Bayes-TrEx: A Bayesian Sampling Approach to Transparency by Example

Serena Booth*  Yilun Zhou*  Ankit Shah  Julie Shah

AAAI 2021
Consider a Corgi/Bread Classifier
Consider a Corgi/Bread Classifier
Consider a Corgi/Bread Classifier
Consider a Corgi/Bread Classifier
Our objective:
Build a *holistic* understanding of the classifier.
Build a *holistic* understanding of the classifier.
1. Develop an understanding of class boundaries.
1. Develop an understanding of class boundaries.

True class

Corgi

Bread

Corgi

Bread

51% Corgi, 49% Bread

48% Corgi, 52% Bread

50% Corgi, 50% Bread

45% Corgi, 55% Bread

Classifier’s prediction

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
2. Predict classifier behavior on new examples.
2. Predict classifier behavior on new examples.

True class

Corgi

Bread

Corgi

Bread
2. Predict classifier behavior on new examples.

True class

Corgi
Bread
Corgi
Bread

Classifier’s prediction

Corgi
Bread
Bread
Corgi

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
3. Predict classifier behavior on novel class instances.
3. Predict classifier behavior on novel class instances.

**True class**

- Cat
- Cake
- Croissant
- Potatoes

**Classifier’s prediction**

- Corgi
- Corgi
- Bread
- Bread

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
How can we build this understanding?
Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.
Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.
Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.

50% Confident? Few images.
Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.

50% Confident? Few images.

Problem: sparse data

Booth and Zhou et al., BayesTrEx, AAAI, 2021
How can we build a holistic understanding?

Input → Explanation pipeline → Saliency Map

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
How can we build a holistic understanding?

Problem: where does the test input come from?
We introduce Bayes-TrEx: an approach to transparency by example
We introduce Bayes-TrEx: an approach to transparency by example
A Corgi/Bread Decision Surface

\[ P(\tilde{y} = \text{Corgi}) \]

\[ P(\text{Corgi}) \approx 1 \]
A Corgi/Bread Decision Surface

\[ P(\bar{y} = \text{Corgi}) \]
A Corgi/Bread Decision Surface

\[ P(\bar{y} = \text{Corgi}) \]

\[ P(\text{Corgi}) \approx 0.5 \]
A Corgi/Bread Decision Surface

\[ P(\text{Corgi}) = 0.5 \text{ Level Set} \]

True Posterior

\[ P(\text{Corgi}) \approx 0.5 \]
We find prediction-matching examples from \textbf{p-level sets}

\[
\text{P(Corgi)} = 0.5 \text{ Level Set} \\
\text{True Posterior} \\
\text{P(Corgi)} \cong 0.5
\]
We find prediction-matching examples from \textbf{p-level sets}

\[ P(\text{Corgi}) = 0.5 \text{ Level Set} \]
\[ \text{True Posterior} \]

\[ P(\text{Corgi}) \cong 0.5 \]
We find prediction-matching examples from $p$-level sets

\[ P(\text{Corgi}) \approx 0.5 \]

$P(\text{Corgi}) = 0.5$ Level Set
True Posterior

\[ P(\text{Corgi}) \approx 0.5 \]
We find prediction-matching examples from p-level sets.

\[ P(\text{Corgi}) \approx 0.5 \]

\[ P(\text{Corgi}) = 0.5 \text{ Level Set} \]

True Posterior

\[ P(\text{Corgi}) \approx 0.5 \]
We want to find a *natural* example $\mathbf{x}$ where the classifier $f(\mathbf{x})$ has confidence $p$

$P(\text{Corgi}) = 0.5$ Level Set
True Posterior

$P(\text{Corgi}) \approx 0.5$
We want to find a *natural* example $\mathbf{x}$ where the classifier $f(\mathbf{x})$ has confidence $p$.

Want to sample from:

$$p(\mathbf{x}|f(\mathbf{x}) = p) \propto p(\mathbf{x}) p(f(\mathbf{x}) = p|\mathbf{x})$$

$P(\text{Corgi}) = 0.5$ Level Set
True Posterior
Want to sample from: \( p(\mathbf{x} | f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x}) p(f(\mathbf{x}) = \mathbf{p} | \mathbf{x}) \)

Problem 1:
\[
\{ \mathbf{x} : f(\mathbf{x}) = \mathbf{p} \}
\]
has small or even zero measure

\[
P(\text{Corgi}) = 0.5 \text{ Level Set}
\]

True Posterior

Booth and Zhou et al., Bayes-TRex, AAAI, 2021
Want to sample from: $p(x|f(x) = p) \propto p(x)p(f(x) = p|x)$

Problem 1:

$\{x : f(x) = p\}$

has small or even zero measure

Problem 2:

$x$ is too high-dimensional
Want to sample from: \[ p(x \mid f(x) = p) \propto p(x) p(f(x) = p \mid x) \]

Problem 1:
\[ \{x : f(x) = p\} \]
has small or even zero measure.

