On the Robustness of Convolutional Neural Networks Regarding Transformed Input Images

Frederik Timme, Jochen Kerdels and Gabriele Peters

Chair of Human-Computer Interaction, Faculty of Mathematics and Computer Science, University of Hagen,
Universitätsstraße 47, 58097 Hagen, Germany

Keywords: Convolutional Neural Networks, Performance Evaluation, Transformations, Data Augmentation.

Abstract: Convolutional Neural Networks (CNNs) have become the dominant and arguably most successful approach for the task of image classification since the release of AlexNet in 2012. Despite their excellent performance, CNNs continue to suffer from a still poorly understood lack of robustness when confronted with adversarial attacks or particular forms of handcrafted datasets. Here we investigate how the recognition performance of three widely used CNN architectures (AlexNet, VGG19 and ResNeXt) changes in response to certain input data transformations. 10,000 images from the ILSVRC2012s validation dataset were systematically manipulated by means of common transformations (*translation*, *rotation*, *color change*, *background replacement*) as well as methods like *image collages* and jigsaw-like *puzzles*. Both the effect of single and combined transformations are investigated. Our results show that three of these input image manipulations (*rotation*, *collage*, and *puzzle*) can cause a significant drop in classification accuracy in all evaluated architectures. In general, the more recent VGG19 and ResNeXt displayed a higher robustness than AlexNet in our experiments indicating that some progress has been made to harden the CNN approach against malicious or unforeseen input.

1 INTRODUCTION

In the last decade Convolutional Neural Networks (CNNs) have successfully replaced traditional keypoint based detection methods like SIFT (Lowe, 2004) or SURF (Bay et al., 2006) as a standard approach for object recognition and image classification (Zheng et al., 2018). On the one hand, these networks repeatedly set new records regarding classification accuracy and performance on well established test sets while, on the other hand, they still tend to show limited robustness when being exposed to input data that differs from their original training and test data (Hendrycks and Dietterich, 2019) (Recht et al., 2019). Several techniques have been proposed to overcome this drawback including techniques for, e.g., data augmentation (Shorten and Khoshgoftaar, 2019) in order to generate a larger collection of more diverse training data. However, it is an open question to what extent these efforts can lead to an increased robustness regarding unexpected input data.

In this work, we systematically evaluate the performance of three popular CNN architectures (AlexNet, VGG19 and ResNeXt) in regard to their robustness against a broad range of input data transformations. The transformations we use are comprised of well known transformations traditionally used for training data augmentation as well as a number of novel, more complex transformations like collages or jigsaw-like puzzles.

The main contribution of this work is the systematic analysis of established and newly introduced transformation methods, which allows a clear comparison of their effect on the accuracy of three widely used CNN architectures.

2 RELATED WORK

Research on the performance degradation of CNNs due to altered input data can be broadly divided into two areas. One area concerns the design of *adversarial attacks* (Wiyatno et al., 2019). Here it was shown that precise, minimal changes to input images both globally (Goodfellow et al., 2014) and locally (Su et al., 2019) allow to fully control the classification output of a CNN. The other area concerns the exploration of datasets that are structurally different from the training dataset in subtle ways, using these datasets to investigate how they affect the performance of CNNs. It was shown that specifically crafted datasets (Barbu et al., 2019) as well as careful

Timme, F., Kerdels, J. and Peters, G.
On the Robustness of Convolutional Neural Networks Regarding Transformed Input Images.
DOI: 10.5220/0010107203960403
In *Proceedings of the 12th International Joint Conference on Computational Intelligence (IJCCI 2020)*, pages 396-403
ISBN: 978-989-758-475-6

selections of existing images (Hendrycks et al., 2020) can cause large drops in classification accuracy.

