Curriculum Learning by Optimizing Learning Dynamics

Tianyi Zhou*, Shengjie Wang*, Jeff A. Bilmes

University *of* Washington, Seattle

MELODI
MachinE Learning, Optimization, & Data Interpretation @ UW

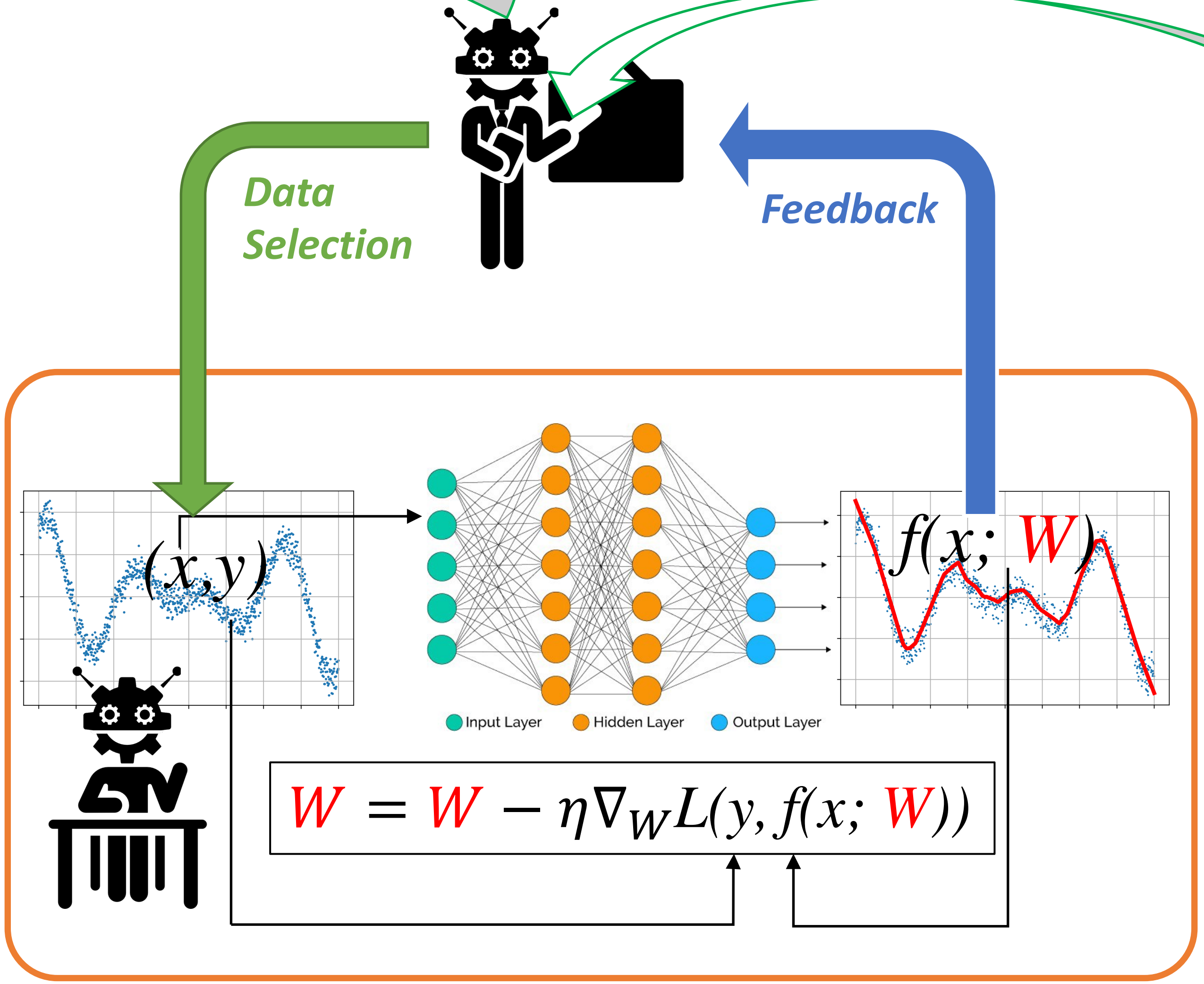
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY *of* WASHINGTON

