**Identification of Model Weakness with Adversarial Examiner**

Michelle Shu, Chenxi Liu, Weichao Qiu, Alan Yuille
Johns Hopkins University

**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:
  \[ E = E_{x \sim P} [L(f(x), y(x))] \approx \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y(x_i)) \]

- Evaluation Metric for Adversarial Examiner:
  \[ E_{\text{examiner}} = E_{z \sim Q} \max_{s \in S} L(f(g(z, s)), y(z)) \]
  \[ \approx \frac{1}{N} \sum_{i=1}^{N} \max_{s \in S} L(f(g(z_i, s_i)), y(z_i)) \]

**Algorithm 1: Adversarial Examiner Procedure**

- **Input:** \( N \) samples \( z_i \sim Q \) and their true labels \( y(z_i) \); Maximum number of examination steps \( T \); Loss function \( L \); Model \( f \); Function \( g \); Space \( S \).
- **for** \( i = 1 \) to \( N \) **do**
  - Initialize \( \text{examiner} \) with \( S \).
- **for** \( t = 1 \) to \( T \) **do**
  - \( s_t^i = \text{examiner}.\text{generate()} \)
  - \( l_t^i = L(f(g(z_i, s_t^i)), y(z_i)) \)
  - \( \text{examiner}.\text{update} (s_t^i, l_t^i) \)
- **return** \( E_{\text{examiner}} = \frac{1}{T} \sum_{i=1}^{N} l_t^i \)

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

**Reinforcement Learning as AE**

- **Definitions:**
  - Space \( S \): Cartesian product of \( C \) factors \( S = \Psi^1 \times \Psi^2 \times \cdots \times \Psi^C \)
  - The candidate \( s_t^i \): composed of \( \psi_i^{t,i} \), \( \psi_{i,t}^{i} \), \( \psi_{i,t}^{1} \), \( \psi_{i,t}^{C} \), where \( \psi_{i,t}^{c} \in \Psi^c \)
  - The probability of generating \( s_t^i \):
    \[ P(s_t^i) = \prod_{c=1}^{C} P(\psi_{i,t}^{c} | \psi_{i,t}^{c-1}) \]

- **Implementation Details:**
  - A LSTM is used to parameterize conditional probabilities.
  - Reward Signal \( R \) is \( L(f(g(z, s_t^i)), y(z)) \)
  - Optimize the rewards using policy gradient:
    \[ \nabla_{\theta} E \left[ R \right] \approx \frac{1}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} \nabla_{\theta} \log P(\psi_{i,t}^{c} | \psi_{i,t}^{c-1}) R_b \]

**Bayesian Optimization as AE**

- **Definitions:**
  - Gaussian Process (GP) is used to maximize \( L(f(g(z, s_t^i)), y(z)) \)
  - The candidate \( s_t^i \): point proposed by the acquisition function \( a : S \rightarrow \mathbb{R}^+ \)

- **Implementation Details:**
  - By the end of examination, the candidates \( \{ s_t^i \in S \}_{t=1}^{T} \) are points that induce the most up-to-date posterior multivariate Gaussian distribution on \( S \).
  - For each iteration \( t = 1, 2, \ldots, T \), we select the next candidate by:
    \[ s_t^i = \arg \max_{s \in S} a(s) \]

**Various Comparisons**

- RL Examiners and BO Examiners are Complementary:
  - Discrete vs. Continuous.
  - Maintaining Sampling Distribution on \( S \) vs. Maintaining Function Value on \( S \).
  - Longer Iteration Regime vs. Shorter Iteration Regime.

- Adversarial Examiner and Adversarial Attacks:
  \[ E_{\text{attack}} \approx \frac{1}{N} \sum_{i=1}^{N} \max_{h_i \in \Delta} L(f(x_i + h_i), y(x_i)) \]
  - Underlying Form \( (z) \) vs. Surface Form \( (s) \)
  - Start with Entire Space vs. “Canonical” Starting Point
  - Non-differentiable Settings vs. Differentiable Settings

**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.