One approach to mitigate this lack of robustness consists in the augmentation of the training data by a broad range of different image transformations. These include simple (mostly affine) transformations like translation (Krizhevsky et al., 2012), rotation and cropping (Taylor and Nitschke, 2017), or color and brightness adjustments (Howard, 2013). More elaborated approaches use, e.g., Generative Adversarial Networks to generate style transferred images or to construct entirely new images (Mikołajczyk and Grochowski, 2018).

Evaluation of such approaches with a focus on different types of transformations and their effect on model accuracy – mostly dealing with a single type of transformation and a narrow parameter range – were performed by, e.g., (Engstrom et al., 2018) and (Azulay and Weiss, 2018).

3 MATERIALS AND METHODS

We evaluated three widely known CNNs regarding their robustness against modified input data: *AlexNet* (Krizhevsky et al., 2012), *ResNeXt50_32x4d* (Xie et al., 2017), and *VGG19* (Simonyan and Zisserman, 2015). In our experiments we modified the input data by a set of both established and novel types of transformations. Table 1 shows the transformations used as well as their parameters and value ranges. The selected value ranges are comparable to those found in other publications, but have been extended in some cases to be able to estimate the performance of the CNNs with regard to heavily modified input data. The more traditional transformations used include *translation*, *rotation*, *color change*, and *background replacement*. As more novel transformations we implemented *puzzle* and *collage* transformations.

The translation transformation moves an input image in vertical and/or horizontal directions and fills the resulting empty sections of the image with zeros (black). The rotation transformation rotates an input image around its center and fills resulting empty sections with zeros. The color change transformation alters the proportion of the green color channel of an input image via an *intensity* parameter. For values less than 1 the intensity values of all pixels in the green color channel are multiplied by this factor, for values greater than 1 the intensity values of the pixels in the other two channels are divided by this value. The background transformation replaces the non-object part of an input image with a constant color.

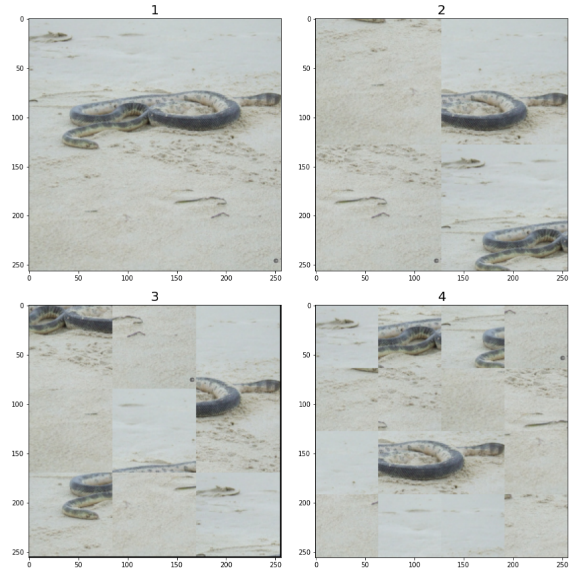

Figure 1: Puzzle transformations. The four images show examples of the puzzle transformation applied to an input image. 1: original image, 2: 2×2 puzzle, 3: 3×3 puzzle, 4: 4×4 puzzle. Tiles are shuffled randomly.

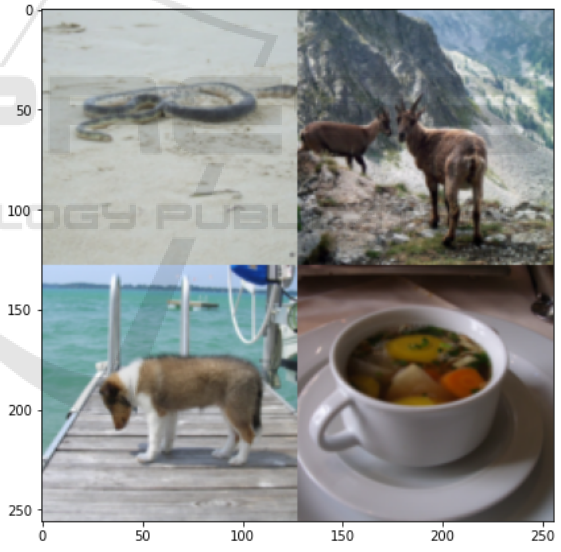

Figure 2: Collage transformation. The image shows a collage build out of four source images each belonging to a different image class.

