Active Learning with Partial Feedback

Under review at ICLR-2019

Peiyun Hu[1], Zachary C. Lipton[1,3], Anima Anandkumar[2,3], Deva Ramanan[1]

[1]Carnegie Mellon University

[2]California Institute of Technology

[3]Amazon AI

Contents

Introduction

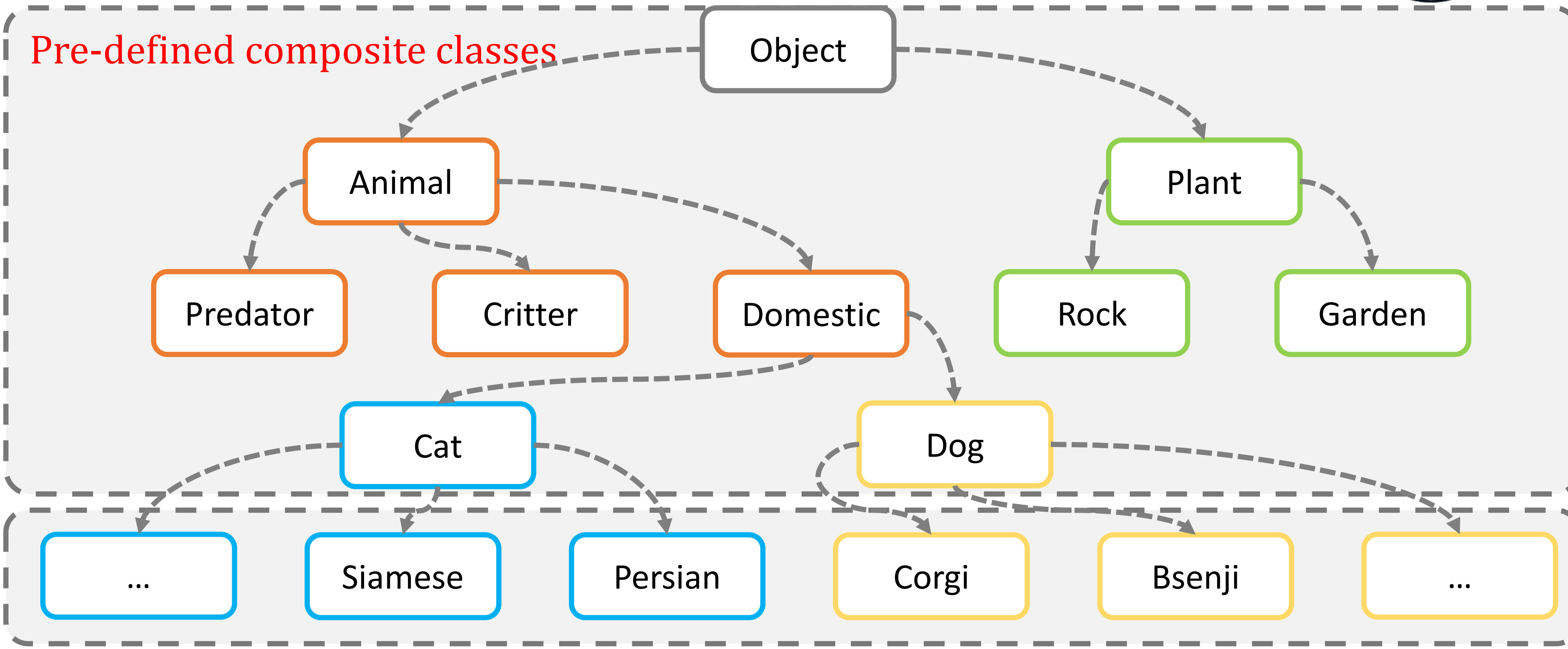
- Given a large set of unlabeled images, and a budget to collect annotations, how can we learn an accurate image classifier most economically?
- Typically, AL treats the labeling process as atomic: every annotation costs the same and produces a correct label.
- However, large-scale multi-class annotation is seldom atomic. We can't simply ask a crowd-worker to select one among 1000 classes.

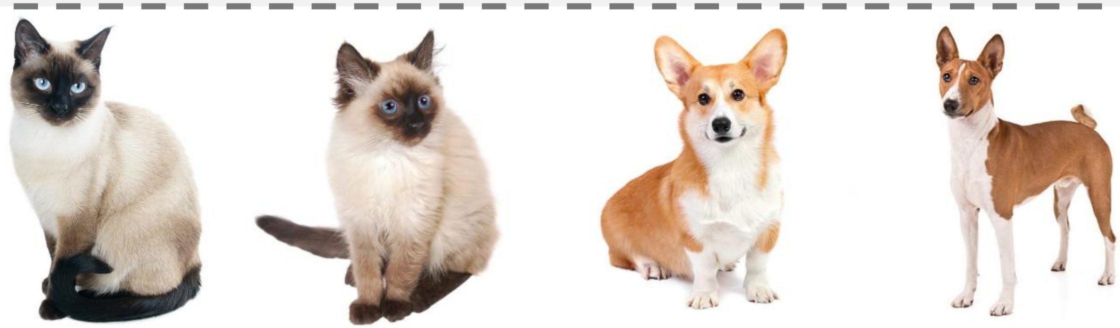
...	Siamese	Persian	Corgi	Bsenji	...




