# Near-Optimal Machine Teaching via Explanatory Teaching Sets

Yuxin Chen  Oisin Mac Aodha  Shihan Su  Pietro Perona  Yisong Yue

Caltech

## 1 Summary

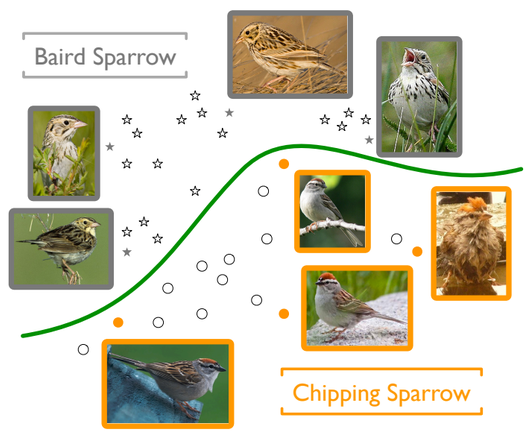
### Motivating Applications

Citizen science, crowdsourcing services, medical diagnosis

**Illustrative Example**

*Field Guide for Naturalists*

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

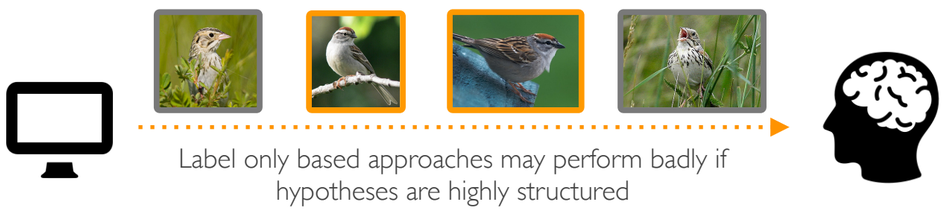



- [ ] Label for Baird Sparrow
- [ ] Label for Chipping Sparrow

### Existing Work

**Label-based** Machine Teaching

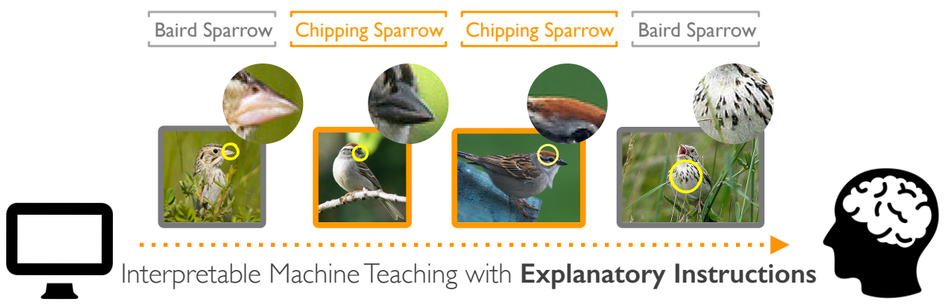
[Goldman & Kearns 1995; Zhu 2013; Singla et al. 2014; Johns et al. 2015; Chen et al. 2016; Liu et al. 2017; …]



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

### Our Approach

**Explanation-based** Machine Teaching




Interpretable Machine Teaching with **Explanatory Instructions**

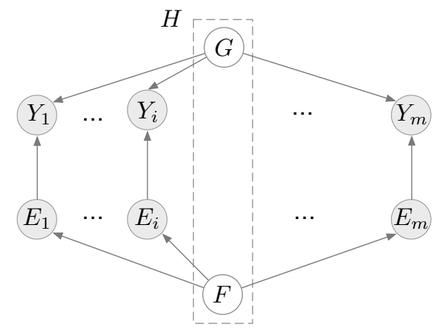
Student's ability to learn a new concept can be greatly improved by providing them with clear and interpretable **explanations** from a knowledgeable teacher

## 2 Explanation-based Machine Teaching

### Problem Formulation

**Teaching set** $\{(x, y, \mathbf{e})\}_{1:m}$

- $x_i \in \mathfrak{X}$ image
- $y_i \in \mathfrak{Y}$ label for $x_i$
- $\mathbf{e}_i \in \mathbb{B}^k$ explanation for $(x_i, y_i)$: $k$-d Boolean vector, each $e_{ij}$ indicates the importance of the $j^{th}$ attributes $a_{ij} \in \mathbb{A}$.
- $\boldsymbol{S}$ indices of selected teaching examples




**Attention model** $F$

Each $f \in \mathfrak{F}$: $\mathfrak{X} \to \mathbb{B}^k$.