The *puzzle* transformation (see Fig. 1) cuts an input image into $n \times n$ rows or columns of equal height or width. The resulting n^2 pieces are then randomly rearranged. The *collage* transformation divides the image area into 2×2 rows and columns of equal height and width. Four input images, each from a different input class, are then inserted into one of the four subareas (see Fig. 2). This way a collage with four objects of different target classes is created al-

Table 1: Transformations with parameters and values. The transformations applied include traditional transformations (translation, rotation, background replacement, color change) as well as novel ones (collage, puzzle). For each transformation, the parameters and value ranges as well as the step size is shown.

Transformations	Parameters	Values	Step size
Translation	Distance x	$[0\%, 40\%]$	20%
	Distance y	$[0\%, 40\%]$	20%
Rotation	Rotation angle	$[-180°, +180°]$	30°
Background replacement	Fill color	{original, black, gray, white, red}	-
Color change	Intensity	$[0.5, 1.5]$	0.25
Collage	Number of rows and columns	2	-
Puzzle	Number of rows and columns	$[1, 4]$	1

lowing to investigate a CNNs' ability to recognize objects from several target classes at the same time.

All transformations were examined individually as well as in combination. In the latter case, two transformations were applied consecutively to an input image. In these cases, background replacement was applied with black background only to reduce the number of possible combinations.

About 10,000 images of the ILSVRC2012 validation set (Russakovsky et al., 2015) and their corresponding bounding boxes served as input data. These covered 200 classes represented by 50 images each.

To ensure that all transformations were identity preserving with respect to the input data, we applied them on a sample of images. *Rotation*, *color change*, *background replacement* and *collage* turned out to be unproblematic. The *translation*, however, carried the danger of shifting the main parts of the object(s) out of the image area. Therefore, we evaluated the part of the object that was still visible after applying the translation. With a translation distance of 20% (simultaneously in x and y direction), for more than 90% of the objects contained in the dataset, the amount of pixels still visible exceeded 40%, which we assume to be easily recognizable for humans. Whether the *puzzle* transformation keeps the identity of the image class and should still be recognized by a neural network is a matter of discussion. However, this transformation can give hints about whether a network recognizes objects based on their microstructure or their macrostructure.

The analysis was performed on a Python environment within the package manager software *Anaconda*. Apart from Python3, the packages numpy, Pytorch, Torchvision, PIL, matplotlib and scimage were used. The analysis was run on an Intel Core i5-6400 CPU (2.7 GHz) with 8 Gigabyte RAM. From Torchvision we took the pretrained CNN architectures as well as the build-in transformations *translation* and *rotation*. The remaining transformations were implemented by the authors.

All pretrained CNN architectures from Torchvision had been trained on the same data, which is the training subset from the ImageNet dataset. Similarly, the same data augmentations had been applied to all models during training: first a crop of the input image with random size (between 0.08 and 1.0 of the original image size) and of random ratio (between 3/4 and 4/3) had been resized to a size of 224 pixel. Afterwards a horizontal flip had been applied with a probability of 50%. Because of this similar training procedure we assume that differences in the accuracy results reported below are caused by the differences in the model architectures only.

4 RESULTS

4.1 Traditional Transformations

Figure 3 shows the results for the *translation* transformation. All models show a stronger degradation of classification accuracy for translations in the horizontal x-axis than in the vertical y-axis. This difference is particularly noticeable for *AlexNet*, where a shift of 40% in the x-direction yields about six percent worse results than in the y-direction. The combination of both directions again significantly reduces the performance of all models: The decrease here ranges from 18 (ResNeXt) up to 43 percentage points (AlexNet). The partial disappearance of objects due to translation discussed above may explain the overall decrease to some extend, but the differences between the architectures remain remarkable.