Thousands of classes

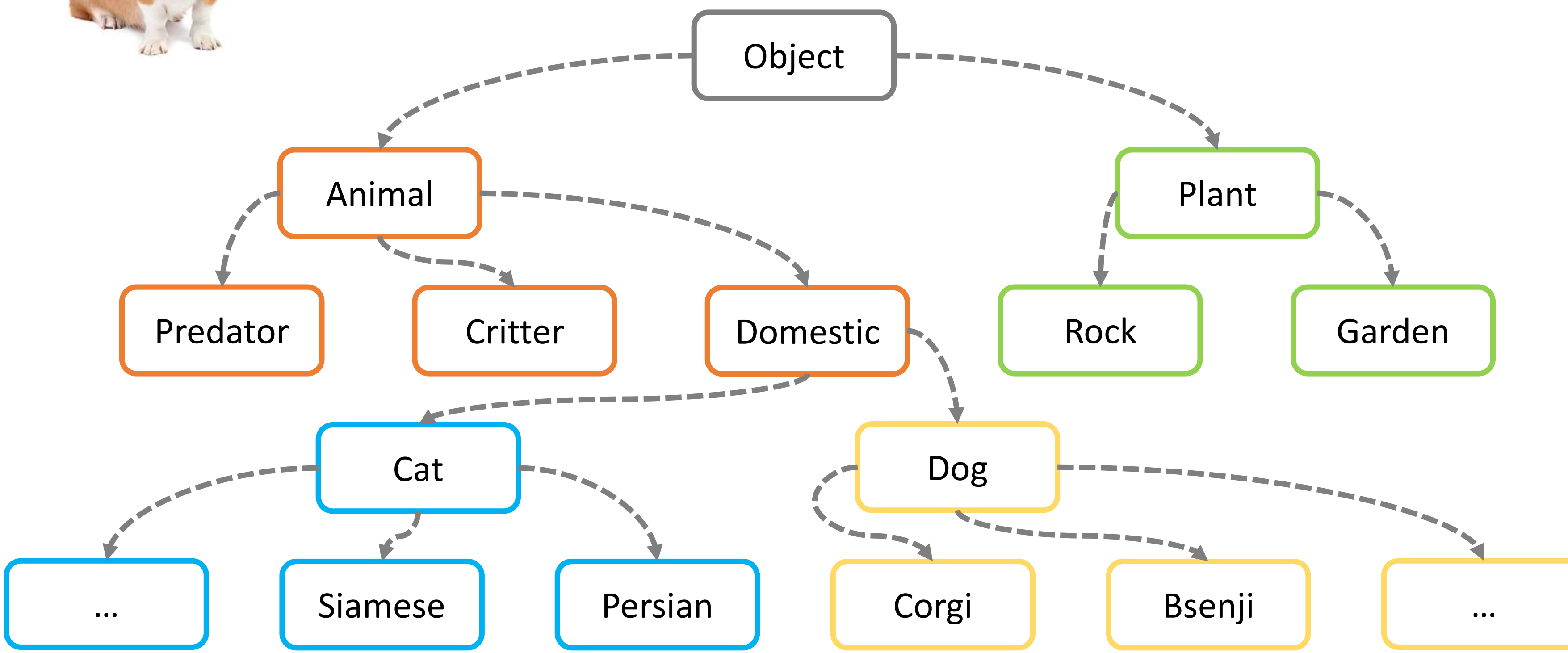
Introduction





Introduction

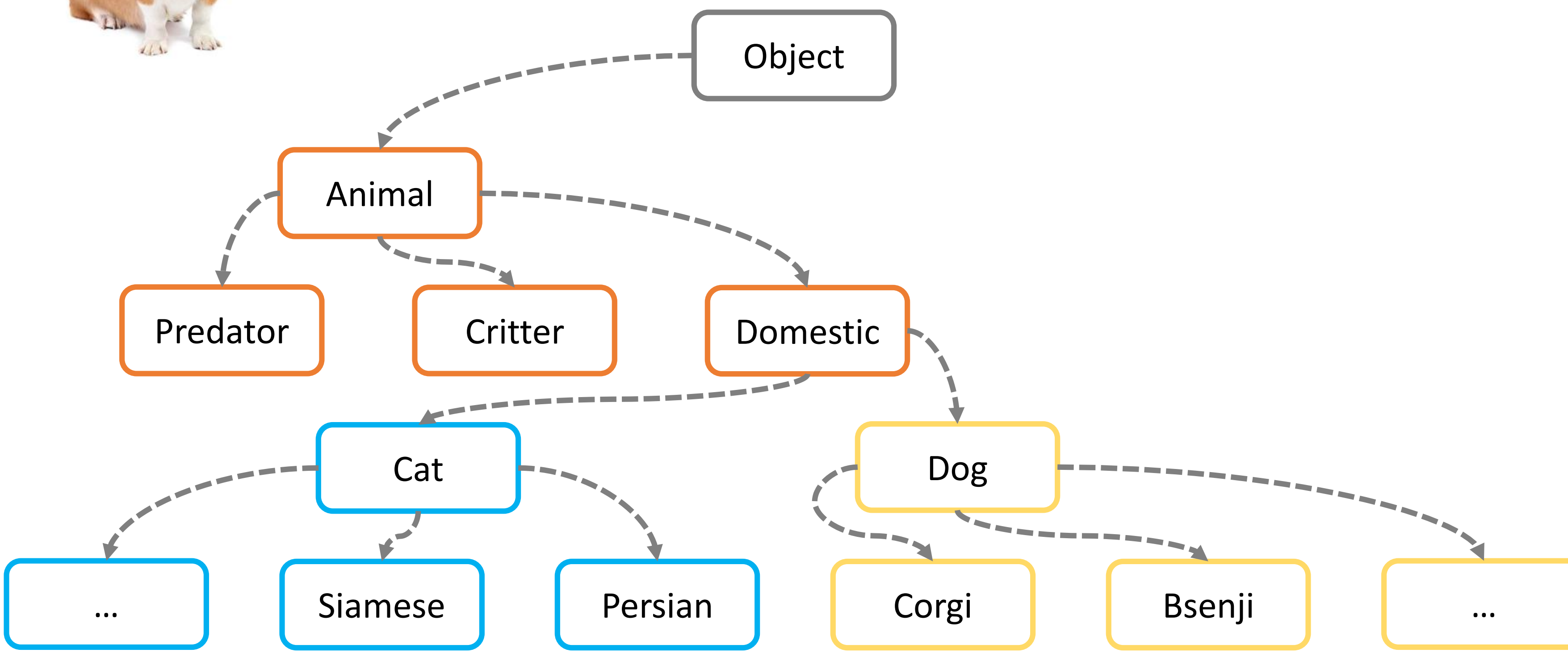
ML: Is it a plant? Human: No!



Introduction

ML: Is it a plant? Human: No!

ML: Is it domestic? Human: Yes!

