Background Invariance by Adversarial Learning

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

\textsuperscript{1} INESC TEC \hspace{1em} \textsuperscript{2} University of Porto \hspace{1em} \textsuperscript{3} Federal University of São Francisco Valley \hspace{1em} \textsuperscript{4} Federal University of Bahia

Corresponding author: rpcruz@inesctec.pt

Motivation

- CNNs fail to distinguish foreground (object) from background.
- Notice how a model with accuracy of 97.3\% can drop to as low as 11\% (random) just by changing the background.

![Figure 1: Accuracy (%) of vanilla CNN trained for MNIST.](Image)

**Figure 1:** Accuracy (%) of vanilla CNN trained for MNIST.

Related Work

Little literature exists on making CNNs background invariant. One work proposes an attention mechanism to avoid artifacts, particularly irregular borders [1].

- Two classifiers: a global CNN $G$, and a local CNN $L$.
- $G$ is trained to classify the entire image $x$.
- A activation maps from a truncated $G'$ are then used to find a bounding box of the object and crop $x'$.
- $L$ is then trained using the smaller $x'$

![Figure 2: Attention mechanism diagram.](Image)

**Figure 2:** Attention mechanism diagram.

Two disadvantages are immediate:

- $L$ operates on a rectangular cropped version of the image and therefore is still influenced by artifacts that remain inside that rectangle.
- Model $G$ is still influenced by artifacts because it did not have the benefit of being trained against the artifacts. While such artifacts are not presented in the training set, they could be generated in a controlled fashion, as our method now proposes:

- The goal is to (during training) be able to place the object in a multitude of contexts (backgrounds).
- Focus on “what” the object is rather than “where” the object is.
- The backgrounds are generated in an adversarial manner.
- However, the insertion of adversarial backgrounds in the image cannot be allowed to destroy the concept (class) one is trying to learn.

- A model $f$ is optimized to minimize a loss $L(y, f(x))$.
- A mask generator $f_m$ is trained to produce a mask $m \in [0,1]$.
  - This U-Net is trained unsupervisedly by finding the best mask that minimizes the previous loss, $L(f(x \odot f_m(x)), y)$.
- A background generator $f_b$ transforms noise $z$ into a background $b$ image.
  - The trick: it (adversarially) maximize the loss $L$.

Summary: Model $f$ tries to minimize a loss while $f_b$ tries to find backgrounds that maximize it:

$$\min_{f_m} \max_{f_b} \sum_{i=1}^{N} L\left(f\left(m_i \odot x_i + b_j \odot (1 - m_j)\right), y_i\right).$$

This is inspired by literature in adversarial training and GANs.

![Figure 4: Background augmentation along the epochs.](Image)

**Figure 4:** Background augmentation along the epochs.

Experiments

- The attention mechanism only negligibly improves on the baseline classifier.
- The proposed method is resilient to a wide range of testing backgrounds.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>97.3</th>
<th>38.0</th>
<th>24.3</th>
<th>61.4</th>
<th>32.9</th>
<th>19.7</th>
<th>11.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla CNN</td>
<td></td>
<td>97.3</td>
<td>38.0</td>
<td>24.3</td>
<td>61.4</td>
<td>32.9</td>
<td>19.7</td>
<td>11.2</td>
</tr>
<tr>
<td>Attention</td>
<td></td>
<td>93.4</td>
<td>28.1</td>
<td>26.8</td>
<td>57.3</td>
<td>40.1</td>
<td>29.3</td>
<td>25.1</td>
</tr>
<tr>
<td>Proposal</td>
<td></td>
<td>94.9</td>
<td>92.3</td>
<td>76.8</td>
<td>93.1</td>
<td>93.7</td>
<td>70.8</td>
<td>86.2</td>
</tr>
</tbody>
</table>

Table 1: Accuracy (%) using MNIST.

Conclusion

- Sometimes it is easier to collect data inside a studio rather than in the real world – for example when training a drone. Unfortunately, a CNN does terrible when used in new backgrounds.
- An adversarially trained data augmentation method is proposed. The proposed method can be used for classification, regression, segmentation, reinforcement learning, etc.

References


This work is financed by ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation – COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e Tecnologia within project POCI-01-0145-FEDER-028857. Furthermore, Ricardo Cruz was supported by Ph.D grant SFRH/BD/122249/2016, also provided by FCT.