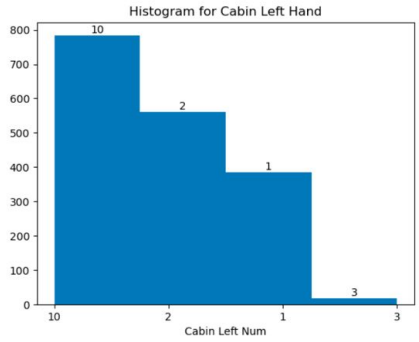
Shaky Ground Truth Presentation

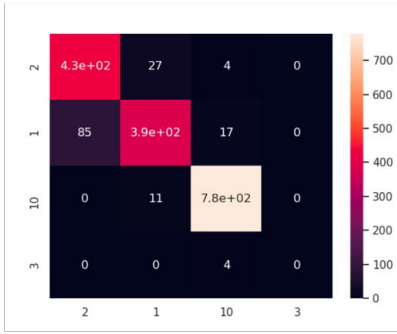
Keywords: Shaky Ground Truth, Uncertainty in ML, human data

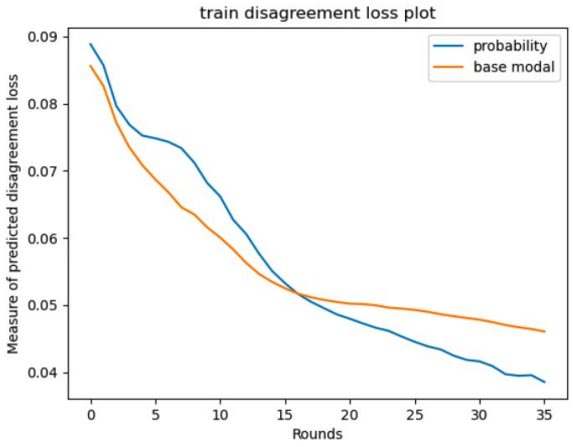
Hangrui Cao
Michigan

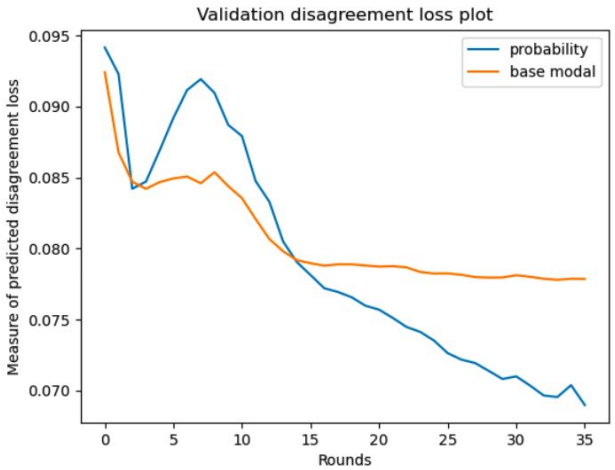
Label difference for labelers among 5 coders!

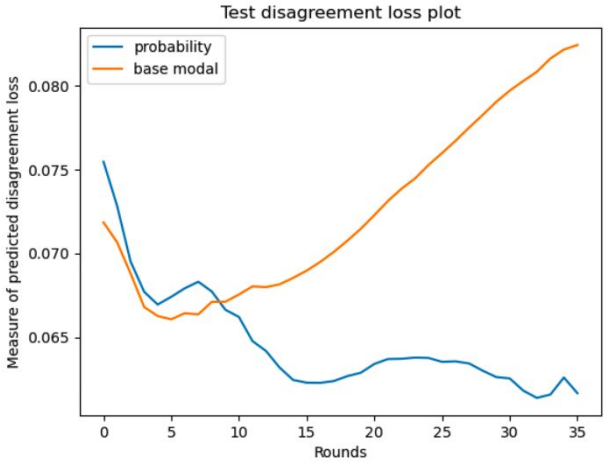
Label: 1, 1, 1, 1, 1, Label: 2, 1, 3, 3, 2











