RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] Label-Free Explainability for Unsupervised Models

Eric Robin Langezaal[1,2], Jesse Belleman[1,2], Tim Veenboer[1,2], and Joeri Noorthoek[1,2]
[1]Equal contributions – [2]FNWI, University of Amsterdam, Amsterdam, The Netherlands

Edited by
Koustuv Sinha,
Maurits Bleeker,
Samarth Bhargav

Received
04 February 2023

Published
20 July 2023

DOI
10.5281/zenodo.8173656

Reproducibility Summary

Scope of Reproducibility — This study is an analysis of the reproducibility of the paper *Label-Free Explainability for Unsupervised Models* written by Jonathan Crabbe and Mihaela van der Schaar. The goal of this study is to verify the three main claims of the paper, which state that their new framework for label-free explainability is capable of extending (i) linear feature importance as well as (ii) example importance methods to the unsupervised setting, whilst guaranteeing crucial properties, such as completeness and invariance with respect to latent symmetries. Finally, they use their framework to (iii) challenge some common beliefs about the interpretability of disentangled VAEs.

Methodology — The paper came with an extensive codebase containing all the necessary scripts to replicate the experiments in the paper. When comparing latent representations we also experimented with a different feature importance algorithm. Furthermore, we extended an experiment with the addition of a state-of-the-art encoder.

Results — We find that the three main claims of the paper hold true as we were able to successfully reproduce each corresponding experiment. Our results differ in some minor aspects, but they do support the validity of the three main claims made in the paper. Furthermore, we also demonstrated that the framework proposed is expandable in new contexts, thus providing further support for its utility and applicability.

What was easy — The authors provided a substantial and thorough explanation within their appendix about the mathematical concepts of their work. Their repository was also supplemented with rigorous documentation which gave an excellent explanation of how to carry out various experiments.

What was difficult — Several experiments in the paper take multiple hours to execute, with one reproduction taking over 32 hours on the low-cost computational setup used. Moreover, a few setup bugs were found, which meant that reproducing the experiments was a more strenuous task than simply executing a series of command-line statements.

Communication with original authors — We initiated communication with the authors and inquired about specific experimental decisions made in their work. The authors provided a comprehensive response via email, addressing all questions raised.

Correspondence should be addressed to Eric Robin Langezaal (eric.langezaal@gmail.com)
The authors have declared that no competing interests exist.
Code is available at https://github.com/EricLangezaal/Re-Label-Free-XAI – DOI https://doi.org/10.5281/zenodo.7885638. – SWH swh:1:dir:9a38c921b461cf9af72bc8f878d52da442fd39c1.
Open peer review is available at https://openreview.net/forum?id=bBVZ3pY4z8p.

1 Introduction

Contemporary machine learning models are notorious for not being transparent in their decision-making. These models are often referred to as black-boxes within the research literature [1, 2]. In high-pressure environments where the risk of serious complications is continually present, the opaqueness of these models actively impedes their usage because their bias can lead to unfair and even dangerous affairs [3, 4]. The lack of interpretability of deep models attracted scores of researchers to the field of *explainable artificial intelligence* (XAI) [5, 6, 7].
A fundamental methodology applied within XAI is *post-hoc explainability*. This collection of methods offers insight into complex model decisions in terms of interpretability and explainability. This study concentrates on two specific varieties of post-hoc explainability: *feature importance* and *example importance*. Feature importance highlights which features the black-box model attends to in order to make its decision. Example importance examines which samples from the training process influence the decision-making on newly observed test examples.
The majority of post-hoc explainability research has been performed in a supervised learning setting. Supervised learning exclusively concerns data where there is an explicit relationship between the input space $\mathcal{X}$ and the label space $\mathcal{Y}$; the model in question attempts to document the correspondence between these two spaces such that $f : \mathcal{X} \mapsto \mathcal{Y}$. There have been a broad number of studies that have focused on uncovering the black-box in a supervised setting [8, 9], especially in the medical field [10, 11]. However, analysis of black boxes in an unsupervised setting has remained largely unexplored. In an unsupervised setting, the model under examination maps the input space $\mathcal{X}$ to a hidden representation or latent space $\mathcal{H}$, such that $f : \mathcal{X} \mapsto \mathcal{H}$.
In their paper, Crabbé and Schaar[12] propose several adjustments to supervised explainability methods to conduct an analysis of explainability in the realm of unsupervised models. This study is a review and reproduction of their research while also implementing a few new features. Their work can be dissected into a few key claims which will be discussed in the following section together with the scope of reproducibility of their research. Thereafter follows a section in which the methodology of the execution of the experiments will be explored. This is followed by a results section which is accompanied by an interpretation of the outcomes. Lastly, a discussion section will reflect upon the reproducibility, the approach taken in this study and possible future endeavours.

