Gradient-based Algorithms for Machine Teaching
Supplementary Materials

Pei Wang
UC, San Diego
pew062@ucsd.edu

Kabir Nagrecha
UC, San Diego
kabir.nagrecha@gmail.com

Nuno Vasconcelos
UC, San Diego
nuno@ucsd.edu

A. Appendix

A.1. Proof of Corollary 1

Proof Under the optimal student assumption, the predictor learned by the student at iteration t is

$$f^t = \arg\min_f \mathcal{R}_{\mathcal{L}^t}[f] = \arg\min_f \sum_{(x_i,y_i)\in\mathcal{L}^t} \phi(y_i f(x_i)). \tag{30}$$

If the teacher selects at least one new example per iteration, $\mathcal{L}^t$ increases with t, i.e. $\mathcal{L}^{t-1} \subset \mathcal{L}^t$. Since $\mathcal{D}$ has finite size n, $\exists k \leq n$ s.t. $\mathcal{L}^k = \mathcal{D}$. It follows that, if $\zeta \geq |\mathcal{D}|$, the student will eventually learn from $\mathcal{L}^k$. From (30) and (1) it follows that $f^k = f^*$. ∎

A.2. Proof of Lemma 1

Proof Assume without loss of generality that $\mathcal{A} = \{(x_1, y_1), \ldots (x_m, y_m)\}$ and $\mathcal{B} = \{(x_{m+1}, y_{m+1}), \ldots (x_n, y_n)\}$ for any $1 < m < n$. Then, it follows from (10) that

$$\nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f) = (w_1, \ldots, w_m, w_{m+1}, \ldots, w_n)^T \tag{31}$$

$$= \left(\nabla^T_{\Psi(\mathcal{A})} R_{\mathcal{A}}(f), \nabla^T_{\Psi(\mathcal{B})} R_{\mathcal{B}}(f)\right) \tag{32}$$

$$= \left(\nabla^T_{\Psi(\mathcal{A})} R_{\mathcal{A}}(f), 0\right) + \left(0, \nabla^T_{\Psi(\mathcal{B})} R_{\mathcal{B}}(f)\right) \tag{33}$$

$$= \nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{A}}(f) + \nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{B}}(f) \tag{34}$$

and (18) follows from (8). ∎

A.3. Proof of Lemma 2

Proof Assume, without loss of generality, that $\mathcal{L}^{t-1}$ contains examples $\{x_i\}_{i=1}^k$ and $\mathcal{D}^{t-1}$ examples $\{x_i\}_{i=k+1}^n$, for some $1 < k < n$. Then

$$\nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{L}^{t-1}}(f^t) = \Big(\nabla^T_{\Psi(\mathcal{L}^{t-1})} R_{\mathcal{L}^{t-1}}(f^t), \nabla^T_{\Psi(\mathcal{D}^{t-1})} R_{\mathcal{L}^{t-1}}(f^t)\Big) \tag{35}$$

$$= \left(\nabla^T_{\Psi(\mathcal{L}^{t-1})} R_{\mathcal{L}^{t-1}}(f^t), 0\right). \tag{36}$$

Since the student is optimal, (30) holds and, using (13), $\nabla_{\Psi(\mathcal{L}^{t-1})} R_{\mathcal{L}^{t-1}}(f^t) = 0$. Hence, $\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}^{t-1}}(f^t) = 0$ and, from (8), $\partial_g R_{\mathcal{L}^{t-1}}(f^t) = 0$. Since, from Lemma 1,

$$\partial_g R_{\mathcal{D}}(f^t) = \partial_g R_{\mathcal{L}^{t-1}}(f^t) + \partial_g R_{\mathcal{D}^{t-1}}(f^t), \tag{37}$$

(19) follows. ∎

A.4. Proof of Theorem 1

Proof For any $g = \sum_{x_i\in\mathcal{D}} \alpha_i \delta(x - x_i)$, $||g|| = 1$ if and only if $||\alpha|| = 1$ and, from (8),

$$\partial_g R_{\mathcal{D}}(f^t) = \left\langle \nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t), \alpha \right\rangle \geq -||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t)||\,||\alpha|| = -||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t)||. \tag{38}$$

Since equality is achieved when α is the direction

$$\alpha^* = -\frac{1}{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t)||} \nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t), \tag{39}$$

the steepest descent solution of (15) is

$$g^* = \sum_{x_i\in\mathcal{D}} \alpha_i^* \delta(x - x_i) \tag{40}$$

Similarly, the steepest descent direction of (17) is

$$h^*(\mathcal{L}) = \sum_{x_i\in\mathcal{L}} \nu_i^* \delta(x - x_i) \tag{41}$$

with

$$\nu^* = -\frac{1}{||\nabla_{\Psi(\mathcal{L})} R_{\mathcal{L}}(f^t)||} \nabla_{\Psi(\mathcal{L})} R_{\mathcal{L}}(f^t), \tag{42}$$