Problem 2:
\[ x \] is too high-dimensional.

\[ \text{P(Corgi) = 0.5 Level Set} \]
\[ \text{True Posterior} \]
Make inference tractable

Relax the formulation by “widening” the level set

$P(\text{Corgi}) = 0.5$ Level Set
Relaxed Posterior
Relaxed Formulation

Introduce a random vector:
\[ \mathbf{u} | \mathbf{x} \sim \mathcal{N}(f(\mathbf{x}), \sigma^2) \]

\[ \text{P(Corgi)} = 0.5 \text{ Level Set} \]

Relaxed Posterior

Booth and Zhou et al., Bayes-TREx, AAAI, 2021
Relaxed Formulation

Introduce a random vector:
\[ \mathbf{u} | \mathbf{x} \sim \mathcal{N}(f(\mathbf{x}), \sigma^2) \]

And sample from the new posterior:
\[ p(\mathbf{x} | \mathbf{u} = \mathbf{u}^*) \propto p(\mathbf{x})p(\mathbf{u} = \mathbf{u}^* | \mathbf{x}) \]
\[ \mathbf{u}^* = \mathbf{p} \]
But, how can we sample an image $x$?
But, how can we sample an image $x$?

We sample $z$ from a latent space $Z$, instead.

\[ x = g(z) \]
But, how can we sample an image $x$?

We sample $z$ from a latent space $Z$, instead.

$$x = g(z)$$

$$u | z \sim N(f(x), \sigma^2)$$

$$p(z | u = u^*) \propto p(z)p(u = u^* | z)$$
To use Bayes-TrEx, we need 3 requirements:
To use Bayes-TrEx, we need 3 requirements:

1. A classifier which outputs class probabilities
To use Bayes-TrEx, we need 3 requirements:

1. A classifier which outputs class probabilities

2. A user-specified confidence target
To use Bayes-TrEx, we need 3 requirements:

1. A classifier which outputs class probabilities
2. A user-specified confidence target
3. A data distribution we can sample from
Experiments

CLEVR

Scene graph

MNIST

VAEs, GANs

Fashion-MNIST

VAEs, GANs
Smoke Test: High Confidence Examples \((p_i=1, \ p_{-i}=0)\)

\(\mathbb{P}_C, \text{ classifier training distribution, red}\)
Smoke Test: High Confidence Examples \((p_{+1} = 1, \ p_{-1} = 0)\)

\(\mathbb{P}_C\), classifier training distribution, red

\(\mathbb{P}_D\), Bayes-TrEx’s data distribution, yellow
Smoke Test: High Confidence Examples \((p_i=1, \ p_{-i}=0)\)

\(\mathbb{P}_C\), classifier training distribution, red

\(\mathbb{P}_D\), Bayes-TrEx’s data distribution, yellow

\(\mathbb{P}_D \approx \mathbb{P}_C\)
Smoke Test: High Confidence Examples \((p_i = 1, \ p_{\neg i} = 0)\)

\(P_C\), classifier training distribution, red

\(P_D\), Bayes-TrEx’s data distribution, yellow

\(P_D \approx P_C\)
Smoke Test: High Confidence Examples \((p_i = 1, \ p_{\neg i} = 0)\)

CLEVR: 5 Spheres
Smoke Test: High Confidence Examples ($p_i=1$, $p_{-i}=0$)

CLEVR: 5 Spheres

$P_D \approx P_C$
Smoke Test: High Confidence Examples ($p_i=1$, $p_{-i}=0$)

CLEVR:
5 Spheres

MNIST

Fashion-MNIST

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
Smoke Test: High Confidence Examples ($p_i=1$, $p_{\neg i}=0$)

\[ \mathbb{P}_D \approx \mathbb{P}_C \]

\[ \approx 0.943 \quad \text{CLEVR: 5 Spheres} \]

\[ \approx 0.999 \quad \text{MNIST} \]

\[ \approx 0.998 \quad \text{Fashion-MNIST} \]

