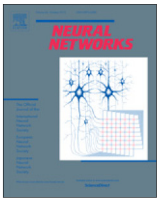
Neural Networks 104 (2018) 40–49

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Using a model of human visual perception to improve deep learning[☆]

Michael Stettler, Gregory Francis [*]

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Purdue University, Department of Psychological Sciences, 703 Third Street, West Lafayette, IN 47906, United States

ARTICLE INFO

Article history:
Received 28 August 2017
Received in revised form 4 January 2018
Accepted 8 April 2018
Available online 17 April 2018

Keywords:
Segmentation
Visual perception
Deep learning

ABSTRACT

Deep learning algorithms achieve human-level (or better) performance on many tasks, but there still remain situations where humans learn better or faster. With regard to classification of images, we argue that some of those situations are because the human visual system represents information in a format that promotes good training and classification. To demonstrate this idea, we show how occluding objects can impair performance of a deep learning system that is trained to classify digits in the MNIST database. We describe a human inspired segmentation and interpolation algorithm that attempts to reconstruct occluded parts of an image, and we show that using this reconstruction algorithm to pre-process occluded images promotes training and classification performance.

1. Introduction

Deep learning algorithms achieve human-level performance in many pattern recognition tasks; but these abilities appear to be the result of mechanisms very different from humans (Volokitin, Roig, & Poggio, 2017). In particular, deep learning algorithms typically need large amounts of data to perform well, while humans can often learn a classification task with relatively small data sets and examples (e.g., Erickson & Kruschke, 1998; Helie & Ashby, 2012). We hypothesize that the differences between humans and deep learning algorithms reflect, in part, the representation of information. Deep learning systems develop formats that promote learning for a particular situation. In contrast, millions of years of evolution have guided the development of the human brain to represent relevant information in a format that promotes efficient and effective learning. For situations that correspond to those evolutionary pressures, we anticipate that machine learning systems could benefit by implementing aspects of human information processing.

To demonstrate the issue, in this paper we consider classification of hand written numbers in the MNIST database (LeCun, Bottou, Bengio, & Haffner, 1998); a task that has been largely solved by deep learning algorithms (Cireşan, Meier, Gambardella, & Schmidhuber, 2010; Cireşan, Meier, & Schmidhuber, 2012). Example images from the MNIST database are provided in the "Original" column of Figs. 1 and 2 (we treated all zero values in the MNIST images as black and all nonzero values as a middle gray color). The black curve with rectangle symbols in Fig. 3 shows performance on a test set of 10,000 images for a convolutional neural network (CNN) using the same architecture as the TensorFlow library tutorial (Abadi et al., 2016; TensorFlow, 2017). Each point corresponds to performance after training on the indicated number of images. The performance essentially replicates the tutorial by showing accuracy above 90% correct for small training sets and a score of 99% after training on the full set of 55,000 training images. (Performance here is a bit lower than for the tutorial, presumably because we digitized the images into black and gray and thus lost some grayscale information.)

The "Occluded" column of Figs. 1 and 2 show the same MNIST images with a set of four occluding white bars. It is clear that the bars hide some information about the digits, but it is also clear that humans can often "see through" the bars to identify the occluded number. What about the CNN? The blue line with upright triangle symbols in Fig. 3 shows accuracy on a test set when the CNN is trained on occluded images and tested on original (unoccluded) images. Accuracy for the CNN peaks at around 80%, which indicates that the occlusion disturbs the CNN training process. One might suspect that the reduced performance, relative to the no-occlusion condition, is because the occluding elements remove information that is needed to discriminate between the MNIST digits. While there is some truth to this suspicion, it does not fully account for the magnitude of the performance drop. The gray dashed line with inverted triangle symbols shows CNN performance when the training and test images are both occluded by a set of three horizontal bars. Performance is only slightly worse than when there is no occlusion in either training or test sets. Thus, training on occluded images is sufficient to classify occluded digits, but the

[☆] Fully documented implementation is available on GitHub: Human-Inspired Segmentation and Interpolation.

[*] Corresponding author.
E-mail address: gfrancis@purdue.edu (G. Francis).

https://doi.org/10.1016/j.neunet.2018.04.005
0893-6080/© 2018 Elsevier Ltd. All rights reserved.

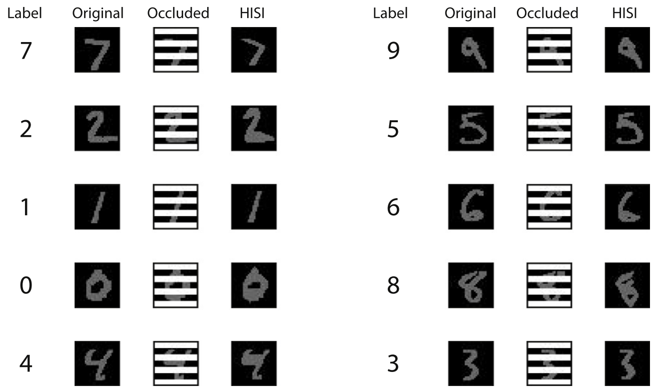


Fig. 1. Example images from the MNIST database that can be reconstructed fairly well when processed by the human-inspired segmentation and interpolation (HISI) algorithm.

