PROJECT 5

DEEPFAKE IMAGE DETECTION

ERIC HANSEN

FLATIRON DS; SELF PACED

2/4/2023

BACKGROUND - CONTEXT

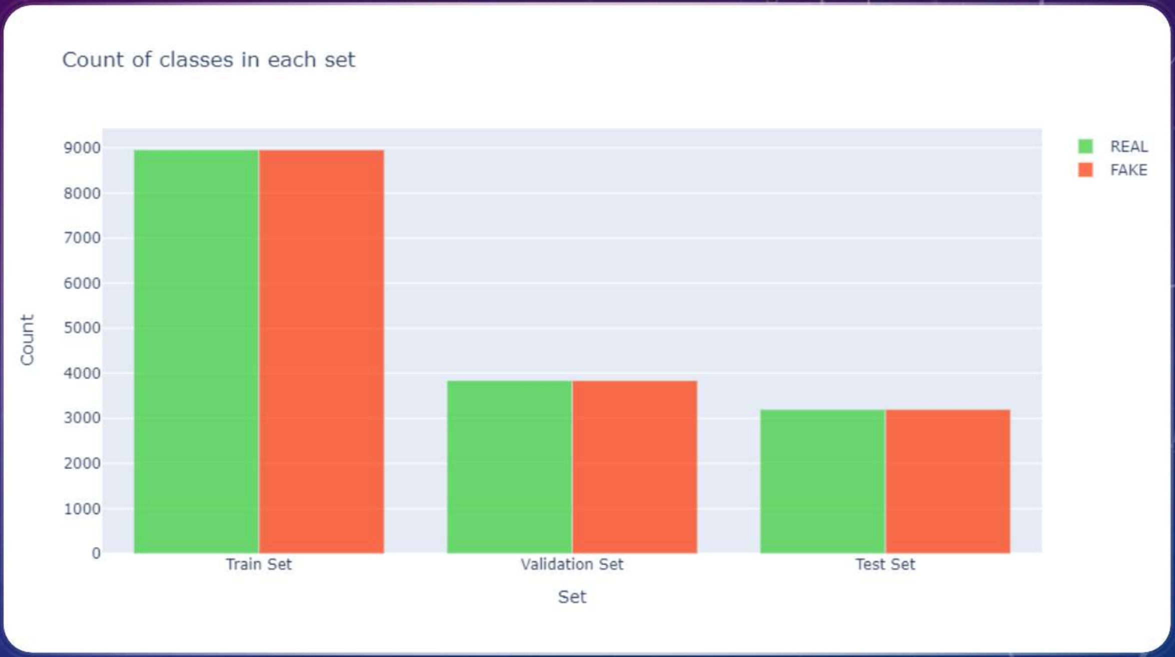
- Deepfakes (a portmanteau of "deep learning" and "fake") are an artificial image or video that replaces an existing entity with another entity's likeness.
- Creation of such media has some benign applications in art and academia, but in general, the capacity for misleading deepfakes is of great concern to many stakeholders due to its potential to be used for harmful ends.

BACKGROUND - STAKEHOLDERS

- Stakeholders in this process include (but aren't limited to)
  - Media producers and distribution platforms
  - Public figures
  - Individual citizens

