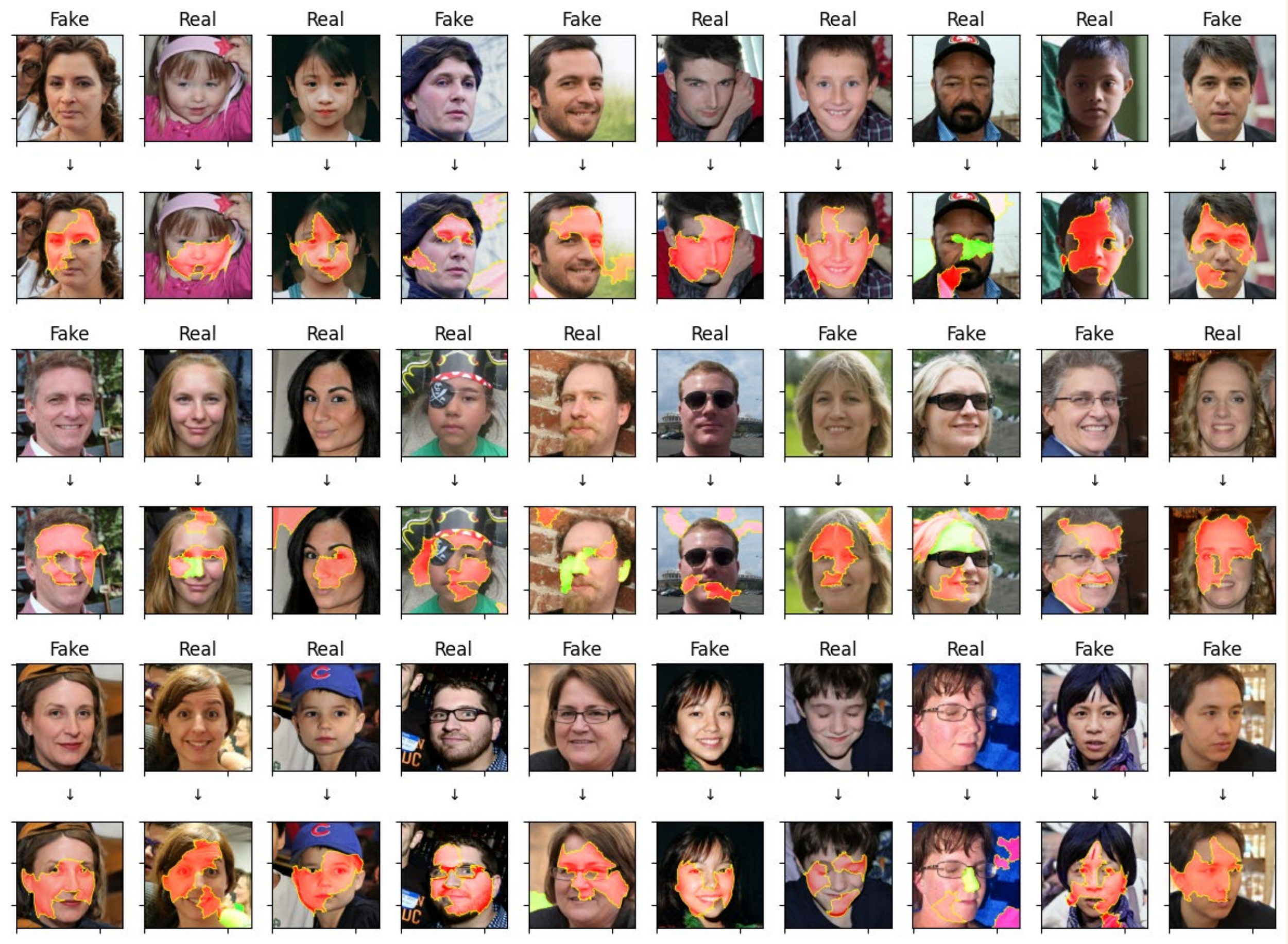




Results Summary & Saliency Maps

Maximum Validation Accuracy:	93.00%
Minimum Validation Loss:	0.1736
Maximum Training Accuracy:	90.61%
Minimum Training Loss:	0.2238

Overall, the project was a success. A 93% validation accuracy was achieved, which exceeded the 90% goal. Further research may focus on improving accuracy and investigating robustness against common adversarial attacks.



References

[1] H. Hao, E. R. Bartusiak, D. Güera, D. Mas, S. Baireddy, Z. Xiang, et al., "Deepfake Detection Using Multiple Data Modalities" in Handbook of Digital Face Manipulation and Detection From DeepFakes to Morphing Attacks Series on Advances in Computer Vision and Pattern Recognition, Springer, vol. 1, pp. 191-212, March 2022.

[2] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "GAN-generated Faces Detection: A Survey and New Perspectives," arXiv:2202.07145 [cs.CV], 2023.

[3] J. Wang, B. Tondi, and M. Barni, "An Eyes-Based Siamese Neural Network for the Detection of GAN-Generated Face Images," Frontiers in Signal Processing, vol. 2, pp. 1-14, 2022, doi: 10.3389/frsip.2022.918725.

[4] xhulu, "70k real faces (from Flickr) and 70k fake faces (GAN-generated)," May 8, 2023. [Online]. Available: https://www.kaggle.com/xhlulu/70k-real-and-70k-fake-faces-gan. [Accessed: May 8, 2023].

[5] R. Can Malli, "keras-vggface," GitHub, 2016. [Online]. Available: https://github.com/rcmalli/keras-vggface. [Accessed: May 8, 2023].

Acknowledgements

The authors extend a gracious acknowledgment to Professor Tompkin for his instruction and support throughout the semester, as well as TA Kelly Patel, who was likewise instrumental in the completion of this project.