# Identifying Model Weakness with Adversarial Examiner

Michelle Shu  Chenxi Liu  Weichao Qiu  Alan Yuille
Johns Hopkins University

JOHNS HOPKINS WHITING SCHOOL *of* ENGINEERING

## MOTIVATION

Problem Description:

- **Despite high benchmark performance, we still believe humans are superior in many machine learning masks**
- The current testing strategy is overly optimistic
- The model evaluations focus on average case and are typically fixed in size.

Our Goal:

- **Adversarial Examiner: to dynamically select the next testing sample based on testing history**
  - Worst case instead of average case
  - Dynamic test set based on test history instead of fixed test set

## EVALUATION PROTOCOL

**Standard Loss Function for Classification:**

$$E = \mathbb{E}_{x \sim \mathcal{P}}[L(f(x), y(x))] \approx \frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y(x_i))$$

**Evaluation Metric for Adversarial Examiner:**

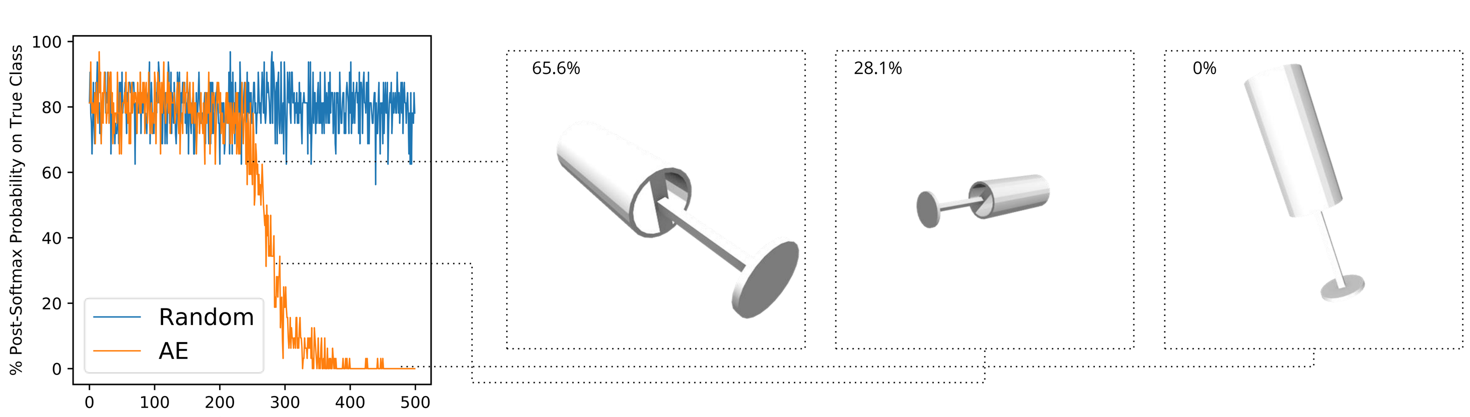
$$E_{\text{examiner}} = \mathbb{E}_{z \sim \mathcal{Q}}[\max_{s \in \mathcal{S}} L(f(g(z,s)), y(z))]$$
$$\approx \frac{1}{N}\sum_{i=1}^{N} \max_{s_i \in \mathcal{S}} L(f(g(z_i, s_i)), y(z_i))$$

**Algorithm 1:** Adversarial Examiner Procedure

**Input:** $N$ samples $z_i \sim \mathcal{Q}$ and their true labels $y(z_i)$; Maximum number of examination steps $T$; Loss function $L$; Model $f$; Function $g$; Space $\mathcal{S}$.

**for** $i = 1$ **to** $N$ **do**
  Initialize `examiner` with $\mathcal{S}$
  **for** $t = 1$ **to** $T$ **do**
    $s_i^t =$ `examiner.generate()`
    $l_i^t = L(f(g(z_i, s_i^t)), y(z_i))$
    `examiner.update(`$s_i^t, l_i^t$`)`
**return** $E_{examiner} = \frac{1}{N}\sum_{i=1}^{N} l_i^T$

