Towards Visually Plausible Explanations

Reproducibility Summary

Introduction

The goal of this assignment is to firstly reimplement the work from the paper Liu et al. (2020) and secondly extend their work. Liu et al. develops a new technique which visually explains VAE by generating attention maps from the learned latent space. Then the paper applies the VAE into two applications: anomaly localisation and latent space disentanglement. This paper reimplements Liu's experiments and compares our results with the results from the paper under restricted computational resources. The acquired results will be analyzed and compared to the original paper.

Scope of Reproducibility

The paper Liu et al. (2020) claims that their research takes steps towards visually explaining generative models by using a new method visual attention maps conditioned on the latent space of a variational autoencoder (VAE). Furthermore, the attention maps can be used to demonstrate the localisation of anomalies in images. This localisation method for anomaly detection, resulted in state-of-the-art performance in the MVTec-AD dataset. Moreover, a new learning objective was formed: attention disentanglement loss. This resulted in better performance on the Dpsrites dataset compared to state-of-the-arts methods.

Methodology

For the first part of the experiment the author's code was used, and for the second experiment, the proposed method from (Kim and Mnih, 2019) was used with the addition of the disentanglement metric mentioned in the original paper. The total training lasted around six hours for the anomaly localisation and for the disentanglement it lasted approximately 40 to 80 hours. Slight hyperparameter tuning was needed for some of the datasets.

Results

The results showed that for the most part the anomaly localisation claims seem to be reproducible (except for the UCSD pedestrians dataset). But the disentanglement was not reproducible.

What was easy

The easiest reproducible parts for this experiment was the part for which the code and documentation was available.

What was difficult

There was a wide range of obstacles, when reproducing this code. Much of the information was dependent on the interpretation of the reader. There is no explicit documentation about the implementation in the paper; hyper-parameters, experiment set-up and model descriptions are not present. Also, the full second part of the implementation was missing from the code base. And many equations or formulations were not sufficiently explained.

Communication with original authors

There was little communication, only an opened issue on their Github repository, giving a solution to a known issue.

1 Introduction

The goal of this assignment is to firstly reimplement the work from the paper Liu et al. (2020) and secondly extend their work. Liu et al. (2020) state that their research takes steps towards visually explaining generative models. This is done by using a new method; visual attention maps conditioned on the latent space of a variational autoencoder (VAE). Attention maps aim at intuitively representing different regions of normal and abnormal images. The paper applies the VAE in two experiments: anomaly localisation and latent space disentanglement. Anomaly localisation means to detect and localise the abnormal phenomena in an image. This is performed through the cues in the attention maps. Latent space disentanglement focuses on learning a decomposed representation, meaning it describes each piece of data x by a set of statistically independent factors z. This has several benefits, such as explaining the underlying decision making process in unsupervised learning (Nitzan et al., 2020). The novel disentanglement loss (mentioned in the original paper) helps utilize the information from the attention maps to improve the performance of Factor VAE (Liu et al., 2020). We reimplemented their experiments and compare our results with the result from the paper. Moreover, we analyze the claims of the paper and assess their validity.

2 Scope of reproducibility

The work presented in this report aligns with our views of having visual systems that are more explainable. The ability to extract insights about the decisions made by Generative models is very important given that they are targets for the new trend of deepFakes (Nguyen et al., 2020). Hence, we think investigation in this area should be developed and applied in the industry and academia.

According to the study from Liu et al. (2020), the proposed attention mechanism is able to highlight anomalous areas in the images, and consequently identify the features in the underlying latent space that cause the abnormality.

Regarding the Anomaly detection, (i) good qualitative results are found for the MNIST dataset, meaning that the attention maps' highlighted regions correlate with high precision to the localisation of a single anomaly, (ii) the qualitative and quantitative results for the UCSD Pedestrians dataset perform better than the baseline. (iii) And, the MVTec-AD dataset out-performed other state-of-the-arts (Bergmann et al., 2019) (for more information on evaluation see section 3.5) Regarding, the attention disentanglement (iv) a new learning objective is proposed that improves the trade-off between reconstruction loss and disentanglement metric over existing methods. Additionally, (v) the paper asserts that the attention disentanglement loss helps to separate the high-response pixel regions which promise better results quantitatively and qualitatively in comparison to the state-of-the-arts (Liu et al., 2020).

