

Uncertainty in Recurrent Decision Tree Classifiers

Stefan Wezel

Supervised by Stephan Alaniz, Zeynep Akata

Explainable Machine Learning

November 3, 2020

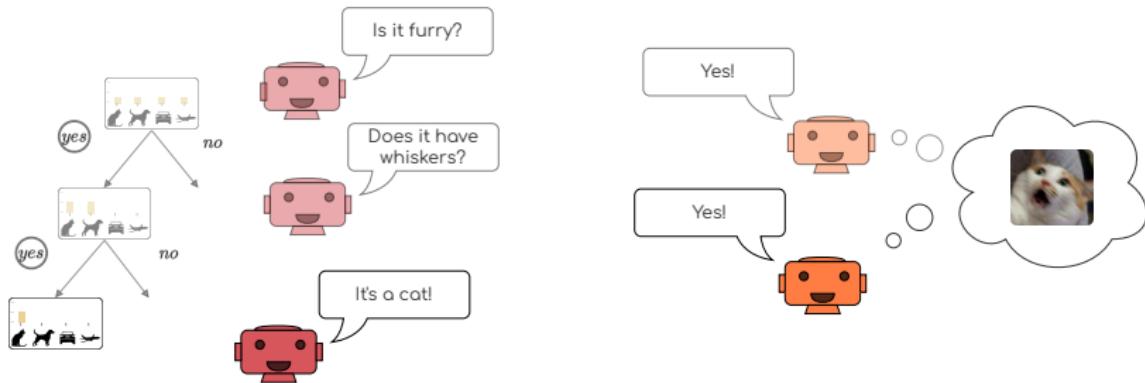
What?

Setting

- Building on RDTC model by Alaniz et al. [1]
- Image classification: many powerful architectures out there
- Prominent example: ResNet
- Popular models only yield classification
- No reasoning behind classification

What?

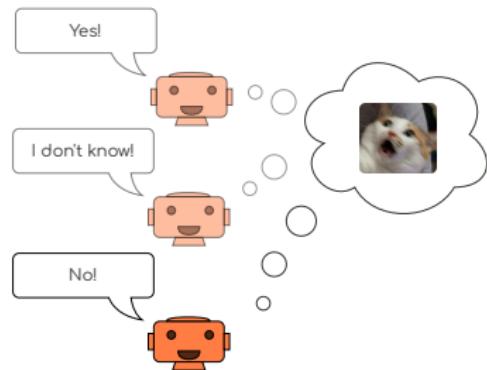
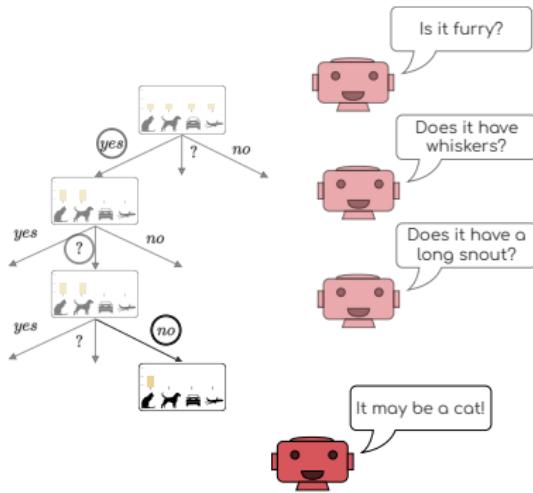
What is a Recurrent Decision Tree Classifier?



- Alaniz et al. [1] propose RDTC
- Two communicating agents
- Left: ask questions — right: look at data and answer them
- Unfolding tree reveals reasoning behind classification

What?

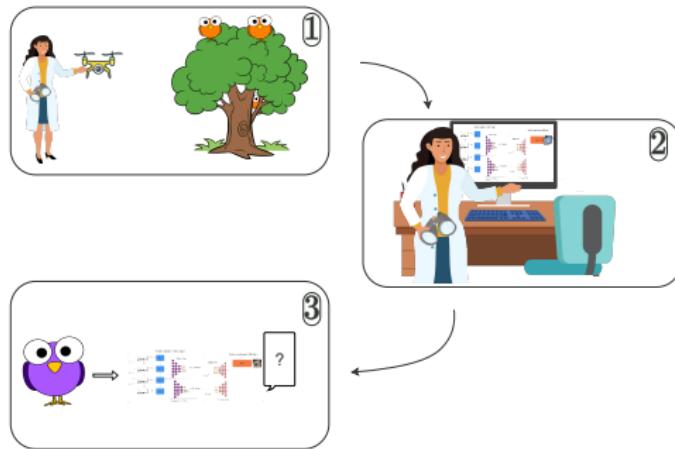
Introducing Uncertainty to RDTC



- Right agent is aware of uncertainties
- Communicates this to left agent

Why do we need uncertainty?

