A Benchmark of in-the-Wild Distribution Shift over Time

Huaxiu Yao¹*  Caroline Choi¹*  Bochuan Cao²  Yoonho Lee¹  Pang Wei Koh¹  Chelsea Finn¹

¹Stanford University, ²Pennsylvania State University
Temporal Distribution Shift

Distribution shift that arise from the passage of **time**.

Predicting flu incidence from search queries

2013: **predicting double** the incidence

0.97 mean correlation with CDC data

Distribution shift over time

Training Distribution (acc: 97.99%)

Test Distribution (acc: 79.50%)

Time

1930s 1960s 2000s 2010s

... ...
Challenges in Existing Datasets

- Don’t focus on natural temporal distribution shifts
Challenges in Existing Datasets

No clear temporal performance drops

Criteria for Selecting Dataset

1. Temporal distribution shift with performance drops

   - Training 2004 - 2009: 91.66%
   - Test 2010 - 2014: 89.66%

2. Gradual temporal distribution shifts

   - MIMIC-Mortality
   - Drug-BA
# Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Yearbook</th>
<th>FMoW</th>
<th>MIMIC-IV</th>
<th>HuffPost</th>
<th>arXiv</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input (x)</strong></td>
<td>yearbook photos</td>
<td>satel. image</td>
<td>diagnosis, treatment (ICD9)</td>
<td>article headline</td>
<td>paper title</td>
</tr>
<tr>
<td><strong>Prediction (y)</strong></td>
<td>gender</td>
<td>land use</td>
<td>readmission, mortality</td>
<td>news tag</td>
<td>primary category</td>
</tr>
<tr>
<td><strong># Examples</strong></td>
<td>37,189</td>
<td>118,886</td>
<td>270,617</td>
<td>63,907</td>
<td>2,057,952</td>
</tr>
</tbody>
</table>

## Train Example
- **Female**
- **Residential**
- Diagnosis: 560, 998, 788, 278, E878, 311, V88, V10, 266, 272
- Treatment: 456, 545
- Readmission: No; Mortality: No

## Test Example
- **Female**
- **Park**
- Diagnosis: 155, 456, 452, 572
- Treatment: 423, 549, 990, 990
- Readmission: Yes; Mortality: Yes

---

- Killer Fail: How Romney's Broken Orca App Cost Him Thousand of Votes
- The Limitations of Deep Learning in Adversarial Settings
- Possible Autopilot Use Probed After Tesla Crashes at 60mph
- Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation
Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Yearbook</th>
<th>FMoW</th>
<th>MIMIC-IV</th>
<th>HuffPost</th>
<th>arXiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>yearbook</td>
<td>satel. image</td>
<td>diagnosis, treatment (ICD9)</td>
<td>article headline</td>
<td>paper title</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>gender</td>
<td>land use</td>
<td>readmission, mortality</td>
<td>news tag</td>
<td>primary category</td>
</tr>
<tr>
<td># Examples</td>
<td>37,189</td>
<td>118,886</td>
<td>270,617</td>
<td>63,907</td>
<td>2,057,952</td>
</tr>
</tbody>
</table>

Train Example
- Female
- Residential
- Diagnosis: 560, 998, 788, 278, E878, 311, V88, V10, 266, 272
- Treatment: 456, 545
- Readmission: No; Mortality: No
- Killer Fall: How Romney's Broken Orca App Cost Him Thousand of Votes
- The Limitations of Deep Learning in Adversarial Settings

Test Example
- Female
- Park
- Diagnosis: 155, 456, 452, 572
- Treatment: 423, 549, 990, 990
- Readmission: Yes; Mortality: Yes
- Possible Autopilot Use Probed After Tesla Crashes at 60mph
- cs.CR
- Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation
- TECH
Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Yearbook</th>
<th>FMoW</th>
<th>MIMIC-IV</th>
<th>HuffPost</th>
<th>arXiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>yearbook photos</td>
<td>satel. image</td>
<td>diagnosis, treatment (ICD9)</td>
<td>article headline</td>
<td>paper title</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>gender</td>
<td>land use</td>
<td>readmission, mortality</td>
<td>news tag</td>
<td>primary category</td>
</tr>
<tr>
<td># Examples</td>
<td>37,189</td>
<td>118,886</td>
<td>270,617</td>
<td>63,907</td>
<td>2,057,952</td>
</tr>
</tbody>
</table>

Train Example
- Female
- Residential
- Diagnosis: 560, 998, 788, 278, E878, 311, V88, V10, 266, 272
- Treatment: 456, 545
- Readmission: No; Mortality: No
- Killer Fall: How Romney's Broken Orca App Cost Him Thousand of Votes
- The Limitations of Deep Learning in Adversarial Settings