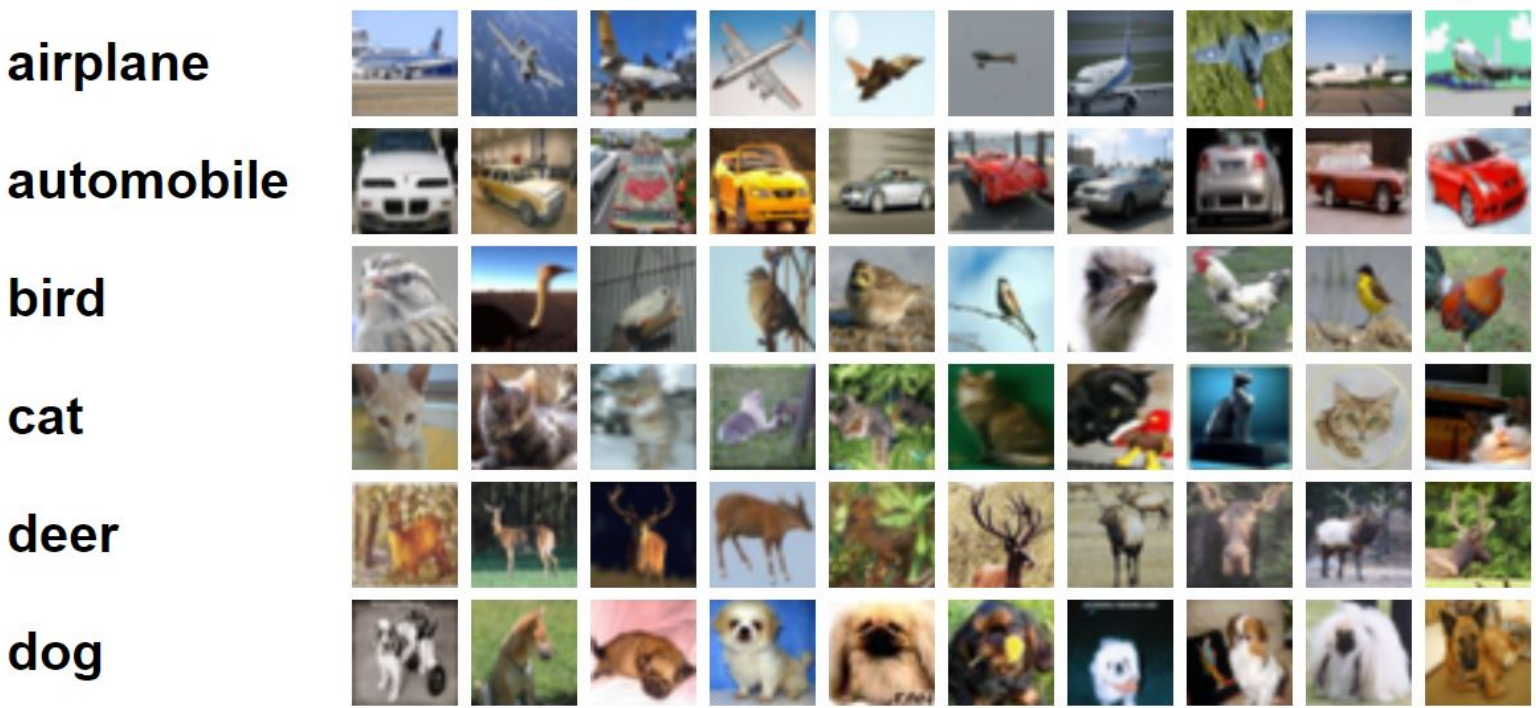
Outline

- Background
- Data Exploration
- Current Methods & Results
- Future directions

Background

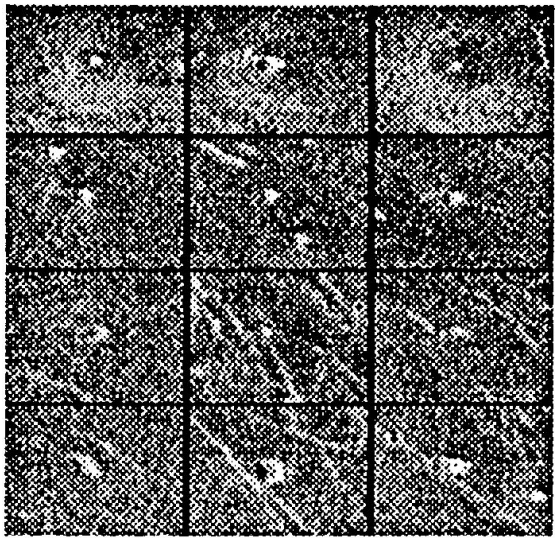
Ground Truth

Ground Truth is the information that is known to be real/true, the true label



Shaky Ground Truth: Origin

- In paper “Knowledge Discovery in Large Image Databases: Dealing with **Uncertainties in Ground Truth**”, mentions the problem of finding volcanoes