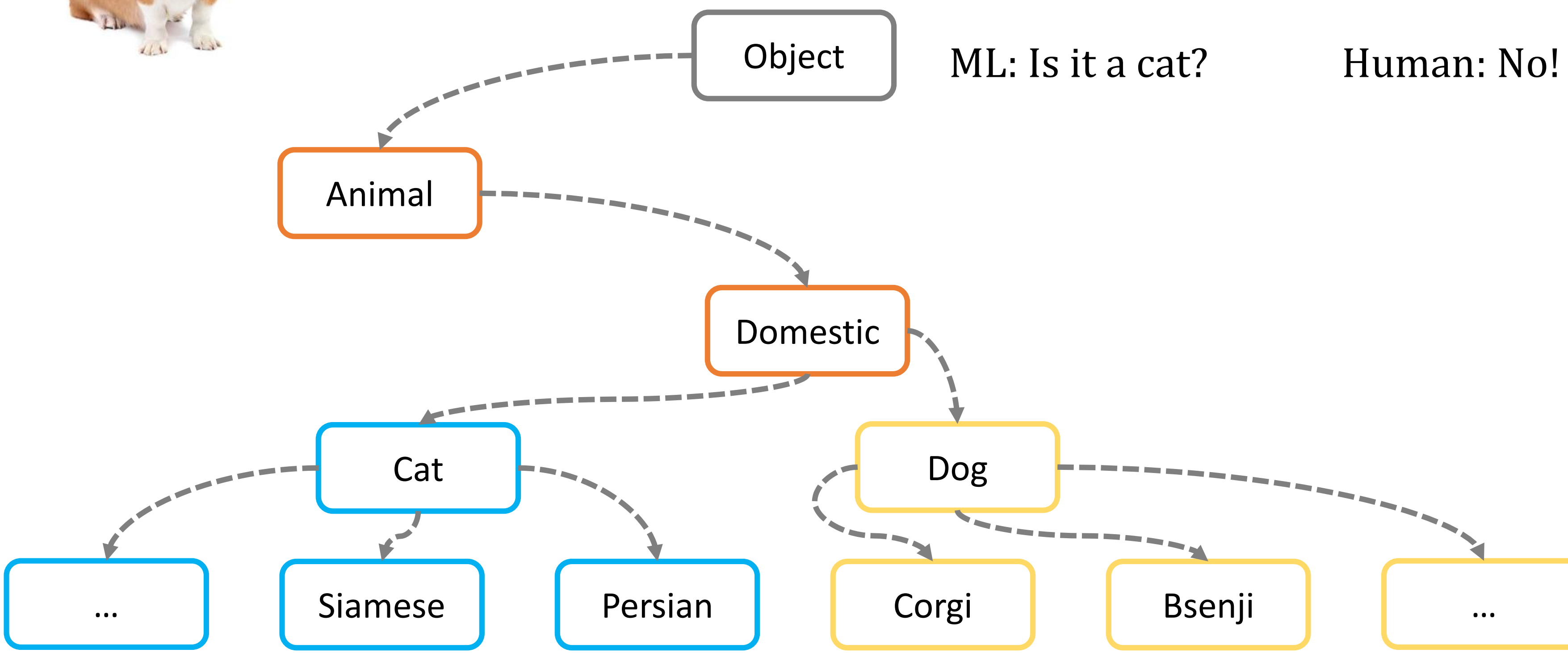


Introduction

ML: Is it a plant? Human: No!

ML: Is it domestic? Human: Yes!

ML: Is it a cat? Human: No!



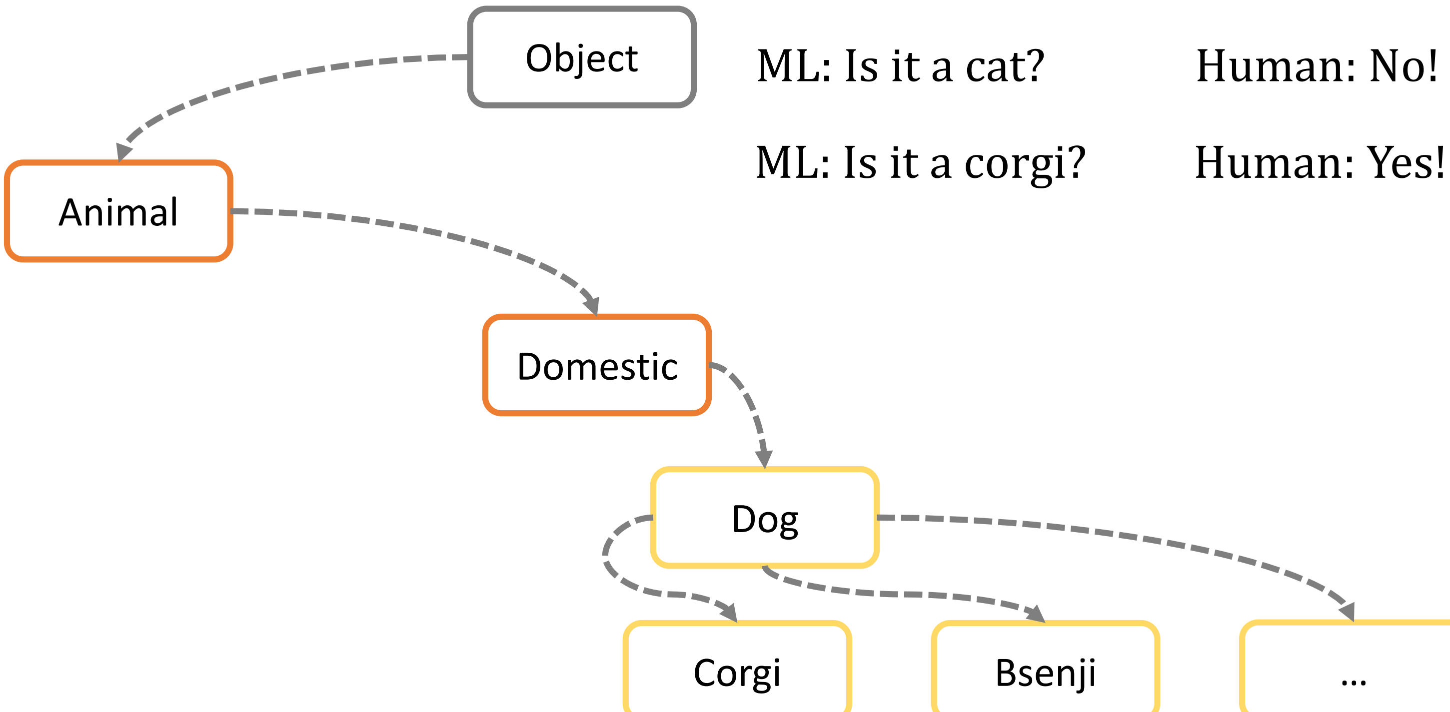
Introduction

ML: Is it a plant? Human: No!

ML: Is it domestic? Human: Yes!

ML: Is it a cat? Human: No!

ML: Is it a corgi? Human: Yes!



