Top-Down Networks: A coarse-to-fine reimagination of CNNs Supplementary material

Ioannis Lelekas Nergis Tomen Silvia L. Pintea Jan C. van Gemert

Computer Vision Lab
Delft University of Technology, Netherlands

1. Exp 2: Adversarial robustness

The entire set of results for the adversarial robustness experiment is provided in figure 1. "ShiftsAttack" is a variant of Spatial attack [2], introducing only spatial shifts. TD networks exhibit enhanced robustness against attacks introducing correlated/uncorrelated noise, as well as against blurring attacks.

Figure 2 presents the robustness results for the CIFAR10-augmented and the ResNet32 architecture variants. Clearly, TD_{uni} and TD_{rev} variants exhibit enhanced robustness against spatial attacks, however, they also have similar to the BU behaviour against other attacks. This can be attributed to the increased number of filters at greater depth of the network, or equivalently increased scale of feature maps, thus greater contribution of the finer scales to the final output. However, finer scales are much more vulnerable against attacks. All in all, the reversal of the BU network for the extraction of the TD variant is not solely efficiency driven, keeping a roughly fixed computational complexity across layers, but also contributes to the network's robustness as well. Finally, we need to mention that the respective figure for the non-augmented CIFAR10 case tells the same story.

Next, figure 3 presents the respective results for reintroducing the perturbation to two of the inputs of the network. Clearly, the highest and medium scale inputs are the most vulnerable ones, except for the simpler case of the MNIST dataset. The absence or scarce information in the high frequency region, yields the medium and lowest scale inputs as the ones with the highest impact.

2. Exp 3.(a): Explainability

2.1. Imagenette training

Imagenette [5] is a 10-class sub-problem of ImageNet [1], allowing experimentation with a more realistic task, without the high training times and computational costs required for training on a large scale dataset. A set of examples, along with their corresponding labels are provided in figure 4. The datasets contains a total of 9469, 3925 training and validation samples respectively. Training samples were resized to 156×156, from where random 128×128 crops were extracted; validation samples were resized to 128×128.

We utilized a lighter version[1] of the ResNet18 architecture introduced in [3] for Imagenette training, as this is a 10-class sub-problem, incorporating the pre-activation unit of [4]. Additionally, the stride s and the kernel extent k of the first convolution for depth initialization were set to $s = 1$ and $k = 3$ respectively. Regarding training, a 128×128 crop is extracted from the original image, or its horizontal flip, while subtracting the per-pixel mean [6]; the color augmentation of [6] is also used. For the BU network a batch size of 128 is used and the network is trained for a total of 50 epochs with a starting learning rate of 0.1. As for the TD, increased memory footprint led to the reduction of the batch size to 64 and the adaptation of the starting learning rate and the total epochs to 0.05 and 80. We trained with SGD with momentum of 0.9 and a weight decay of 0.001; we also adopted a 3-stage learning rate decay scheme, where the learning rate is divided by 10 at 50% and 80% of the total epochs. Regarding performance, BU outperformed the TD variant by roughly 4%. Grad-CAM is finally utilized for generating class-discriminate localization maps of the most informative features.

2.2. Grad-CAM heatmap visualizations

Figure 5 displays some additional Grad-CAM visualizations. The visualizations are obtained by using a ResNet18 architecture for the BU networks and its corresponding TD variant. The original images are taken from the Imagenette dataset [5].

The TD model provides localized activations, focusing on certain informative aspects of the image, while the BU model focuses on large connected areas. Because of this

[1] dividing the filters of the original architecture by 2.