Volcano Plot

Data Exploration & Setting

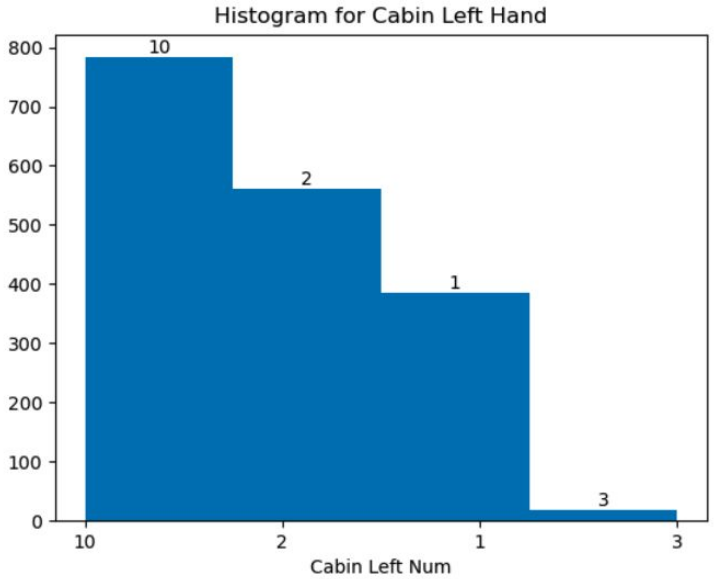
Dataset

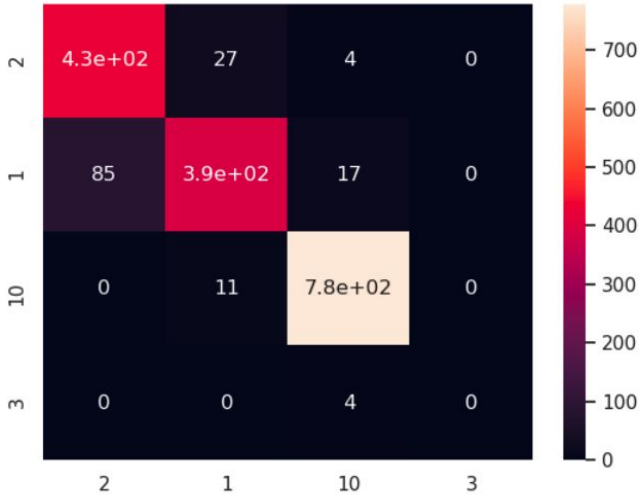
- Consists of 1746 images
- Each image input: 384x288
- Thanks for five coders: Ruma, Nithin, Haomeng, David, and Andrea labeled them
- Three main issues:
  - Coders' answer differ on the result
  - The label result is not balanced
  - No modal value in some circumstances (like 1, 2, 2, 3, 3, modal value 2 or 3?)∫

Metrics Defined - Cabin

- For example, manipulation: 1 (yes), 2 (no), 3 (indiscernible), 10 (cannot be determined)
- Use rubric version 5 for training
- Rubrics in: https://docs.google.com/spreadsheets/d/1EkjbG4UAQQrjFG5Kpk1b-_TOQT1JQZJogu17DkZtXq0/edit#gid=500947917

Label Distribution - Manipulation (Example)





Variation do exist!
AUROC score for five users: 0.906

Method

Three Methods

- Different people give different label on
- For example, for one frame, five coders label it as: [0, 1, 1, 2, 1], the modal value is 1, and the probability for each label is [0.2, 0.6, 0.2]

Method	**Ground Truth**
Base (modal value)	One image, with label of the modal value 1
Duplicate(choice 1, make the dataset five times)	Five images, three with label 1, one with label 2, one with label 0
Duplicate(choice 2: keep the dataset the same)	Five images, three with label 1, one with label 2, one with label 0 (When train, load randomly ⅕ of them to keep them the same)
Probability (Bayesian, only)	One image, with label of [0.2, 0.6, 0.2], each refers to the probability of each category

Initial Challenges

- Train dataset not sufficient (Originally try prediction of the eye gaze)
- The accuracy is low, as it is hard for CNN to predict the eye gaze direction
- Finally pick the metrics “manipulation”, “left hand” and “right hand” to perform machine learning training in the cabin dataset.