2 Scope of reproducibility

In their work, Crabbé and Schaar propose several extensions to transfer already existing feature and example importance methods from the supervised to the unsupervised domain. Making use of their newly proposed label-free explainability methods, they challenge the interpretability of disentangled representations. In this study we focus on the reproducibility of their paper and aim to verify the following claims:

- **Claim 1 - Label-free feature importance (LFI)**: The extension that allows computing feature importance scores in the label-free setting provides sensible scores and consequently, selecting pixels with higher importance scores (according to the various explainability methods) to be perturbed yields significantly larger latent shifts than perturbing random pixels.

- **Claim 2 - Label-free example importance (LEI)**: The authors postulate two types of example importance: loss-based and representation based. The similarity rate is calculated between the ground truth label of the examples and the label of the test example. This respective similarity rate is higher for the most important examples compared to the least important examples, which indicates the effectiveness of the label-free example importance methods.

- **Claim 3 - Interpretability of disentangled VAEs (IDV)**: The saliency maps of individual latent units in disentangled VAEs cannot be meaningfully interpreted on their own. The authors show that increasing the disentanglement in the VAE does not decrease the correlation between the saliency maps of different latent units. This is quantified by a study of the Pearson correlation coefficient between the saliency maps of different latent units.

3 Methodology

The authors provided an open-source implementation of their experiments on GitHub [1]. The repository is well-structured and contains a comprehensive README file, which explains the installation process for all necessary packages, as well as the steps required to run the scripts and reproduce the results reported in the paper. Aside from a few minor technical adjustments, namely the rectification of an import error and the correction of the order of two function arguments, the code supplied by the authors functioned as intended. As a result, we were able to generate an output for every experiment that the authors described in their research paper.

3.1 Model descriptions

For the consistency checks of the label-free proposed feature and example importance methods, three different models are fit on three different datasets. The hardware is described in Appendix E and the datasets are described in Appendix B. A denoising autoencoder CNN is fit on the MNIST dataset [13], an LSTM reconstruction autoencoder on the ECG5000 time series dataset [14] and a SimCLR [15] neural network with a ResNet-18 [16] backbone on the CIFAR-10 dataset [17]. We extract an encoder, denoted by f_e, from each model for interpretation of their respective latent spaces.
For the analysis of the disentangled VAEs, a β-VAE [18] and a TC-VAE [19] are set up and trained on the MNIST and dSprites [20] datasets. The number of latent units (d_H) used is 6 for MNIST and 3 for dSprites. For both VAEs, 5 models are trained for each $\beta \in \{1, 5, 10\}$. This results in 6 different VAE configurations, with 30 models being trained in total.

3.2 Hyperparameters

The pre-established hyperparameter settings from the source code and corresponding paper were used in our experimentation. However, the paper indicated that 20 runs of each VAE configuration were conducted, while the codebase only incorporated 5 runs. For the reproducibility study, we opted for 5 runs due to time constraints, see Table 3 in Appendix E.

