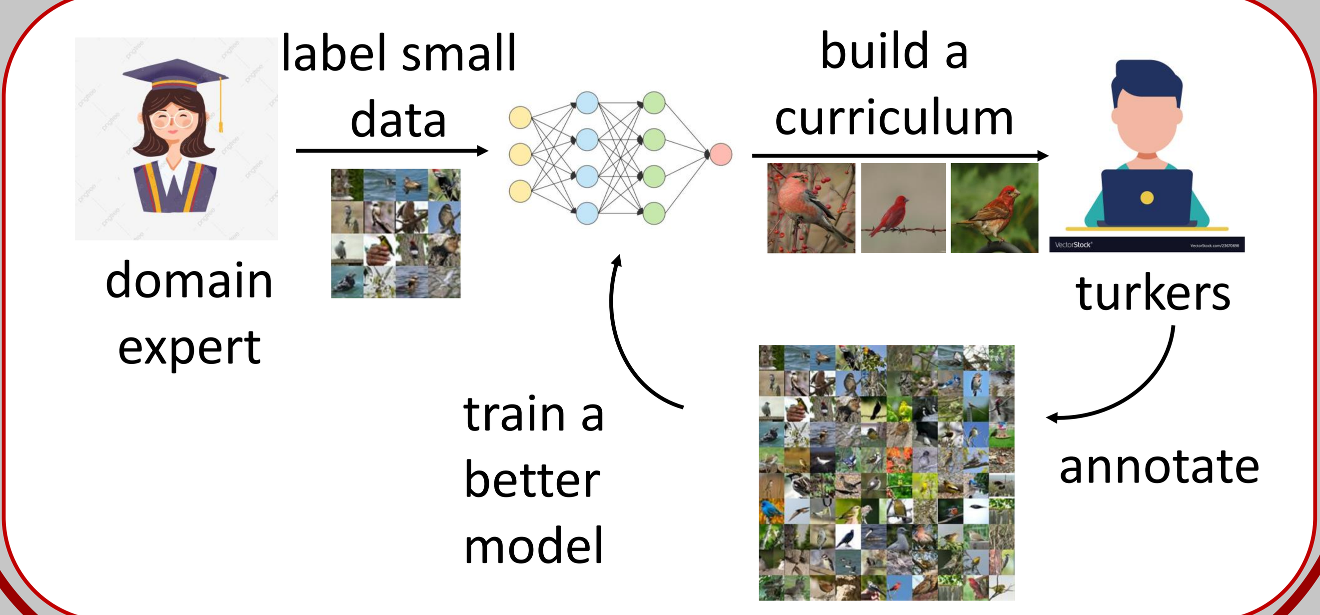
Gradient-based Algorithms for Machine Teaching

Pei Wang   Kabir Nagrecha   Nuno Vasconcelos

SVCL, ECE, University of California, San Diego

Motivation

- Although crowd-sourcing can scalably annotate everyday objects, actions, or scenes data, It is hard to do it on fine-grained expert domain, because annotations require highly specialized and domain specific knowledge.
- Annotation by specialists is usually too expensive and rarely feasible at a large scale.
- Less label-intensive forms of learning, including few-shot learning, transfer learning, semi-supervised learning and self-supervised learning, still underperform supervised learning.
- We use machine teaching algorithms to train crowdsource annotators to label data from specialized domains and make scalable supervised learning possible.