To push the boundaries of the proposed visual explanations we extend their approach to Restricted Boltzmann Machines, hypothesising similar claims to the original paper; expecting significant performance qualitative and quantitatively .

3 Methodology

The research is divided into three different part: the anomaly detection, the latent space disentanglement and the ablation study.

3.1 Datasets

The datasets used in this research are the MNIST, UCSD Ped 1, MVTec-AD and Dsprites datasets. The MNIST dataset contains 70,000 grey level images of handwritten digits (60,000 in training and 10,000 in test set) that are size-normalized and centered in a fixed-size image of 28x28 (LeCun et al., 1998).

The MVTec-AD Dataset contains 5354 high-resolution images. These images are divided into fifteen different objects and textures with various defects. It includes pixel-wise annotations. (Bergmann et al., 2019). The images are resized to 256x256 (as is done in the original paper). The data is also augmented by performing random flips and rotations. For more information on the distribution for this dataset and the MNIST dataset see figure 9 & 8 in the supplementary materials.

The UCSD Ped 1 anomaly detection dataset is compromised of images (frames from videos) from a stationary camera overlooking a pedestrian walkway. The anomalies are in the form of non-pedestrian entities (e.g. cart, wheelchair). Peds1 consists of 5500 frames training test and 3400 anomalous test frames. The frames are resized to 100x100 pixels (same as the original paper). Some manually generated pixel-level binary anomaly masks are available. (Li et al., 2013).

The Dsprites datasets consists of 737280 64x64 black and white images of 2D shapes. The images are augmented by varying six ground truth independent latent factors (color, shape, scale, rotation and x and y positions).(Higgins et al., 2016) (Matthey et al., 2017).

3.2 Anomaly detection

The anomaly localisation with attention maps applies a vanilla VAE (one-class VAE). This essentially is an autoencoder trained with a reconstruction objective and a variational term objective that attempts learning a standard normal latent space distribution. The variational term objective is generally implemented with a Kullback-Leibler distribution metric, which is based on a computation featuring the latent space distribution and a standard Gaussian. Additionally you can refer to (Liu et al., 2020)(Kingma and Welling, 2014). The attention maps are generated by computing a score from the latent space and back-propagating on this score to a particular convolutional layer to get the gradients of the score with respect to the activations of that layer. The paper introduced a way to generate attention maps for anomaly localisation, by computing a score through taking the element-wise sum of the mean vector, μ, from the encoder (for more detailed information see (Liu et al., 2020)).

There are no specifics on the model architectures for the anomaly localisation. However, the Github repository from the authors provides a simple VAE architecture for MNIST, which provides decent qualitative results. This model was adapted to fit the resized UCSD images, by adding extra convolutional layers. For the MVTec-AD dataset according to the original paper a ResNet-18 is used. To build the architecture for this dataset, a ResNet-18 VAE was used by replacing the last two layers of the standard ResNet-18 architecture to the latent dimension space, in this case 32, to build a ResNet-18 encoder. For the decoder, transpose convolutions and upsampling layers are used instead of convolutional layers (for more information on the architectures supplementary material B.1).

The hyperparameter settings can be found in the Table 1, the hyperparameters for the MNIST dataset were already predefined and sufficient to achieve comparable results to the original paper, so these functions were used. These also served as a starting point for (manual) tuning the other two models on the UCSD and MVTec-AD dataset. However, for the MVTec-AD it was changed slightly to accommodate our computational power.

