Interpretable Machine Teaching via Feature Feedback

Shihan Su Yuxin Chen Oisin Mac Aodha Pietro Perona Yisong Yue

Caltech

1 Summary

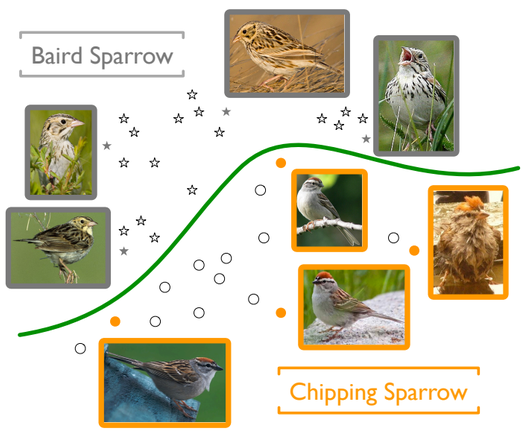
Motivating Applications

Citizen science, crowdsourcing services, medical diagnosis

Illustrative Example

Field Guide for Naturalists

e.g., teaching a novice to classify bird images

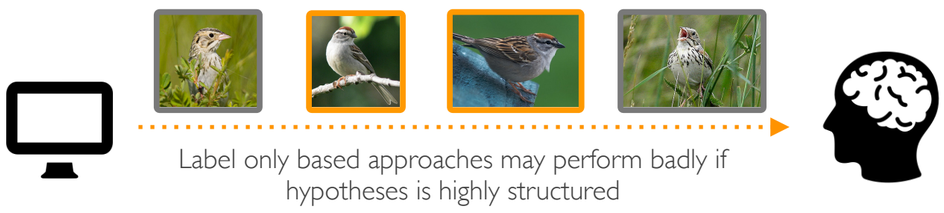


- ☐ Label for Baird Sparrow
- ☐ Label for Chipping Sparrow

Existing Work

Label-based Machine Teaching

[Goldman & Kearns 1995; Zhu 2013; Singla et al. 2014; Johns et al. 2015; Chen et al. 2016; Liu et al. 2017; …]



Label only based approaches may perform badly if hypotheses is highly structured

Our Approach

Explanation-based Machine Teaching

Interpretable Machine Teaching with **Feature Feedback**

Student's ability to learn a new concept can be greatly improved by providing them with clear and interpretable **explanations** from a knowledgeable teacher

2 Machine Teaching via Feature Feedback

Problem Formulation

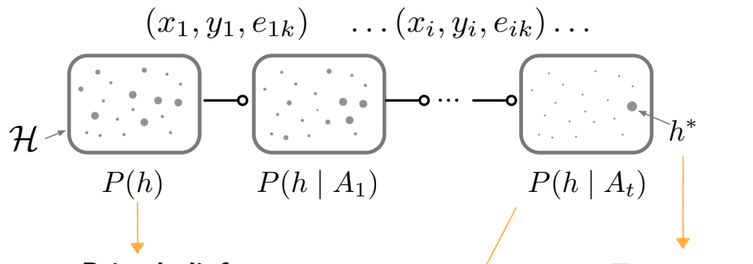
Teaching image set: $\{(x, y, e)\}_{1:m}$
- x_i image
- y_i label for x_i
- e_{ik} importance of f_{ik} selected from up to n interpretable features ($f_{i1}, \ldots, f_{in}$)
- $\boldsymbol{A_t}$ *selected teaching set up to t*

Hypothesis space: H

Each $h \in H$: $x_i \rightarrow y_i$
It also tells how important each feature is in predicting its label

Bayesian model of the learners

Learner's Belief is a prob. distribution over H
It is updated upon receiving teaching image



Prior belief of learners' hypotheses

Target hypothesis

Learner's posterior at t

$$P(h \mid A_t) = \frac{1}{Z_t} \cdot P(h) \cdot \underbrace{\prod_{\substack{x_i \in A_t \\ y_i \neq \mathrm{sgn}(h(x_i))}} P(y_i \mid h, x_i) \cdot \prod_{e_{ik} \in A_t} P(e_{ik} \mid h, x_i)}_{V(h \mid A_t)\text{: } \textbf{Unnormalized} \text{ posterior at } t}.$$

likelihood of teaching labels — *likelihood of explanations*

Normalization factor

Likelihood of observations

$$P(y_i \mid h, x_i) = \mathrm{expit}(-\alpha h(x_i) y_i)$$
$$P(e_{ik} \mid h, x_i) = \exp(-\beta w_k |e_{ik} - \mathrm{imp}(h, f_{ik})|)$$

h's estimate of the importance of feature k for image i

Learner's progress, measured by the expected classification error
$\mathbb{E}[\mathrm{err}(h) \mid A_t] = \sum_h P(h \mid A_t)\, \mathrm{err}(h)$

Our goal is to find the optimal teaching set achieving expected error at most ϵ

$$A^* = \underset{A \in T}{\mathrm{argmin}} |A|, \ \text{s.t.} \ \mathbb{E}[\mathrm{err}(h) \mid A] \leq \epsilon$$

The Greedy Teaching Algorithm

Input Teaching image set $\{(x, y, e)\}_{1:m}$; hypotheses H; likelihood params α, β; prior $P(h)$; tolerance ϵ