Test Example
- Female
- Park
- Diagnosis: 155, 456, 452, 572
- Treatment: 423, 549, 990, 990
- Readmission: Yes; Mortality: Yes
- Possible Autopilot Use Probed After Tesla Crashes at 60mph
- Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation
Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Yearbook</th>
<th>FMoW</th>
<th>MIMIC-IV</th>
<th>HuffPost</th>
<th>arXiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>yearbook</td>
<td>satel. image</td>
<td>diagnosis, treatment (ICD9)</td>
<td>article headline</td>
<td>paper title</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>gender</td>
<td>land use</td>
<td>readmission</td>
<td>mortality</td>
<td>news tag</td>
</tr>
<tr>
<td># Examples</td>
<td>37,189</td>
<td>118,886</td>
<td>270,617</td>
<td>63,907</td>
<td>2,057,952</td>
</tr>
</tbody>
</table>

Train Example
- Female
- Residential
- Diagnosis: 560, 998, 788, 278, E878, 311, V88, V10, 266, 272
- Treatment: 456, 545
- Readmission: No; Mortality: No

Test Example
- Female
- Park
- Diagnosis: 155, 456, 452, 572
- Treatment: 423, 549, 990, 990
- Readmission: Yes; Mortality: Yes

Killer Fall: How Romney’s Broken Orca App Cost Him Thousand of Votes
The Limitations of Deep Learning in Adversarial Settings
Possible Autopilot Use Probed After Tesla Crashes at 60mph
Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation
Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Yearbook</th>
<th>FMoW</th>
<th>MIMIC-IV</th>
<th>HuffPost</th>
<th>arXiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>yearbook photos</td>
<td>satel. image</td>
<td>diagnosis, treatment (ICD9)</td>
<td>article headline</td>
<td>paper title</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>gender</td>
<td>land use</td>
<td>readmission</td>
<td>mortality</td>
<td>news tag</td>
</tr>
<tr>
<td># Examples</td>
<td>37,189</td>
<td>118,886</td>
<td>270,617</td>
<td>63,907</td>
<td>2,057,952</td>
</tr>
</tbody>
</table>

Train Example
- Female
- Residential
- Diagnosis: 560, 998, 788, 278, E878, 311, V88, V10, 266, 272
- Treatment: 456, 545
- Readmission: No; Mortality: No

Test Example
- Female
- Park
- Diagnosis: 155, 456, 452, 572
- Treatment: 423, 549, 990, 990
- Readmission: Yes; Mortality: Yes

Killer Fail: How Romney's Broken Orca App Cost Him Thousand of Votes
- TECH
- cs.CR

The Limitations of Deep Learning in Adversarial Settings
- Possible Autopilot Use Probed After Tesla Crashes at 60mph
- TECH
- cs.LG

Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation
Two Evaluation Strategies

• **Eval-Fix**

• **Eval-Stream**
Gaps between ID and OOD performance

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ID</th>
<th>OOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yearbook</td>
<td>97.99%</td>
<td>79.50%</td>
</tr>
<tr>
<td>FMoW</td>
<td>60.88%</td>
<td>51.99%</td>
</tr>
<tr>
<td>MIMIC-Readmission</td>
<td>69.90%</td>
<td>58.51%</td>
</tr>
<tr>
<td>MIMIC-Mortality</td>
<td>90.86%</td>
<td>69.74%</td>
</tr>
<tr>
<td>Huffpost</td>
<td>79.40%</td>
<td>68.71%</td>
</tr>
<tr>
<td>arXiv</td>
<td>53.78%</td>
<td>45.94%</td>
</tr>
</tbody>
</table>
Experiments – Eval-Fix

Most continual learning and invariant learning approaches do not show substantial improvements compared to ERM and Fine-tuning.
Wild-Time Package

Load Dataset

```python
>>> import argparse
>>> from WildTime import dataloader, baseline_trainer
# Load the corresponding config for a specific baseline and dataset
>>> from WildTime.configs.eval_fix.configs_fmow import configs_fmow_ewc
>>> configs = argparse.Namespace(**configs_fmow_ewc)

# If you only need data, you only need the get_data method
>>> fmow_data = dataloader.getdata("fmow", configs)

# If you need to run a baseline, use the following method
>>> baseline_trainer.train(configs)
```
A Benchmark of in-the-Wild Distribution Shift over Time

Code, paper and contact info at https://wildtime.github.io