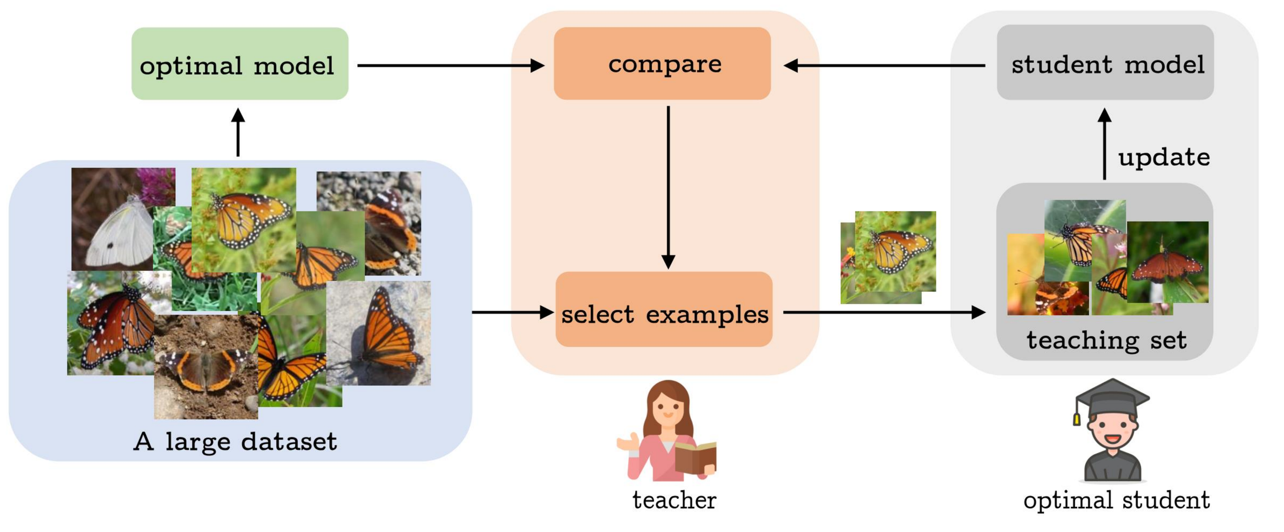


Preliminaries

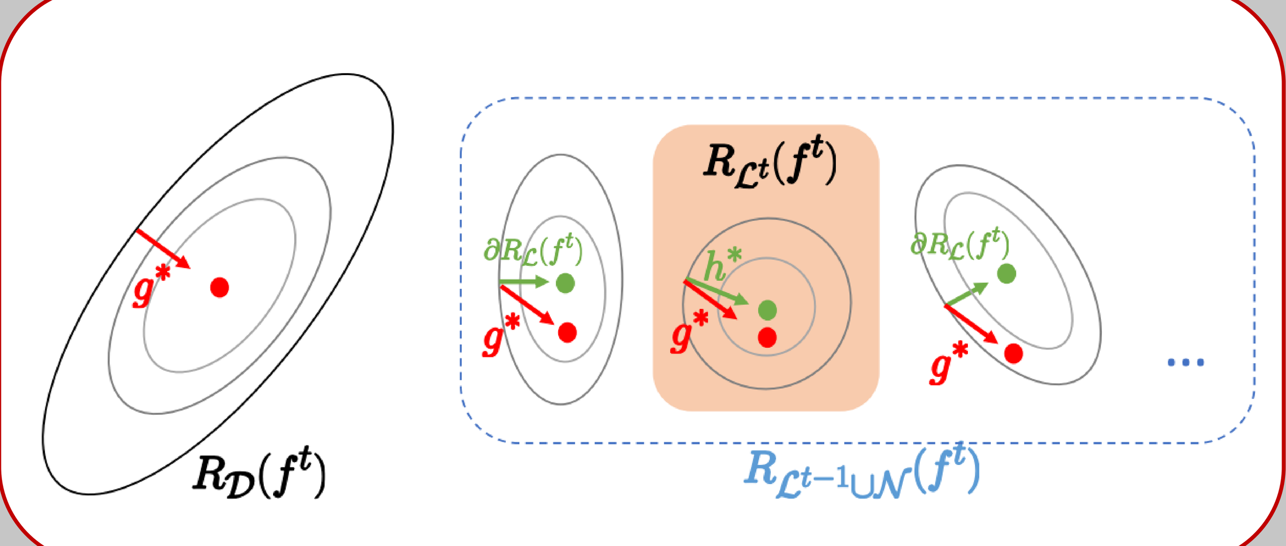
- **Machine teaching**
  - It is assumed that the teacher can access to a much larger example dataset $\mathcal{D} = \{(x_1, y_1), \dots (x_N, y_N)\}$
  - From this dataset, the teacher assembles a small ordered subset $\mathcal{L} = \{(x_1^l, y_1^l), \dots (x_K^l, y_K^l)\}$
  - The student learns on the selected $\mathcal{L} \subset \mathcal{D}$ and tries to achieve the best trade-off between
    - Learning the optimal predictor f^* for $\mathcal{D}$
    - Spending the least amount of effort, usually measured by the cardinality of $\mathcal{L}$
- **Sub-optimal student assumption**
  - The student has limited capacity and memory.