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
Class Boundaries \((p_i=0.5, p_j=0.5, p_{-i,-j}=0)\)
Class Boundaries \((p_i=0.5, p_j=0.5, p_{-i,-j}=0)\)

\[ P_D \approx P_C \]

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
Class Boundaries \((p_i=0.5, p_j=0.5, p_{-i,-j}=0)\)

\[ \mathbb{P}_D \approx \mathbb{P}_C \]
Class Boundaries \((p_i=0.5, p_j=0.5, p_{-i,-j}=0)\)
Class Boundaries \((p_i = 0.5, p_j = 0.5, p_{-i, -j} = 0)\)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
</tr>
<tr>
<td>1</td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
<td><img src="image36.png" alt="Image" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="image37.png" alt="Image" /></td>
<td><img src="image38.png" alt="Image" /></td>
<td><img src="image39.png" alt="Image" /></td>
<td><img src="image40.png" alt="Image" /></td>
<td><img src="image41.png" alt="Image" /></td>
<td><img src="image42.png" alt="Image" /></td>
<td><img src="image43.png" alt="Image" /></td>
<td><img src="image44.png" alt="Image" /></td>
<td><img src="image45.png" alt="Image" /></td>
</tr>
<tr>
<td>5</td>
<td><img src="image46.png" alt="Image" /></td>
<td><img src="image47.png" alt="Image" /></td>
<td><img src="image48.png" alt="Image" /></td>
<td><img src="image49.png" alt="Image" /></td>
<td><img src="image50.png" alt="Image" /></td>
<td><img src="image51.png" alt="Image" /></td>
<td><img src="image52.png" alt="Image" /></td>
<td><img src="image53.png" alt="Image" /></td>
<td><img src="image54.png" alt="Image" /></td>
</tr>
<tr>
<td>6</td>
<td><img src="image55.png" alt="Image" /></td>
<td><img src="image56.png" alt="Image" /></td>
<td><img src="image57.png" alt="Image" /></td>
<td><img src="image58.png" alt="Image" /></td>
<td><img src="image59.png" alt="Image" /></td>
<td><img src="image60.png" alt="Image" /></td>
<td><img src="image61.png" alt="Image" /></td>
<td><img src="image62.png" alt="Image" /></td>
<td><img src="image63.png" alt="Image" /></td>
</tr>
<tr>
<td>7</td>
<td><img src="image64.png" alt="Image" /></td>
<td><img src="image65.png" alt="Image" /></td>
<td><img src="image66.png" alt="Image" /></td>
<td><img src="image67.png" alt="Image" /></td>
<td><img src="image68.png" alt="Image" /></td>
<td><img src="image69.png" alt="Image" /></td>
<td><img src="image70.png" alt="Image" /></td>
<td><img src="image71.png" alt="Image" /></td>
<td><img src="image72.png" alt="Image" /></td>
</tr>
<tr>
<td>8</td>
<td><img src="image73.png" alt="Image" /></td>
<td><img src="image74.png" alt="Image" /></td>
<td><img src="image75.png" alt="Image" /></td>
<td><img src="image76.png" alt="Image" /></td>
<td><img src="image77.png" alt="Image" /></td>
<td><img src="image78.png" alt="Image" /></td>
<td><img src="image79.png" alt="Image" /></td>
<td><img src="image80.png" alt="Image" /></td>
<td><img src="image81.png" alt="Image" /></td>
</tr>
<tr>
<td>9</td>
<td><img src="image82.png" alt="Image" /></td>
<td><img src="image83.png" alt="Image" /></td>
<td><img src="image84.png" alt="Image" /></td>
<td><img src="image85.png" alt="Image" /></td>
<td><img src="image86.png" alt="Image" /></td>
<td><img src="image87.png" alt="Image" /></td>
<td><img src="image88.png" alt="Image" /></td>
<td><img src="image89.png" alt="Image" /></td>
<td><img src="image90.png" alt="Image" /></td>
</tr>
</tbody>
</table>

\(P_D \approx P_C\)
High Confidence Failures

$P_C$, classifier training distribution, red
High Confidence Failures

\( P_C \), classifier training distribution, red
High Confidence Failures

$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx’s data distribution, yellow