Experiment setting

- In total 1764 sample images of cabin dataset
- Initially, design the dataset to fully random choice with probabily (0.8, 0.1, 0.1)
- For the issue of **fairness**, I resample the test, train, validation dataset, to ensure the fairness of labeling

Comparison Without Probability

- For manipulation, sample:

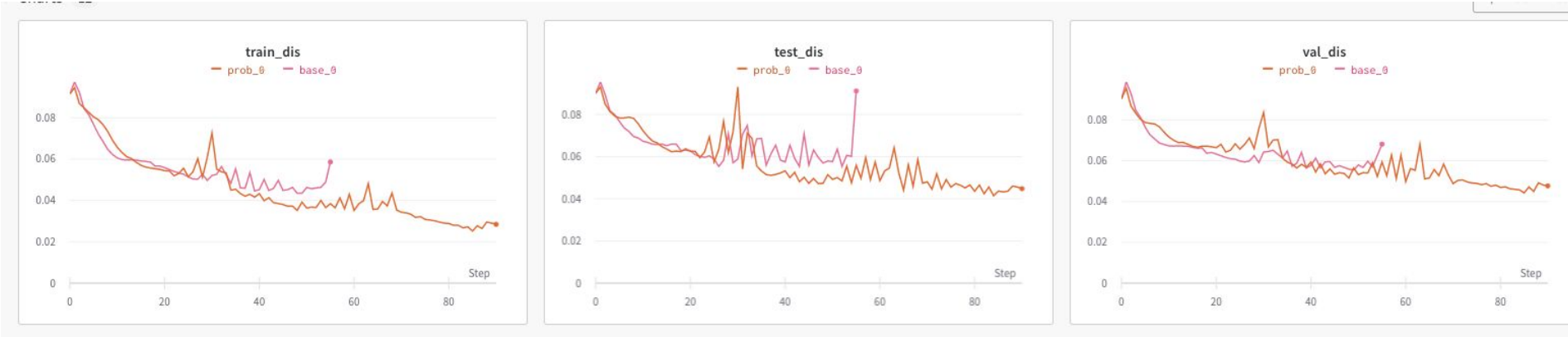
	Test Accuracy (At convergence)	**Round to target accuracy (0.90)**
Base_modal	0.9655	8
Duplicate	0.9026	10
Duplicate (random pick 1 out of 5)	0.9044	51

For duplicate one: the model is likely to be “obfuscated” for some input

Refined Model Detail

- Set train label as the probability vector
- The loss function: MSEloss instead of Cross Entropy to measure the loss between the probability vector
- The backward process of the training will involve the softmax layer, that is the model will be optimized based on the probability vector

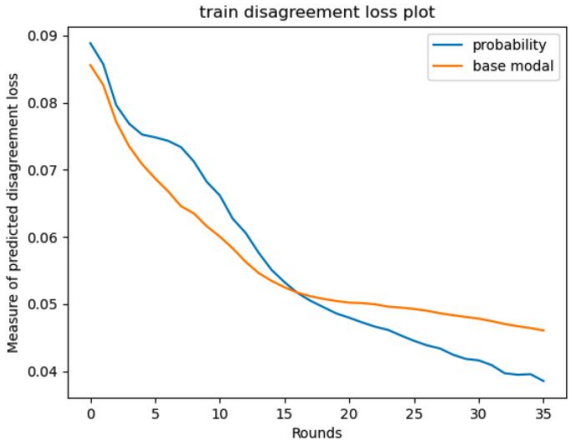
On Other Metrics

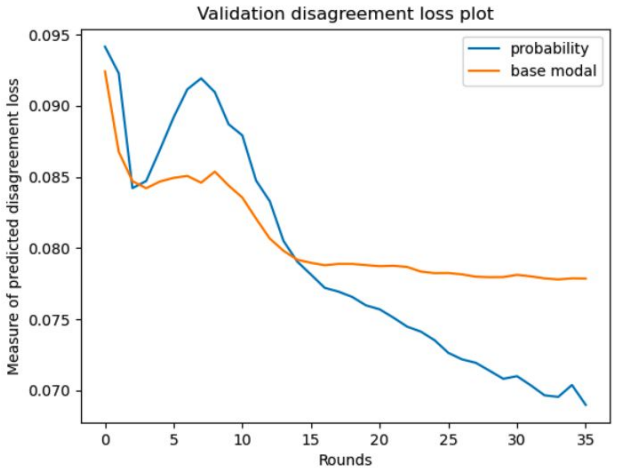


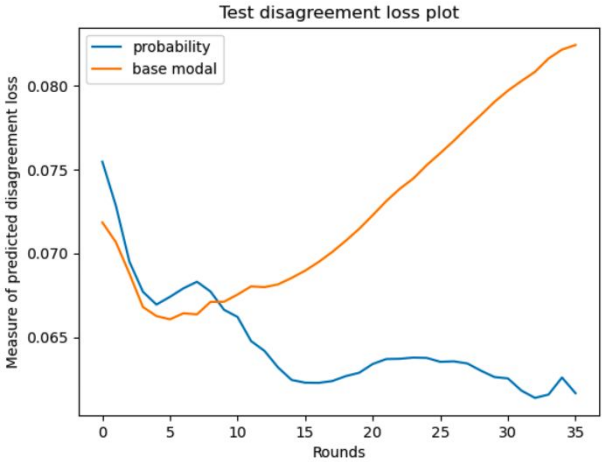
It is able to predict with less user disagreeent