3.3 Experimental setup and code

The experiments outlined in the original paper are briefly discussed in this section. Subsequently, supplementary experiments to their study are explained. For a thorough comprehension of the used mathematical notation we refer the reader to Appendix A.

[1] Code published on GitHub at https://https://github.com/vanderschaarlab/Label-Free-XAI

Consistency checks — In the first experiment, consistency checks are conducted to ascertain the validity of the proposed example and feature importance methods for unsupervised models. To evaluate the feature importance explanations and verify the LFI claim, we compute for each image $\boldsymbol{x}$ the label-free feature importance score $b_i(\boldsymbol{f}_e, \boldsymbol{x})$, making use of the existing feature importance methods Gradient Shap [21], Integrated Gradients [22] and Saliency [23]. The top M features, as determined by these scores, are masked using a mask $\boldsymbol{m} \in \{0,1\}^{d_X}$, where d_X is the dimensionality of the input space. The latent shift resulting from replacing these most important features with a baseline value, represented as $\bar{x}$, is subsequently measured as follows: $||\boldsymbol{f}_e(\boldsymbol{x}) - \boldsymbol{f}_e(\boldsymbol{m} \odot \boldsymbol{x} + (1-\boldsymbol{m}) \odot \bar{x})||$. The baselines used for different models can be found in Appendix C. The average shift across the entire test set is reported for various values of M and different feature importance methods. If the LFI claim is valid, we expect the latent shift to be higher when more important features are masked.

To evaluate the example importance methods and verify the LEI claim, a thousand training examples are sampled. For each training example the label-free example importance score $c^n(\boldsymbol{f}_e, \boldsymbol{x})$ is computed using several loss-based (Influence Functions [24] and TracIn [25]) and representation-based (Deep-KNN [26] and SimplEx [27]) example importance methods (the definition of these methods can be found in Appendices A.2 and A.3). To verify the saliency of high-scoring examples, the M most important examples $(x^{n_1}, .., x^{n_M})$ are selected and their ground truth labels $(y^{n_1}, .., y^{n_M})$ are compared to the label y of the to be explained test image x. Subsequently, the similarity rates are computed: $\frac{1}{M}\sum_{m=1}^{M} \delta_{y,y^{n_m}}$, where δ is the Kronecker delta. The same experiment is conducted for the M least important examples. The distribution of the similarity rates across a thousand test examples is computed for various values of M. If the LEI claim holds, the similarity between the most important examples is expected to be higher than for the least important examples.

Comparing representations from different pretext tasks — To test the efficacy of the feature and example importance scores, Crabbé and Schaar introduce a specific use case. Neural models, in this particular case the autoencoder, can be adapted to be suitable for different pretext tasks on the MNIST dataset. The aim of this experiment is to establish how the latent representations from different pretext tasks compare to each other. The denoising autoencoder is used as described in Section 3.1. Furthermore, two additional pretext tasks with their autoencoders are considered: reconstruction and inpainting [28]. Lastly, a classifier is trained that includes the same layers as the encoder, as well as an additional linear layer with an input dimension of 4 and an output dimension of 10, which uses a Softmax activation function to transform the latent representations into class probabilities. The hidden representations of the classifier are extracted from the penultimate layer. For each encoder $\boldsymbol{f}_e$, the label-free Gradient Shap feature importance method (introduced in the consistency checks paragraph) is used to extract saliency maps from the feature importance scores $b_i(\boldsymbol{f}_e, \boldsymbol{x})$ for the test images. For the comparison of the saliency maps produced by the different models, the average Pearson correlation coefficient [29] is computed across 5 runs.

We use the label-free Deep-KNN example importance method (introduced in the consistency checks paragraph) to compute the example importance $c^n(\boldsymbol{f}_e, \boldsymbol{x})$ of a thousand training examples for a thousand test images. Again, the average Pearson coefficient across 5 runs is used to compare the example importance scores created by the different encoders.