**Decision model** $G$

Each $g \in \mathfrak{G}$: $(\mathbb{A} \times \mathbb{B})^k \to \mathfrak{Y}$.

**Hypothesis** $H$

Each $h = g \circ f$: $\mathfrak{X} \to \mathfrak{Y}$ is a composite of the attention function and decision function

**Bayesian model of the learners**

$P(f), P(g)$

$P(\mathbf{e}_i \mid f, x_i) = \nu_i^f$ if $f(x_i) \neq \mathbf{e}_i$

$P(y_i \mid \mathbf{e}_i, g, x_i) = \nu_i^g$ if $g \circ f^*(\mathbf{e}_i) \neq y_i$

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

**Our goal** is to find the optimal teaching set achieving expected error $\epsilon$ : $\mathcal{S}^* = \underset{\mathcal{S} \subseteq \{1,\ldots,m\}}{\mathrm{argmin}} |\mathcal{S}|,\ \text{s.t. } \mathbb{E}[\mathrm{err}(h) \mid \mathbf{e}_{\mathcal{S}}, y_{\mathcal{S}}] \leq \epsilon$

### The NOTES Algorithm

**Quantifying uncertainty**

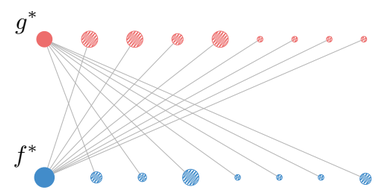
For each pair of attention/decision functions, we define edge weight

initial edge weight; error induced by following attention f; error induced by making decision g

$$w(\{f,g\}) = \mathbb{P}(f) \cdot \mathbb{P}(g) \cdot (\mathrm{err}(g^* \circ f) + \mathrm{err}(g \circ f^*))$$

$$w(\{f,g\} \mid \mathcal{S}) = w(\{f,g\}) \prod_{i \in \mathcal{S}} \mathbb{P}(\mathbf{e}_i, y_i \mid f, g)$$

edge weight after observing $S$




Bipartite graph constructed by NOTES.
Size of nodes represents prior prob.
Edges are drawn between $f^*$ (resp. $g^*$) and all $g$'s (resp. all $f$'s).

**Surrogate objective function:**

remaining weight of the bipartite graph upon observing set $S$:

$$r(\mathcal{S}) = \sum_{\{f,g\} \in \mathcal{E}} w(\{f,g\} \mid \mathcal{S})$$

**Thm** The worst-case cost of NOTES achieving error $\epsilon$ is within a logarithmic factor of the worst-case cost of the optimal algorithm achieving error of at least $P(f^*)P(g^*)\epsilon/2$

**Near-Optimal Teaching via Explanatory Sets (NOTES)**

**Input** Teaching image set $\{(x, y, e)\}_{1:m}$; hyp. $\{h(x):=g \circ f(x)\}$; noise params $\{\nu^f, \nu^g\}_{1:m}$; prior $P(f)$, $P(g)$; tolerance $\epsilon$

**Output** Selected images to teach, $S$

---

**Start** $S \leftarrow \varnothing$;

**Loop** $i^* = \mathrm{argmin}_i\, r(S \cup \{i\})$;

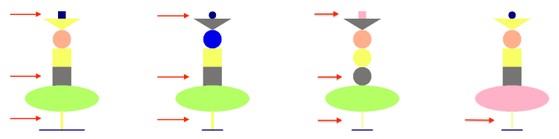
$S \leftarrow S \cup \{i\}$;

**Until** $r(S) \leq P(f^*)P(g^*)\,\epsilon$

## 3 Experimental Results

### Datasets

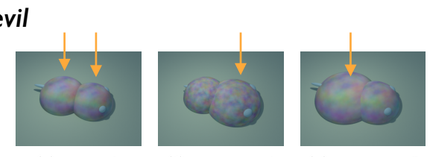
***Jupiter vs Mars***



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

Objects from Mars have blue top, square middle and thick base
Objects with any other combination belong to Jupiter

***Vespula vs Weevil***

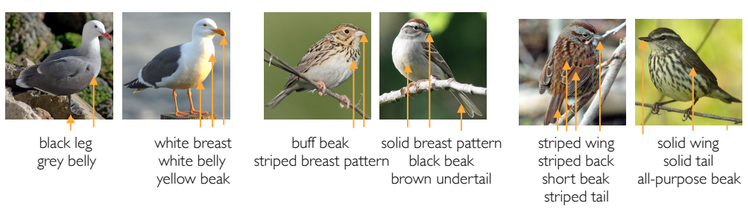


(e) Weevil (f) Vespula 1 (g) Vespula 2

Weevil has a mid-sized body and head, while Vespula does not.