**Evaluating a model's ability to recognize a *lamp* instance in ShapeNet:**




## REINFORCEMENT LEARNING AS AE

**Definitions:**

- Space $\mathcal{S}$: Cartesian product of $C$ factors $\mathcal{S} = \Psi^1 \times \Psi^2 \times \cdots \times \Psi^C$
- The candidate $s_i^t$: composed of $\psi^1_{(i,t)}, \psi^2_{(i,t)}, \ldots, \psi^C_{(i,t)}$, where $\psi^c_{(i,t)} \in \Psi^c$
- The probability of generating $s_i^t$:

$$P(s_i^t) = \prod_{c=1}^{C} P(\psi^c_{(i,t)} | \psi^{c-1:1}_{(i,t)})$$

**Implementation Details:**

- A LSTM is used to parameterize conditional probabilities
- Reward Signal $R$ is $L(f(g(z_i, s_i^t)), y(z_i))$
- Optimize the weights $\theta$ using policy gradient:

$$\nabla_\theta \mathbb{E}_{P(s_i^t;\theta)}[R] \approx \frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} \nabla_\theta \log P(\psi^c_{(i,t)} | \psi^{c-1:1}_{(i,t)}) R_b$$

## BAYESIAN OPTIMIZATION AS AE

**Definitions:**

- Gaussian Process (GP) is used to maximize $L(f(g(z_i, s_i^t)), y(z_i))$
- The candidate $s_i^t$: point proposed by the acquisition function $a : \mathcal{S} \to \mathbb{R}^+$

**Implementation Details:**

- By the end of examination, the candidates $\{s_i^t \in \mathcal{S}\}_{t=1}^T$ are points that induce the most up-to-date posterior multivariate Gaussian distribution on $\mathcal{S}$.
- For each iteration $t = 1, 2, \ldots, T$, we select the next candidate by:

$$s_i^t = \arg\max_{s \in \mathcal{S}} a(s)$$

## VARIOUS COMPARISONS

**RL Examiner and BO Examiner are Complementary:**

- Discrete vs. Continuous
- Maintaining Sampling Distribution on $\mathcal{S}$ vs. Maintaining Function Value on $\mathcal{S}$
- Longer Iteration Regime vs. Shorter Iteration Regime

**Adversarial Examiner and Adversarial Attacks:**

$$E_{\text{attack}} \approx \frac{1}{N}\sum_{i=1}^{N} \max_{\delta_i \in \Delta} L(f(x_i + \delta_i), y(x_i))$$

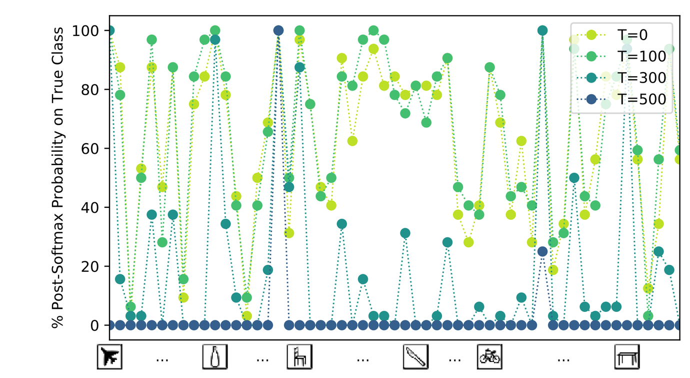
- Underlying Form ($z$) vs. Surface Form ($x$)
- Start with Entire Space vs. "Canonical" Starting Point
- Non-differentiable Settings vs. Differentiable Settings

## EXPERIMENTS & RESULTS

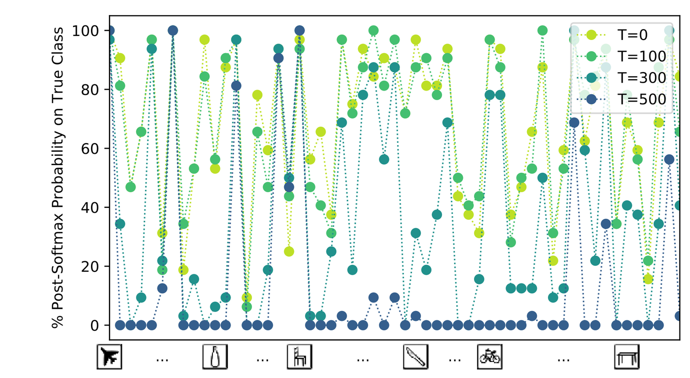
**Evaluation Model Details**

- ResNet34 and AlexNet training: 12 factors, 10 images per 3D object
- RL(LSTM): 9 continuous factors discretized to 100 choices
- BO(GP): 2 random examples, Gaussian Process upper confidence bound (UCB)