Besides the quantitative experiments above, qualitative experiments were conducted in the original paper. The top examples of the various encoders for a particular test image are displayed alongside its saliency maps to convey the qualitative differences between the autoencoders.

Extensions of the pretext use case — In the results part of Section 4.1, Crabbé and Schaar state that 'Label-free Integrated Gradients outperform other methods for each model', which is also confirmed in Figure 1. Yet in the pretext task experiment described above, Gradient Shap is used, which the authors motivate by its better computational efficiency. However, in our experience the Integrated Gradients feature importance method actually required less computational time compared to Gradient Shap (Table 3 in Appendix E). Since the former distinctly outperforms the latter (Figure 1), we opted to therefore also repeat the pretext experiment from Section 4.2 of the original paper with the Integrated Gradients feature importance method.
Furthermore, in their discussion section the authors propose that the analysis of the label-free feature and example based importance scores can be extended in several ways. One of the suggested methods is the addition of state-of-the-art (SOTA) autoencoders which can be utilised in comparison with the other pretext models. Examining the latent representation of a SOTA autoencoder in relation to the hidden representations of other autoencoders might yield interesting results. We opted for experimenting with the Stacked Capsule Autoencoder [30] (SCAE), which is detailed in Appendix D.1.

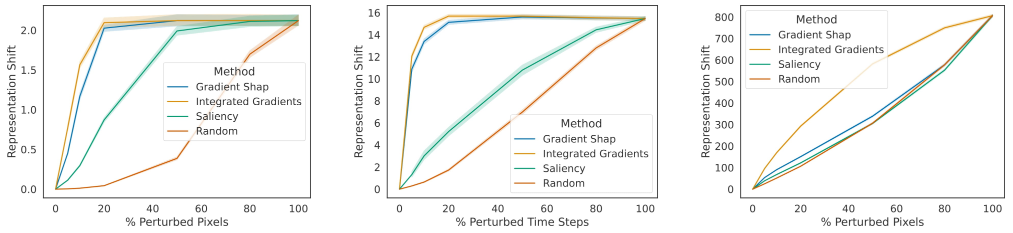
Disentangled VAEs — In the original paper the interpretability of the latent units in disentangled VAEs is analysed using their saliency maps generated by a feature importance method. These units are exclusively sensitive to a single data generative factor, allowing for interpretable meanings. The aim of their experiments is to determine whether it is possible to identify the associated generative factor of each latent unit by analysing their generated saliency maps. To answer this question and verify the IDV claim, qualitative and quantitative experiments have been set up.
The VAE models as described in Section 3.1 are evaluated with Gradient Shap to get an importance score $a_i(\mu_j, \boldsymbol{x})$ for each pixel x_i from an image $\boldsymbol{x}$ to predict the latent units $\mu_j \in [d_H]$. For a quantitative result, the Pearson correlation between the saliency maps of different latent pairs is averaged over 5 runs. A low Pearson correlation indicates that the latent units attend to different parts of the image. Therefore, we pick the VAE configuration with the lowest average correlation for both datasets. For these selected VAEs, the saliency maps are shown for 2 test images on which we can perform a qualitative analysis. Lastly, the influence of β (which regulates the degree of disentanglement in the VAEs) on the Pearson correlation between the different latent units is evaluated by making boxplots of the correlations for various β values. If the IDV claim holds, the Pearson correlation should not necessarily decrease as β increases.

4 Results

4.1 Results reproducing original paper

The following section will enumerate the obtained results from reproducing the experiments done by Crabbé and Schaar. The results received from the additional experiments will be displayed subsequently. Each subsection will examine the original claim presented by the authors and show our obtained results.