The results of the *rotation* transformation[1] are shown in figure 4. Here, the rotation of the input data leads to a similar pattern of performance degradation in all three models. As expected, the classification accuracy has its global maximum at a rotation of 0° while local maxima are noticeable at 90°, −90° and 180°. Between these maxima there is a significantly reduced accuracy with minima at around −120° and

[1] The given angles refer to *counterclockwise* rotations.

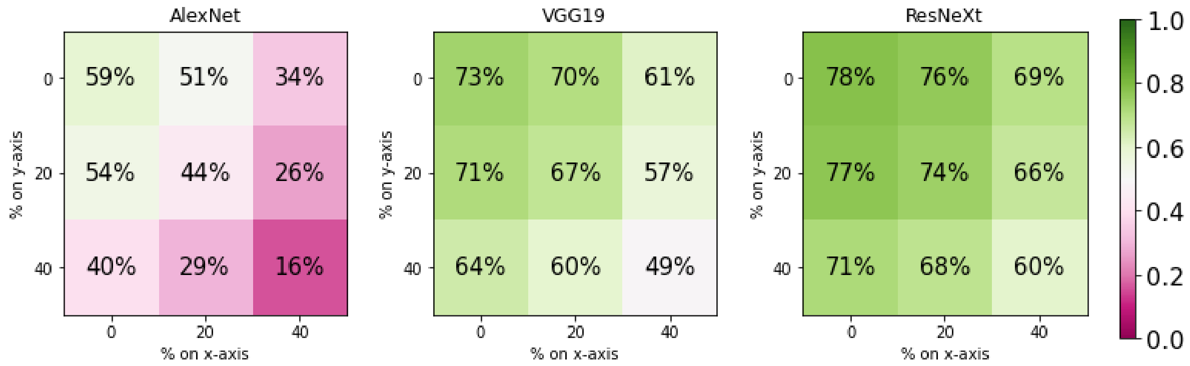

Figure 3: Change of accuracy when input images are translated. Each matrix shows the accuracy for one model in regard to the translation distance in x- and y-direction as percentage of image size.

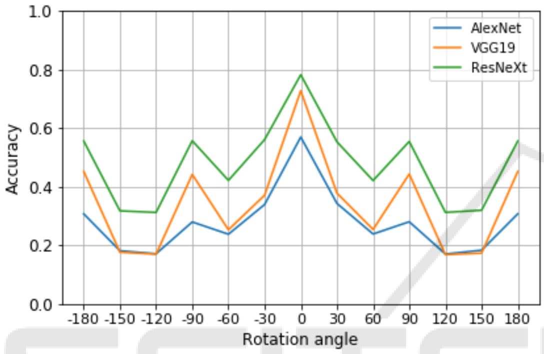

Figure 4: Change of accuracy when input images are rotated. The plot shows the classification accuracy of each model regarding input images rotated by the given angle.

150° respectively. The decrease in detection performance ranges from 55% (VGG19) to 39% (AlexNet).

Figure 5 shows the results of modifying the input images with a *color change* transformation. For all evaluated models the classification accuracy drops with higher intensity values. Note that the performance degradation is not uniform. A decrease in the intensity of the green color channel – i.e., a red-blueish tint of the images – affected the recognition performance more than an increase of the green color channel. The overall change in accuracy across intensity values from 0.5 to 1.5 ranges from 8% (ResNeXt) to 23% (AlexNet).

The changes in classification accuracy due to a replacement of the input images' background are shown in figure 6. Here, VGG19 and ResNeXt appear to be less affected by the *background replacement* transformation than AlexNet. The former are largely invariant to the specific color used and show a $\approx 10\%$ decrease of accuracy across all cases. In contrast, the results for AlexNet display clear differences in the degree of performance degradation depending on the specific background color. Note the particularly strong degradation in case of a red background color.