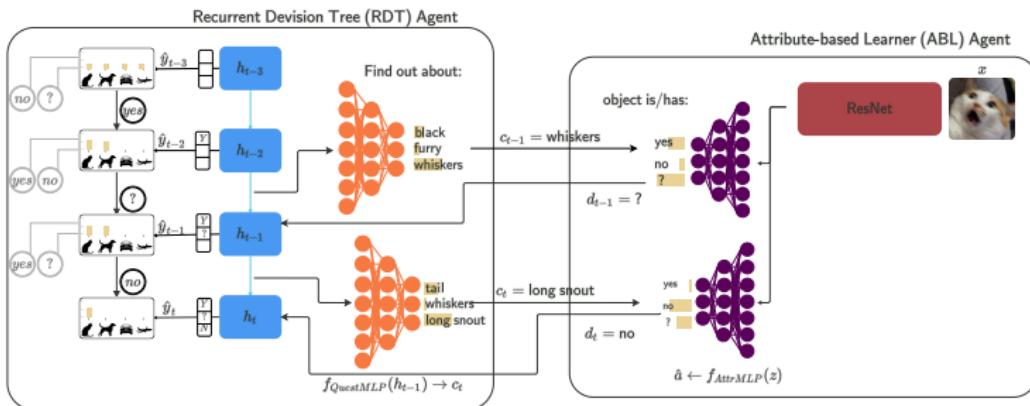
A Practical Example...



- Ornithologist surveys area using drone and CV software
- Classification is automated with our model
- Bird species unknown to model yield high uncertainty
- Those can be classified manually

How?

Architecture

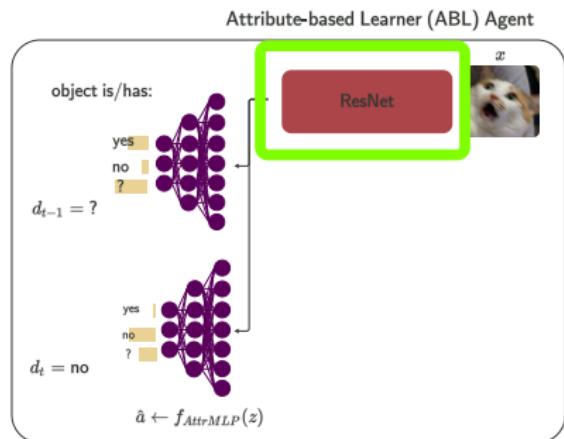


- RDT can not see image but ask if attribute is present
- AbL can see image and answer RDT's questions

Attribute-based Learner

Extracting features: ResNet

- Extract features using ResNet
- Then pass to $f_{AttrMLP}$

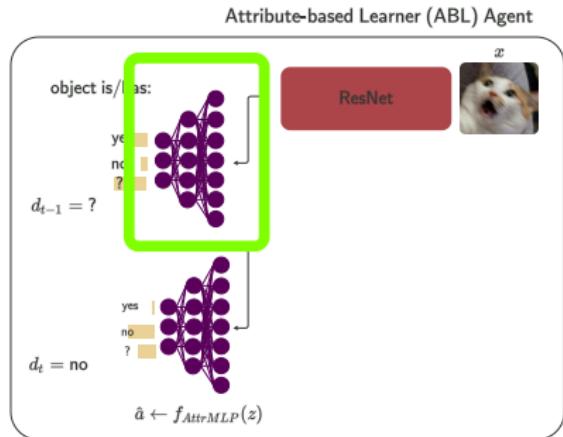


Attribute-based Learner

Mapping features to attributes: Attribute MLP

- Map features to answers
- Yes, No, ? for each attribute
- Discrete answers with TempSoftmax

