Towards a Safer and More Reliable Selective Classifier

With Human Knowledge and Value Incorporated

Xinyue Chen, August 2022

Towards a Safer and More Reliable Selective Classifier

With Human Knowledge and Value Incorporated

Which one do you reject?

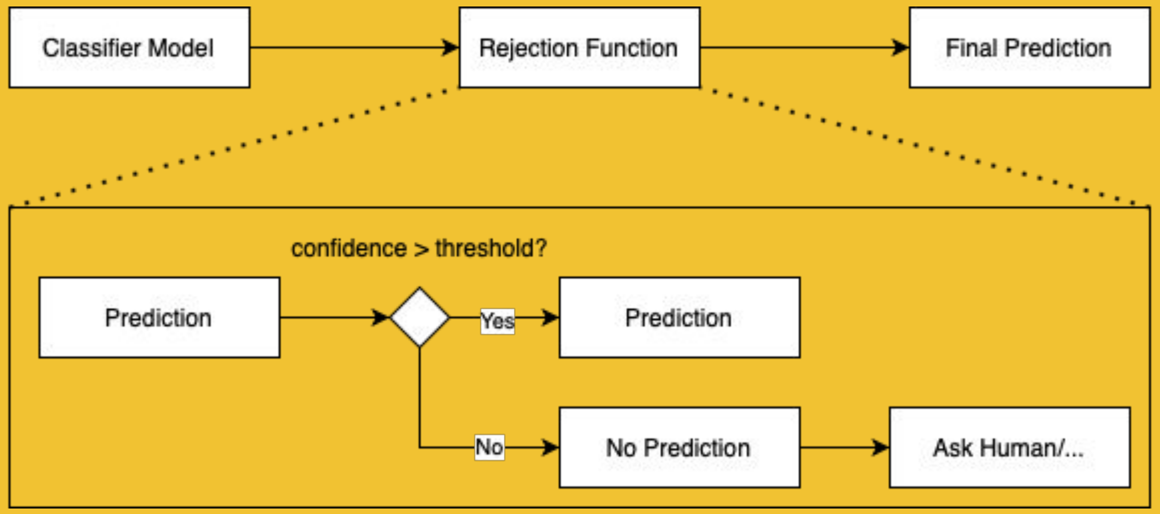
Prediction Dog Cat

Which one do you reject?

Prediction Confidence	Dog 98%	Cat 60%

A Traditional Rejector

(Confidence-based)



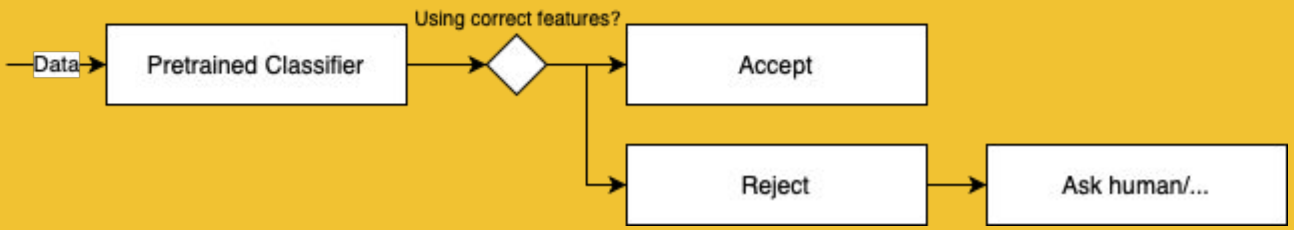
However...

Heavily dependent on model calibration

Unknown unknowns, i.e. high-confidence errors

What if we...

(Feature-based)



Thus...