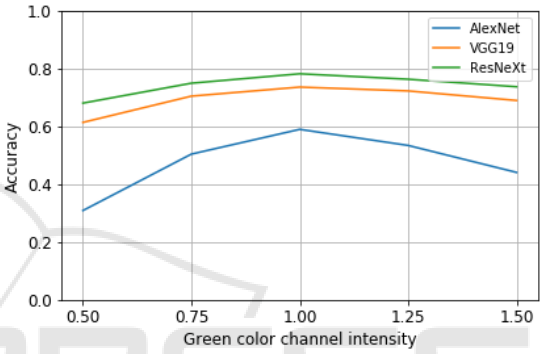

Figure 5: Change of accuracy when input images have altered colors. The plot shows the classification accuracy of each model for input images with varying intensity values of the green color channel. A value of 1 represents an unaltered image.

4.2 Novel Transformations

Figure 7 shows the recognition performance of all three models with respect to the *collage* transformation. It shows how well each network was able to recognize one, two, three, or all four of the image classes present in the collage images. For this purpose the k classes with the highest probability values were considered as guesses of the network, with $1 \leq k \leq 9$. Comparing the top1-accuracy ($k = 1$) with that for unaltered input images, the percentage of correctly classified images[2] is significantly lower for AlexNet and VGG19 (from 59% and 73% to 9% and 42%, respectively). In case of ResNeXt the percentage of correctly recognized class labels drops by 8% from 78% to 70%. The collage transformation results also offer some insight into the ability of each model to recognize more than one object class at a time. Even with increasing values of k AlexNet is not able to recognize

[2] An image counts as correctly classified as soon as one of the four classes in the collage is the top1.

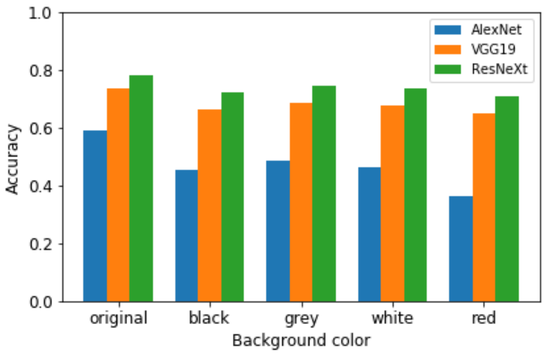

Figure 6: Change of accuracy when input images have an altered background. The graph shows the accuracy of each model for input images with replaced backgrounds. Unaltered input images are indicated by “original”. In all other cases the non-object part of the image has been replaced by the respective color.

more than one object class per collage image in most cases. VGG19 appears to be limited to recognize two out of the four object classes present in the collage. In contrast, ResNeXt seems to have fewer difficulties to recognize multiple object classes in a collage image. In more than 20% of the input images two of the four object classes are in the top2 predictions ($k = 2$). For $k > 3$, ResNeXt manages to put all four image classes among the predictions. For $k = 10$ ResNeXt detects one or more of the four object classes in over 98% of the images.

Figure 8 shows the accuracy of the evaluated networks with respect to the *puzzle* transformation. The x-axis shows n, the number of rows and columns used for constructing the puzzle, i.e., $n = 1$ refers to an unchanged input image, $n = 2$ refers to a 2×2 puzzle, and so on. All three models show a decreasing accuracy with an increase in image fragmentation. Overall, AlexNet appears to be significantly more affected by the puzzle transformation than VGG19 and ResNeXt.

4.3 Combined Transformations

Figure 9 shows the results of combining the rotation transformation with a black color background replacement. In contrast to the sole application of rotations evaluated above, in which areas of different size have to be filled in with black due to different rotation angles, the combination of rotation and background replacement results in the same ratio of useful information (the object) and pixels filled in (the background) independent of the rotation angle[3]. The combination of both transformations results in a similar pattern as seen for the rotation transformation above. The recognition performance continues to be significantly higher at multiples of ±90° than for other rotation angles, while the overall accuracy is further reduced compared to the sole application of the rotation transformation.

