Training a CNN to be Background Invariant

Ricardo Cruz and Jaime S. Cardoso

INESC TEC – Faculty of Engineering, University of Porto

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.

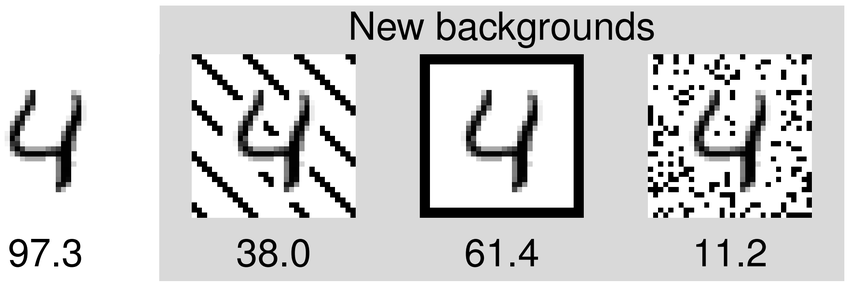

Figure 1: Accuracy (%) of vanilla CNN trained for MNIST.

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^T are then used to find a bounding box of the object and crop x'.
- L is then trained using the smaller x'

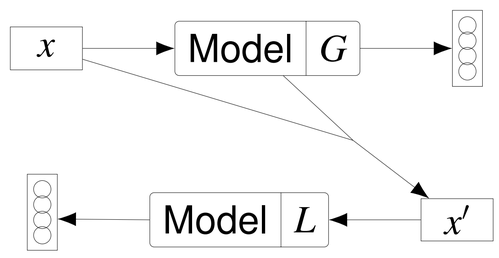

Figure 2: Attention mechanism diagram.

Two disadvantages are immediate:

i. L operates on a rectangular cropped version of the image and therefore is still influenced by artifacts that remain inside that rectangle.

ii. 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:

Proposed Method

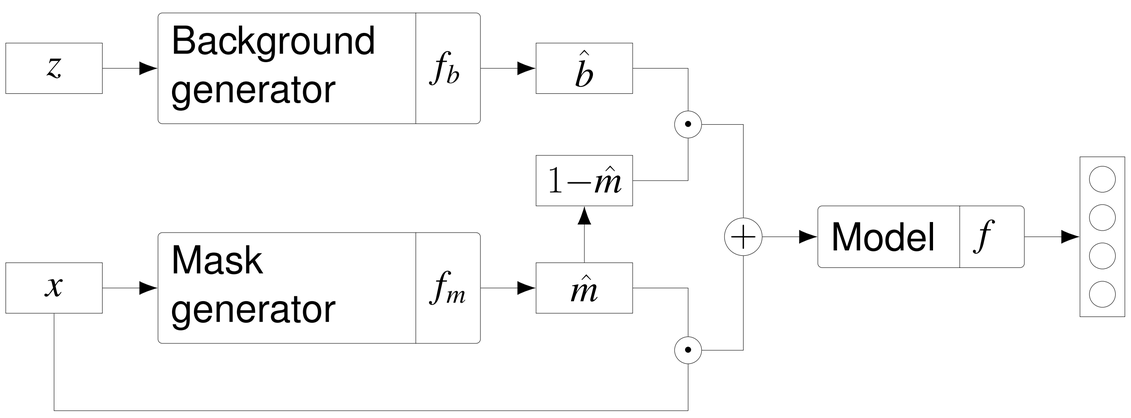

Figure 3: Proposed method.

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

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

$$\min_{f, f_m} \max_{f_b} \sum_{i=1}^{N} \mathcal{L}\left(f\left(\hat{m}_i \odot x_i + \hat{b}_j \odot (1 - \hat{m}_i)\right), y_i\right).$$

This is inspired by literature in adversarial training and GANs.

(a) $t = 1$ (b) $t = 10$ (c) $t = 20$ (d) $t = 30$ (e) $t = 40$

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.

Vanilla CNN	97.3	38.0	24.3	61.4	32.9	19.7	11.2
Attention	93.4	28.1	26.8	57.3	40.1	29.3	25.1
Proposal	94.9	92.3	76.8	93.1	93.7	70.8	86.2

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

[1] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv*, 2018.

RECPAD 2020, 30 October 2020, Portugal

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/122248/2016, also provided by FCT.