3.3 Latent space disentanglement

The factor disentanglement adapts the Factor VAE which attempts to learn a factorial distribution for the latent variables distributions by using a Multilayer Perceptron. The Factor VAE loss (L_{FV}) component takes into account the batch size N, the reconstruction loss, the Kullback-Leibler metric and the total correlation penalty. The γ hyperparameter controls the importance of the discriminator cross-entropy loss. For details of the Factor VAE refer to (Kim and Mnih, 2019). The innovative contribution in the disentanglement combines the information from the attention maps to create the disentanglement loss component

$$L_{AD} = 2 * \frac{\sum_{ij} \min(A^1_{ij}, A^2_{ij})}{\sum_{ij} A^1_{ij} + A^2_{ij}} \tag{1}$$

where the model loss is $L = L_{FV} + \lambda L_{AD}$ and λ is a hyperparameter. The last loss component promotes the divergence between the two attention maps and takes into account their lowest response pixel regions. The attention maps selection mechanism used in all experiments pairs the attention maps and averages the disentanglement loss from equation 1 over all pairs. For example, in the case of 4 attention maps a, b, c, d we average the loss over the pairs $(a, b), (b, c), (c, d)$. One drawback is that by forcing the attention between consecutive maps to diverge it might lead to one of them becoming closer to another map, however these trade-offs are beyond the scope of the current experiments and future work could explore different attention maps selection mechanisms.

The disentanglement metric employed was proposed by (Kim and Mnih, 2019). It is a majority vote linear classifier that aims at being robust over hyperparameter changes and different training schedules.

Since the attention disentanglement implementation was missing from the author's codebase, we used this Github repository for the implementation. Moreover, the disentanglement metric used in the experiments by Kim and Mnih (2019) was adapted from the following Github repository. The architecture for the experiments is described in the supplementary material C.1. The influence of the attention disentanglement loss (defined by equation 1) was compared through varying the values of gamma in the Factor VAE and the AD-Factor VAE (Table 1). Moreover, the latent dimensions that collapsed to the prior were removed by using a threshold of 0.01 (in the metric score calculation). The amount of epochs and values of gamma are based on the FactorVAE original paper (the exact values of the hyperparameters are shown in Table 1). Across the experiments for each model we use 10 latent variables and the number of iterations was half of the iterations mentioned in Kim and Mnih (2019). For more details on this decision refer to supplementary material C.2.

Hyperparameters	Epochs	Batch Size	Learning rate	β_1	β_2	ϵ	
MNIST VAE	200	128	0.001	0.9	0.999	1×10^{-8}	
UCSD Ped1 VAE	300	128	0.001/0.0005/0.0001	0.9	0.999	1×10^{-8}	
MVTec-AD VAE	300	16	0.0005	0.9	0.999	1×10^{-8}	
Hyperparameters	Iterations	Batch Size	Learning rate	β_1	β_2	γ	λ
Experi1 Vani F. VAE	150000	64	0.0001	0.9	0.999	10/20/30/40/50	-
Experi1 AD-F. VAE	150000	64	0.0001	0.9	0.999	10/20/30/40/50	1.0
Experi2 AD-F. VAE	150000	64	0.0001	0.9	0.999	5/15/25/35/45	1.0
Experi3 AD-F. VAE	150000	64	0.0001	0.9	0.999	10/20/30/40/50	0.5/1.5
Discriminator	-	-	0.0001	0.5	0.9	-	-

Table 1: Hyperparameters for the experiments: the first part is for the anomaly localisation experiment and the second part is for the disentanglement experiment + ablation study (the optimizer used was Adam).

3.4 Ablation study

Two distinct types of ablation studies were performed. The first type is to analyse discrepancies and ambiguities in the code. The second type is to extend the approach of the original paper to a different type of Generative model.

3.4.1 Hyper parameters and discrepancies

The aim was to reproduce the paper from the available code and description in the original paper. However, due to discrepancies between the author's codebase and paper, additional experiments were performed. First of all, in equation 2 in the original paper there is a ReLU activation on the attention maps, which is absent from the codebase. But, this barely affected any of the attention maps and quantitative results, so this was deemed unnecessary. Moreover, the codebase uses L2 normalization on the gradients, $\frac{dz}{dA}$, before computing the weights α for generating the attention maps. This was not mentioned in the paper, but also hardly affected the generated attention maps and AUROC scores. Lastly, the visualization of the attention maps as shown in the paper was accomplished by interpolating the attention maps to the input size and overlaying them on top of the input image (also not explained in the paper).