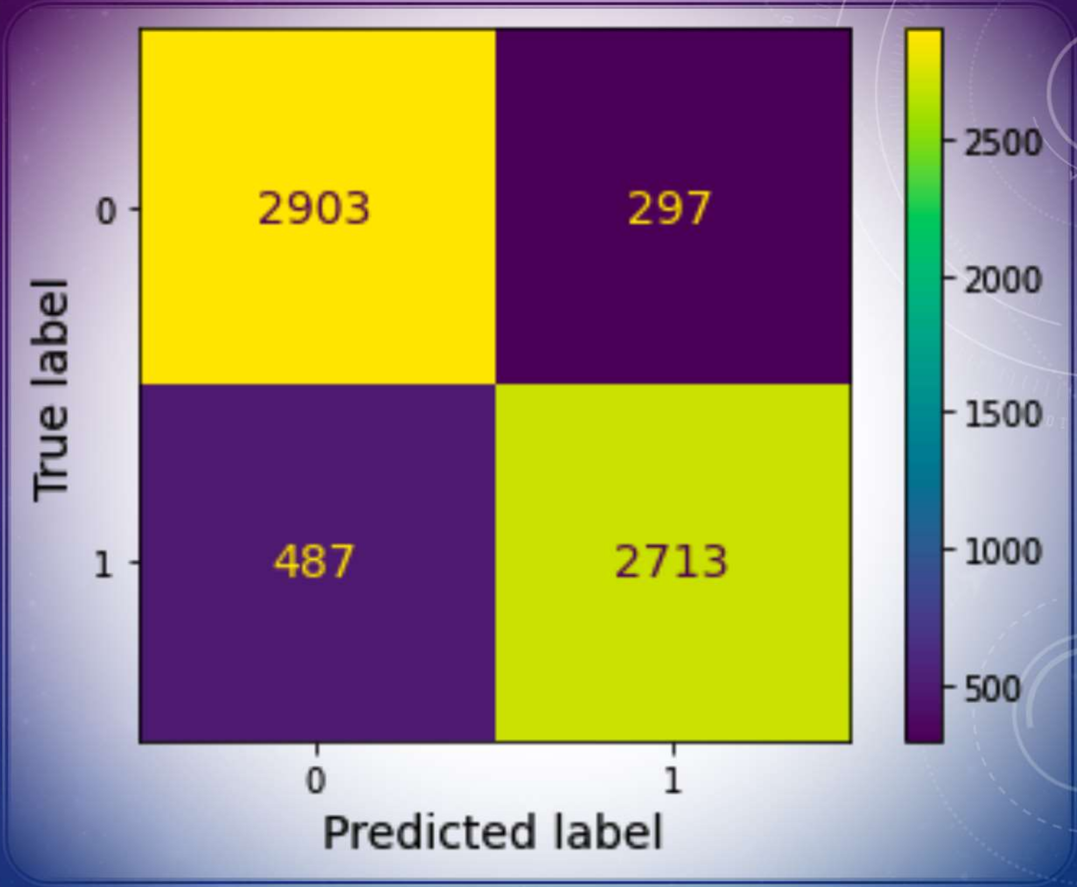
BACKGROUND – DATA

- Deepfake Detection Challenge (part of a Kaggle-hosted challenge)
  - https://www.kaggle.com/competitions/deepfake-detection-challenge
- Extracted 95K jpeg images from
  - https://www.kaggle.com/datasets/dagnelies/deepfake-faces
- 16,000 Real images and 16,000 Fake images were used in this analysis



MODELS

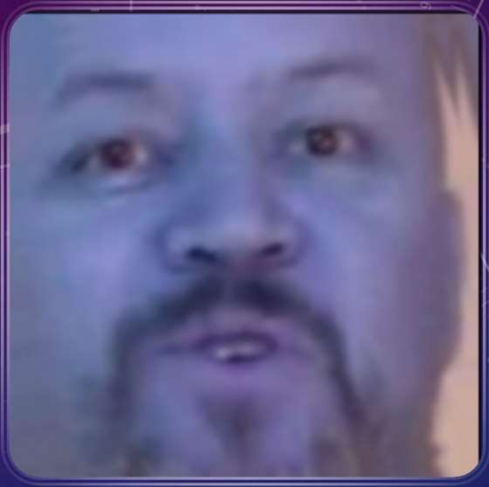
- Two from-scratch models and five pre-trained transfer learning models were tuned for classification
- The model with highest accuracy (87.8%) was Xception
  - 90% precision (tp/predp)
  - 85% recall (tp/actualp)
- 0 corresponds to Real images, 1 corresponds to Fake images

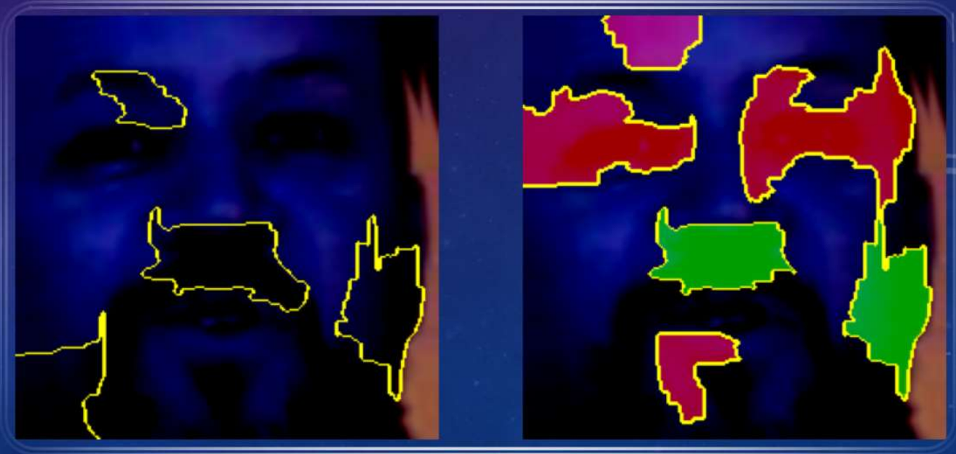


INTERPRETABILITY USING LIME

MEANINGFUL FEATURES FOR A “REAL” SAMPLE IMAGE ARE REPRESENTED VISUALLY

GREEN ARE “REAL” FEATURES AND RED ARE “FAKE”