Figure 10 shows the results of combining the *puzzle* and *background replacement* transformations. The results display a high degree of consistency with the sole application of the puzzle transformation (Fig. 8). Adding the background replacement transformation reduced the performance by 10% to 20% across the parameter range.

5 DISCUSSION

The presented results reveal a clear trend regarding the comparison of the three models: across almost all of the evaluated transformations, the accuracy of ResNeXt remains higher than that of the VGG19 network, which in turn has a higher accuracy than AlexNet. There are some minor exceptions to this observation. For instance, for some rotational transformations VGG19’s accuracy drops below that of AlexNet.

Overall, there is a clear gap between AlexNet’s performance compared to the other two networks. This gap manifests itself not only in terms of absolute performance, but also regarding the robustness against the degree with which the individual transformations were applied. This difference is particularly prominent in the background replacement and color change transformations: even with heavily modified input data, the more recent models VGG19 and ResNeXt continue to show relatively high recognition rates that remain higher than the performance of AlexNet for unaltered input material. Interestingly, AlexNet’s robustness seems to fail especially with regard to red color components, which is visible in both the background replacement and color change transformations.

In general, it appears that the increased network complexity and the increased detection performance of VGG19 and ResNeXt correlates with a higher robustness against different types of transformations. This holds true especially in case of the color change, background replacement, and translation transformations. Thus, it might be the case that VGG19 and ResNeXt perform better than AlexNet in real-world applications where backgrounds or object environments change a lot or objects are shifted in the im-

[3] This does not hold true for the rare cases in which the object to be recognized lies very close to the images edge and its rotation results in parts of the object becoming invisible.

