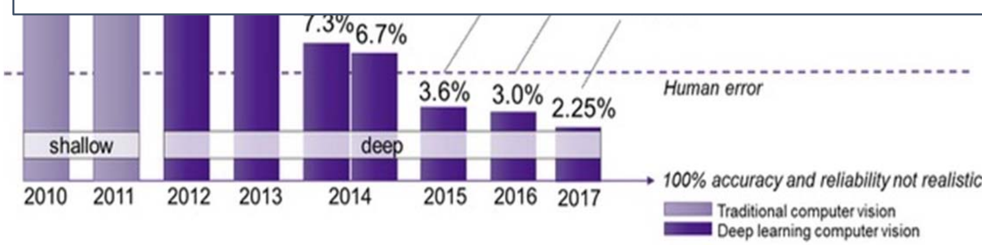
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

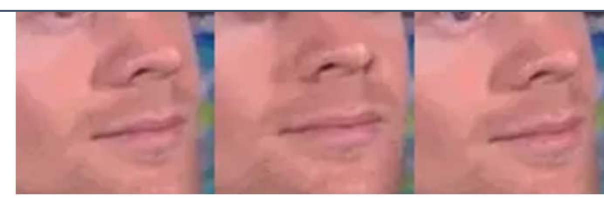
Identifying Model Weakness with Adversarial Examiner

Michelle Shu, Chenxi Liu, Weichao Qiu, Alan Yuille
Oct 11th, 2019
Johns Hopkins University

Motivation:

Why is there a mismatch?





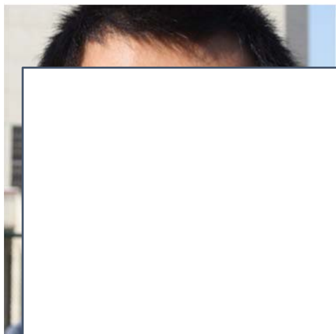
https://news.stanford.edu/2018/05/15/how-ai-is-changing-science/

Motivation:

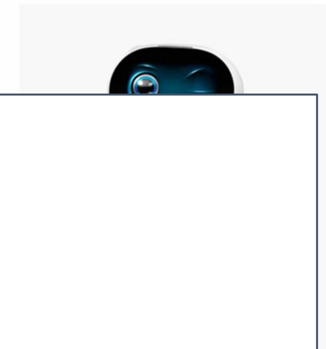
Turing test

Johns Hopkins Whiting School of Engineering

Identifying Model Weakness with Adversarial Examiner



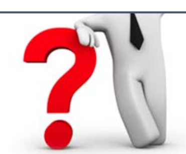
80% Accuracy -> **99.9% Accuracy**



Lesson 1:

The test should focus more on *worst case* than average case.

999/1000



1/100

Lesson 2:
The test should be *dynamic* instead of *fixed*

999/1000

1/100

Solution: Adversarial Examiner (AE)

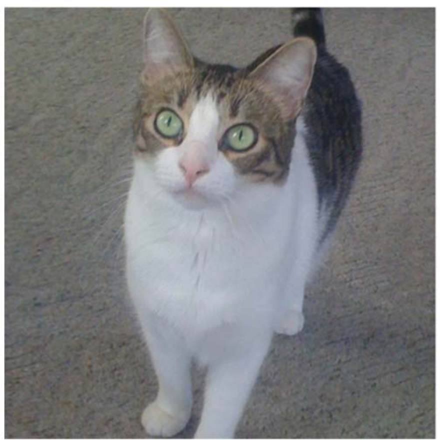
- Worst case instead of average case
- Dynamic test set based on test history instead of fixed test set





Solution: Adversarial Examiner (AE) Definitions

Underlying Form z



3D object: Cat

Additional Information s

Is it bleping: *yes*
Viewing distance: *close-up*
...
State: *cute*

Surface Form x = g(z, s)

2D image: Cat

Solution: Adversarial Examiner (AE)

In standard classification tasks:

Standard evaluation metric vs. AE's evaluation metric

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

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

$L(\cdot,\cdot)$ Loss function

$f(x_i)$ Predicted Label

$y(x_i)$ Groundtruth Label

P Underlying Distribution for x

Q Underlying Distribution for z

$\boldsymbol{S}$ Information to transform z to x

$g(z_i, s_i)$ transform function

Relationship: AA and AE

In standard classification tasks:

Adversarial Attack (AA) vs. Adversarial Examiner (AE)

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

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

$L(\cdot,\cdot)$ Loss function

$f(x_i)$ Predicted Label

$y(x_i)$ Groundtruth Label

P Underlying Distribution for x

Q Underlying Distribution for z

$\mathcal{S}$ Information to transform z to x

$g(z_i, s_i)$ transform function

1. AE deals with underlying form z while AA deals with surface form x.
2. There is a "canonical" starting point for AA but AE starts with the entire space S.

Solution: Adversarial Examiner (AE)

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`)`

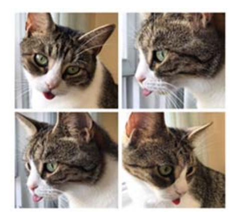
return $E_{examiner} = \frac{1}{N} \sum_{i=1}^{N} l_i^T$

Underlying Form z

Additional Information s

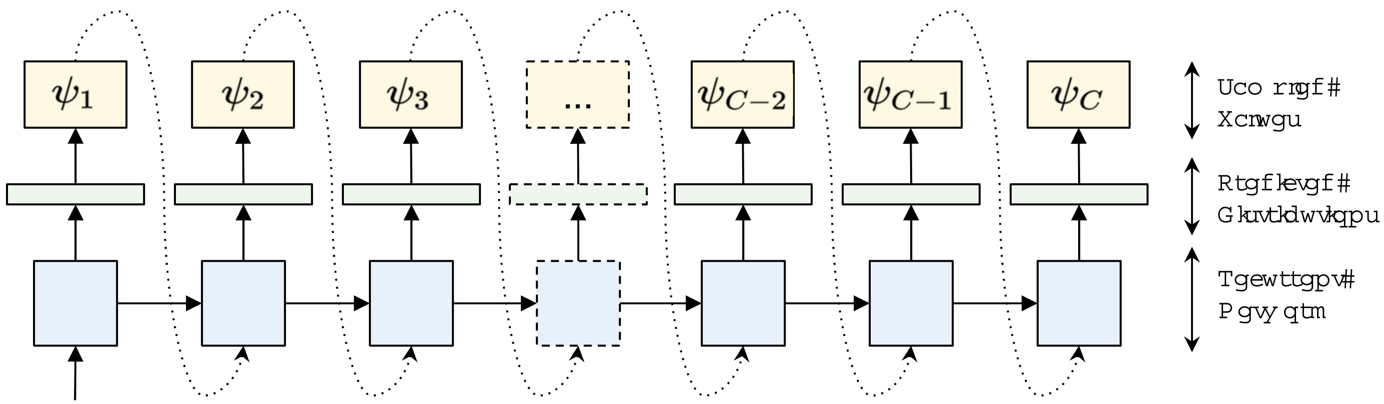
Is it bleping: *yes*
Viewing distance: *close-up*
...
State: *cute*

Surface Form x = g(z, s)



Deep Learning Based AE (LSTM + Reinforcement Learning):

Let space $\mathcal{S}$ be the Cartesian product of C factors $\mathcal{S} = \Psi^1 \times \Psi^2 \times \cdots \times \Psi^C$



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

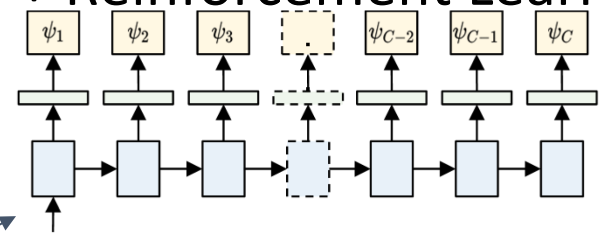
Deep Learning Based AE (LSTM + Reinforcement Learning):

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 $E_{examiner} = \frac{1}{N}\sum_{i=1}^{N} l_i^T$