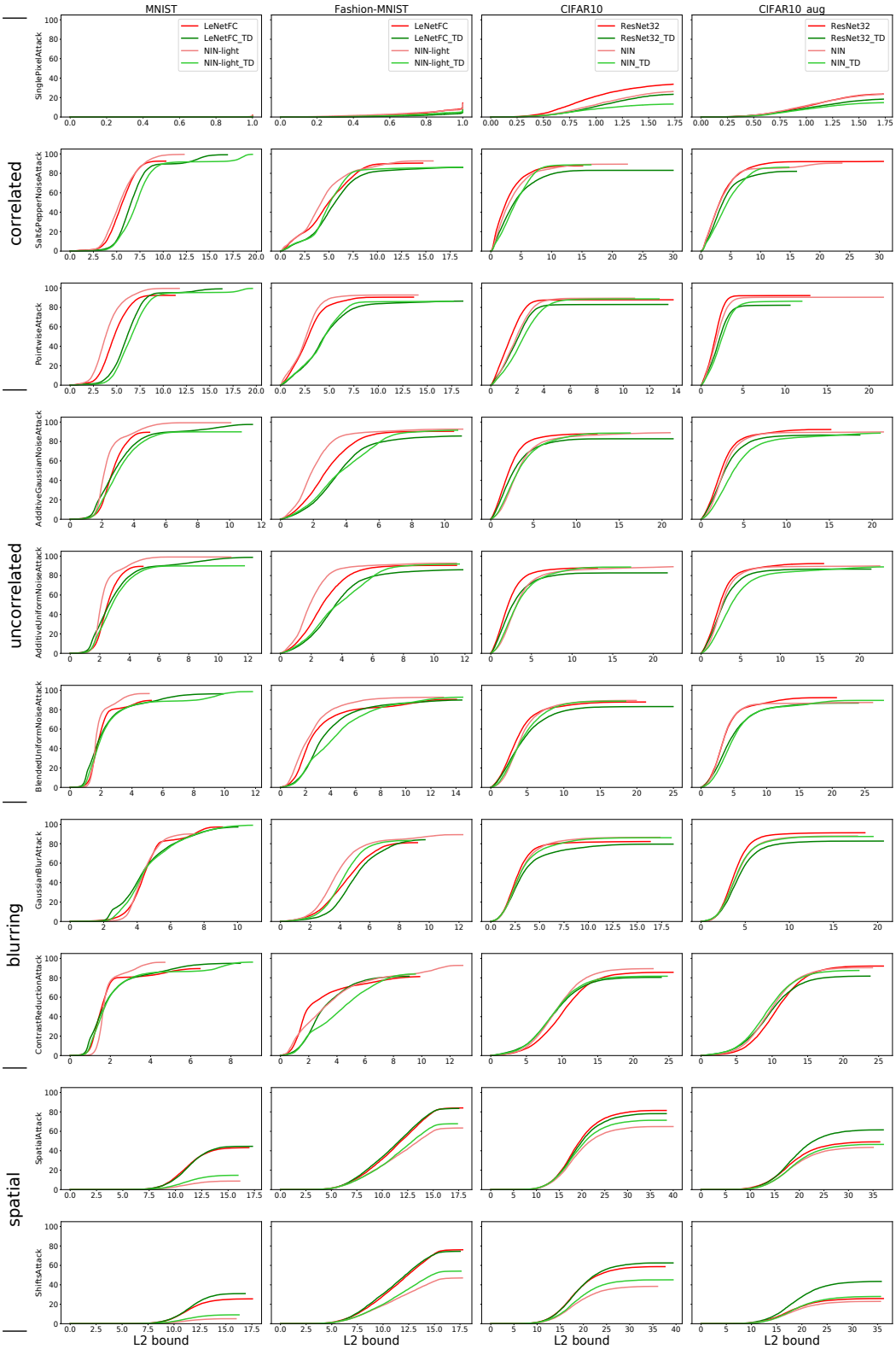

Figure 1. **Exp:2** Complete set of results for the second experiment. Plots of test accuracy loss versus the $L2$ distance between original and perturbed input, where each column corresponds to a different task. TD networks exhibit enhanced robustness against correlated/uncorrelated noise and blurring attacks.