On the Attention Disentanglement section there were also some lacking documentations. On formula 6 there is no information regarding the λ hyperparameter, so the most suitable value was searched in an ablation study.

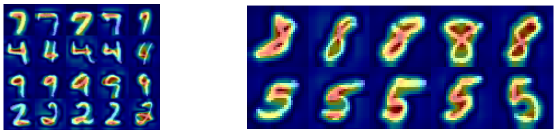
3.4.2 Applying Stacked RBM to Anomaly Detection

To extend the work of the original paper, the novelty of this paper is applied to the stacked restricted boltzman machine (stacked RBM) (Van et al., 2017). In order to perform anomaly detection, an attention map is generated for the RBM. A RBM is composed of visible and hidden units. It is parameterized by a bias vector in the visible layer, a bias vector in the hidden layer and a weight matrix that does transformations between these two layers. A RBM is an energy-based model that aims at minimizing the energy score of the network and maximizing the likelihood of observed data. The optimal weights are found by applying contrastive divergence (CD) with Gibbs Sampling. This means that a RBM draws samples alternatively from two conditional probabilities until the CD converges. We stack multiple RBMs together to form a deep RBM network. The deep RBM network is trained from bottom RBM to top RBM. To generate the attention map, we take the output of the layer that we specify, and compute the attention map as is explained in section 3.2. Moreover, the hidden units are constraint to either 1 or -1, which makes it impossible to apply the reparametarization trick used in the original paper to generate the attention maps. In order to approximate an equivalent continuous distribution we used the Gumbel-Max Trick (Maddison et al., 2017).

3.5 Evaluation

The evaluation is performed in the same fashion as in the original paper. For anomaly detection, the performance of all quantitative results are evaluated pixel-level segmentation through the area under the receiver operating characteristic curve score(ROC AUC or AUROC). The true positive rate (TPR) is defined as the amount of pixels across the testing set correctly classified as anomalous pixels. On the other hand, the False positive rate (FPR) is the amount of pixels wrongly classified as anomalous across the testing set. Moreover, the best threshold is picked based on the thresholds returned in the ROC curve and picking the best one based off the intersection-over-union (IoU).

Since the UCSD and MVTec-AD datasets include target masks, these are evaluated quantitatively and qualitatively, wheareas the MNIST datset is only evaluated qualitatively. For the UCSD, the attention mechanism is also tested with different spatial solutions by back-propagating the three penultimate convolutional layers in the model.



(a) Anomaly attention maps for model trained on digit 1, evaluated on digits 7, 4, 9 and 2

(b) Anomaly attention maps for model trained on digit 3, evaluated on digits 8 and 5

Figure 1: Qualitative results on MNIST dataset

For the disentanglement experiment, two different random seeds were used to smooth the results for all the experiments and reduce instabilities. Quantitative results were obtained by evaluating the performance through the training metrics of reconstruction error, true total correlation (from VAE) and the test metric of disentanglement. Here the baselines were the Factor VAE and β-VAE.

3.6 Computational requirements

The experiments were run on two different systems. The anomaly localisation was performed on Nvidia RTX 2070 Super (8GB VRAM). The total amount of time for running the model constituted around six hours and 18 minutes. However, this varied between datasets, since they are of different lengths. On the other hand, the attention disentanglement was performed with a Nvidia GTX 1080 Ti GPU (11GB VRAM). The experiment 1 and 2 took around 40 hours each to be completed, while the third took about 75 hours.

4 Results

This section describes the results that were obtained from the experiments. 4.1 shows the results regarding the claims made by the original paper. And section 4.2 shows the results from the ablation studies performed on the original paper. For visualizing the figures with more detail, please refer to the following Github repository.

