Background Invariance by Adversarial Learning
ICPR 2020, 10–15th January 2021

Ricardo Cruz 1,2
Ricardo M. Prates 3,4
Eduardo F. Simas Filho 4
Joaquim F. Pinto Costa 2
Jaime S. Cardoso 1,2

1 INESC TEC, 2 University of Porto
3 Federal University of São Francisco Valley
4 Federal University of Bahia
Motivation

**Goal:** classify electrical insulators.

**Problem:** training set background (indoors) $\neq$ testing set background (outdoors).
Motivation

Unfortunately, a model with accuracy of 97.3% can drop to as low as 11% (random) just by changing the background.

- A good CNN can drop precipitously, when the object background is changed.
- Notice how a model with accuracy of 97.3% can drop to as low as 11% (random) just by changing the background.
Proposed Method: Overview

Traditional model.
Proposed Method: Overview

- **Background generator**: creates new backgrounds.
- **Mask generator**: helps injecting the new backgrounds.
Proposed Method: Details

\[
\min_{f, f_m} \max_{f_b} \text{Loss}
\]

Step 1. **Model** \( f \) is optimized to minimize a loss \( \mathcal{L}(y, f(x)) \) using an image \( x \) as input with label \( y \) as the ground-truth.

Step 2.

**Mask generator** \( f_m \) is trained to produce a mask \( \hat{m} \) (\( \hat{m} \in [0, 1] \)).

\begin{itemize}
  \item U-Net is used as the architecture.
  \item The image is segmented through a element-wise product, \( x' = x \odot \hat{m} \).
\end{itemize}

Step 3. **Background generator** \( f_b \) transforms noise \( z \) into a background image \( \hat{b} \).

\begin{itemize}
  \item Model \( f \) is also optimized during Step 3.
  \item In the case of monochrome images, a constrain term is added to disallow the background from filling over half the pixels.
\end{itemize}
Proposed Method: Details

- The goal is to (during training) be able to place the object in a multitude of contexts (backgrounds).
- Try to find backgrounds that “fool” the model, thus making it robust.

Examples of the dynamic background along the epochs.
Experiments

- Datasets: MNIST (10 digits) and Fashion-MNIST (10 pieces of clothe).
  - (artificially enhanced by introducing backgrounds as illustrated in the figure)

(a) Original (b) Stripes (c) Board (d) Border (e) Circles (f) Clock (g) Random

Backgrounds introduced for MNIST and Fashion-MNIST.
Results: Accuracy (%)  

<table>
<thead>
<tr>
<th></th>
<th>MNIST</th>
<th>Fashion-MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>97.3</td>
<td>90.1</td>
</tr>
<tr>
<td>Attention</td>
<td>93.4</td>
<td>81.2</td>
</tr>
<tr>
<td>Proposal</td>
<td>94.9</td>
<td>70.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>38.0</th>
<th>21.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.4</td>
<td>24.3</td>
<td>24.6</td>
</tr>
<tr>
<td>61.4</td>
<td>36.9</td>
<td>61.5</td>
</tr>
<tr>
<td>32.9</td>
<td>36.9</td>
<td>66.5</td>
</tr>
<tr>
<td>19.7</td>
<td>28.5</td>
<td>60.9</td>
</tr>
<tr>
<td>11.2</td>
<td>29.6</td>
<td>60.8</td>
</tr>
</tbody>
</table>

▶ Interestingly, the attention mechanism results only negligibly improve on the baseline classifier.
  ▶ This mechanism works by cropping the image and, not surprisingly, it was found to perform best in the border case (with over 50% accuracy).
▶ The proposed method is resilient to a wide range of testing backgrounds.
Results for the Case Study

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>71.9</td>
</tr>
<tr>
<td>Attention</td>
<td>45.8</td>
</tr>
<tr>
<td>Proposal</td>
<td>88.7</td>
</tr>
</tbody>
</table>

Impact of background generator

<table>
<thead>
<tr>
<th>Background</th>
<th>Validation Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>76.3</td>
</tr>
<tr>
<td>Noise</td>
<td>59.6</td>
</tr>
<tr>
<td>Proposal</td>
<td>88.7</td>
</tr>
<tr>
<td>Real backgrounds</td>
<td>93.8</td>
</tr>
</tbody>
</table>

Sensitivity analysis – baseline

Sensitivity analysis – proposal
Conclusion

1. Sometimes it is easier to collect data inside a studio rather than in the real world – for example when training a drone.
2. Unfortunately, convolutional neural networks’ performance degrades terribly when used in new backgrounds.
3. An adversarially trained model is proposed where the model tries to minimize its loss while a generator injecting new backgrounds to maximize the loss.
4. The proposed method was evaluated for the task of classification, but it could potentially be used for other tasks:
   ▶ regression problems
   ▶ segmentation
   ▶ reinforcement learning tasks.
Final Thoughts

Typically adversarial training has been used to improve generators that produce content for the user. But at INESC TEC, we have been using adversarial training to improve the classifier itself. If you like the idea, you may also want to consult:


Thank you!

Ricardo Cruz <ricardo.p.cruz@inesctec.pt>
Ricardo M. Prates, Eduardo F. Simas Filho
Joaquim F. Pinto Costa, Jaime S. Cardoso