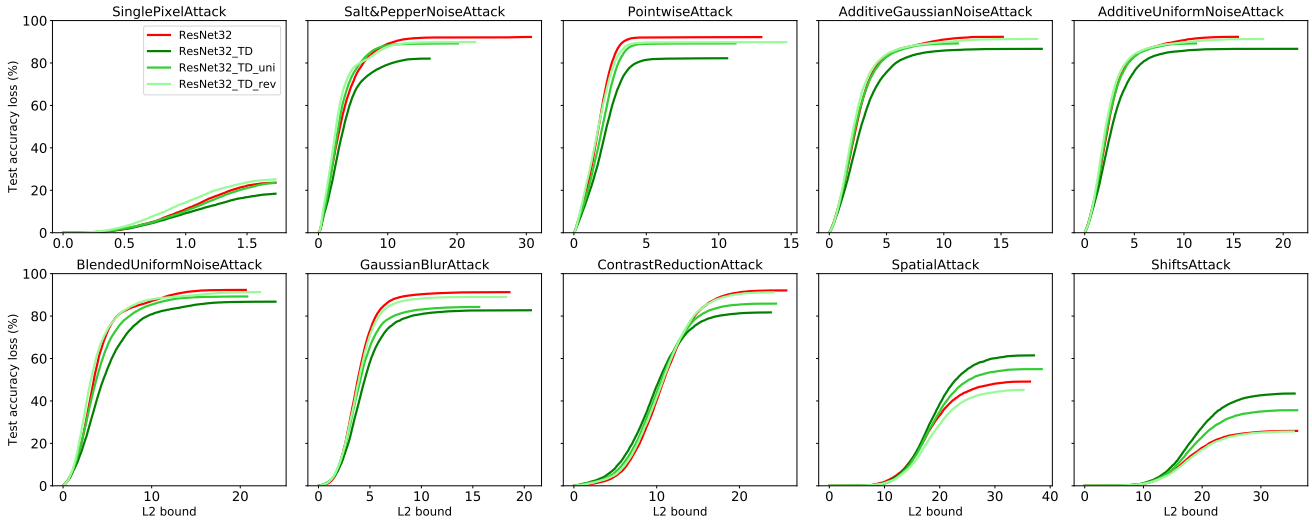

Figure 2. **Exp 2:** Test accuracy loss versus the $L2$ distance between original and perturbed input, for the CIFAR10-augmented and the ResNet32 architectures. Robustness is enhanced for the spatial attacks, but in general TD_{uni} and TD_{rev} variants exhibit similar behaviour to the BU baseline, which can be attributed to the increased filters at deeper layers.

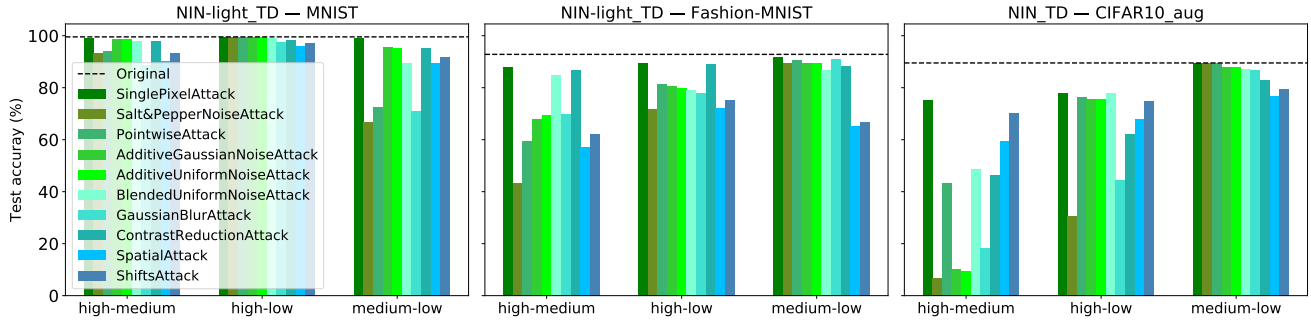

Figure 3. **Exp 2:** Reintroducing perturbations to two of the inputs of TD model when using a NIN-light backbone for MNIST and Fashion-MNIST, and the NIN backbone for CIFAR10. Clearly, perturbing the two highest scale inputs, "high-medium" has the highest impact. Regarding the case of the simpler MNIST and the information gathered in the low to mid frequency region, the medium and the lowest scale input have the highest impact instead.

difference we believe the TD model may be more precise than the BU model for tasks such as weakly-supervised object detection.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[2] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. *CoRR*, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on Computer Vision*, pages 630–645, 2016.

[5] FastAI Jeremy Howard. The imagenette dataset. `https://github.com/fastai/imagenette`.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

Figure 4. **Exp 3.(a):** Validation samples from the Imagenette dataset [5], along with their corresponding ground truth labels. Samples are resized to 128×128.