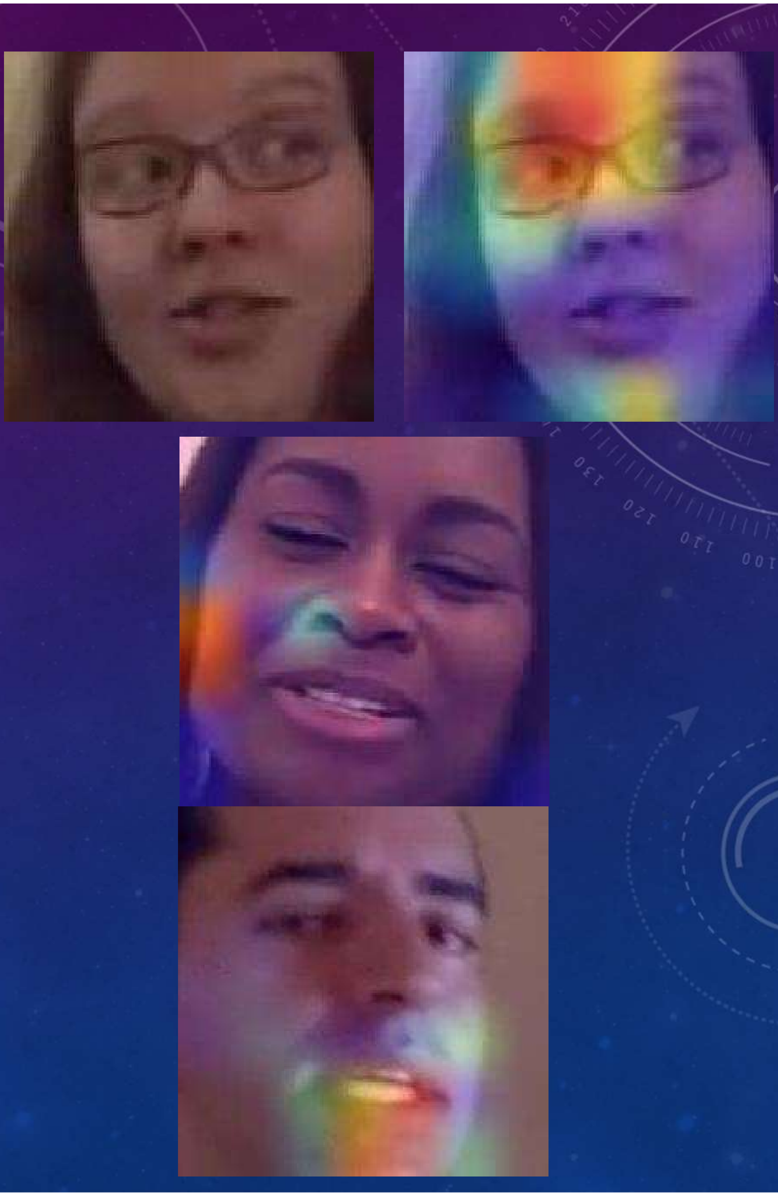
INTERPRETABILITY USING GRAD-CAM

MEANINGFUL FEATURES FOR “FAKE” CLASSIFICATION ARE SHOWN WITH COLORED AREA

FIRST IMAGE IS “FAKE” AND CORRECTLY CLASSIFIED

NEXT IS INCORRECTLY CLASSIFIED AS REAL

THE FINAL IMAGE IS “REAL” AND CORRECTLY CLASSIFIED



CONCLUSIONS/QUESTIONS

- Deepfake images that were created using the method of the deepfake detection challenge can be classified with at least 87.8% accuracy
- Some visual interpretability is explored, but the models are sufficiently complex that it is challenging for humans to learn from their performance

CONCLUSIONS/QUESTIONS

- Questions?

CLOSING

- Thank you!