Output Selected images to teach, A

Start $A \leftarrow \varnothing$;

Loop $i^*, k^* = \mathrm{argmin}_{i,k} \sum_h V(h \mid A \cup \{(x_i, y_i, e_{ik})\})\, \mathrm{err}(h)$;
$A \leftarrow A \cup \{(x_{i^*}, y_{i^*}, e_{i^*k^*})\}$;

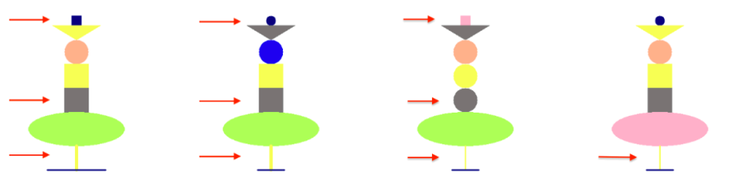
Until $\sum_h V(h \mid A)\, \mathrm{err}(h) \leq P(h^*)\epsilon$

Thm: The worst-case cost of our greedy strategy achieving error ϵ is within a logarithmic factor of the worst-case cost of the optimal algorithm achieving error of at least $P(h^*)\epsilon/2$

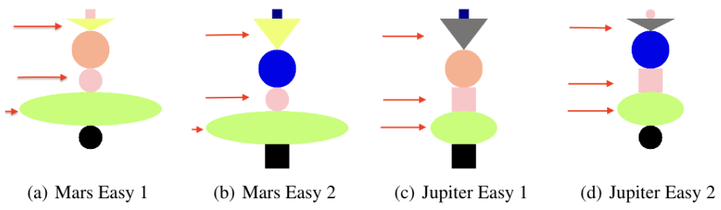
3 Results

Datasets

- Real human participants tested on two synthetic datasets: hard and easy
- Each with 128 images of two balanced classes: Jupiter and Mars

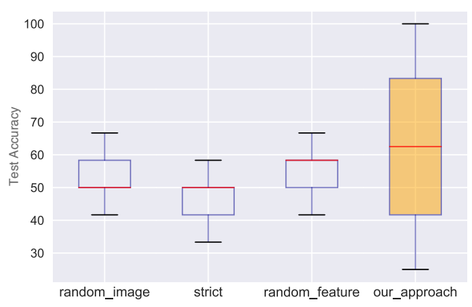


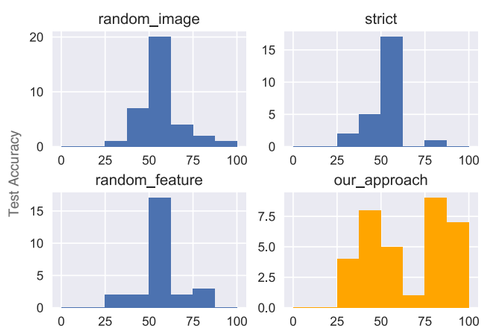
(a) Mars Hard 1 (b) Mars Hard 2 (c) Jupiter Hard 1 (d) Jupiter Hard 2

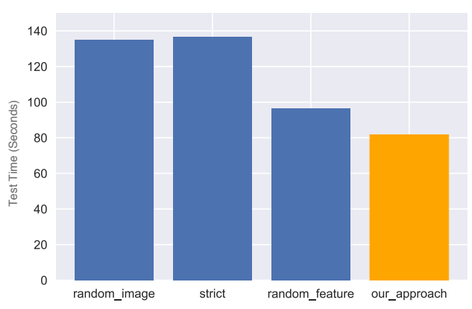


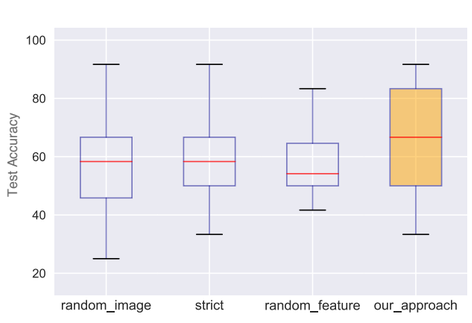
(a) Mars Easy 1 (b) Mars Easy 2 (c) Jupiter Easy 1 (d) Jupiter Easy 2

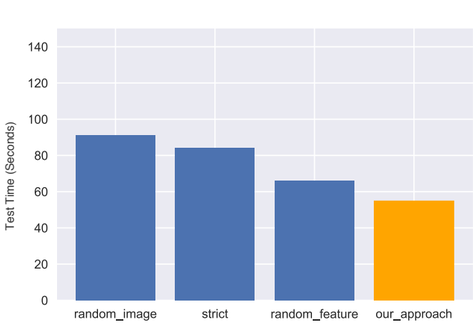
Results











Decision Rule

- Predictive features: parts indicated by arrows
- Hard Mars: blue top, square middle and thick base
  Jupiter: any other combination
- Easy Mars: yellow top, circle middle and large ellipse near bottom
  Jupiter: any other combination

Baseline

Random:	Random images with no explanations
Random-feature:	Random images with random explanations
STRICT:	Label-based greedy approach

Conclusion

With explanation based machine teaching, students achieve
- Better accuracy
- Faster question answering at test time