IIP to production

Goal: make the IIP more robust

Sean Abreau

Data

- Aggregate statistics
  - Ensuring our dataset covers the full spectrum of age, ethnicity and gender
  - Gather stats about backgrounds and lighting conditions
- Create list of global feature importances
  - Ranks dataset by most challenging examples
  - Improves cleaning and training

Fairness

- Balance or weigh (age, gender, ethnicity) features to balance training dataset
- Rate constraint to guarantee recall is at least n% on a subset of our data
- Min diff to penalize model for differences in predicted distributions

Data Drift Detection

- Our data will change over time it’s important for us to monitor how it’s changing
- We can monitor our statistics to see which feature distributions are changing and by what quantity
- We can predict how our models performance will be impacted by the changing data distribution
    - Using techniques like oversampling and undersampling to test our models

Non-iris/partial iris detection

- Many of the images in our dataset are non-iris or partial iris
- The first model in our pipeline should quickly identify these images and filter them out before reaching our semantic segmentation model
- These images should then be sent to our systems to help improve our Non-iris/partial iris model, with labeling or identifying system weaknesses
- These could be considered an edge case

Edge cases

- If we created embeddings of our images we can run a clustering model to quickly find edge cases
- We can create a front end interface and api (like task 2) for our data team to easily test and visualize our model performance on those images

Model Performance

- Our model performance will change over time we need to monitor how
  - We can compare our models current performance to historical data performance
- Metrics
  - F1
  - AUC
  - MAP
  - IOU
- Use wandb to monitor experiments and hyperparameter sweeps
- Create model registry to maintain versioning

Robustness

- Create robustness benchmark
  - Create subsets of data separate and different from our training (in known ways) dataset based on (geography, ethnicity etc.)
  - Audit to monitor model bias
- Many image augmentations
  - Blur
  - Rotations
  - Color filters.. Etc.
- Monitor compression/robustness tradeoff
  - Compression disproportionately affects underrepresented features

A/B testing

- To maintain a high level of confidence in our IIP we can deploy 2 models at a given time
  - Testing the performance of each and if a new model is performing poorly we can quickly switch to the other