MaxGrad

- **Optimal student assumption**
  - Mainly focus on crowd sourcing context;
  - The teaching set must be small;
  - Humans are good at few-shot learning scenery;
  - free-willing participants rated by their performance;
- **Iterative machine teaching**



- **Select new examples by MaxGrad**



- At iteration t, the teacher has access to the population risk $\mathcal{R}_{\mathcal{D}}(f^t)$ and corresponding steepest descent direction;
- The student can only learn from the teaching set $\mathcal{L}^{t-1}$ of iteration $t-1$ and newly selected examples $\mathcal{N}$;
- MaxGrad selects N so that the steepest descent direction on $\mathcal{L}^t = \mathcal{L}^{t-1} \cup \mathcal{N}$ is closet to g^*.

Algorithm 1 MaxGrad

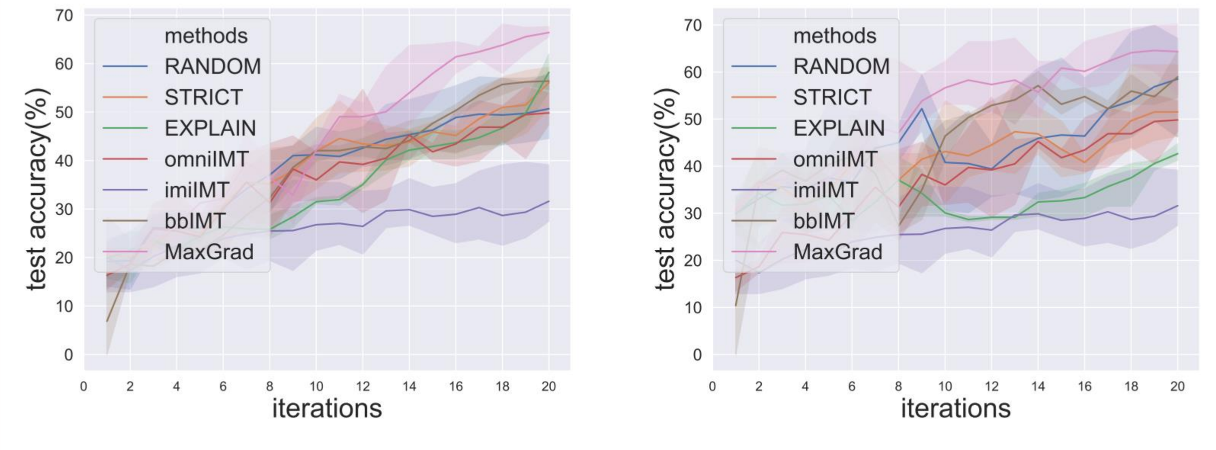
Input Data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, codewords $\mathcal{Y}$, max iter. T, effort τ.

1: **Initialization:** $\mathcal{L}^0 \leftarrow \emptyset$, f^1, $\mathcal{D}^0 \leftarrow \mathcal{D}$.
2: **for** $t = \{1, \dots, T\}$ **do**
3: compute ξ_i for all examples in $\mathcal{D}^{t-1}$.
4: order examples by decreasing ξ_i and select top τ to create $\mathcal{N}^t$.
5: teaching set update: $\mathcal{L}^t \leftarrow \mathcal{L}^{t-1} \cup \mathcal{N}^t$
6: student update: $f^{t+1} = f^*(\mathcal{L}^t)$.
7: $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \setminus \mathcal{N}^t$
8: **end for**