Making sure the model is using the features it should use

“Value”

$$V(m, D, k) = (1 - \rho_\tau)(\alpha_\tau - k(1 - \alpha_\tau))$$

τ: rejection threshold
ϱ: percentage of rejection
α: accuracy
k: ratio of Vw/Vc
Assuming perfect calibrated models, t=k/(k+1)

- Accuracy - α
- Coverage - ϱ
- Value
  - Figures (k, value)
  - VOB = $\int$ (v(SceneRejector) - v(Baseline))/10

Why?

Stop counting,

Start accounting for social values.

Research Question

How can we effectively improve the reliability and social value of a pretrained classifier, by building a rejector that inspects the features used for prediction?

- Compared with baseline rejectors, how much improvement can it bring?
- In what situation is it suitable to use our proposed rejector, and what is not?

How to implement this function?

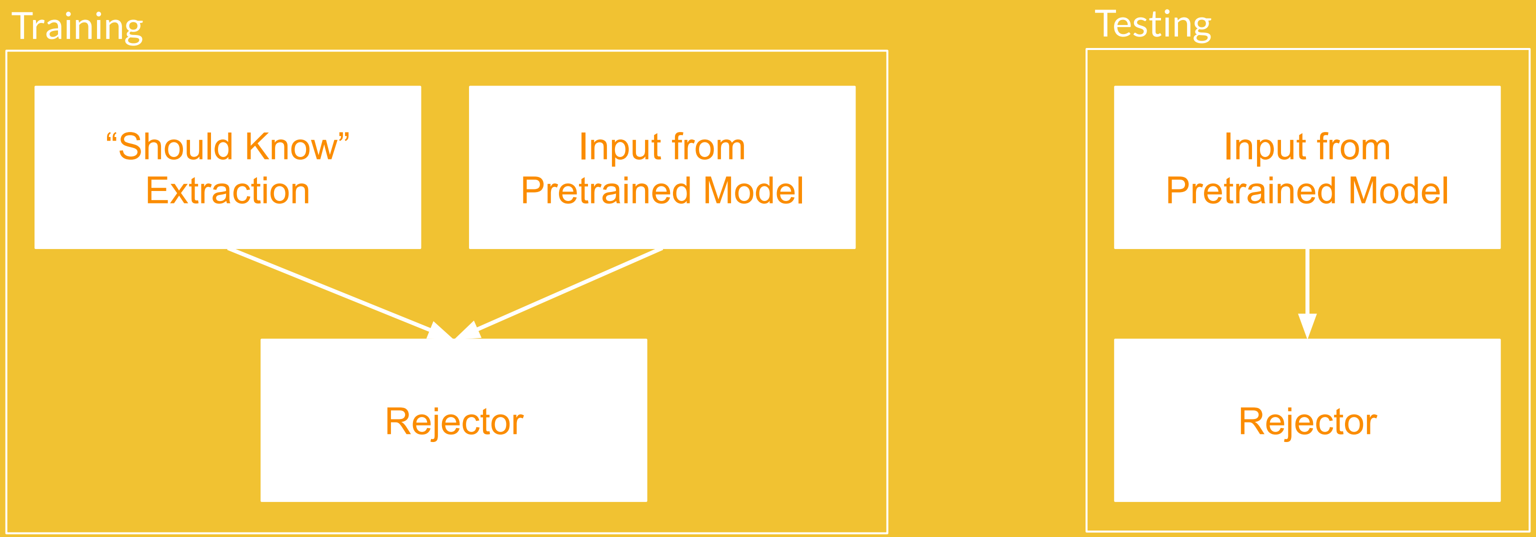
How to explain the behavior of the rejector in different situations?

How to measure the improvement?