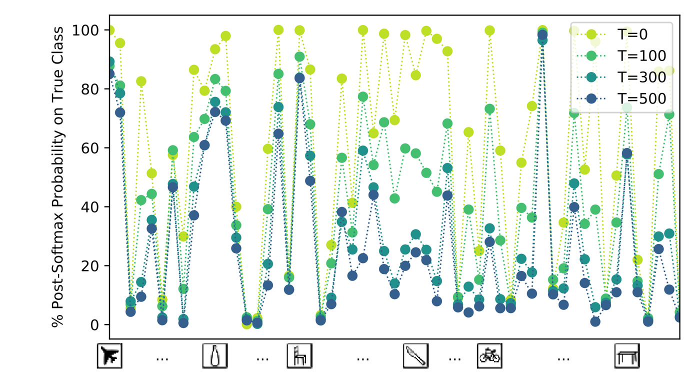
**Evaluating Model Performance with AE:**



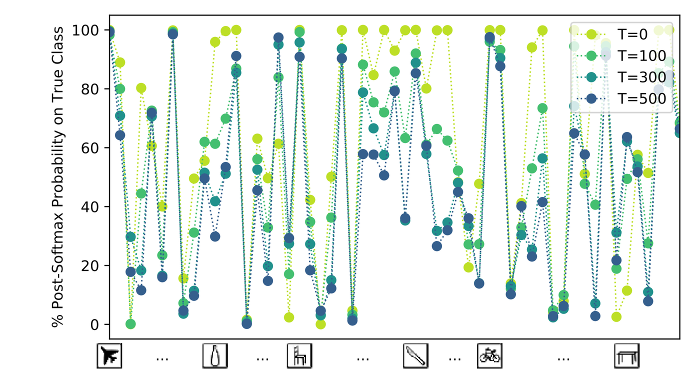

(a) RL Examiner on AlexNet




(b) RL Examiner on ResNet34




(c) BO Examiner on AlexNet



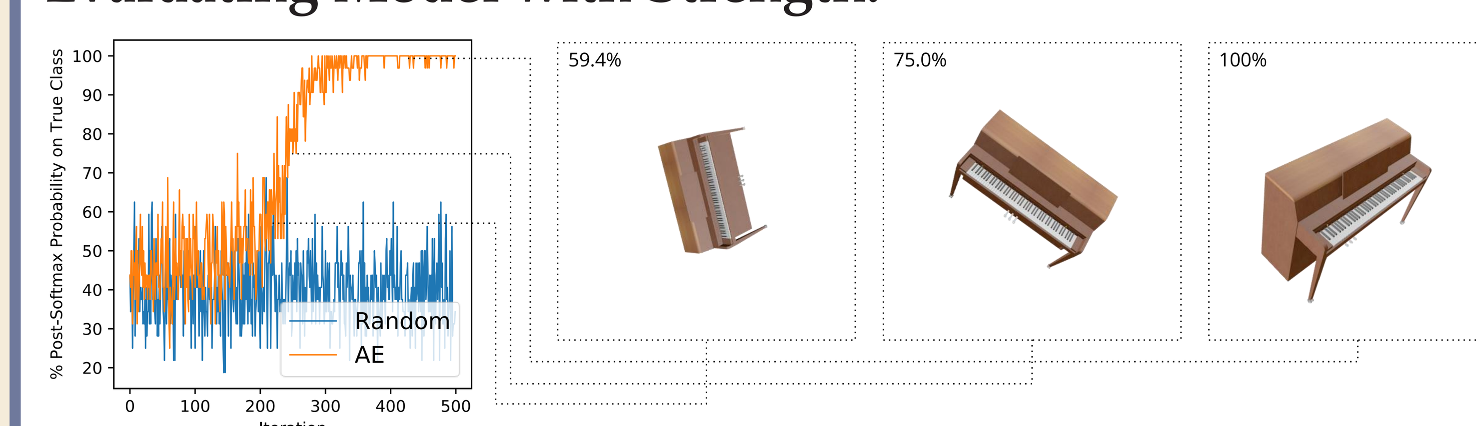

(d) BO Examiner on ResNet34

**Examining Models Trained with Less Data:**

| | $m = 10$ | $m = 5$ | $m = 2$ | $m = 1$ |
|---|---|---|---|---|
| RL | 63.81% | 57.43% | 35.05% | 18.92% |
| BO | 49.79% | 43.06% | 22.19% | 10.92% |

**Evaluating Model with Artificial Weaknesses and Order Change:**

**Evaluating Model with Strength:**




**Conclusion**

- We advocate for a new testing paradigm for machine learning models, where more emphasis is placed on the worst case instead of reporting the average case performance.
- We hope to extend to other domains (e.g. language) and see more ubiquitous usage of our general adversarial examination framework.