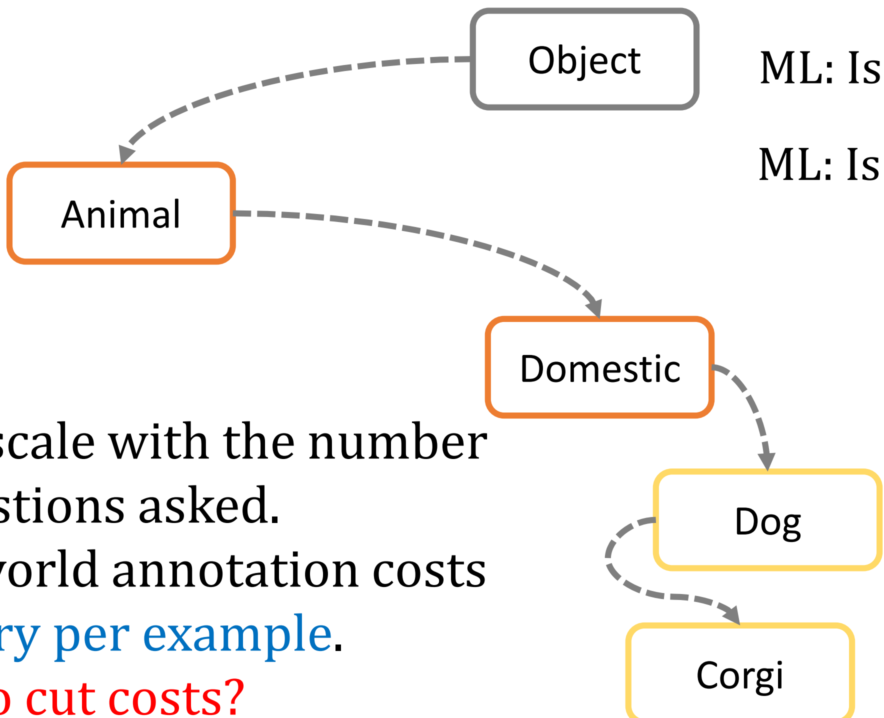
Introduction

ML: Is it a plant? Human: No!

ML: Is it domestic? Human: Yes!

ML: Is it a cat? Human: No!

ML: Is it a corgi? Human: Yes!

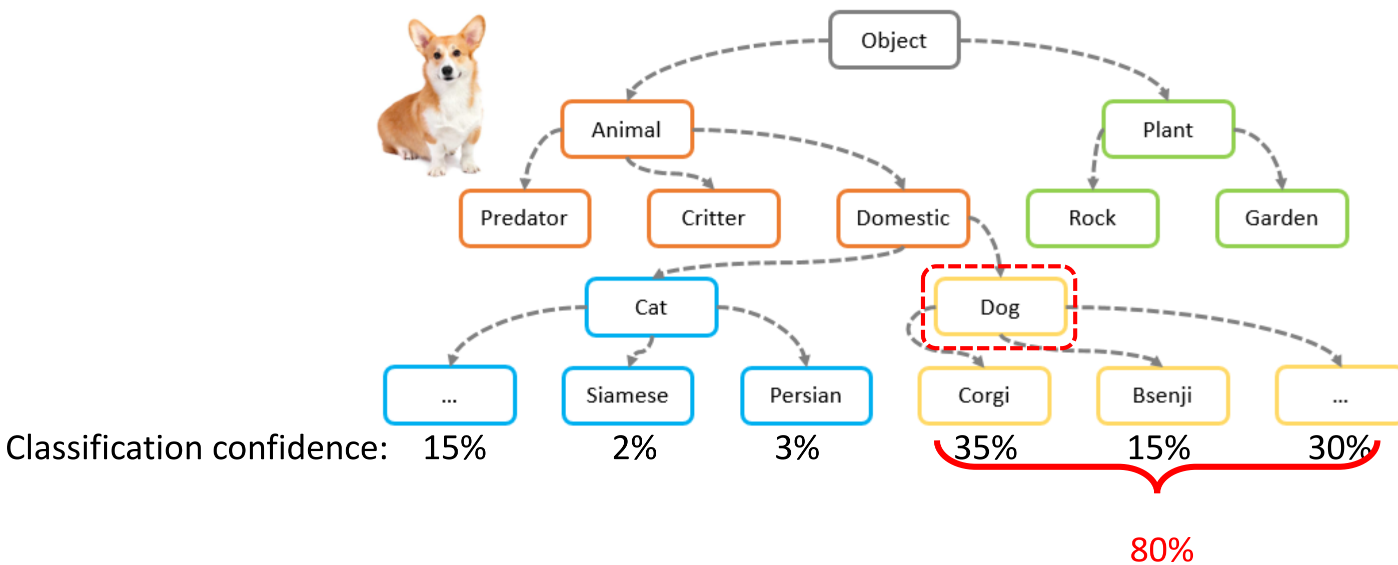


- Costs scale with the number of questions asked.
- Real-world annotation costs can vary per example.
- How to cut costs?

Introduction

- Why start at the top of the tree – "is this an artificial object?" – when we can cut costs by jumping straight to dog breeds?

  (i) Good strategies for choosing (example, class) pairs.



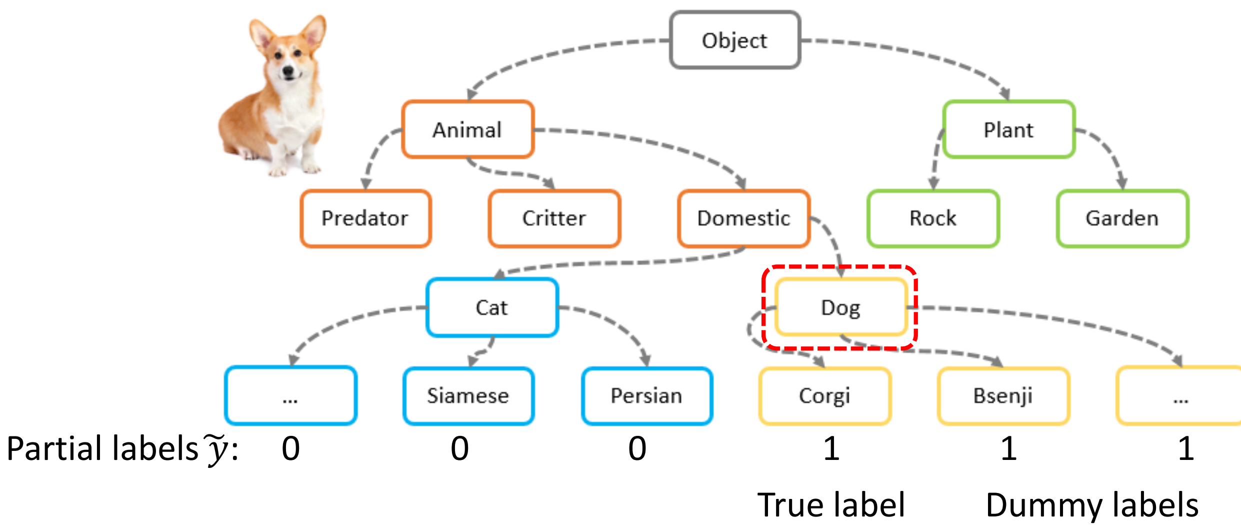
Introduction

- Why start at the top of the tree – "is this an artificial object?" – when we can cut costs by jumping straight to dog breeds?