Scope

- Pretrained classifier
- Neural Networks - usually not well calibrated (Guo et al., 2017)
- Computer Vision - Scene classification

Methods



Training
"Should Know" Extraction
Input from Pretrained Model
Rejector
Testing
Input from Pretrained Model
Rejector

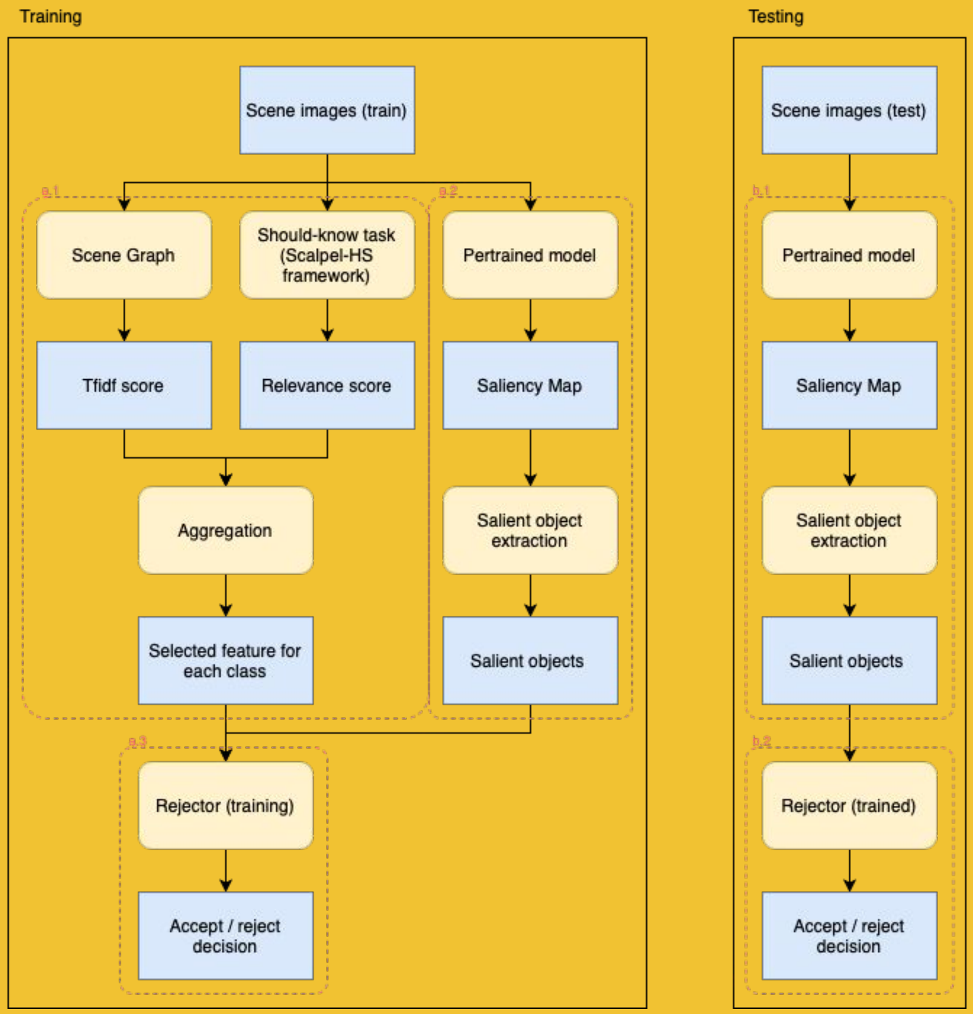
Methods

Training for “kitchen”