4.1 Results reproducing original paper

4.1.1 Anomaly detection

Figure 1 illustrates qualitative results from the MNIST dataset. Figure 1a show attention maps for a model trained on inlier digit 1, and evaluated on digit 7, 4, 9 and 2, as done in the original paper. Figure 1b shows the attention maps from a model trained on digits 3, and evaluated on digit 5 and 8. We achieve similar anomaly maps to the original paper, for instance horizontal bars are highlighted as anomalous regions when trained on digits 1.

Figure 2 compares the qualitative results from the original paper with our results, with similar examples shown for the anomaly attention maps and target masks. We observe that though very similar, our attention maps are less precisely localised around the anomaly. Additionally, the AUROC our model achieved was lower (0.8731 as opposed to 0.92) than reported in the paper. Furthermore, for all convolutional layers to which we back-propagate, we achieved the exact same AUROC.

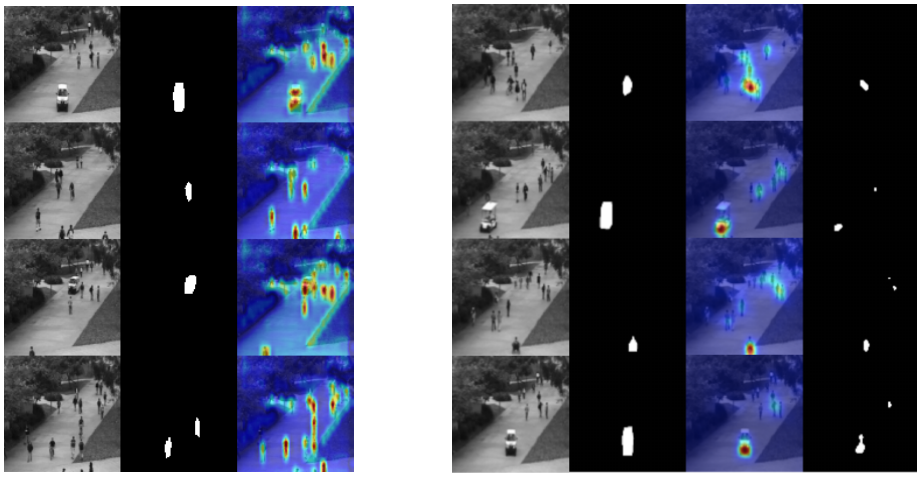
The MVTec-AD AUROC and IoU scores are similar to that of the paper (see Table 2). In figure 3 various sample comparisons are shown side-by-side with samples from the original paper. We achieved AUROC scores that were close enough, however, the IoU scores on average were much higher. Perhaps this is due to a different reduction method used to calculate the overall IoU of the test set, or different test sets that were used. Qualitatively, the anomaly maps look very similar to the original results for most categories, especially as hazelnut, leather and wood.

4.1.2 Latent space disentanglement

The first attention disentanglement experiment results are depicted in figure 4. The trade-off between the reconstruction error and disentanglement metric for the FactorVAE and the AD-FactorVAE is assessed. Given our assumption about the attention map selection mechanism, the author's claim regarding the improvement of the disentanglement metric over state-of-the-arts is not reproducible. Over training iterations, the disentanglement metric does not offer much insights since in both approaches the fluctuations are minimal after 100,000 iterations. In figure 5 the lowest reconstruction losses are observed in AD-FactorVAE with $\gamma = 30$ and FactorVAE with $\gamma = 20$.

Category	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule
Ours Orig.	0.63 / 0.51 0.78 /0.1	0.69 / 0.52 0.73 / 0.02	0.80 / 0.55 0.95 / 0.24	0.73 / 0.54 0.80 / 0.23	0.70 / 0.54 0.77 / 0.14	0.70 / 0.46 0.87 / 0.27	0.79 / 0.59 0.90 / 0.18	0.82 / 0.50 0.74 / 0.11
Category	Hazelnut	Metal	Pill	Screw	Toothbrush	Transistor	Zipper	
Ours Orig.	0.90 / 0.60 0.98 / 0.44	0.64 / 0.43 0.94 / 0.49	0.78 / 0.48 0.83 / 0.18	0.91 / 0.50 0.97 / 0.17	0.83 / 0.49 0.94 / 0.14	0.72 / 0.48 0.93 / 0.30	0.67 / 0.48 0.78 / 0.06	

Table 2: AUROC and best IoU (AUROC/IoU) for each object in the MVTec-AD dataset. Our results (ours) in comparison to the results from the original paper (orig.)