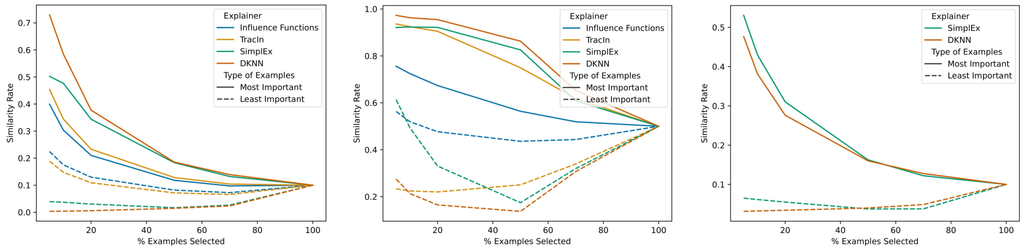
Label-Free Feature Importance — The results of this experiment are presented in Figure 1, which illustrates that our LFI consistency checks for the MNIST and ECG5000 datasets are consistent with those of the original study. However, upon conducting LFI consistency checks for the CIFAR10 dataset, a deviation in trend was observed for all four feature importance methods, with a side-by-side comparison being depicted in Figure 6 in Appendix F.1. Nevertheless, the phenomenon where Integrated Gradients and Gradient Shap outperform the random baseline is similar to what was observed in the original paper, albeit with a different curve shape.



(a) MNIST (b) ECG5000 (c) CIFAR10

Figure 1. Consistency check for label-free feature importance (average and 95% confidence interval).

Label-Free Example Importance — Figure 2 showcases the results of this reproducibility experiment, which demonstrates that the LEI consistency checks for both the MNIST and ECG5000 dataset are consistent with those of the original study. However, similar to the observations made in the reproducibility experiment of LFI, the LEI consistency checks for the CIFAR10 dataset in our study deviate from the trend that was observed for the LEI consistency checks for the CIFAR10 dataset in the original paper. The main difference is the scale of the y-axis, the similarity rate, which is greater in our experiment, as illustrated side-by-side in Figure 7 in Appendix F.2.



(a) MNIST (b) ECG5000 (c) CIFAR10

Figure 2. Consistency check for label-free example importance (only representation-based methods apply to SimCLR).

Pretext use case comparison — Table 1 contains information from two distinct experiments, namely the reproduction of the Pearson correlations displayed in the original paper acquired using Gradient Shap feature importance **and** our extension of applying the Integrated Gradients method, which will be discussed together with our addition of SCAE in Section 4.2.

Grad. Shap Int. Grad.	Recon.	Denois.	Inpaint.	Classif.	SCAE
Reconstruction		$.40 \pm .01$	$.33 \pm .03$	$.45 \pm .01$	$.39 \pm .01$
Denoising	$.47 \pm .06$		$.30 \pm .02$	$.39 \pm .02$	$.37 \pm .03$
Inpainting	$.45 \pm .07$	$.43 \pm .05$		$.32 \pm .01$	$.41 \pm .03$
Classification	$.42 \pm .01$	$.38 \pm .04$	$.35 \pm .03$		$.43 \pm .02$
SCAE	$.35 \pm .01$	$.37 \pm .03$	$.42 \pm .02$	$.42 \pm .01$	

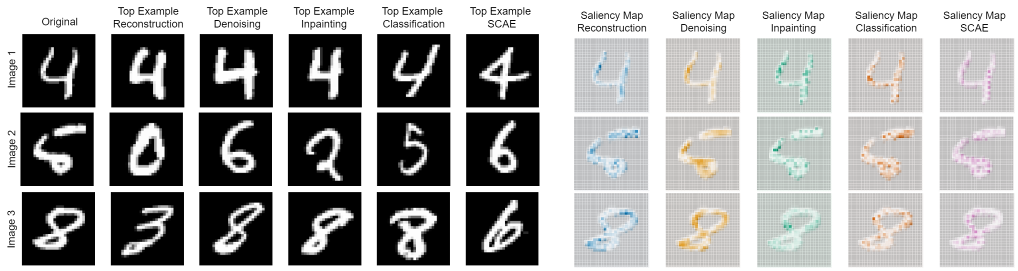
Table 1. Pearson correlation for saliency maps (avg +/- std) for feature importance methods Gradient Shap (Reconstruction) and Integrated Gradients.