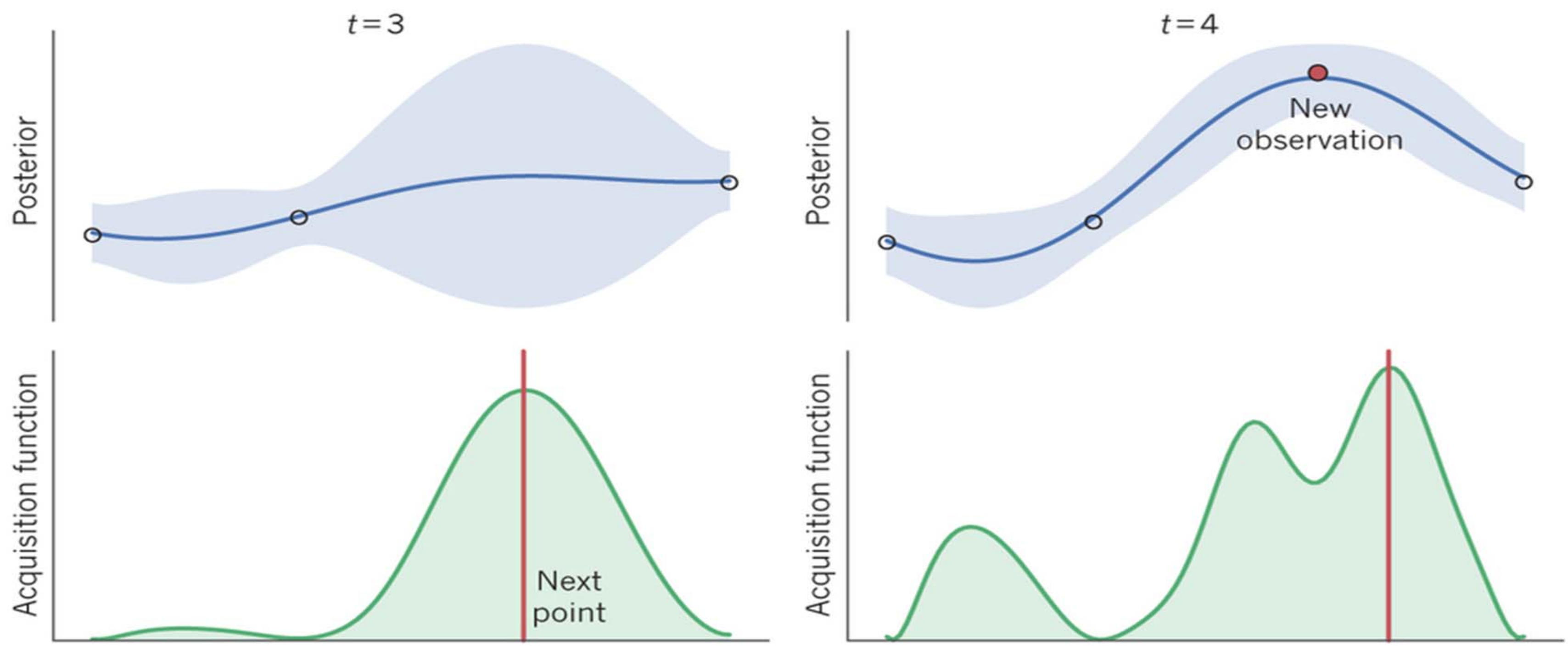


R Reward signal

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

Bayesian Optimization Based AE:

$$s_i^t = \operatorname*{argmax}_{s \in \mathcal{S}} a(s)$$



Bayesian Optimization Based AE:

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**

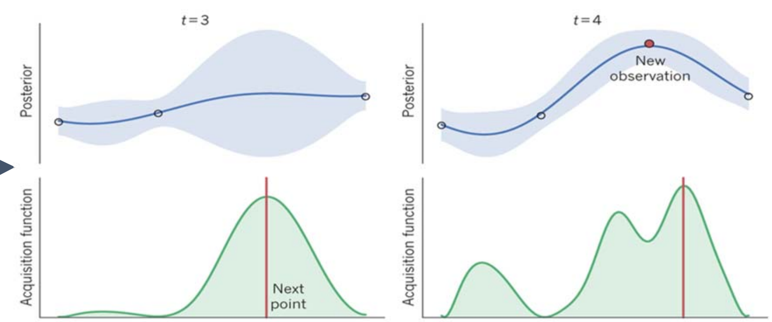
```
    Initialize examiner with S
    for t = 1 to T do
        s_i^t = examiner.generate()
        l_i^t = L(f(g(z_i, s_i^t)), y(z_i))
        examiner.update(s_i^t, l_i^t)
```