(a) Our reproduced results on the UCSD dataset, showing input image, ground-truth masks and anomaly attention maps

(b) The original paper's results on the UCSD dataset, showing input image, ground-truth masks, anomaly attention maps and produced localisation maps

Figure 2: Results from the anomaly localisation in the UCSD dataset

The qualitative results present the attention maps of the above highlighted models configurations with the lowest reconstruction loss and disentanglement metric in figure 4 (left). The first row shows three input images and the next four rows describe the first and second highest response attentions maps for FactorVAE and AD-FactorVAE. The FactorVAE attention maps show the highest response in similar regions where the sprite is visible. However, in two situation this is not the case and the attention is spread. On the AD-FactorVAE attention maps, the attention is mostly in the same region indicating that they did not disentangle. Given the insights from the anomaly experiments, we know that the layer selected for computing the attention maps has a considerable influence. In this case we used the last convolutional layer of the encoder, which turned out to be detrimental.

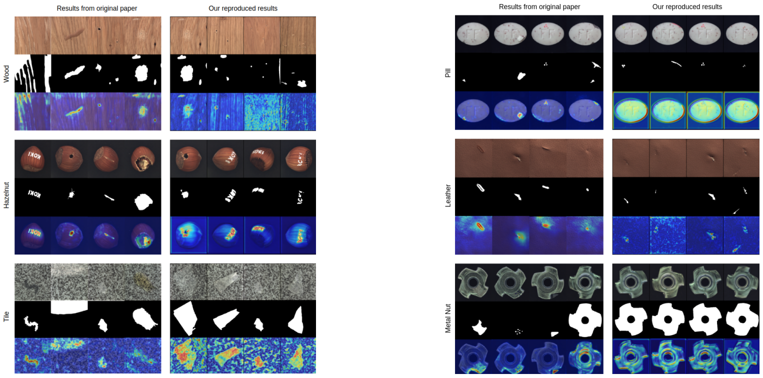


Figure 3: Results from the anomaly localisation in the MVTec-AD dataset

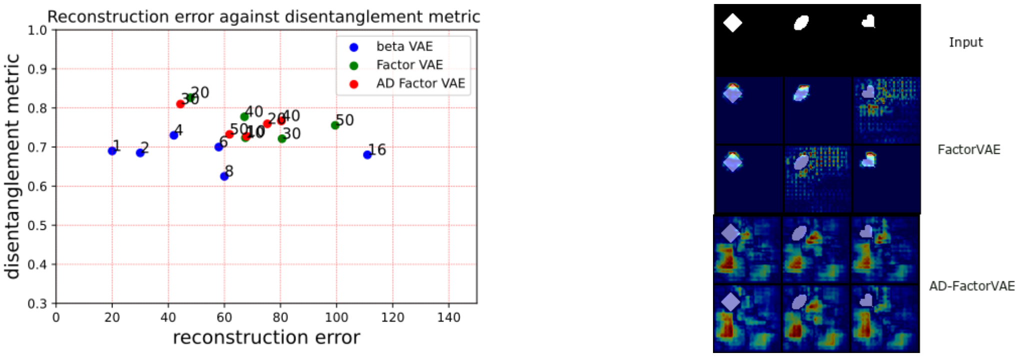


Figure 4: Left: Trade-off between disentanglement metric and reconstruction loss for trained FactorVAEs and AD-FactorVAE with 150000 iterations and $\lambda = 1.0$. The label is the γ value. Right: The first row is the input image, 2nd and 3rd are the FactorVAE attention maps while 4th and 5th are AD-FactorVAE attention maps.