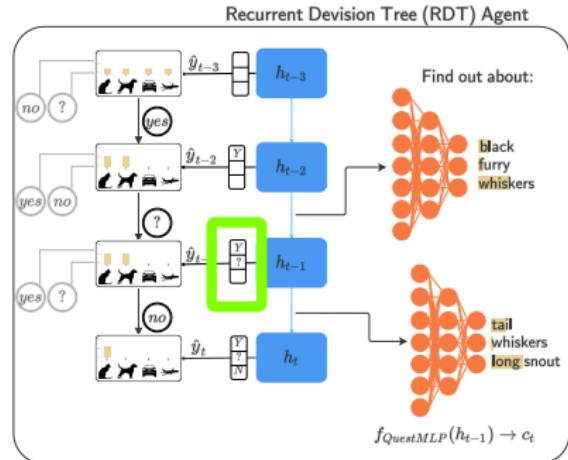
$$\frac{\exp((\log \pi_i)/\tau)}{\sum_{j=1}^K \exp((\log \pi_j)/\tau)}$$
$$= \hat{a}$$



Recurrent Decision Tree

Storing decisions: Explicit memory

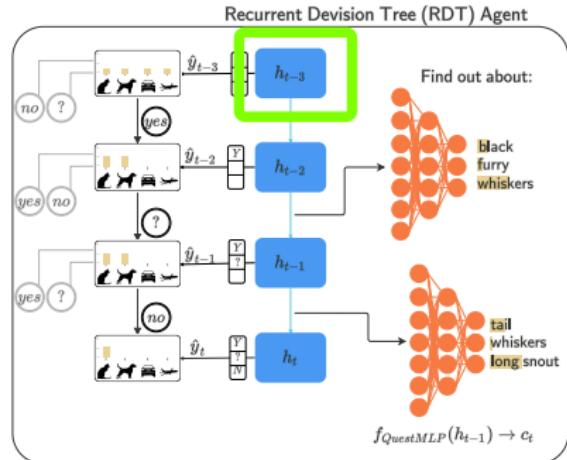
- Store questions (indices)
- and answers (Y,N,?)
- Iteratively increases



Recurrent Decision Tree

Keeping track: LSTM

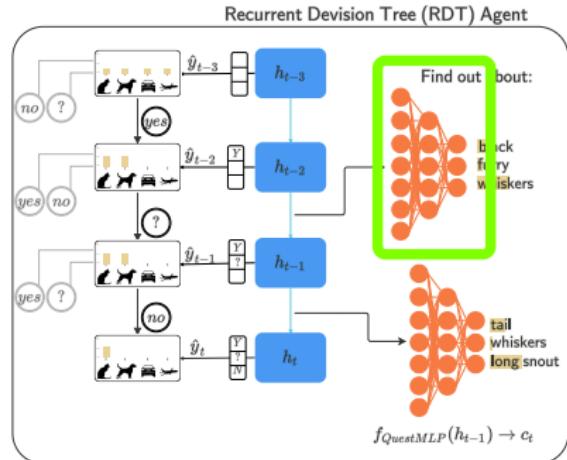
- Hidden state based on memory
- Basis for next question



Recurrent Decision Tree

Choosing questions: Question MLP

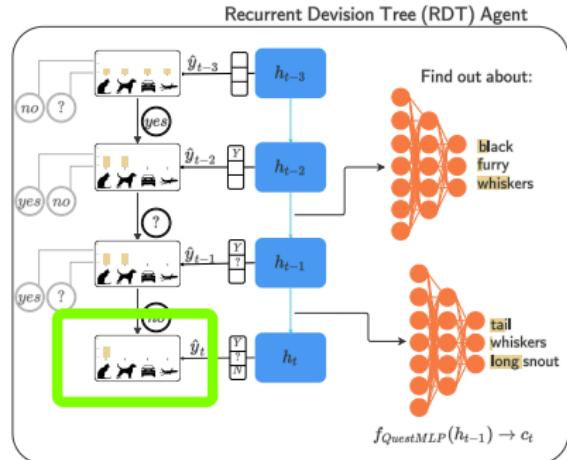
- o Pose new question c_t
- o $d_t = \hat{a}[c_t]$
- o Sample from log probabilities
- o $c_t \sim p(c_t)$
- o With GumbelSoftmax