  (i) Good strategies for choosing (example, class) pairs.

- Should we necessarily label each example to completion?

  (ii) Techniques for learning from the partially-labeled data that results when labeling examples to completion isn't required.



Contents

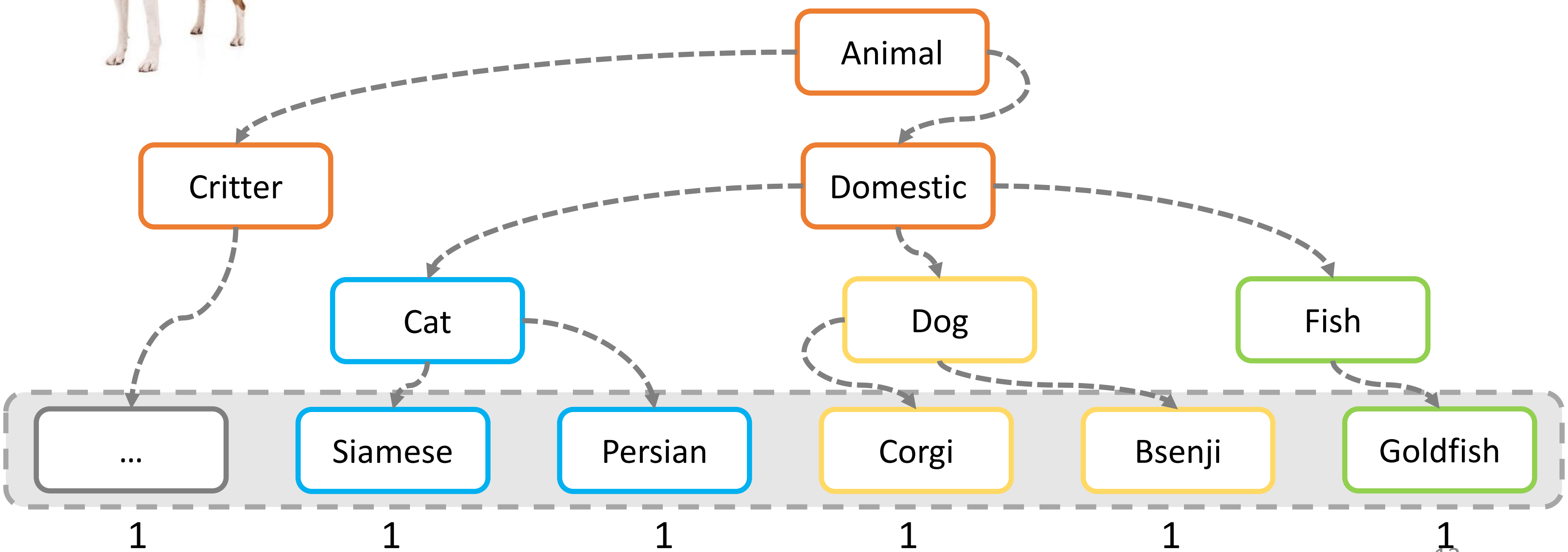
- Introduction
- Methods
- Experiments
- Conclusions

Methods

- Pick a question $q = (x_i, c_j)$ and ask the annotator, does x_i contain a c_j?

$$\tilde{y}^{(t+1)} = \begin{cases} \tilde{y}^{(t)} \setminus c & \text{if } \alpha = 0 \\ \tilde{y}^{(t)} \setminus \overline{c} & \text{if } \alpha = 1 \end{cases}$$

ML: Does it contain a cat? (c=cat) Human: No! (α=0)



Partial labels ($\tilde{y}^{(t)}$)

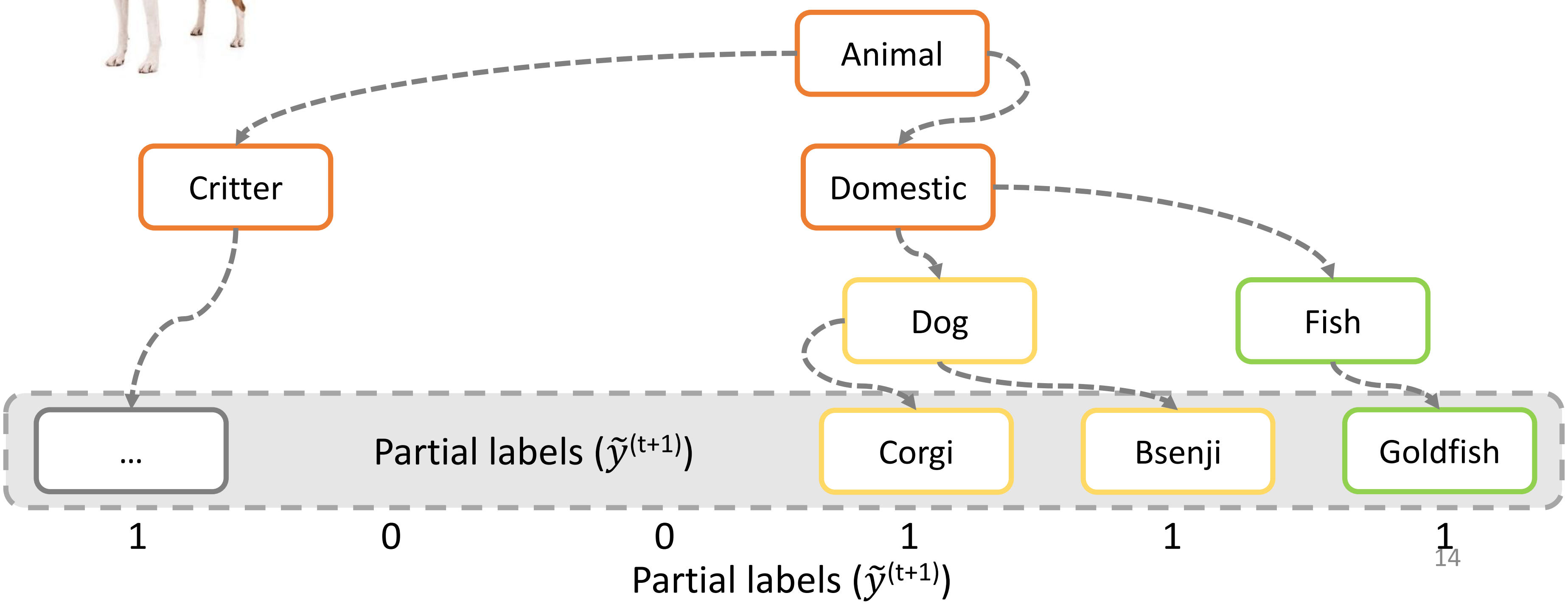
Methods

- Pick a question $q = (x_i, c_i)$ and ask the annotator, does x_i contain a c_j?