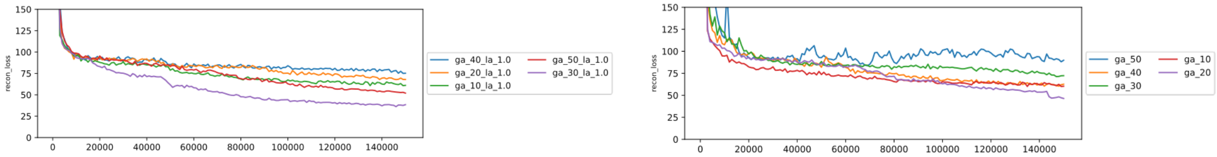


Figure 5: The top plot corresponds to the reconstruction loss over training iterations and the bottom to the Total Correlation over training iterations. On left is the AD-FactorVAE statistics and on the right for the FactorVAE.

4.2 Results beyond original paper

4.2.1 Ablation study

The results for the second disentanglement experiment are shown in figure 6. The disentanglement metric depicted on the left shows that all the gamma values are very similar in performance. By observing the reconstruction loss it becomes visible that $\gamma = 25$ and $\gamma = 30$ have lower reconstruction errors, at least with $\lambda = 1.0$. Additionally, the true total correlation loss fluctuates close to 0.

Figure 7 shows the results for experiment 3. On the disentanglement metric there is a separation where the $\gamma = 1.5$ shows higher disentanglement. This trend is verified by the lower reconstruction loss for $\gamma = 20, \lambda = 1.5$ and $\gamma = 30, \lambda = 1.5$. There appears to be a smaller total correlation fluctuation compared to the $\lambda = 1.0$ counterparts. These are no definitive results and further experiments with higher λ values should be performed to build stronger premises.

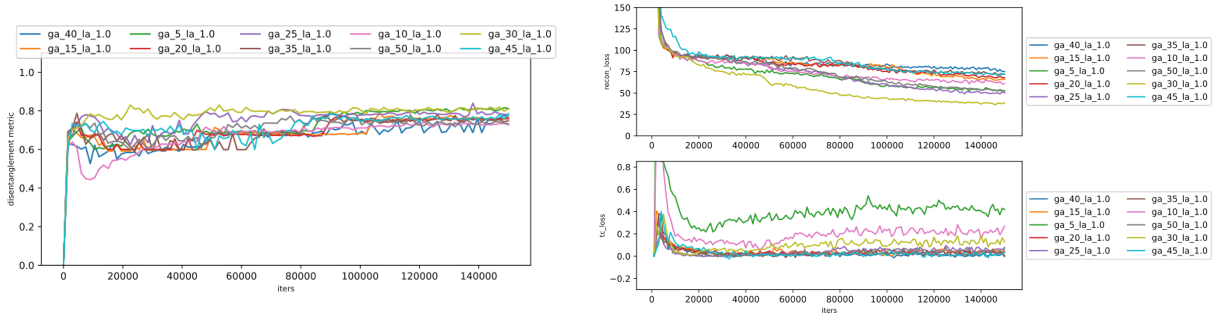


Figure 6: Left - Disentanglement metric over training iterations for the AD-FactorVAE γ combinations tested with $\lambda = 1.0$. Right - The top plot corresponds to the reconstruction loss over training and the bottom to the Total Correlation over training.