Recurrent Decision Tree

Making a classification: Class MLP

- Based on explicit memory
- Make classification
- In every communication step



Training

Joint Objective

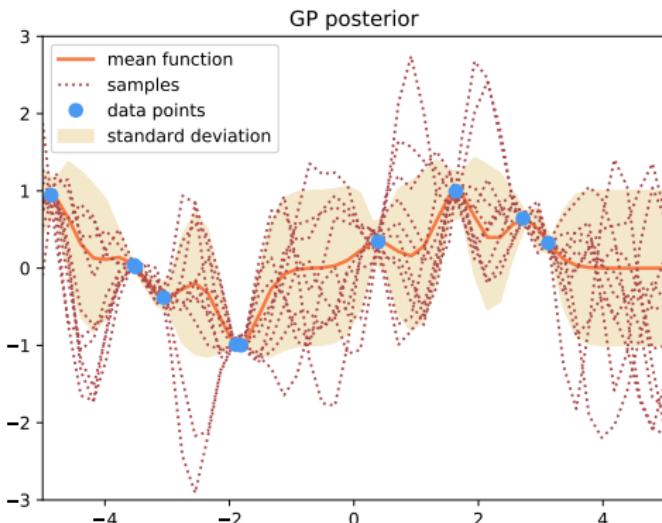
- Optimize for class (and attribute accuracy)

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^T \left[(1 - \lambda) \underbrace{\mathcal{L}_{CE}(y, \hat{y}_t)}_{Class} + \lambda \underbrace{\mathcal{L}_{CE}(\alpha_{y,ct}, \hat{\alpha}_{ct})}_{Attr} \right]$$

- λ can be used to balance the two terms
- For all of our experiments, we use $\lambda = 0.2$
- Discourage deep trees

Background

A small excursion to Gaussian Processes (GP)



- Data can be described by (infinitely) many functions
- GP is PDF over these functions
- Parameterized by mean function and covariance function
- Variance corresponds to model uncertainty

Background

Dropout Uncertainty Estimation

- Why care about GP's?
- Neural net is set of weighted linear functions, activated by non-linearity
- PDF over each weight → finite GP
- Neural net \approx GP
- → Extract model uncertainty from a neural net
- Proof by Gal and Ghahramani [2]
- → Theoretical foundation for estimating uncertainty in RDTC

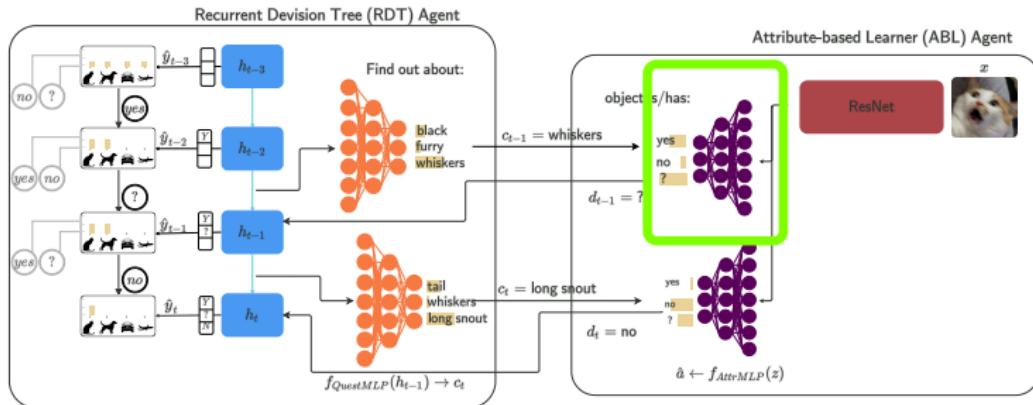
Background

Dropout Uncertainty Estimation

- Posterior over functions requires computing integrals
- Methods of variational inference
 - GP objective → minimization objective
- For covariance function → Monte-Carlo integration
- Rewrite approximate GP objective as dropout net objective
- Variance arising from dropout → model uncertainty

How?

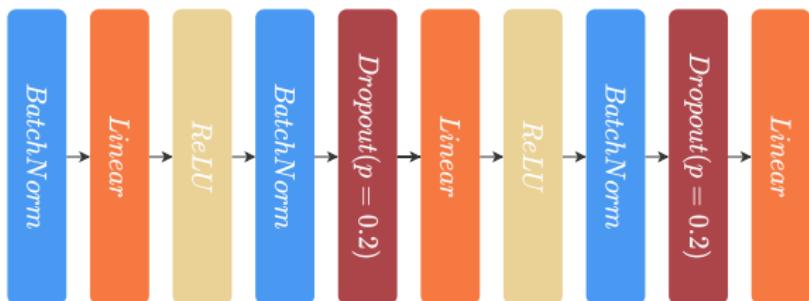
How do we get uncertainty information?