return $E_{examiner} = \frac{1}{N} \sum_{i=1}^{N} l_i^T$

$$s_i^t = \operatorname*{argmax}_{s \in \mathcal{S}} a(s)$$

Newly Observed Point

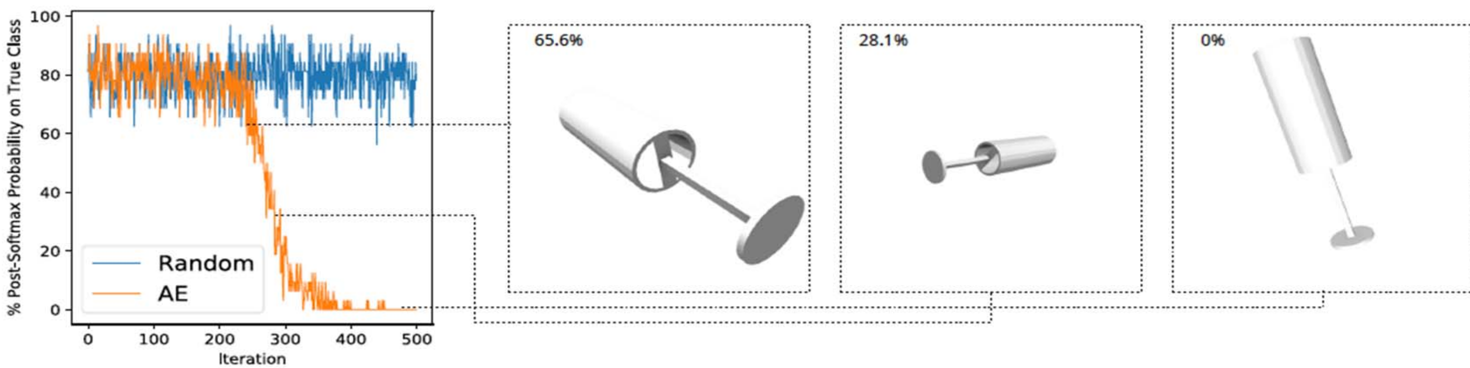
$$(s_i^t, L(f(g(z_i, s_i^t)), y(z_i)))$$



Experiments on ShapeNet:

- Model Type: ResNet34 vs. AlexNet
- Training Set: Varied training set size
- Multiple Weakness: Artificial Weakness
- Reversed Examination: Identify Model Strength

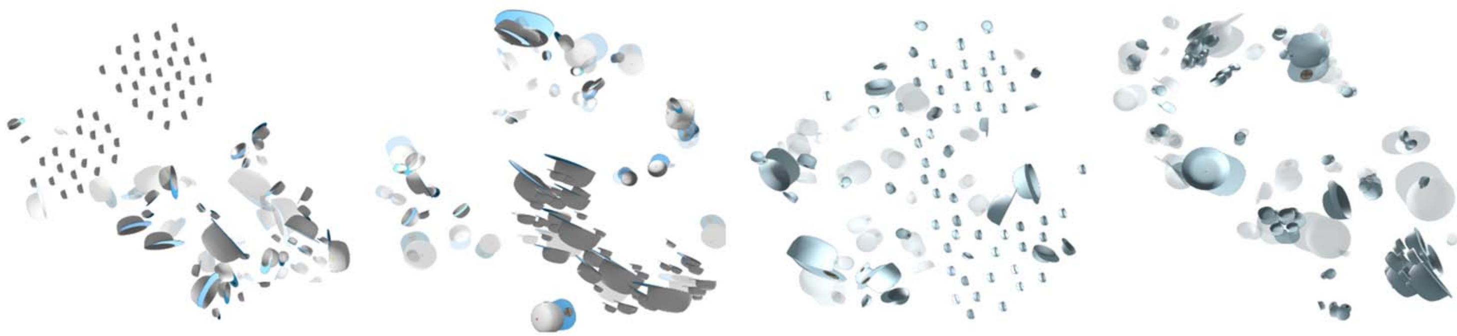
Experiments on ShapeNet:



	α_o	β_o	ζ_o	Γ_o	Γ_l	r_l	A_l	U_l	r_v	A_v	U_v	θ_v
UB	2π	2π	2π	5	1	20	360	90	5	180	90	360
LB	0	0	0	0	0.3	8	0	-90	1	0	-90	0

Table 1: Upper bound (UB) and lower bound (LB) of factors for s: sun rotation angles ($\alpha_o, \beta_o, \zeta_o$), sun energy ($\Gamma_o$), point light energy ($\Gamma_l$), point light distance (r_l), point light location (A_l, U_l), viewpoint distance (r_v), viewpoint location (A_v, U_v), viewpoint angle (θ_v).

Experiments on ShapeNet: ResNet34 vs. Alexnet



(a) RL on AlexNet (b) BO on AlexNet (c) RL on ResNet34 (d) BO on ResNet34

Model	Examiner	$T = 0$	$T = 100$	$T = 300$	$T = 500$
AlexNet	RL	63.98%	65.91%	18.92%	2.27%
	BO	60.05%	43.58%	29.98%	25.43%
ResNet34	RL	69.03%	68.58%	38.86%	13.13%
	BO	64.19%	54.89%	48.07%	45.55%

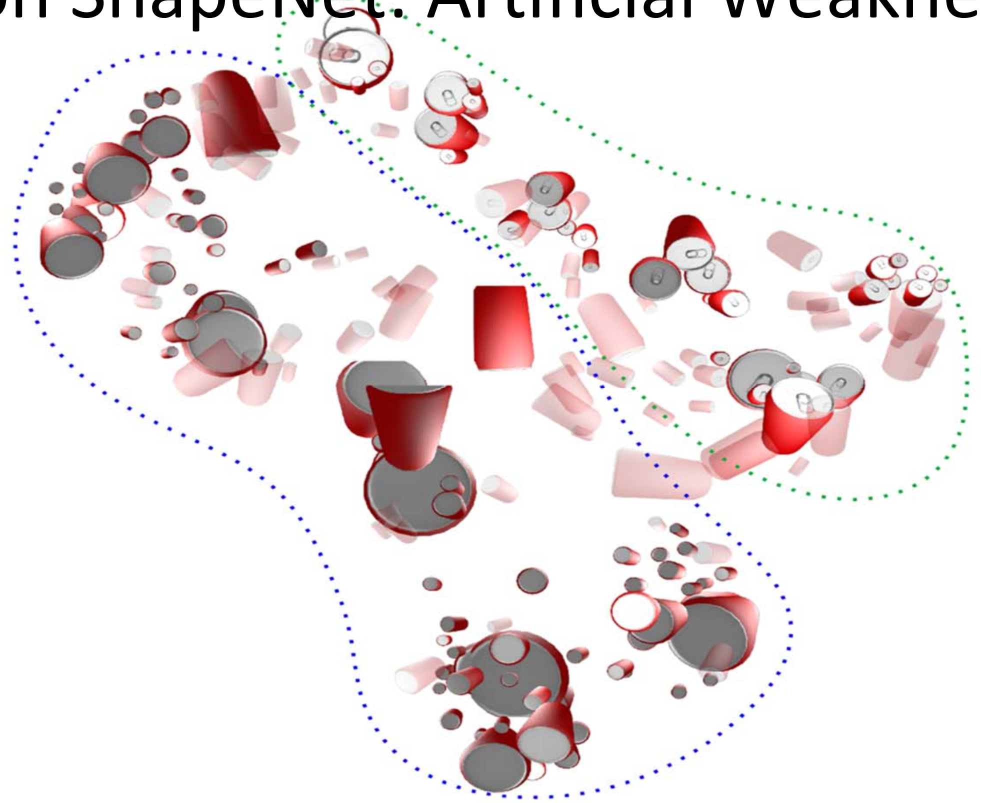
Experiments on ShapeNet: Varied Training Size