Ground truth: kitchen

Prediction: non-kitchen

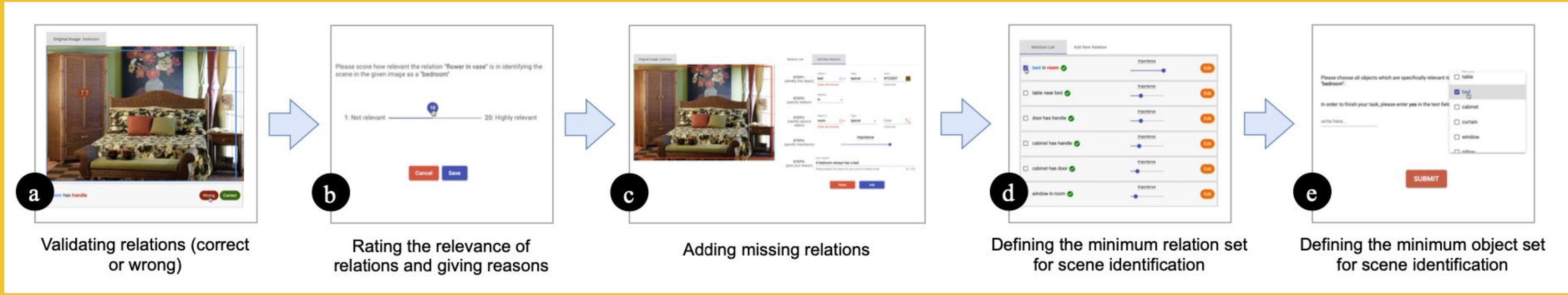
Rejector


Training
Scene images (train)
a.1
Scene Graph
Should-know task (Scalpel-HS framework)
Tfidf score
Relevance score
Aggregation
Selected feature for each class
a.2
Pertrained model
Saliency Map
Salient object extraction
Salient objects
a.3
Rejector (training)
Accept / reject decision
Testing
Scene images (test)
b.1
Pertrained model
Saliency Map
Salient object extraction
Salient objects
b.2
Rejector (trained)
Accept / reject decision

“Should Know” Extraction

Scalpel-HS Framework

- Task and results inherited from “What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition” (Sharifi, 2022)
- Take objects from crowdsourced results (not relation)
- Normalize the relevance score of each object in one class



“Should Know” Extraction

TFIDF Importance

- Scene graph detects all the objects in one image
- Calculate the tfidf score for each object in one class

Oven
Door
Pot
Sink

Oven
Door
Dishwasher

Oven
Door
Table
Person

Plate
Food
Door

Kitchen

Bed
Door
Table

Bed
Books
Table
Chair
Door

Bed
Basket
Cat

Bed
Lamp
Door

Bedroom

Input from Pretrained Model

Salient Object Extraction

Rejector Training

Data Preparation

Shower	Sink	Towel
0	0	0

- X: the aggregated “should know” features matched with “really know”, one hot encoding
- Y: accept or reject
    - Correct prediction - accept
    - Wrong prediction - reject
    - Balanced distribution of correct prediction (CP) and wrong prediction (WP)