- Use Gal and Ghahramani's [2] proof to estimate uncertainty
- Keep ResNet
- Make $f_{AttrMLP}$ a dropout MLP
- After extracting features → compute $Var(n \text{ forward passes})$

Getting Uncertainty Information

Estimating Uncertainty in the AbL



- We include dropout layers in $f_{AttrMLP}$
- We tested different configurations
- Combination of batchnorm and dropout worked best

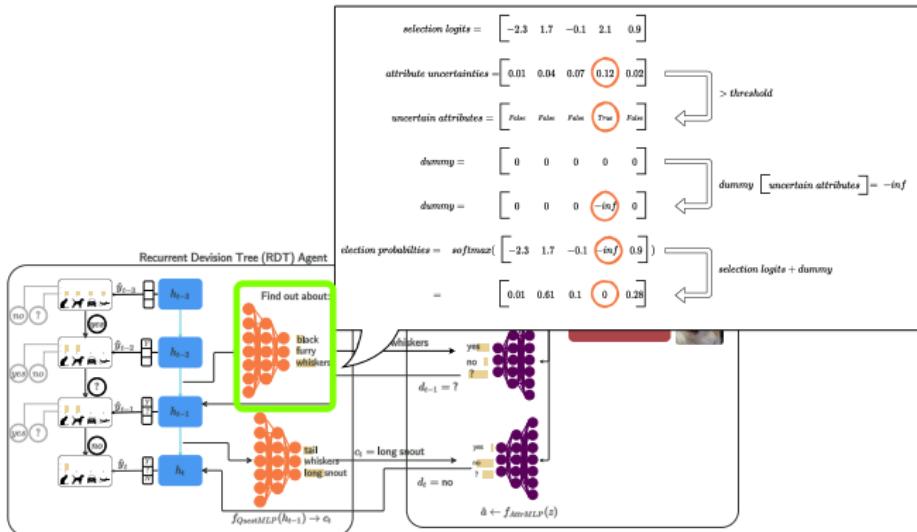
Using Uncertainty Information

Uncertainty Information as Inductive Bias

- We use uncertainty in two different strategies
 - Prevent model from asking about uncertain attributes
→ remRDTC
 - We give the model the ability to answer with 'I don't know'
→ extRDTC
- Attribute uncertain if above hyperparameter threshold

Using Uncertainty Information

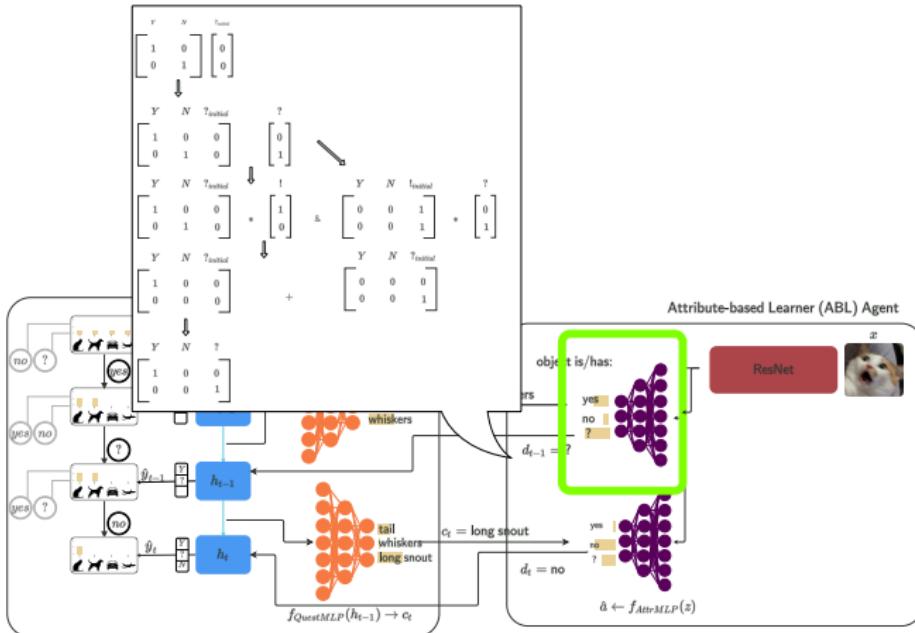
remRDTc - Removing uncertain attributes



- Idea: Skip uncertain questions
- Reminder: output from $f_{QuestMLP}$ is attribute index
- If attribute deemed uncertain, replace logit with $-\infty$
→ Gumbel softmax cannot pick these attributes

