ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Don't skip the skips: autoencoder skip connections improve latent representation discrepancy for anomaly detection

Anne-Sophie Collin, Cyril de Bodt, Dounia Mulders and Christophe De Vleeschouwer*

UCLouvain - ICTEAM/ELEN

Abstract. Reconstruction-based anomaly detection typically relies on the reconstruction of a defect-free output from an input image. Such reconstruction can be obtained by training an autoencoder to reconstruct clean images from inputs corrupted with a synthetic defect. Previous works have shown that adopting an autoencoder with skip connections improves reconstruction sharpness. However, it remains unclear how skip connections affect the latent representations learned during training. Here, we compare internal representations of autoencoders with and without skip connections. Experiments over the MVTec AD dataset reveal that skip connections enable the autoencoder latent representations to intrinsically discriminate between clean and defective images.

1 Introduction

Detecting defective samples in a production line based on visual inspection is of great importance in industrial applications. The scarcity of anomalies and variety in appearances make a supervised problem formulation inappropriate [1]. Instead, a training set of exclusively clean, i.e. defect-free images, and a test set including both clean and defective images is generally considered.

Reconstruction-based approaches rely on an autoencoder trained on clean images to perform an identity mapping [2, 3, 4]. During inference, such network is expected to reconstruct exclusively clean structures, leading to large reconstruction residuals when defective images are fed to the network [5]. In a previous study [6], we considered an autoencoder architecture equipped with long skip connections trained over images corrupted with a homemade synthetic defect model. To avoid convergence towards a pure identity operator, we corrupted training images with synthetic defects, adding stains of variable size and random colour over the original image. This approach showed significant improvement in detecting anomalies, especially on texture images.

In this paper, we investigate the impact of skip connections on the internal representations constructed by the autoencoder of both clean and defective structures. As represented in Figure 1, adding skip connection leads to the differentiation between representations for clean and corrupted images. Skip connections are known to have the capacity to propagate fine-grained information from the encoder to the decoder [7]. However, it was not intuitive nor

*CdB is supported by Service Public de Wallonie Recherche under grant n°2010235-ARIAC by DIGITALWALLONIA4.AI. DM is a Research Fellow of the F.R.S.-FNRS. CDV is a Research Director with the F.R.S.-FNRS.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

straightforward to predict the impact of those skip connections on the internal representations of clean and defective structures across the autoencoder layers. We observe that the corruption of the training input image with synthetic defects leads to a significant change in the latent representations when using skip connections in the autoencoder architecture. In addition to providing intuitive insights on the workings of trained autoencoders, this analysis proposes a valuable basis for extending the anomaly detection mechanism beyond the sole use of the reconstruction residual.

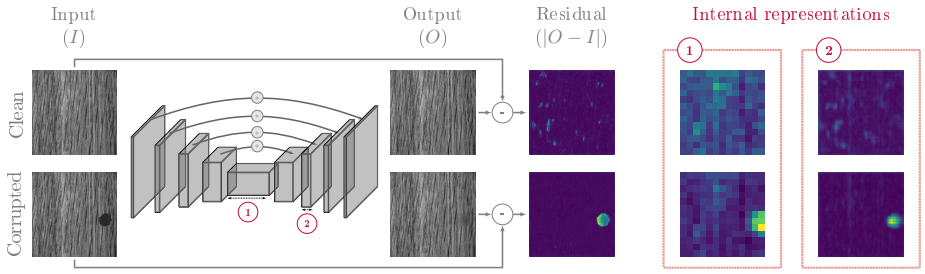


Fig. 1: Discrepancy between latent representations for a clean (first row) and a corrupted (second row) version of an image in an autoencoder with skip connections. Comparison with an autoencoder without skip connections will be presented in a further section. *(Red blocs.)* Mean along the channel axis of the activation tensors in the bottleneck *(1)* and in an intermediate layer *(2)*.

2 Methods

2.1 Network architectures

In this work, the reconstruction of a clean version of any input image is based on the use of a convolutional neural network. Our architecture, referred to as an **Autoencoder with Skip connections (AESc)**, is a variant of U-Net [7]. AESc takes input images of size 256×256 and projects them onto a latent space of dimension $13 \times 13 \times 256$ by means of six consecutive convolutional layers stridden by a factor two. The back projection is performed by six layers of convolution followed by an upsampling operator with a factor of two. All convolutions have a 5×5 kernel. Unlike the original U-Net version, our skip connections perform an addition, not a concatenation, of feature maps from the encoder to the decoder.

As a comparison, we also consider an **Autoencoder (AE)** network which has the same architecture as described above, but without the skip connections.

2.2 Synthetic corruption

Our networks are trained to reconstruct a clean image out of a corrupted version of it. The synthetic corruption that is considered is the addition of an irregular ellipse of variable size and random colour, as described in [6].

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2.3 Notations

Input images. We denote by $\boldsymbol{x} \in \mathbb{R}^{W \times H \times C}$ an input image, where W, H and C stand for the width, height and number of channels of $\boldsymbol{x}$. In this article, the input images are either clean or corrupted with synthetic defect (as defined in [6]). For the MVTec AD dataset, we exclusively consider grayscale images with $W = H = 256$ and $C = 1$.