The Probability

- Analyze the output loss from the softMax layer of the base and probability one







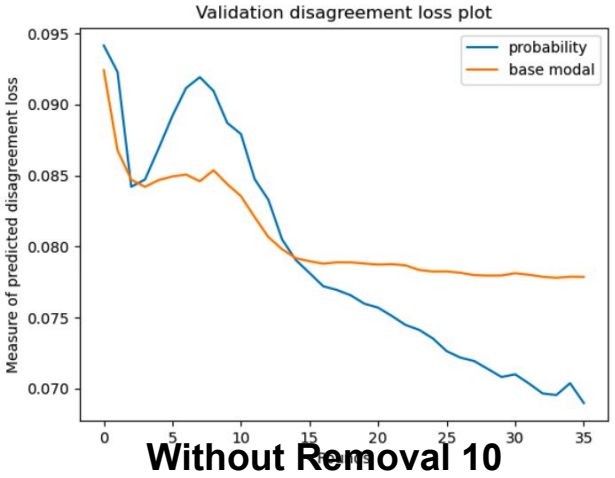
Further Exploration

- Check if unbalanced label will result in lower accuracy in some categories, where the train samples are small.

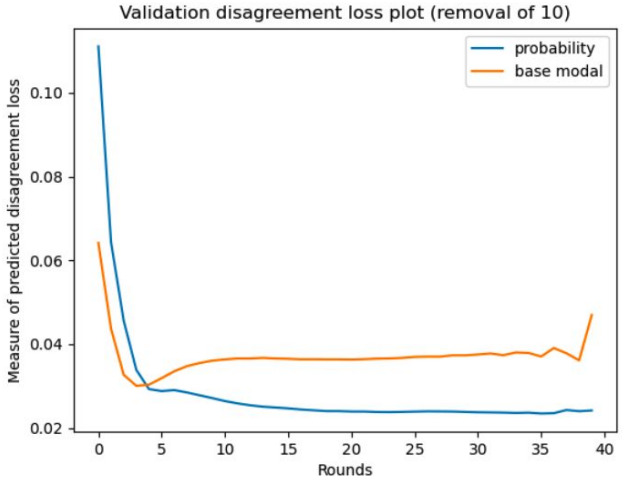
Fairness issue in Prediction for the base model

Adjusted data exploration: Removal of label 10

- If we remove the label 10, in total there will be 914 images left
- The results align with the one without removal, but **the extent of imbalanced is reduced (difference between the probability and base modal becomes smaller)**



Without Removal 10



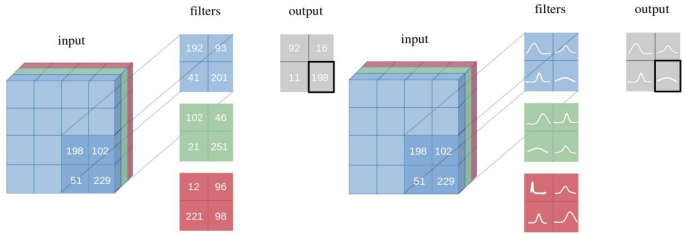
Removal of 10

Future Directions

- Issue: some disagreement over dataset might by **noise**
- Tried the Bayesian CNN, speed low, with similar performance (**uncertainty in the feature extraction)**
- Further interpret the output of layers (explainable of neural networks)
- Questions:
  - How to exploit better data exploit
  - How to explore

Related Works

- **Bayesian CNN:** https://github.com/kumar-shridhar/PyTorch-BayesianCNN



- **Multi-class Labeling:** https://www.uco.es/kdis/mllresources/#Corel16kDesc

  (It is related to our case, but more it focuses on: each image has different characteristics) For our case, it is like hand_right + manipulation_yes + gaze_up+... for one frame. It is

- **Confidence Learning:** estimate uncertainty in dataset, use

References

1. https://arxiv.org/abs/1911.00068 Confidence Learning