Using Uncertainty Information

extRDTC - Extending the vocabulary



- Idea: AbL can say 'I don't know'
- Binary vector with 1s where uncertainty is above threshold
 - → Append to initial answer
- Prevent conflicting answers

Introducing Uncertainty

- Use proof of Gal and Ghahramani [2] to estimate uncertainty in RDTC [1]
- Estimate uncertainty in $f_{AttrMLP}$
- Uncertainty information is used in two strategies
 - remRDTC
 - extRDTC

Experiments

- RDTC is now aware of, and can express uncertainties
- In experiments, we:
 - Investigate uncertainty and its relationship to other variables
 - Test our model on OOD data
 - Test the model's performance on benchmark datasets

Experiments

Datasets with attributes

- Animals with Attributes 2 (AWA2)
 - medium size, coarse grained
- aPY
 - small size, coarse grained
- CUB
 - large size, fine grained

Experiments

Uncertainty and its relationship to other variables

- Accuracy, uncertainty, and usage of attributes

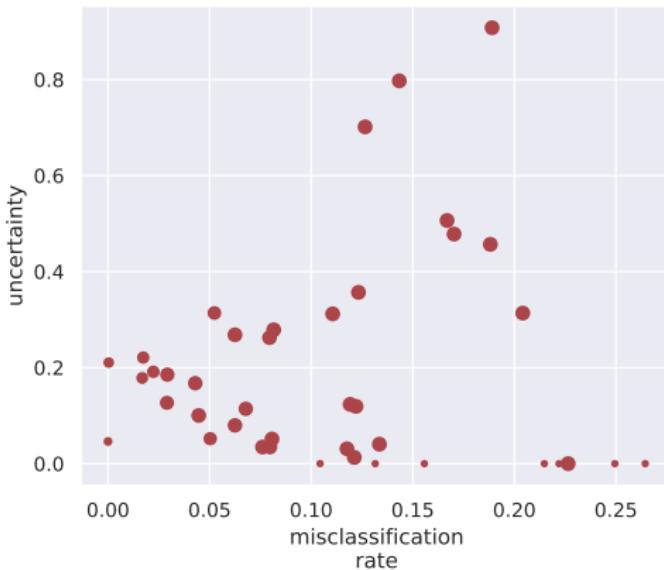


- Positive correlation between usage and accuracy
- No correlation between uncertainty and accuracy

Experiments

Uncertainty and its relationship to other variables

- Uncertainty, misclassification rate, usage ($\text{size} \propto \text{usage}$)

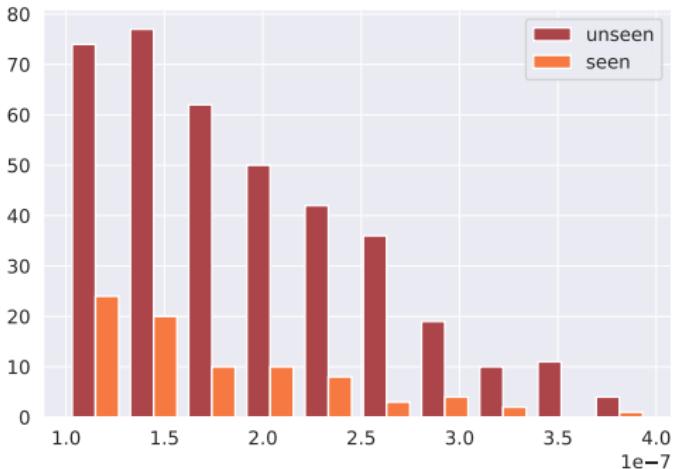


- Uncertain attributes seem to be misclassified more often

Experiments

OOD Detection

- extRDTc in zero shot setting (using CUB)
- Histogram of uncertainty values of seen and unseen classes



- Every bin > 0, more values from unseen classes
- Unseen classes have higher uncertainty values

Experiments

Comparison to other models

- Decision Tree (DT)
 - Split on ResNet features
 - Split until every leaf node corresponds to one class
- Explainable Decision Tree (XDT)
 - Split on $f_{AttrMLP}$ attributes
 - Split until every leaf node corresponds to one class
- dNDF [4]
 - Every inner node is parameterized differentiable function
 - Leaf nodes represent class distributions
 - Learn the optimal route through the tree for each example