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

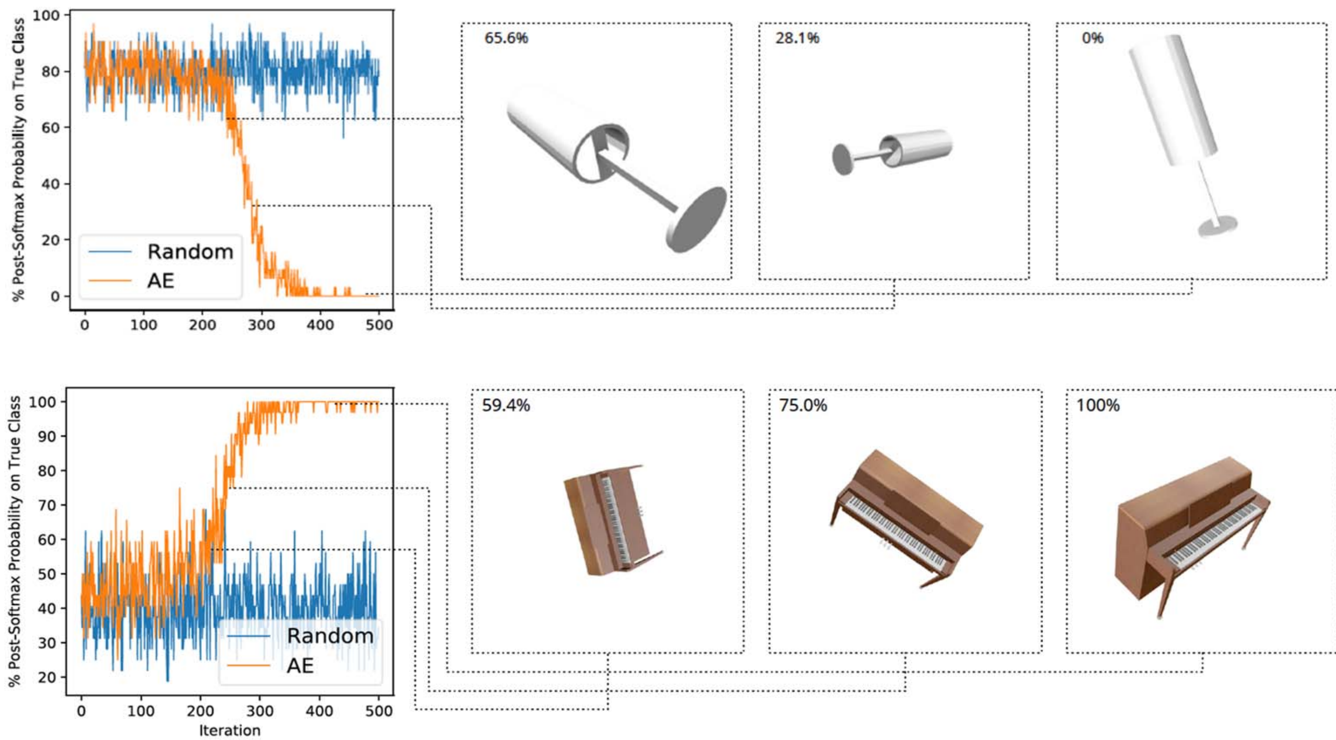
	α_o	β_o	ζ_o	Γ_o	Γ_l	r_l	A_l	U_l	r_v	A_v	U_v	θ_v
UB	2π	2π	2π	5	1	20	360	90	5	180	90	360
LB	0	0	0	0	0.3	8	0	-90	1	0	-90	0

Table 1: Upper bound (UB) and lower bound (LB) of factors for s: sun rotation angles ($\alpha_o, \beta_o, \zeta_o$), sun energy ($\Gamma_o$), point light energy ($\Gamma_l$), point light distance (r_l), point light location (A_l, U_l), viewpoint distance (r_v), viewpoint location (A_v, U_v), viewpoint angle (θ_v).

Experiments on ShapeNet: Artificial Weakness



Experiments on ShapeNet: Identifying Model Strength



Take-Home Message:

Motivated by the mismatch, we try to mimic some aspects of turing test:

- Worst case instead of average case
- Dynamic test set based on test history instead of fixed test set

Some Problems:

- Implicit form z and transform function g(z, s) is hard to obtain in some tasks
- CV People cannot abandon fixed datasets (yet)

Ongoing Experiment:

- Apply AE to 6D Pose Estimation Task:

Thank You!