***Caltech-UCSD Birds (CUB) dataset***

- CUB-1: Western Gull vs Heermann Gull
- CUB-2: Baird Sparrow vs Chipping Sparrow
- CUB-3: Song Sparrow vs Northern Waterthrush

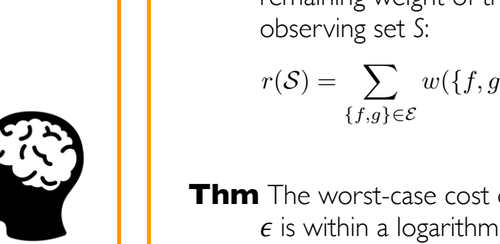


black leg, grey belly | white breast, white belly, yellow beak | buff beak, striped breast pattern | solid breast pattern, black beak, brown undertail | striped wing, striped back, short beak, striped tail | solid wing, solid tail, all-purpose beak

### Explaining Teaching Examples

LIME Explainer (Ribeiro et al., 2016)

LIME aims to reproduce the predictive results of h* in the vicinity of the input via a sparse linear classifier



"X" is classified as the blue class, because its horizon axis is less than the threshold marked by the red line

***Likelihood functions for inconsistent explanations***

learner's ability to adapt to explanations; attention function associated with the linear function $\ell_{(h,x)}$

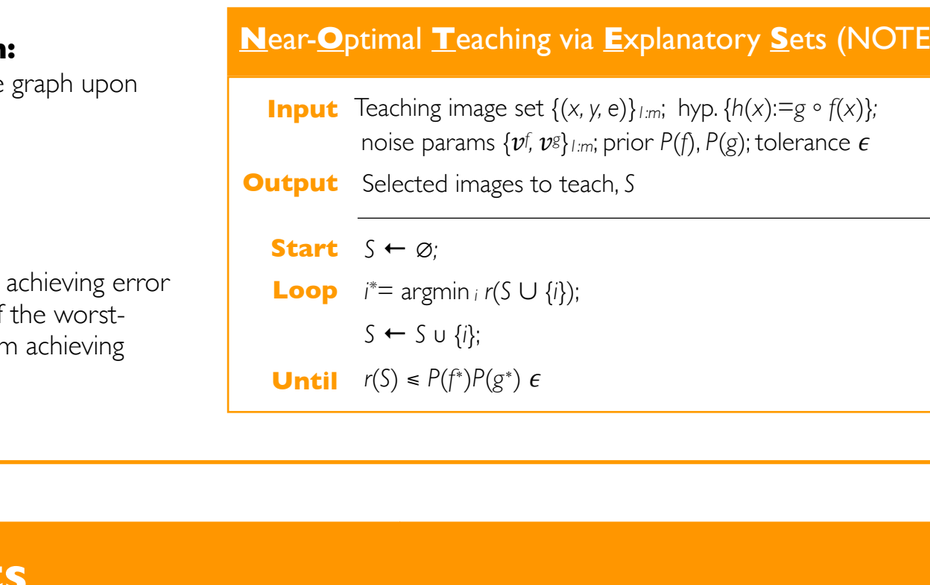
$$\nu_i^f = 2\,\mathrm{expit}\left(\alpha\left(\left|\frac{f_\ell(x_i) \wedge \mathbf{e}_i}{f_\ell(x_i) \vee \mathbf{e}_i}\right| - 1\right)\right)$$

$$\nu_i^g = \mathrm{expit}\left(\beta \ell_{(h,x)}(x) y_i\right)$$

learner's ability to adapt to labels; $\ell_{(h,x)}$ is the parse linear approximation of h at example x

### Results

simulated and real-human learners




… more detailed results (including CUB-2/CUB-3) in paper

***Baselines***:

random: Rand. examples with no explanations
randexp: Rand. examples with rand. explanations
STRICT: Label-based greedy approach

With explanation-based machine teaching, learners achieve

- Better accuracy
- Faster question answering at test time