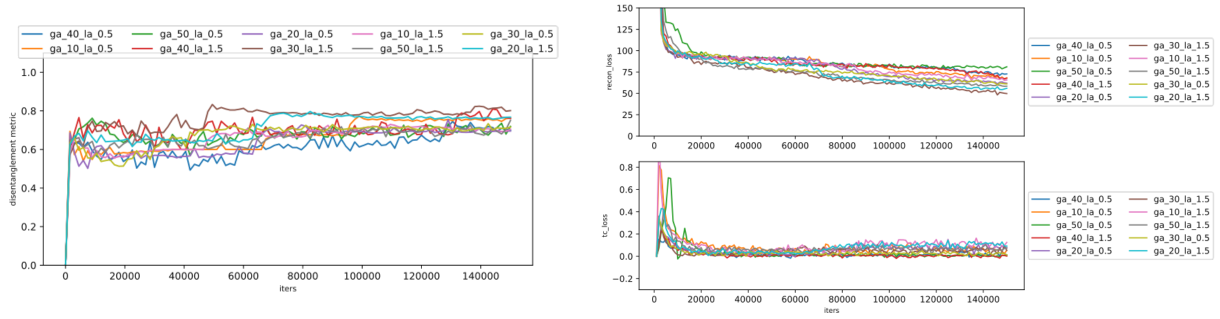


Figure 7: Left - Disentanglement metric over training iterations for the AD-FactorVAE γ and λ combinations tested. Right - The top plot corresponds to the reconstruction loss over training iterations and the bottom to the Total Correlation over training iterations.

4.2.2 Anomaly attention maps with Stacked RBMs on MNIST

The results from the stacked RBM do not seem to show meaningful information regarding potential anomalous regions in the image as with the VAE anomaly attention maps. We suspect this is caused by the use of fully connected layers in which the weights are not spatially correlated, as opposed to convolutional layer. Therefore, the results might not be as interpretable. For more information and some example results see supplementary material D figure 12.

5 Discussion

Reproducing the results over all went seemingly well. Most of the qualitative results for the experiment on the anomaly detection experiments were in line with the results from the original paper. Thus, for anomaly detection we confirm claim (i) the results from the MNIST dataset are quantitatively good. Furthermore, claim (iii) is also confirmed, since our MVTec-AD results were similar to the results in the original paper.

However, claim (ii) for the UCSD dataset, was not reproducible, since the attention maps were qualitatively not localised enough and the quantitative re AUROC scores were more in line with the baseline.

Regarding disentanglement claim (iv) the results appear to be irreproducible given our assumption about the attention map selection strategy, since our results deviate strongly from the results in the original paper. For claim (v) the results cannot be validated, since the convolutional layer used for generating the attention maps in the encoder was not properly selected. In the next sections, the limits of the original paper and the difficulties as well as ease will be discussed.

5.1 What was easy

The easier part of the experiment was the first part, since the code was readily available and similar results were quickly reproduced.

5.2 What was difficult

The most difficult part of reproducing this paper was that much of the information was left up to the interpretation of the reader. For example, there is no explicit documentation about the implementation in the paper. Hyperparameters (e.g. value of λ, number of epochs and the learning rate), experiment set-up, dataset augmentation and model descriptions are not present. The only presence of this information is in a referral to the gradcam paper (Selvaraju et al., 2017), in which the implementation is explained in more detail. Additionally, there were many discrepancies between the paper and the codebase. For example, in the original paper, equation 5 for the attention disentanglement loss, is not clear on the selection mechanism for the attention maps A^1_{ij} and A^2_{ij} in equation 1 (our solution for this is explained in section 3.3). For more information and solutions on this see section 3.4.

Furthermore, in the paper a second approach for computing the score to backpropagate on for the attention maps in anomaly detection is presented. However, this method was not used in their paper and the equation (like many others) was not sufficiently explained; a *u* is mentioned, but there is no mention of what this *u* might refer to.

5.3 Communication with original authors

Some communication with the author was had through the use of issues in the author's Github repository. Some of the code was not running correctly due to different version of pytorch conflicting with each other hence the communication with the authors was made only in that occasion.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework.

Hyunjik Kim and Andriy Mnih. 2019. Disentangling by factorising.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32.

Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. 2020. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. 2017. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2020. Deep learning for deepfakes creation and detection: A survey.

Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. 2020. Face identity disentanglement via latent space mapping.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Nguyen Thanh Van, Tran Ngoc Thinh, et al. 2017. An anomaly-based network intrusion detection system using deep learning. In *2017 international conference on system science and engineering (ICSSE)*, pages 210–214. IEEE.