$\mathbb{P}_D \subsetneq \mathbb{P}_C$
High Confidence Failures

Contains 1 Cube

97.2%
96.0%
94.5%
67.3%
93.5%
High Confidence Failures

0: 0.981 ± 0.027
1: 0.953 ± 0.028
2: 0.968 ± 0.028
3: 0.969 ± 0.027
4: 0.955 ± 0.030

Sandal: 0.986 ± 0.030
Shirt: 0.938 ± 0.032
Sneaker: 0.969 ± 0.028
Bag: 0.967 ± 0.026
Ankle boot: 0.971 ± 0.027
High Confidence Failures

0: 0.981 ± 0.027
1: 0.953 ± 0.028
2: 0.968 ± 0.028
3: 0.969 ± 0.027
4: 0.955 ± 0.030

Sandal: 0.986 ± 0.030
Shirt: 0.938 ± 0.032
Sneaker: 0.969 ± 0.028
Bag: 0.967 ± 0.026
Ankle boot: 0.971 ± 0.027
Novel Class Extrapolation

$P_C$, classifier training distribution, red
Novel Class Extrapolation

$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx’s data distribution, yellow

$\mathbb{P}_D \cap \mathbb{P}_C \neq \mathbb{P}_D \neq \mathbb{P}_C$
Novel Class Extrapolation

Contains 5 Cubes: 89.3%
Novel Class Extrapolation

Contains 5 Cubes

89.3%
81.2%
83.5%
90.4%
90.5%
Domain Adaptation

$P_C$, classifier training distribution, red
Domain Adaptation

$P_D \cap P_C = \emptyset$

$P_C$, classifier training distribution, red

$P_D$, Bayes-TrEx’s data distribution, yellow
Domain Adaptation: ADDA

Overall:
• Baseline: 61% accuracy
• ADDA: 71% accuracy
Domain Adaptation: ADDA

Overall:
- Baseline: 61% accuracy
- ADDA: 71% accuracy
Domain Adaptation: ADDA

Overall:
- Baseline: 61% accuracy
- ADDA: 71% accuracy

High Confidence Examples:
- Baseline: 80% accuracy
- ADDA: 72% accuracy

Booth and Zhou et al., Bayes-TREx, AAAI, 2021
How does Bayes-TrEx compare to the test set?

Test set:
22 Ambiguous Pairings
How does Bayes-TrEx compare to the test set?
How does Bayes-TrEx compare to the test set?

Test set exposes more mislabelings than true classification failures.
How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled  Misclassified

Test set exposes more mislabelings than true classification failures.
How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled Misclassified

Test set exposes more mislabelings than true classification failures.
How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled  Misclassified

Test set exposes more mislabelings than true classification failures.
How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled Misclassified

Test set exposes more mislabelings than true classification failures.

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled Misclassified

Test set exposes more mislabelings than true classification failures.
How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled: 7
Misclassified: Ø

Test set exposes more mislabelings than true classification failures.

60/84 MNIST
42/93 Fashion-MNIST
Bayes-TrEx

Class Boundaries

High Confidence Failures
Limitation 1: Trends & Coverage Measures

\[ P(\bar{y} = \text{Corgi}) = 0.5 \text{ Level Set} \]

Relaxed Formulation

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
Limitation 2: Latent Space Dimensionality

Object 1: {
    shape: \{cube, sphere, cylinder\},
    color: \{cyan, ..., green\},
    material: \{rubber, metal\},
    x-coord: \[0,1\],
    y-coord: \[0,1\],
    r-coord: \[0, 2\pi\]
},
...
Object 5: {
    shape: \{cube, sphere, cylinder\},
    color: \{cyan, ..., green\},
    material: \{rubber, metal\},
    x-coord: \[0,1\],
    y-coord: \[0,1\],
    r-coord: \[0, 2\pi\]
}
Limitation 3: Transparency for Classification Tasks

Input Image
Limitation 2: Transparency for Classification Tasks

In Review & On ArXiv (2012.13615):
Zhou et al., RoCUS: Robot Controller Understanding via Sampling

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021
Bayes-TrEx

Paper, code, & many more experiments: github.com/serenabooth/bayes-trex

Serena Booth*  
@SerenaLBooth

Yilun Zhou*  
yilun@mit.edu

Ankit Shah  
@ankitjs

Julie Shah  
@julie_a_shah

Follow up (robotics): arxiv.org/abs/2012.13615