Model

- Decision tree - faster training, more interpretable
- Optimized for accuracy

Experimental Setup

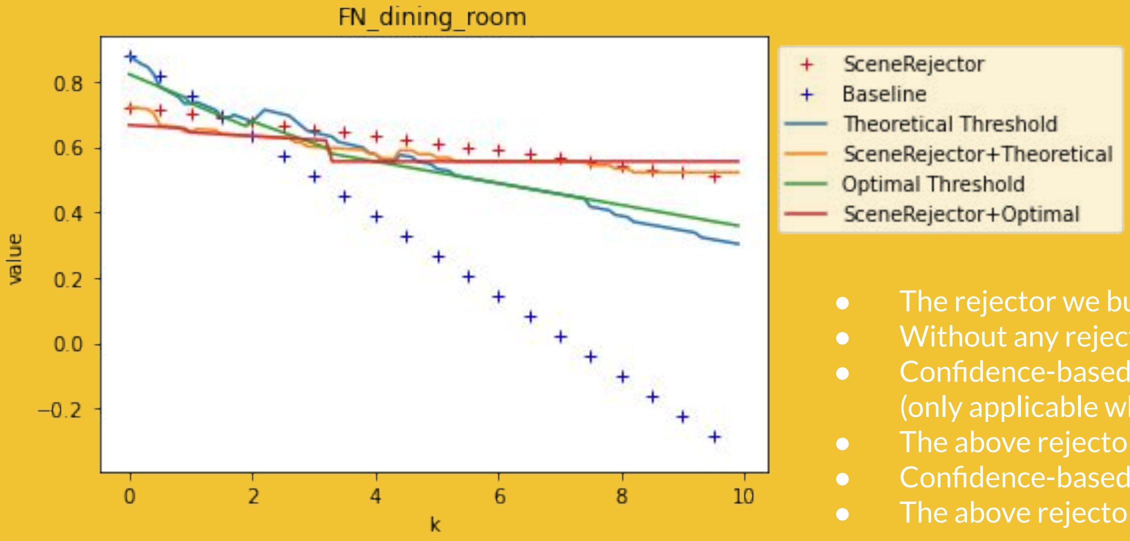
Pretrained Model

- ResNet
- Biased during training (manually injected FP and FN)
- Multi-class → binary classification for simplification as the first step
- 8 binary classifier

Dataset

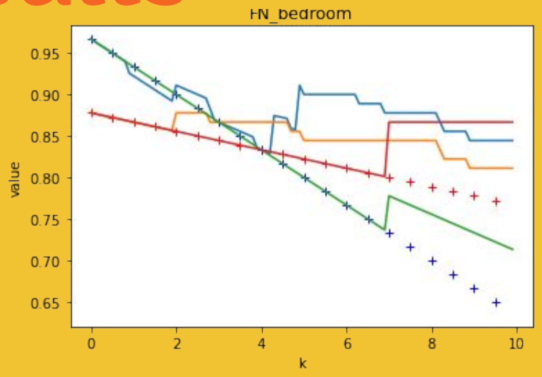
- *PLACES*
- 225 WP and CP for both conditions of FN and FP