Activation tensors. We define the operator $l_i(\cdot) : \mathbb{R}^{W \times H \times C} \to \mathbb{R}^{W_i \times H_i \times C_i}$ as the projection of an input image onto its activation tensor in the i^{th} layer. In this notation, W_i, H_i, and C_i respectively denote the width and height (spatial dimensions) and the number of channels in layer i. According to our setup, $l_1(\boldsymbol{x})$ corresponds to the activation tensor obtained after one convolutional layer while $l_5(\boldsymbol{x})$ is the activation tensor in the autoencoder bottleneck.

Mean of activation tensors along channel axis. Given an activation tensor $l_i(\boldsymbol{x}) \in \mathbb{R}^{W_i \times H_i \times C_i}$, we write $\mu_i(\boldsymbol{x}) \in \mathbb{R}^{W_i \times H_i}$ for the mean of $l_i(\boldsymbol{x})$ along the channel axis. In the figures of this work, $\mu_i(\boldsymbol{x})$ is used to illustrate the latent representation of the autoencoder.

3 Experiments and Discussion

In this section, we analyse the impact of skip connections on the internal representations of an autoencoder performing anomaly detection in the reconstruction-based framework detailed in [6]. To detect defective structures in an input image, it is expected that the network does not generalise the identity mapping (promoted for clean structures during training) to the reconstruction of unseen (defective) structures.

We first provide a qualitative analysis regarding the propagation across layers of both clean and defective structures of the training set, for autoencoders with and without skip connections. Then, we quantify the discrepancy between the internal representations of a pair of clean and defective images. We rely on the MVTec AD dataset [2], which contains 15 image classes. Among these 15 image categories, 5 correspond to textures and 10 to objects. Since our observations tend to differ on textures and objects, we present our results on two images that are representative of each image category.

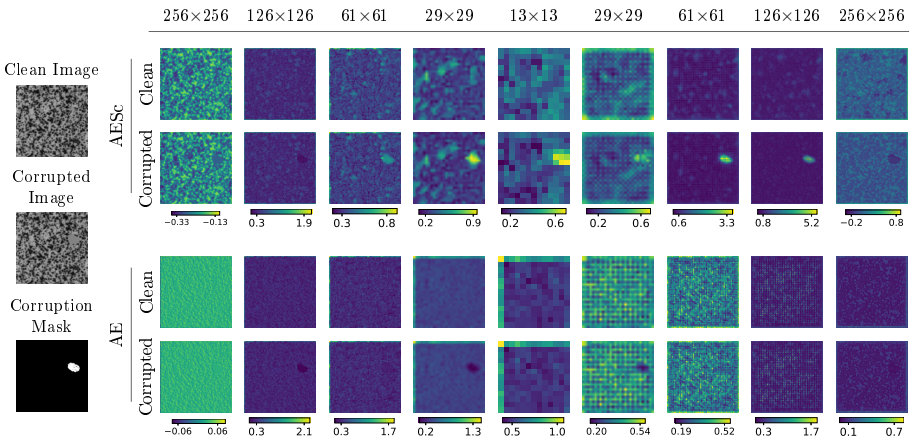
3.1 Visualisation of activation tensors

Figure 2 provides a visual insight into how a clean and defective image propagate through an autoencoder, with and without skip connections. These autoencoders are denoted as AESc and AE, respectively. First, Figure 2a shows the mean of the activation tensors along the channel axis in all layers of AESc for an image from the texture category. Comparing the first and second rows suggests that corrupting a texture input image with a synthetic defect leads to a strong discrepancy between the latent representations of the clean image and a corrupted version of it. Specifically, it leads to significantly larger activation intensities in

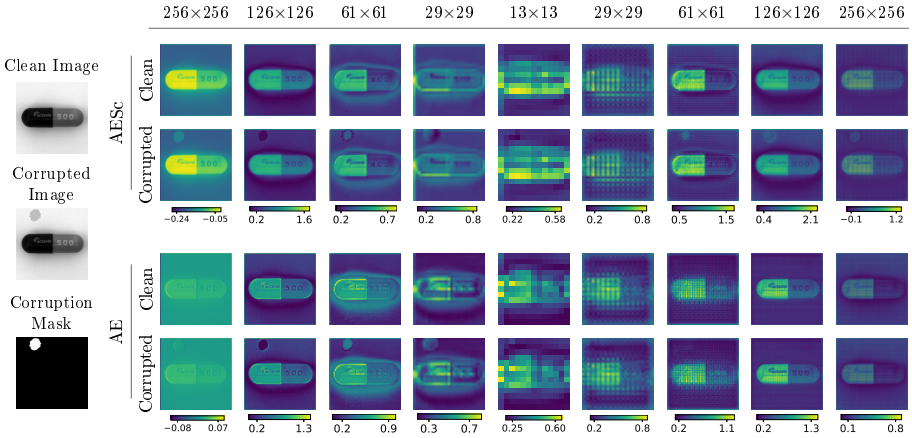
ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

the region covered by the corruption mask. With the AE model, such a large difference is not visible.