$$\tilde{y}^{(t+1)} = \begin{cases} \tilde{y}^{(t)} \setminus c & \text{if } \alpha = 0 \\ \tilde{y}^{(t)} \setminus \bar{c} & \text{if } \alpha = 1 \end{cases}$$

ML: Does it contain a cat? (c=cat)

Human: No! (α=0)

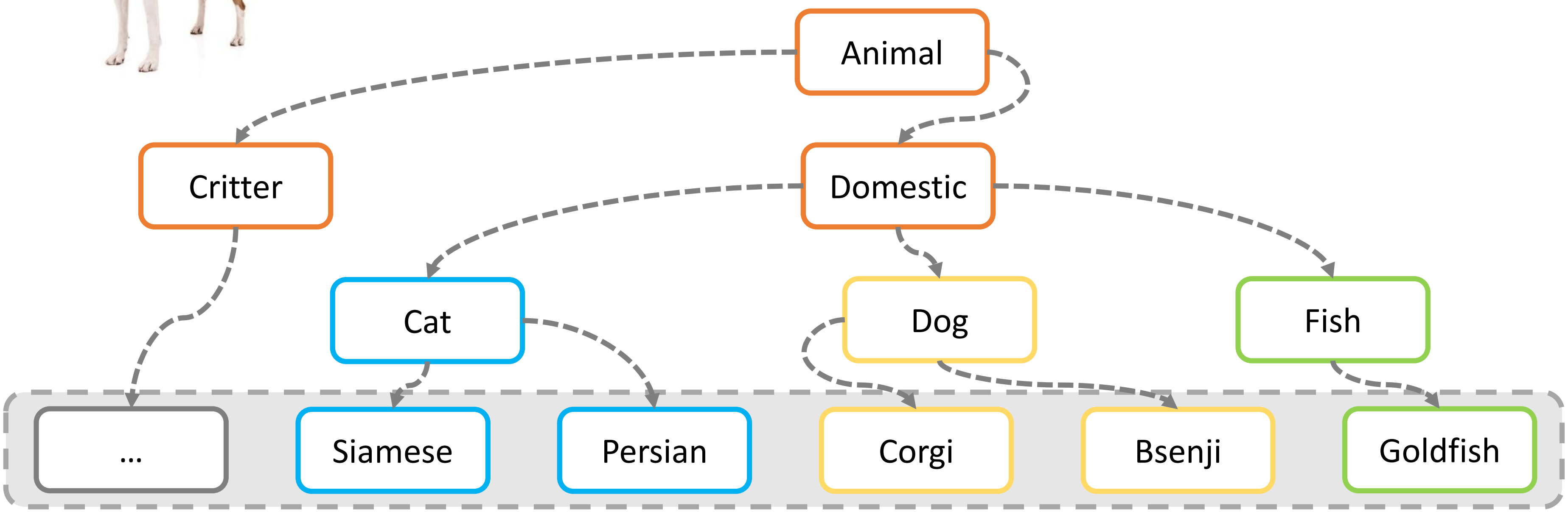


Methods

- Pick a question $q = (x_i, c_i)$ and ask the annotator, does x_i contain a c_j?

$$\tilde{y}^{(t+1)} = \begin{cases} \tilde{y}^{(t)} \setminus c & \text{if } \alpha = 0 \\ \tilde{y}^{(t)} \setminus \overline{c} & \text{if } \alpha = 1 \end{cases}$$

ML: Does it contain a dog? (c=dog)   Human: Yes! (α=1)



1 1 1 1 1 1

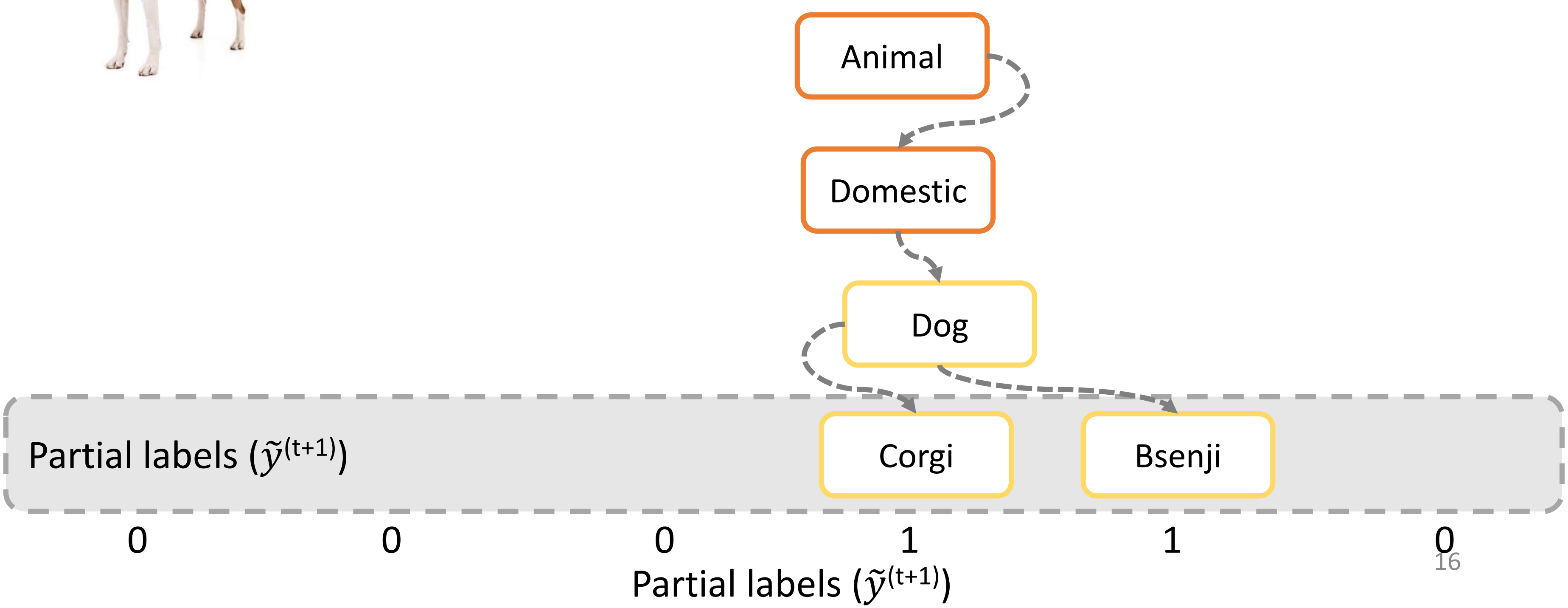
Partial labels ($\tilde{y}^{(t)}$)