Output $\mathcal{L}^t$

Experiment results

- **On the simulated learners**

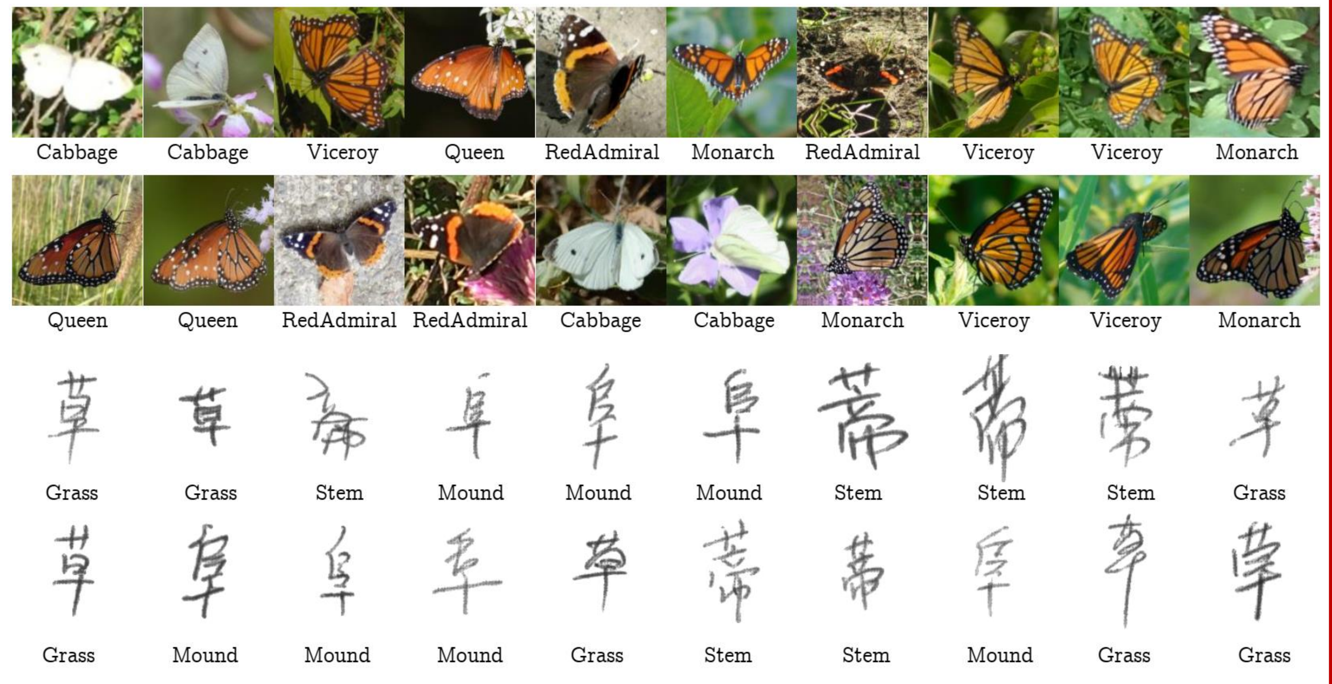


(a) butterflies   (b) Chinese Characters

- **On the real learners**

	Butterflies	Chinese Char.
RANDOM	65.20	47.05
STRICT	65.00	51.51
EXPLAIN	68.33	65.44
omniIMT*	70.07(18.30)	64.36(19.58)
imiIMT*	72.70(17.63)	64.46(23.72)
bbIMT*	76.09(18.05)	64.37(19.57)
RANDOM*	63.15(18.17)	51.53(24.47)
MaxGrad	**80.33**(19.76)	**81.89**(12.93)

- **Selected examples**



Reference

1. Adish Singla, Ilija Bogunovic, Gabor Bartok, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify, *ICML* 2014
2. Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona,and Yisong Yue. Teaching categories to human learners with visual explanations, CVPR 2018
3. Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching, ICML 2017
4. Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James M Rehg, and Le Song. Towards black-box iterative machine teaching, ICML 2018

CVPR VIRTUAL JUNE 19-25

SVCL UCSD