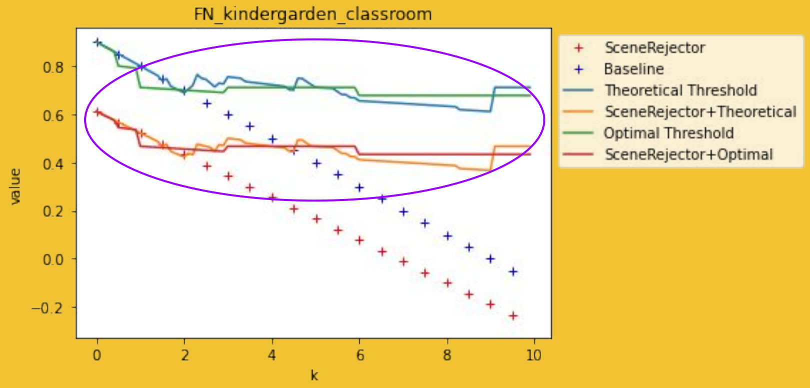
Results

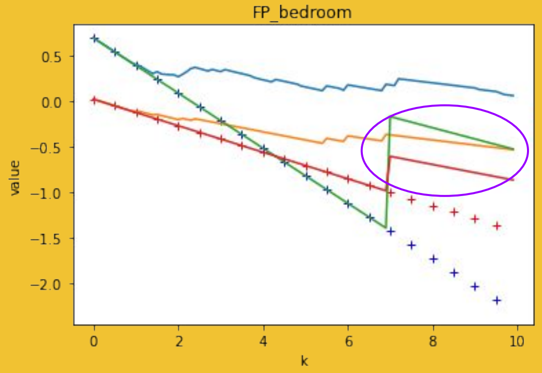


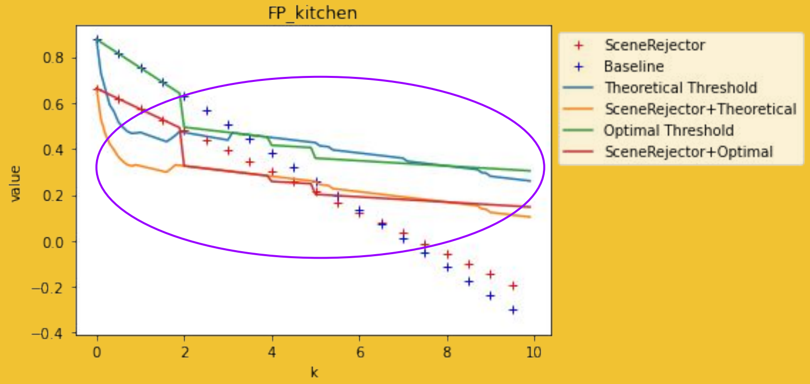
- The rejector we built
- Without any rejector
- Confidence-based rejector with theoretical threshold (only applicable when ECE=0)
- The above rejector + SceneRejector
- Confidence-based rejector with optimal threshold
- The above rejector + SceneRejector

Results


FN_bedroom
value
k


FN_kindergarden_classroom
value
k
SceneRejector
Baseline
Theoretical Threshold
SceneRejector+Theoretical
Optimal Threshold
SceneRejector+Optimal

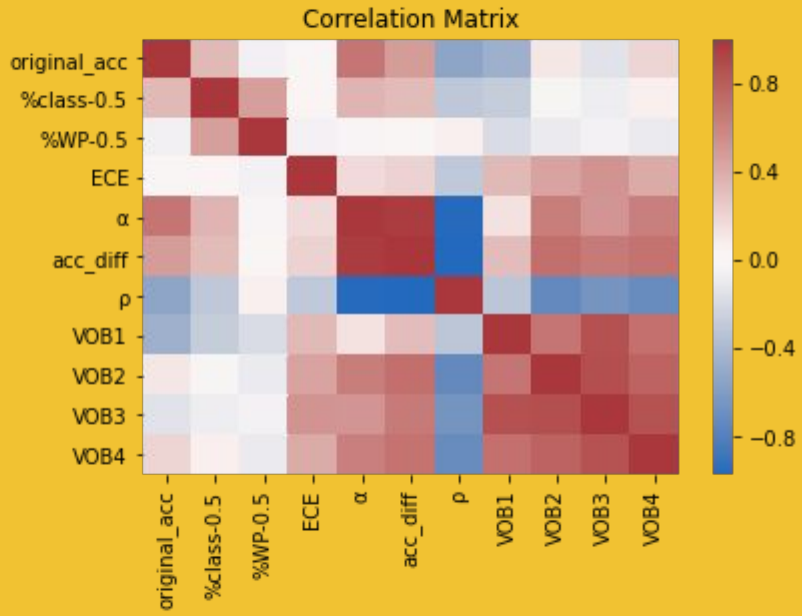

FP_bedroom
value
k


FP_kitchen
value
k
SceneRejector
Baseline
Theoretical Threshold
SceneRejector+Theoretical
Optimal Threshold
SceneRejector+Optimal

Observations

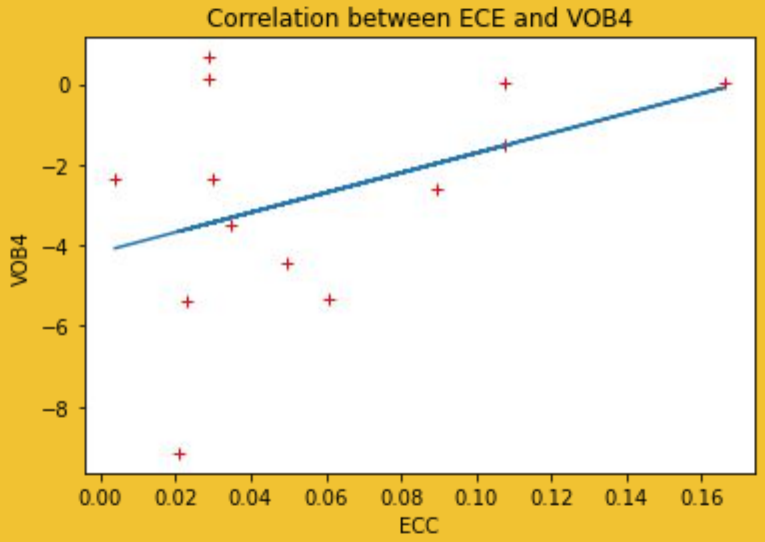
- Compared with baseline, SceneRejector creates better value as the penalty of a wrong prediction gets greater.
- The addition of SceneRejector to a confidence-based rejector does not improve value (already well calibrated)
- Best rejector? Optimal/ Optimal+SR