Second, Figure 2b displays a similar analysis for an image from the object category. In contrast to the texture image, there is no significant visual disparity between the AESc and AE models. Unlike the previous example, higher activation intensities correlated with the corruption are not clearly identifiable.



(a) Example over an image from the texture category: Tile.



(b) Example over an image from the object category: Capsule.

Fig. 2: Mean of the activation tensors along the channel axis $\{\mu_1, \mu_2, \ldots, \mu_9\}$ for a clean image and a corrupted version with synthetic defect (as defined in [6]). In both (a) and (b), the first and second rows show μ_i obtained with the AESc model on the clean and corrupted images, respectively. The third and fourth rows show μ_i obtained with the AE model. Spatial dimensions of the mean activation tensors are provided in the column labels.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

3.2 Quantification of the discrepancy

We quantify the discrepancy between the latent representations of a clean image $\boldsymbol{x}_{\text{cl.}}$ and a corrupted version $\boldsymbol{x}_{\text{corr.}}$ in the i^{th} layer with the cosine similarity applied over the mean of the activation tensors along the channel axis: $\cos(\mu_i(\boldsymbol{x}_{\text{cl.}}), \mu_i(\boldsymbol{x}_{\text{corr.}}))$. Figure 3 compares the distributions of this metric in all layers of AESc (orange) and AE (blue) for the MVTec AD dataset. Overall, the distances between the activation tensors for AESc increase as we go deeper in the network while those for AE decrease. This trend, which is stronger for texture images, shows that AESc learns distinct latent representations for clean and defective images. This observation is in line with observations made in Figure 2a. The latent representations of AESc show higher activation intensities where the image is affected by the corruption mask. In comparison, the latent representations of AE appear as being more similar for clean and corrupted versions of the same image.

While the previous visual experiment (Figure 2b) showed no strong discrepancy between latent representations of the AESc and AE models for the Capsule category, this quantitative analysis suggests that the AESc decoder discriminate between clean and defective structures for several object categories.

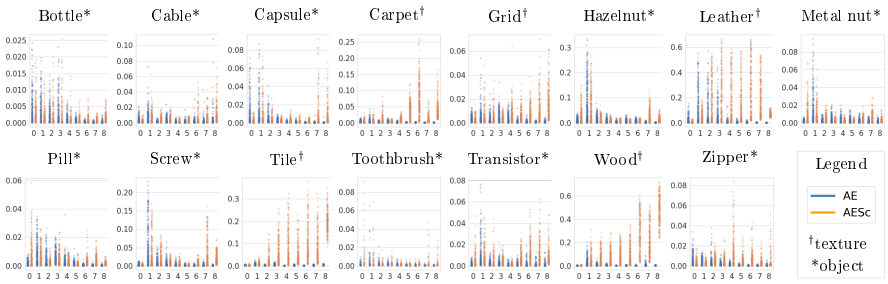


Fig. 3: Distribution of cosine similarity (y-axis) applied over the mean of the activation tensors along the channel axis: $\cos(\mu_i(\boldsymbol{x}_{\text{cl.}}), \mu_i(\boldsymbol{x}_{\text{corr.}}))$ between a clean image $\boldsymbol{x}_{\text{cl.}}$ and a corrupted version $\boldsymbol{x}_{\text{corr.}}$ of the same image. Distributions obtained with the AE (blue) and AESc (orange) models are compared across network layers (x-axis) for all image categories of the MVTec AD dataset.

4 Conclusions and Perspectives

In this paper, we study the impact of autoencoder skip connections on the latent representations of clean and defective images for anomaly detection. Through a visual analysis of the mean of the activation tensors along the channel axis, a strong discrepancy between a clean and a corrupted input texture is observed in nearly all layers of an autoencoder with skip connections. Specifically, activation intensities significantly increase on the corruption's spatial support. Without skip connections, the autoencoder does not adopt this behaviour.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Representative examples for both texture and object image categories are studied. Although a discrepancy is not visible with an object input, the quantitative study reveals that the representation tensors of clean and defective objects are distinguishable from each other when skip connections are added to the autoencoder architecture.

This discrepancy between clean and defective structures across layers in an autoencoder with skip connections is of particular interest to improve the anomaly detection mechanism introduced in [6], which solely relies on the reconstruction residual. The mean of the activation tensors in the decoder provides interesting anomaly maps that can be useful for anomaly localisation. It would be particularly interesting to study the intrinsic properties of this new anomaly localisation information (robustness to noise, generalisation from synthetic to real defects, etc.) to determine the extent to which it can augment existing detection mechanisms. Also, an explicit optimisation constraint during training can be considered to bias towards discrepancy of latent representations between clean and defective structures.

References

[1] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad - a comprehensive real-world dataset for unsupervised anomaly detection. *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600, 2019.

[3] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proc. Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 372–380, 2019.

[4] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. pages 146–157, 2017.

[5] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71, 2022.

[6] Anne-Sophie Collin and Christophe de Vleeschouwer. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2021.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015.