Methods

- Pick a question $q = (x_i, c_i)$ and ask the annotator, does x_i contain a c_j?

$$\tilde{y}^{(t+1)} = \begin{cases} \tilde{y}^{(t)} \setminus c & \text{if } \alpha = 0 \\ \tilde{y}^{(t)} \setminus \overline{c} & \text{if } \alpha = 1 \end{cases}$$

ML: Does it contain a dog? (c=dog) Human: Yes! (α=1)



Methods

Learning Process

- At each round t, the learner selects a pair (x, c) for labeling
- After receiving binary feedback, the agent updates the corresponding partial label $\tilde{y}^{(t)} \rightarrow \tilde{y}^{(t+1)}$
- The agent then re-estimates its model, using all available partial labels and selects another question q.
- In batch-mode, the ALPF learner re-estimates its model once per T queries which is necessary when training is expensive (e.g. deep learning).

Algorithm 1 Active Learning with Partial Feedback

Input: $\mathbf{X} \leftarrow (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $\mathbf{Q} \leftarrow (\mathbf{q}_1, \ldots, \mathbf{q}_M)$, K, T.
Input: $\mathcal{D} \leftarrow [\boldsymbol{x}_i]_{i=1}^{N}$, $\mathcal{C} \leftarrow [c_j]_{j=1}^{M}$, k, T
Initialize: $\tilde{y}_i^{(0)} \leftarrow \{1, \ldots, k\}$, $\theta \leftarrow \theta^{(0)}$, $t \leftarrow 0$
repeat
  Score every $(\boldsymbol{x}_i, c_j)$ with θ
  repeat
    Select $(\boldsymbol{x}_{i^*}, c_{j^*})$ with the best score
    Query c_{j^*} on data $\boldsymbol{x}_{i^*}$
    Receive feedback α
    Update $\tilde{y}_{i^*}^{(t+1)}$ according to α
    $t \leftarrow t + 1$
  until $(t \bmod T = 0)$ or $(\forall i, |\tilde{y}_i^{(t)}| = 1)$
  $\theta \leftarrow \arg\min_\theta \mathcal{L}(\theta)$
until $\forall i, |\tilde{y}_i^{(t)}| = 1$ or t exhausts budget

Methods

Learning from Partial Labels

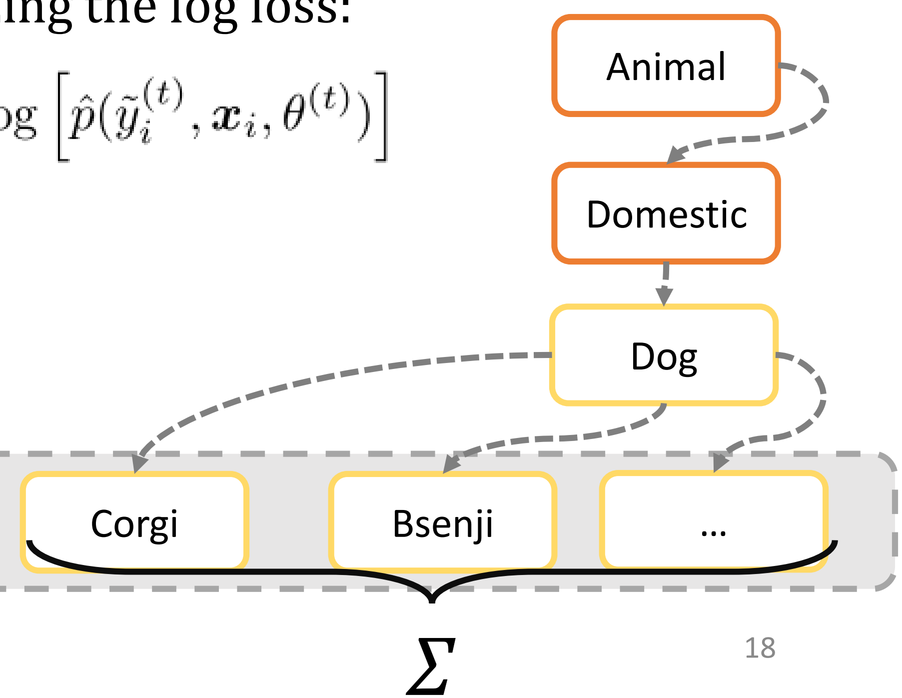
- The probability assigned to a partial label $\tilde{y}$ can be expressed by marginalizing over the atomic classes that it contains:

$$\hat{p}(\tilde{y}^{(t)}, \boldsymbol{x}, \theta^{(t)}) = \sum\nolimits_{y \in \tilde{y}^{(t)}} \hat{y}(y, \boldsymbol{x}, \theta^{(t)})$$

- We optimize our model by minimizing the log loss:

$$\mathcal{L}(\theta^{(t+1)}) = -\frac{1}{n} \sum_{i=1}^{n} \log \left[\hat{p}(\tilde{y}_i^{(t)}, \boldsymbol{x}_i, \theta^{(t)}) \right]$$

Partial labels ($\tilde{y}^{(t)}$)



Methods

Query Strategies

- **Expected Information Gain (EIG):** In our case, each answer to the query yields a different partial label.
- The notation $\hat{y}_0$, and $\hat{y}_1$ denote consequent predictive distributions for each answer (no or yes).
- Generalizing maximum entropy to ALPF by selecting questions with greatest expected reduction in entropy.

$$\arg\max \underbrace{EIG_{(\boldsymbol{x},c)}}_{\text{reduction entropy}} = \underbrace{S(\hat{\boldsymbol{y}})}_{\text{current entropy}} - \underbrace{[\hat{p}(c,\boldsymbol{x},\theta)S(\hat{\boldsymbol{y}}_1) + (1-\hat{p}(c,\boldsymbol{x},\theta))S(\hat{\boldsymbol{y}}_0)]}_{\text{expected entropy}}$$

Methods

Query Strategies

- **Expected Remaining Classes (ERC):** It's a heuristic strategy that suggests arriving as quickly as possible at exactly-labeled examples.
- At each round, ERC selects those examples for which the expected number of remaining classes is fewest:

$$\operatorname{argmin} \ \underbrace{ERC_{(\boldsymbol{x},c)} = \hat{p}(c,\boldsymbol{x},\theta)||\hat{\boldsymbol{y}}_1||_0 + (1-\hat{p}(c,\boldsymbol{x},\theta))||\hat{\boldsymbol{y}}_0||_0}_{\text{expected remaining classes}}$$

- **Expected Decrease in Classes (EDC):** The strategy which we expect to result in the greatest reduction in the number of potential classes.

$$\operatorname{argmax} \ \underbrace{EDC_{(\boldsymbol{x},c)}}_{\text{expected decrease}} = \underbrace{|\tilde{y}^{(t)}|}_{\text{current partial label}} - \underbrace{ERC_{(\boldsymbol{x},c)}}_{\text{expected remaining classes}}$$

Contents

- Introduction
- Methods
- Experiments
- Conclusions

Experiments

Learning from Partial Labels

- Train a standard multi-class classifier with γ(%) exactly labeled training.
- Train another classifier with the remaining (1-γ)% partially labeled at a different granularity(level of hierarchy).