Experiments

Comparison to other models

- aRDTC
 - RDTC with $\lambda > 0$
- randRDTC
 - Random baseline
 - Like remRDTC, but uncertainty vector is generated randomly
- ResNet
 - Not explainable
 - Trained on ImageNet, fine-tuned for specific datasets

Experiments

Results on Benchmark Datasets

| | AWA2 | aPY | CUB |
|---------------|-----------|-----------|-----------|
| ResNet [3] | 98.2± 0.0 | 85.1± 0.6 | 79.0± 0.2 |
| DT | 78.0± 0.4 | 64.3± 0.6 | 19.3± 0.3 |
| dNDF[4] | 97.6± 0.2 | 85.0± 0.6 | 73.8± 0.3 |
| RDTC[1] | 98.0± 0.1 | 85.7± 0.7 | 78.1± 0.2 |
| XDT | 73.9± 0.9 | 59.9± 1.5 | 4.9± 1.3 |
| aRDTC[1] | 98.6 | 86.1 | 77.9± 0.6 |
| remRDTC(ours) | 98.7 | 86.4 | 77.7 |
| extRDTC(ours) | 98.7 | 85.4 | 77.8 |
| randtRDTC | 98.5 | 86.5 | 76.8 |

Experiments

Results on Benchmark Datasets

| | aRDTC [1] | randRDTC | remRDTC(Ours) | extRDTC(Ours) |
|------------|-------------|----------|---------------|---------------|
| CUB | | | | |
| Class | 77.9 | 76.8 | 77.7 | 77.8 |
| Attribute | 84.9 | 83.9 | 87.0 | 86.9 |

Discussion

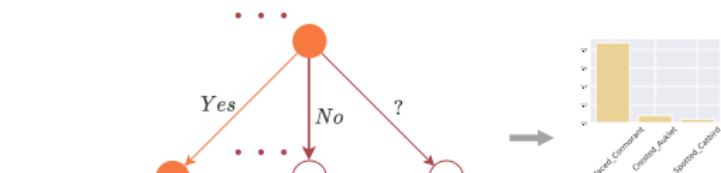
Looking back...

- The right kind of uncertainty?
 - Uncertainty arising from noise or occlusions not considered
 - Dropout Uncertainty Estimation lacks empirical proof
- Other methods of uncertainty estimation
 - Often computationally expensive
 - Open area of research
- Beyond attribute uncertainty
 - Estimate uncertainty in other parts (i.e. $f_{ClassMLP}$)

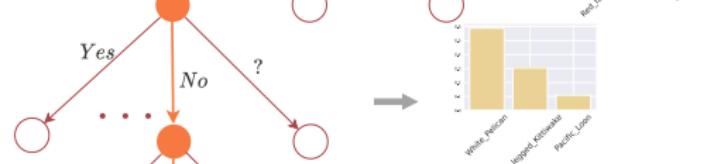
Conclusions

A qualitative Example

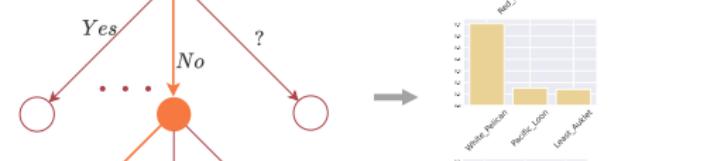
has_size :: small



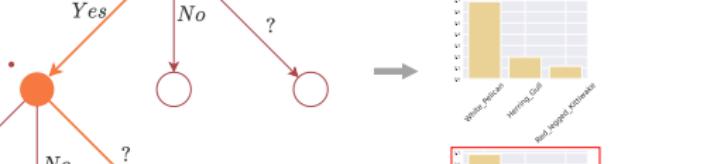
has_wing_color :: olive



has_under_tail_color :: green



has_nape_color :: black



has_belly_pattern :: solid



References

- [1] S. Alaniz, D. Marcos, and Z. Akata. Learning decision trees recurrently through communication. ICML XXAI Workshop, 2020.
- [2] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning, pages 1050–1059, 2016.
- [3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
- [4] P. Kotschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo. Deep neural decision forests. In Proceedings of the IEEE international conference on computer vision, pages 1467–1475, 2015.