UNIVERSITY OF WASHINGTON · LVX SIT · 1861

Curriculum Learning from Human Heuristics

Equip machine learning with human-learning strategies



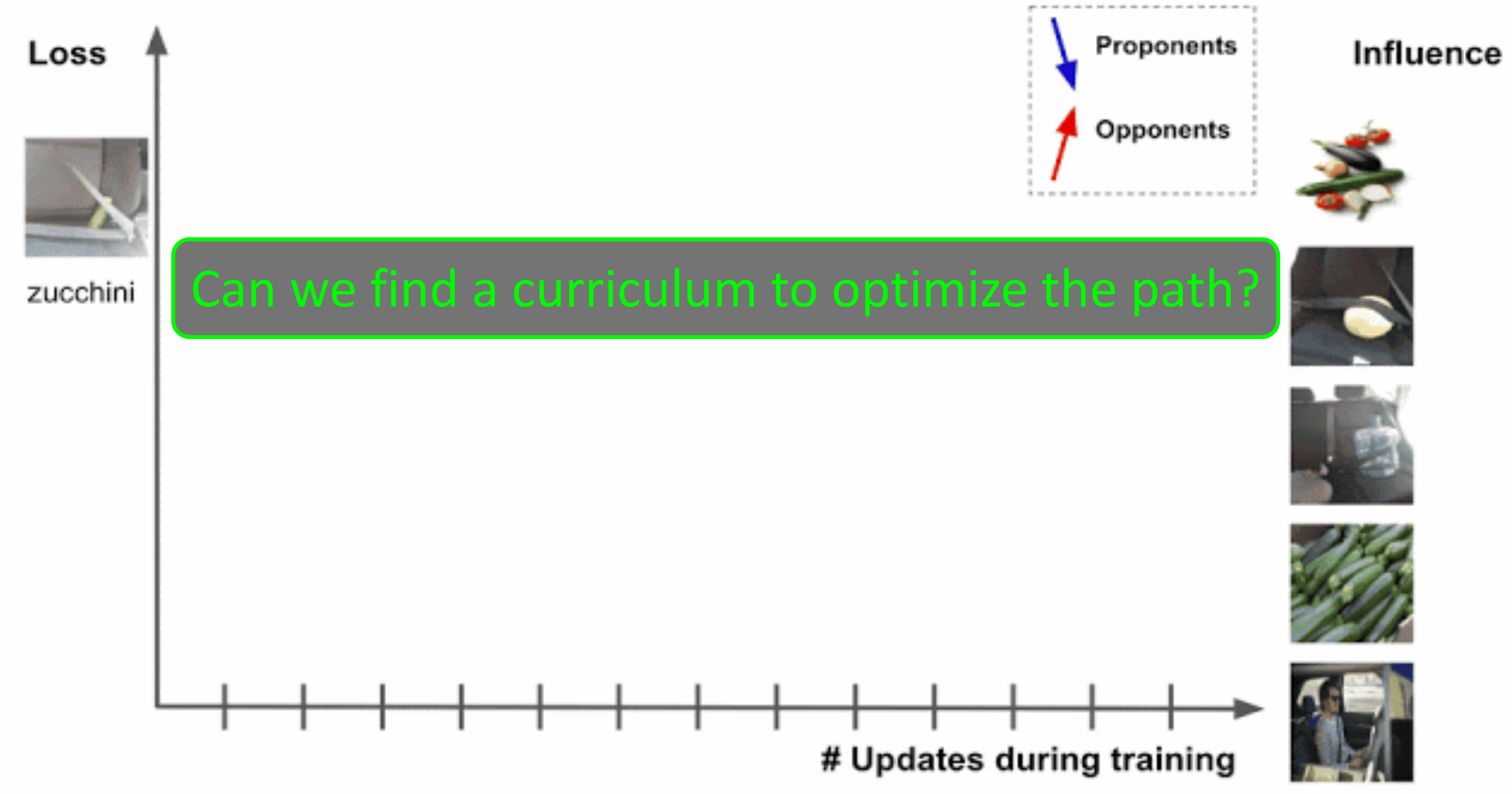
Human heuristics of curriculum:

- From easy to hard;
- From diverse to hard;
- Curiosity in early-stage
- Focus on easily-forgotten data;
- Choose representative data for exemplar-based learning;
- … …

How to justify the effectiveness of human curriculum on machine learning?

Training Dynamics on Individual Samples

Figure credit: Pruthi et al., 2020



Can we find a curriculum to optimize the path?

Optimizing Training Dynamics [Zhou et al., *AISTATS* 2021]

- **Gradient flow** (continuous-time gradient descent) **on a subset S**:

$$\left.\frac{\partial \theta}{\partial t}\right|_S = -\sum_{i \in S} \frac{\partial \ell(x_i)}{\partial \theta} = \sum_{i \in S} -\frac{\partial \ell(x_i)}{\partial f(x_i)} \cdot \frac{\partial f(x_i)}{\partial \theta}$$

- Find S that maximizes the **speed of loss decreasing** (regression):

$$\max_{S \subseteq [n], |S| \le k} \mathbb{E}_{x \sim \mathcal{D}} \left[- \left.\frac{\partial \ell(x)}{\partial t}\right|_S \right] = \mathbb{E}_{x \sim \mathcal{D}} \left\langle y - f(x), \boxed{\left.\frac{\partial f(x)}{\partial t}\right|_S} \right\rangle$$

- The linear dynamics (speed) of model output $f(x)$ is:

$$\left.\frac{\partial f(x)}{\partial t}\right|_S = \frac{\partial f(x)}{\partial \theta} \cdot \left.\frac{\partial \theta}{\partial t}\right|_S = \frac{\partial f(x)}{\partial \theta} \cdot \sum_{i \in S} -\frac{\partial \ell(x_i)}{\partial f(x_i)} \cdot \frac{\partial f(x_i)}{\partial \theta}$$