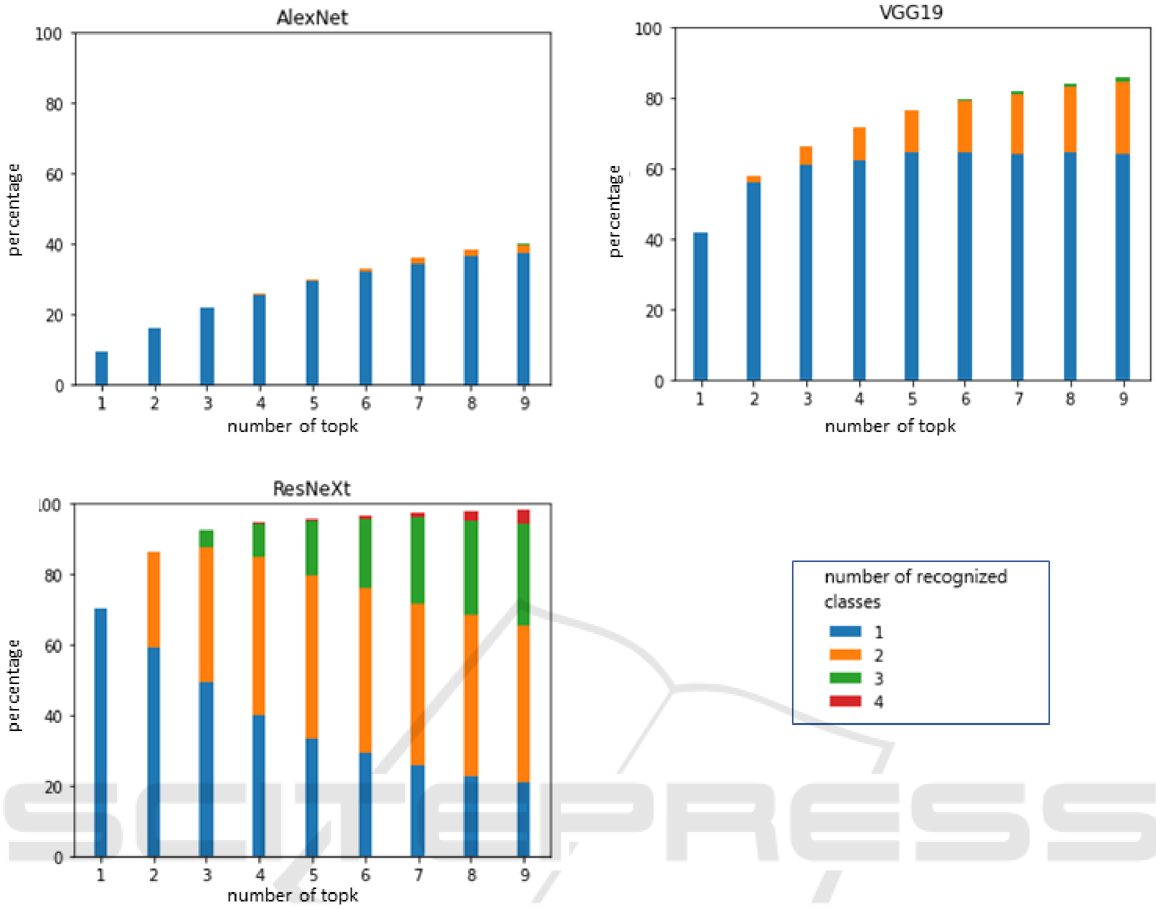

Figure 7: Change of accuracy when collage input images are used. Colors indicate the number of correctly recognized classes per collage (one, two, three, or all four). A class is considered to be correctly classified if it is present in the k classes with highest probability predicted by each model, with $1 \leq k \leq 9$.

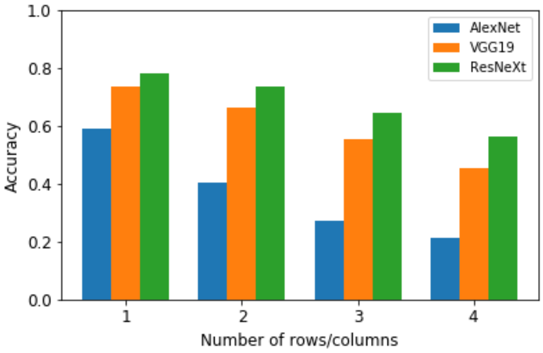

Figure 8: Change of accuracy when input images are transformed by the puzzle transformation. The graph shows the accuracy of each model for an increasing number of tiles. 1 represents the original image, whereas 4 represents a puzzle with $4 \times 4 = 16$ tiles. The tiles are arranged randomly.

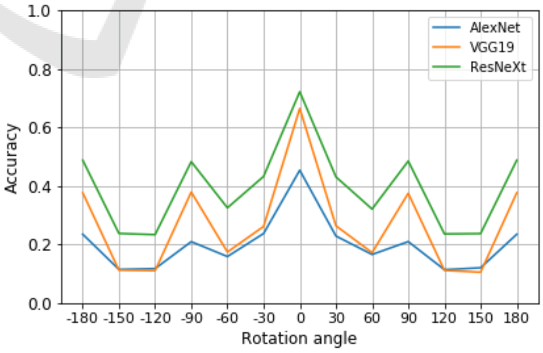

Figure 9: Change of accuracy when input images are transformed by background replacement and rotation. First the non-object part of the images have been replaced by a black color. Afterwards the rotation transformation has been applied with rotation angles ranging from -180 to 180 degrees.

age plane. Future research based on real-world data should be able to illuminate this aspect further.