Assuming, without loss of generality, that $\exists k$ such that $x_i \in \mathcal{L}$ for $i < k$, then

$$h^*(\mathcal{L}) = \sum_{x_i\in\mathcal{D}} \beta_i^* \delta(x - x_i) \tag{43}$$

where

$$(\beta^*)^T = \left(\nu^T, 0\right) = -\frac{1}{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t)||} \nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t), \tag{44}$$

and

$$\langle g^*, h^*(\mathcal{L})\rangle = \langle \alpha^*, \beta^*\rangle \tag{45}$$

$$= \left\langle -\frac{1}{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t)||} \nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t), -\frac{1}{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t)||} \nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t) \right\rangle \tag{46}$$

$$= \frac{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t)||^2}{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t)|| \, ||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t)||} \tag{47}$$

$$= \frac{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t)||}{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t)||}, \tag{48}$$

where we have used the fact that

$$\nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{D}}(f^t) = \left(\nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{L}}(f^t), \nabla^T_{\Psi(\mathcal{D})} R_{\mathcal{D}-\mathcal{L}}(f^t)\right). \tag{49}$$

It follows that the solution of (16) is

$$\mathcal{N}^t = \arg\max_{\mathcal{N}\in\mathcal{P}^t} ||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}^{t-1}\cup\mathcal{N}}(f^t)||^2. \tag{50}$$

$$= \arg\max_{\mathcal{N}\in\mathcal{P}^t} \left\{||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}^{t-1}}(f^t)||^2 + ||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{N}}(f^t)||^2\right\} \tag{51}$$

$$= \arg\max_{\mathcal{N}\in\mathcal{P}^t} ||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{N}}(f^t)||^2 \tag{52}$$

$$= \arg\max_{\mathcal{N}\in\mathcal{P}^t} ||\nabla_{\Psi(\mathcal{N})} R_{\mathcal{N}}(f^t)||^2 \tag{53}$$

where we have used the fact that, from Lemma 2, $||\nabla_{\Psi(\mathcal{D})} R_{\mathcal{L}^{t-1}}(f^t)||^2 = 0$. ∎

B. Other implementation details

Both datasets were subject to standard normalizations. Training images were first randomly resized to 224×224 and then randomly flipped, whereas testing images were first resized to 256×256 and then center-cropped to 224×224. All images were also first converted to $[0.0, 1.0]$ from $[0, 255]$ and then normalized by subtracting the mean $[0.485, 0.456, 0.406]$ and dividing by the standard deviation $[0.229, 0.224, 0.225]$ of each RGB color channel. On both datasets, we use the train-test split of [2]. The data is accessible in [1]. The 512-D output of global average pooling of the ResNet-18 is used for the output of $f(x)$ on the multiclass case. More details are available in our attached code. In real learner evaluation, we require that workers be masters to do our tasks. Additionally, we require non-Chinese speaker on Chinese Characters dataset experiments. Each turker is paid \$1 for the teaching task.

C. Selected teaching examples

We show the selected teaching images of MaxGrad on both datasets in Figure 1 and 2. Also, Figure 3 shows histograms of test time accuracy, at the end of the training. MaxGrad is clearly more effective than RANDOM overall.

References

[1] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. https://github.com/macaodha/explain_teach/tree/master/data.

[2] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3820–3828, 2018.

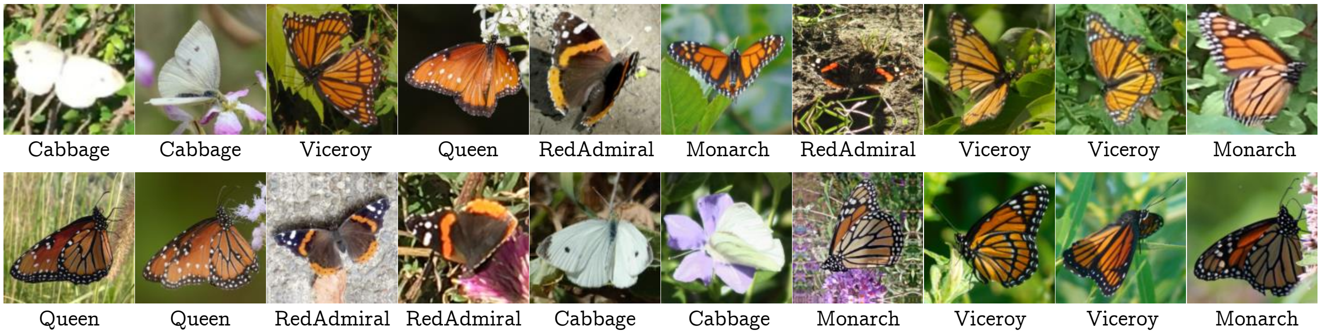

Figure 1: Selected teaching images on Butterflies

Figure 2: Selected teaching images on Chinese Characters

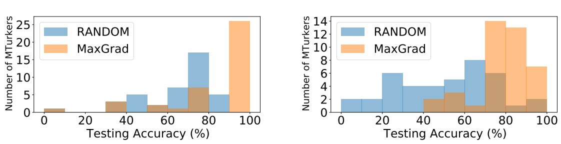

Figure 3: Test performance for human learners: learners binned by test accuracy. Left: Butterflies. Right: Chines chars.