Analysis



Variable	Meaning
original_acc	The accuracy of the pretrained classifier
%class-0.5	The absolute value of the difference between the percentage of samples of the class of interest and 0.5, the percentage in a balanced binary rejector training dataset
%WP-0.5	Within the sample set of the class of interest, the absolute value of the difference between the percentage of WP samples and 0.5, the percentage of WP in a dataset with a balanced WP CP distribution
ECE	The ECE score of the pretrained classifier
α	The accuracy of the accepted set given by SceneRejector
acc_diff	The difference between α and original_acc
ρ	The rejection rate of SceneRejecor
VOB1	$\int_k (value(SceneRejector) - value(Baseline))/10$
VOB2	$\int_k (value(SceneRejector + Theoretical) - value(TheoreticalThreshold))/10$
VOB3	$\int_k (value(SceneRejector + Optimal) - value(OptimalThreshold))/10$
VOB4	$\int_k (value(SceneRejector) - value(OptimalThreshold))/10$

Analysis

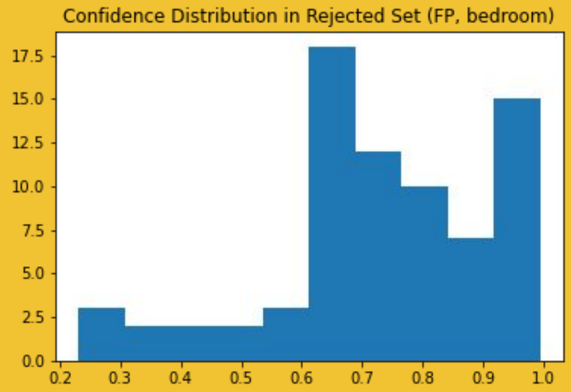


Pretrained model ECE ↑,
Value ↑
r = 0.4075

ECE: Expected Calibration Error
VOB4: how much more value SceneRejector creates than confidence-based rejector (optimal threshold)

Analysis - Properties of Rejected Set

Condition	Class	Accuracy of Rejected Set
FP	bathroom	0.4458
	kindergarden_classroom	0.2813
	bedroom	0.4595
	kitchen	0.1364
FN	dining_room	0.3478
	bedroom	0.2000
	kindergarden_classroom	0.3704
	kitchen	0.2353



Analysis - Properties of Rejected Set

Ground truth: kitchen
Prediction: dining_room

Ground truth: kitchen
Prediction: dining_room

Conclusion

- SceneRejector is a working rejector and validates the concept of feature-based rejectors
- SceneRejector creates better value as the penalty of a wrong prediction increases
- A positive correlation between the ECE score of the pretrained classifier and the value of SceneRejector
- SceneRejector is able to reject high-confidence errors, i.e. unknown unknowns
- Impact: technical and social

Discussion & Future Works

- Scene Graph inaccuracies
    - Include object detection model training in rejector training
- Conversion from multi-class to binary classification
- Only one task, can results be generalized?
- More human involvement
    - Manual aggregation and verification of “should know” features
    - During rejector training, accept/reject label determined by prediction/ground truth match
- Revise the expression of “value”
- Experiment with more ways/architectures to build the rejector
- Interpret rejector behaviors

Towards a Safer and More Reliable Selective Classifier

With Human Knowledge and Value Incorporated

Xinyue Chen, August 2022