Despite the progress shown by VGG19 and ResNeXt, a simple rotation of the input data can still cause a noticeable decrease in performance of all evaluated models. This is particularly surprising since

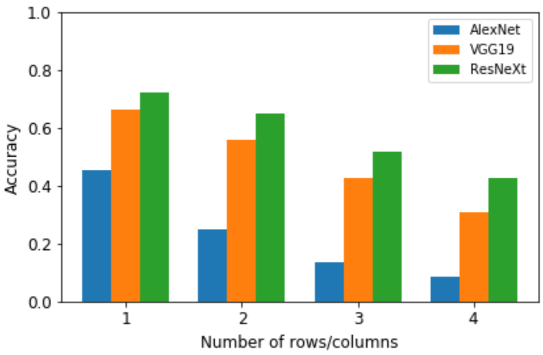

Figure 10: Change of accuracy when input images are transformed by background replacement and puzzle. First the non-object part of the images have been replaced by a black color. Afterwards the puzzle transformation has been applied with puzzle resolutions ranging from 1 (original image) to 4 (4 × 4 puzzle).

rotations are widely used as data augmentation methods. On the other hand, this observation is consistent with the results of (Engstrom et al., 2018), although they evaluated much smaller rotation angles. Given the relatively good performance of the evaluated networks at multiples of ±90° one might suspect that the lack of blank image areas at these angles might be the underlying cause for these irregularities. However, the results obtained from the combination of rotation and background replacement contradict this assumption.

Regarding the results obtained by the collage transformation experiment, ResNeXt's high accuracy suggests that this model has an improved ability to recognize the patterns of an object independent of the object's environment or background. This is consistent with the good performance of ResNeXt in the background replacement experiment. AlexNet on the other hand is hardly able to detect more than one object class simultaneously. This might indicate that AlexNet depends more heavily on patterns that are present in the image background to perform a correct classification.

With regard to the puzzle transformation experiment, the more recent models show a higher robustness as well. This may be related to the fact that they are already more robust against translations of objects in the image plane and therefore have less problems with a changed spatial arrangement of the image. Yet, this could also be an indication for a detection behavior that is more specialized in local patterns. For further insights, puzzles with an even smaller fragmentation of the input images could be constructed, while simultaneously controlling the extend of fragmentation of the object itself.

Further, more extensive experiments may use reduced step sizes for the transformation parameters to facilitate a more fine grained comparison with previous publications. For instance, (Azulay and Weiss, 2018) have shown that translations of even a few pixels can lead to significant performance drops in some architectures. Therefore, using smaller step sizes on the transformations evaluated in our work may facilitate a more detailed reasoning about the robustness of the networks.

6 CONCLUSION

Our results show that the more recent architectures VGG19 and ResNeXt appear to have an increased robustness against many kinds of image transformations including color changes and background replacement. However, invariance towards rotational transformations of the input appears to remain problematic as these transformations cause a significant decrease in recognition performance of all evaluated models.

The collage and puzzle transformations introduced here appear to be suitable benchmarks to further investigate the strengths and weaknesses of CNNs as they were able to reveal markedly different abilities among the evaluated architectures.

REFERENCES

Azulay, A. and Weiss, Y. (2018). Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 9448–9458. Curran Associates, Inc.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In Leonardis, A., Bischof, H., and Pinz, A., editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg. Springer Berlin Heidelberg.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. (2018). A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. (2020). Natural adversarial examples. *arXiv preprint arXiv:1907.07174*.

Howard, A. G. (2013). Some improvements on deep convolutional neural network based image classification. *arXiv:1312.5402*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Mikołajczyk, A. and Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400, Long Beach, California, USA. PMLR.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *arXiv:1409.0575*.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

Su, J., Vargas, D. V., and Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841.

Taylor, L. and Nitschke, G. (2017). Improving deep learning using generic data augmentation. *arXiv:1708.06020*.

Wiyatno, R. R., Xu, A., Dia, O., and de Berker, A. (2019). Adversarial examples in modern machine learning: A review. *arXiv:1911.05268*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. *arXiv:1611.05431*.

Zheng, L., Yang, Y., and Tian, Q. (2018). SIFT meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244.