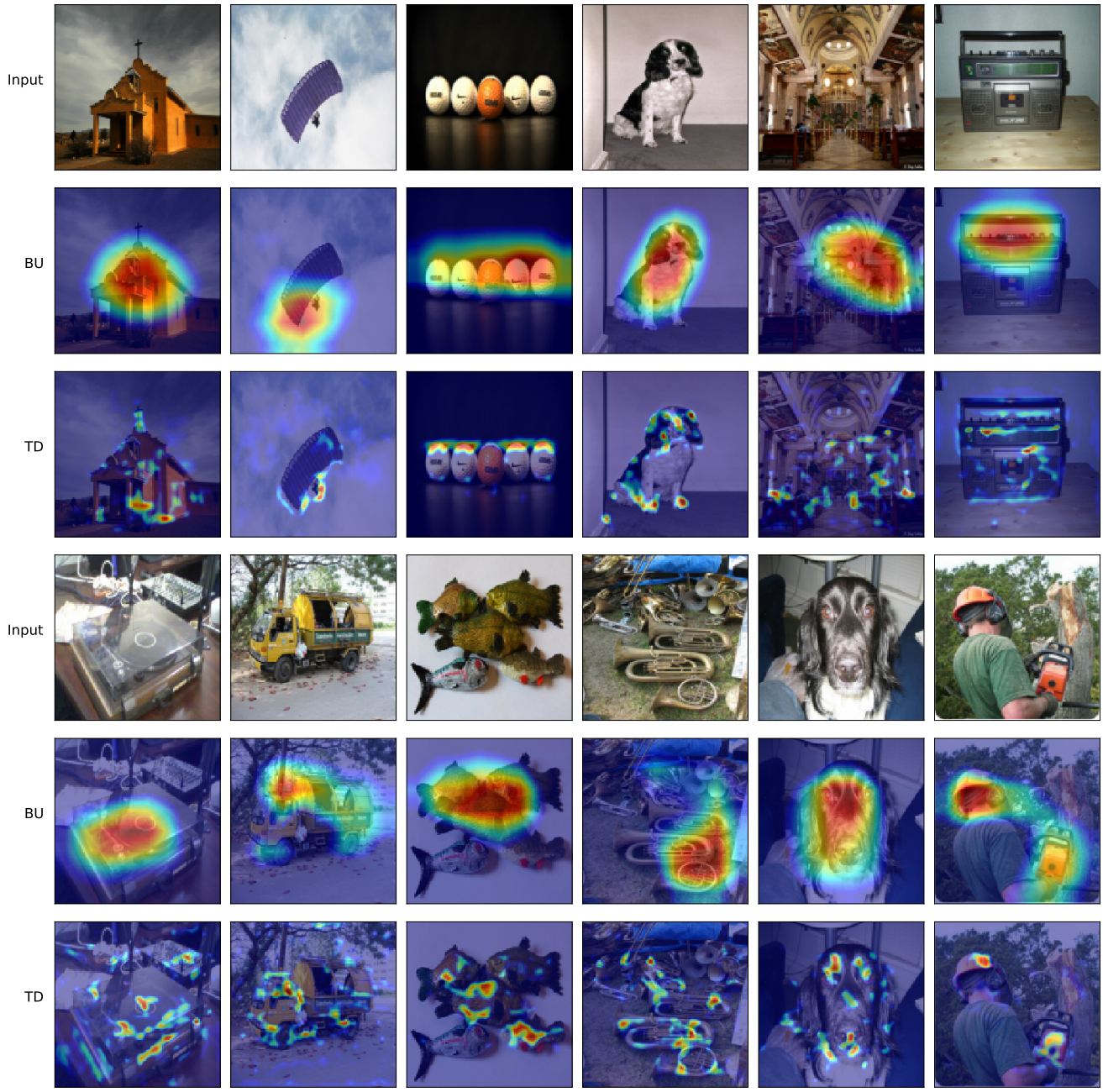

Figure 5. **Exp 3.(a):** Grad-CAM heatmaps visualization on validation images from the Imagenette dataset [5], using a ResNet18 architecture for BU and its corresponding TD variant. All images are correctly classified. Rows 1 and 4 show the original Imagenette images; rows 2 and 5 show the BU heatmaps, while rows 3 and 6 visualize the TD heatmaps. Focusing on local information rather than global information, may help the TD to be more precise for object detection than the BU model.