Key Observations

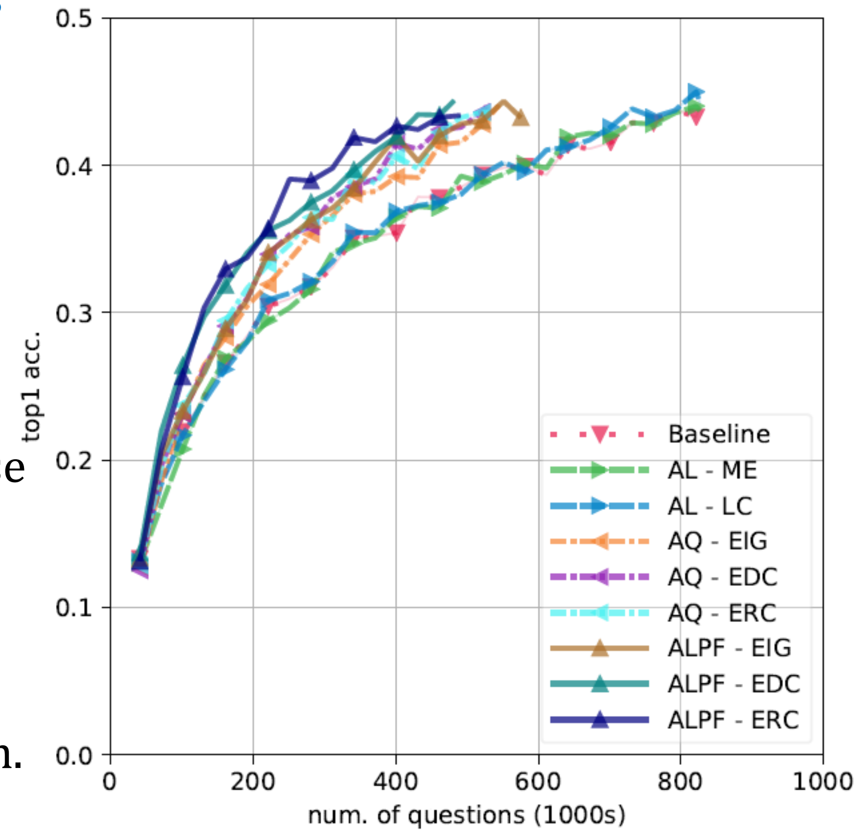
- Additional coarse-grained partial labels improve model accuracy
- As expected, the improvement diminishes as partial label gets coarser.

Table 1: Learning from partial labels on Tiny ImageNet. These results demonstrate the usefulness of our training scheme absent the additional complications due to ALPF. In each row, γ% of examples are assigned labels at the *atomic class* (Level 0). Levels 1, 2, and 4 denote progressively coarser composite labels tracing through the WordNet hierarchy.

$\gamma(\%)$	γ	$(1-\gamma)$		
	Level 0	Level 1	Level 2	Level 4
20	0.285	**+0.113**	+0.086	+0.025
40	0.351	+0.079	+0.056	+0.016
60	0.391	+0.051	+0.036	+0.018
80	0.432	+0.015	+0.017	-0.009
100	0.441	-	-	-

Experiments

- **Baseline**: This learner samples examples at random. Once an example is sampled, the learner choosing the question that most evenly splits the probability mass until that example is exactly labeled.
- **AL**: Selecting examples with uncertainty sampling but selecting questions as baseline.
- **AQ**: Choosing examples at random but use partial feedback strategies, moving on to the next example after finding an example's exact label.
- **ALPF**: ALPF learners are free to choose any (example, question) pair at each turn.



Tiny ImageNet: 100k examples, 200 classes, |C|=304

Experiments

	Annotation Budget (w.r.t. baseline labeling cost)						Labeling Cost
	10%	20%	30%	40%	50%	100%	
TinyImageNet							
Baseline	0.186	0.266	0.310	0.351	0.354	0.441	827k
AL - ME	0.169	0.269	0.303	0.347	0.365	-	827k
AL - LC	0.184	0.262	0.313	0.355	0.369	-	827k
AQ - EIG	0.186	0.283	0.336	0.381	0.393	-	545k
AQ - EDC	0.196	0.291	0.353	0.386	0.415	-	530k
AQ - ERC	0.194	0.295	0.346	0.394	0.406	-	531k
ALPF - EIG	0.203	0.289	0.351	0.384	0.420	-	575k
ALPF - EDC	**0.220**	0.319	0.363	0.397	0.420	-	482k
ALPF - ERC	0.207	**0.330**	**0.391**	**0.419**	**0.427**	-	491k

- Vanilla active learning does not improve over i.i.d. baselines.
- AQ provides a dramatic improvement over baseline. The advantage persists throughout training. These learners sample examples randomly and label to completion (until an exact label is produced) before moving on, differing only in how efficiently they annotate data.
- On Tiny ImageNet, at 30% of budget, ALPF-ERC outperforms AQ methods by 4.5% and outperforms the i.i.d. baseline by 8.1%.

Contents

- Introduction
- Methods
- Experiments
- Conclusions

Conclusions

Reviewers' Opinions

- Reviewer 1: The way of solving both the learning from partial labels and the sampling strategies are not particularly insightful. Also, there is a lack of theoretical guarantees to show value of a partial label as compared to the true label. However, as these are not the main points of the paper (introduction of a novel learning setting), I see these as minor concerns. (**Rating: 7: Good paper, accept**)
- Reviewer 2: My main concern about this work is the lack of theoretical guarantees, which is usually important for active learning paper. it's better to provide some analysis on the efficiency of ALPF to further improve the quality of the paper. (**Rating: 6: Marginally above acceptance threshold**)