Optimize Training Dynamics [Zhou et al., *AISTATS* 2021]

- Draw $D \sim \mathcal{D}$, the dynamics-optimization objective is

$$\mathbb{E}_{x\sim\mathcal{D}}\left[-\left.\frac{\partial \ell(x)}{\partial t}\right|_S\right] \approx \frac{1}{|D|}\sum_{i\in S}\boxed{\left\langle y_i - f(x_i), \left.\frac{\partial f(x_i)}{\partial t}\right|_D\right\rangle}$$

- Select top-k samples with the highest scores $a_t(i)$:

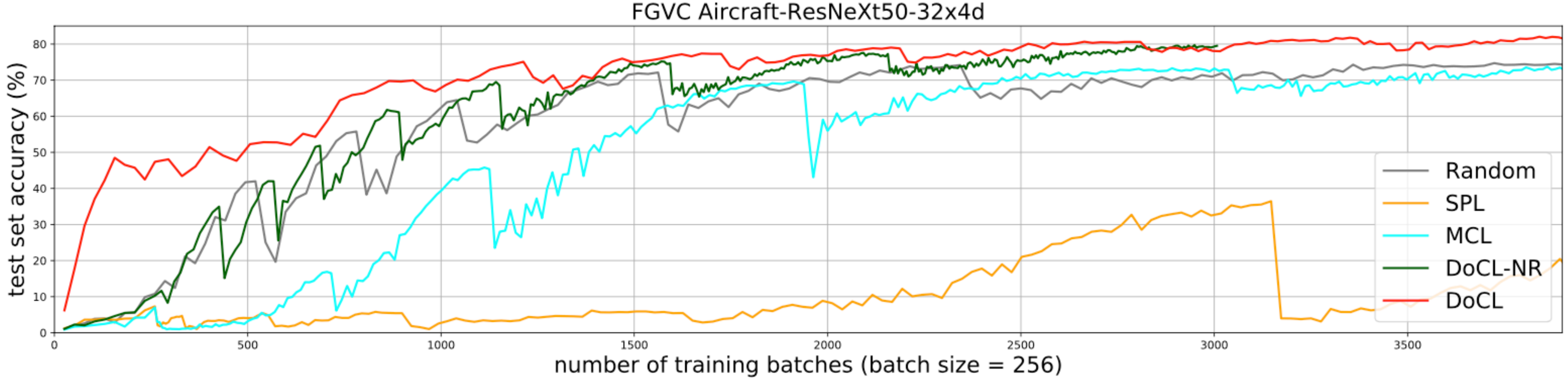
$$a_t(i) \triangleq \left\langle \boxed{y_i - f(x_i;\theta_t),} \boxed{\left.\frac{\partial f(x_i;\theta_t)}{\partial t}\right|_D} \right\rangle$$

  - **Larger residual (loss)**
  - **Output changes faster**

Dynamics-optimized Curriculum Learning (DoCL)

Table 1: The test accuracy (%) achieved by random mini-batch SGD (Random), SPL, MCL, DoCL-NR and DoCL in training DNNs on 9 datasets (without pre-training). In MCL, DoCL-NR and DoCL, we apply lazier-than-lazy-greedy [29] for Eq. (15) on CIFAR10, CIFAR100, SVHN and FMNIST. DoCL achieves the highest test accuracy over all 9 datasets.

Curriculum	CIFAR10	CIFAR100	Food-101	ImageNet	SVHN	FMNIST	Birdsnap	Aircraft	Cars
Random	96.18	79.64	83.56	75.04	96.48	95.22	64.23	74.71	78.73
SPL [24]	93.55	80.25	81.36	73.23	96.15	92.09	63.26	68.95	77.61
MCL [49]	96.60	80.99	84.18	75.09	96.93	95.07	65.76	75.28	76.98
DoCL-NR	96.40	81.42	84.75	75.62	96.80	95.50	66.59	79.72	81.48
DoCL (Ours)	**97.43**	**83.23**	**87.45**	**79.54**	**97.36**	**95.89**	**71.37**	**82.40**	**86.26**



DoCL and Neural Tangent Kernel (NTK)

- Define **residual** and **tangent kernel** (gradient similarity) as:

$$r_i \triangleq \frac{\partial \ell(x_i)}{\partial f(x_i)} = f(x_i) - y_i, \quad H_{i,j} \triangleq \left\langle \frac{\partial f(x_i)}{\partial \theta}, \frac{\partial f(x_j)}{\partial \theta} \right\rangle$$

- **[NTK interpretation] DoCL** score $a_t(i)$ selects samples with
  - **Larger residual** for themselves
  - **Similar gradient as many other samples with large residuals**

$$a_t(i) = \left[\boxed{\frac{1}{|D|} \sum_{j \in D} H_{i,j} r_j} \right] \cdot \boxed{r_i}$$

Thank you!

Poster Session 4:
April 14 (wed) at 12:45-14:45 PDT