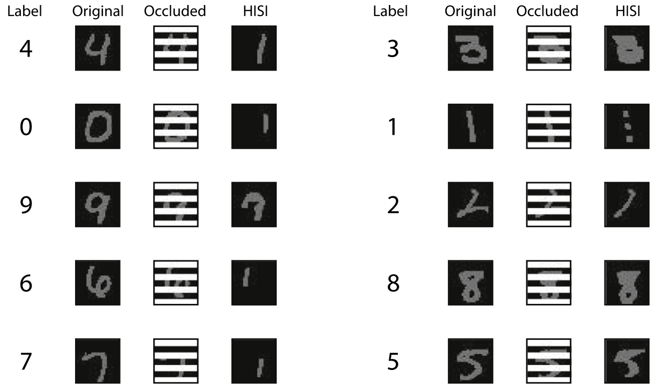


Fig. 2. Example images from the MNIST database that are reconstructed rather badly when processed by the HISI algorithm.

representation of that information is not robust when the training and tests sets differ with regard to occlusion. More generally, we suspect that there will often be situations where a deep learning system is tested on images that have occlusions, or other attributes, that were not part of the training set. Our main idea is that human-related properties of the visual system provide a general purpose way of dealing with these situations.

CNN performance is truly dismal if it is trained on unoccluded original images and tested on images occluded with four bars. The red line with circle symbols in Fig. 3 shows that performance for this condition is around 16%, which is only a little above chance. During training the CNN should develop normally, but the learned patterns are not helpful for classifying the information available in occluded test images.

This situation may seem somewhat artificial (although one can imagine situations where a CNN is trained on high quality data and then used to classify lower quality data); but it demonstrates how the CNN's representation of relevant information derives from the training set. Indeed, such a result is hardly surprising because when training on the original MNIST images, the CNN takes full advantage of the available information to maximize classification performance. Such optimization causes problems when some of the expected information is occluded and therefore unavailable to the CNN. Indeed, when using a CNN to classify occluded images standard practice would be to make sure that the training set includes occluded images. However, the properties of the test images may be unknown at the time of training, so the standard practice restricts the use of CNNs to only certain types of situations.

Moreover, even training with occluded images might not be sufficient in a worst case scenario where there is little overlap of visible regions between the training and test sets. The cyan line with diamond symbols in Fig. 3 demonstrates the CNN's performance when it is trained on occluded images and then tested on images with complementary occlusion, as in Fig. 4. In this worst case situation, classification accuracy is much worse than for the original images, with very poor performance for small training sets and peak performance maxing out just above 45%. Again, such a result is hardly surprising because the test set is dramatically different from the training set in terms of the location and type of information that is presented to the CNN.

While the importance of having a training set that matches the test set is well recognized among scientists who utilize machine learning, what is striking to us is that many of the occluded images can be well identified by human observers. One of the key attributes of human perception is that it represents visual information in terms of "objects" and "groups" (Anderson, Laurent, & Yantis, 2005; Moore, Mordkoff, & Enns, 2007; Moore, Yantis, & Vaughan, 1998). When looking at the images in the Occluded columns of Figs. 1 and 2 human observers generally find it easy to identify the occluded number because the human visual system has evolved to select and segment occluding objects, to connect visible parts of objects, and to reconstruct the full occluded object (Francis, Manassi, & Herzog, 2017; Grossberg, 1994; Grossberg & Mingolla, 1985a, b; Raizada & Grossberg, 2001). By default, a CNN trained on the original MNIST images will have no need for

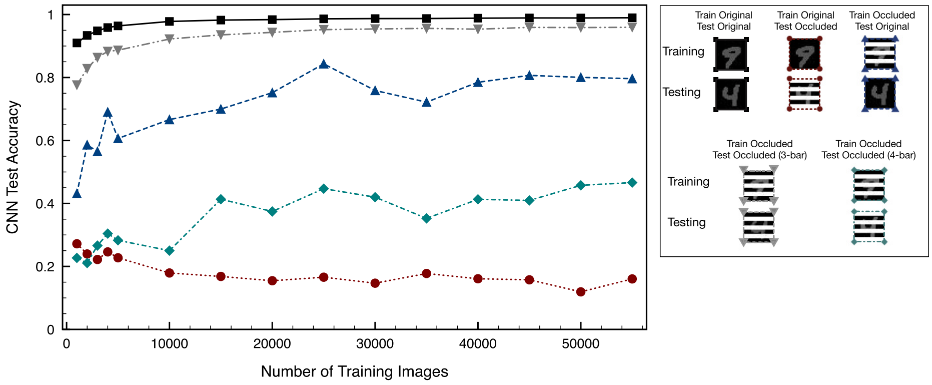


Fig. 3. CNN test set accuracy as a function of the number of training images for horizontal occluding bars. The different curves correspond to the different training/testing conditions. The CNN performs very well on the original MNIST images (black squares). Training and testing with a 3-bar occluder (gray inverted triangles) slightly impairs CNN accuracy. Training on partially, 3-bar, occluded images and testing on unoccluded images (blue upright triangles) somewhat reduces performance. Training on partially, 3-bar, occluded images and then testing on images with complementary, 4-bar, occlusion (cyan diamonds) also produces rather poor performance. Training on the original image and testing on partially, 4-bar, occluded images (red circles) reduces performance to almost the level of random guessing.