Table 2 contains Pearson correlations for the Deep-KNN example importance method, which just like the Gradient Shap results from Table 1 are consistent with the correlation scores reported in the original study.

	Reconstruction	Denoising	Inpainting	Classification	SCAE
Reconstruction					
Denoising	$.12 \pm .06$				
Inpainting	$.15 \pm .06$	$.17 \pm .09$			
Classification	$.08 \pm .03$	$.07 \pm .03$	$.08 \pm .04$		
SCAE	$.08 \pm .02$	$.10 \pm .03$	$.10 \pm .02$	$.05 \pm .01$	

Table 2. Reproduced Pearson correlation (avg +/- std) for the Deep-KNN example importance method.

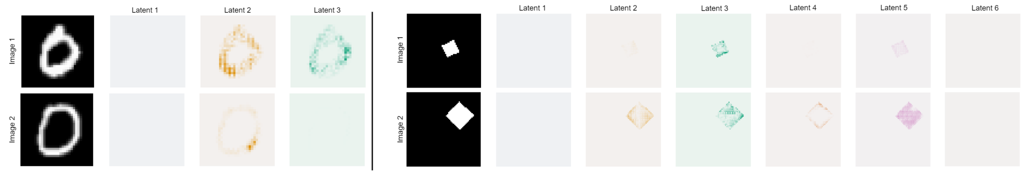
The reproducibility results for the qualitative analysis of the label-free explainability framework are presented in Figure 3 and are coherent with the results of the original study. Specifically for feature importance, the saliency maps for different pretext tasks vary heavily, which is consistent with the original study. Furthermore, the top examples are rarely similar, as is also suggested by the quantitative analysis done by Crabbé and Schaar.



(a) Top Examples **(b)** Saliency Maps

Figure 3. Label-Free explanations for varying pretext tasks.

Disentangled VAEs – To qualitatively comment on the saliency maps of different latent units, we pick the combination of VAE type and corresponding β value that lead to the lowest average Pearson correlation to display. Differently from the original study, the lowest correlation (Appendix F.3) was achieved by a TC-VAE with $\beta = 10$ for MNIST, and a β-VAE with $\beta = 5$ for the dSprites dataset.

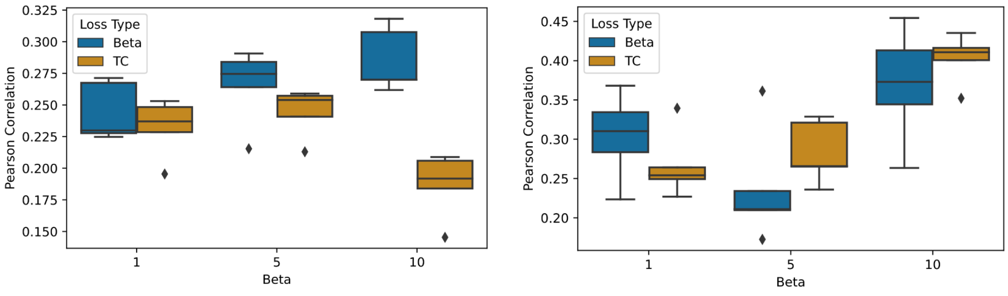


(a) MNIST (TC-VAE, $\beta = 10$) **(b)** dSprites (β-VAE, $\beta = 5$)

Figure 4. Saliency maps of the models with lowest Pearson Correlation.

Figure 4 displays the feature importance per latent unit for two different images for both datasets, using the VAE configurations above. We observe that it is difficult to meaningfully interpret the saliency maps for a given latent unit. Looking at the MNIST results, latent unit 3 has moderate to strong feature activation for the first test image while having no feature activation for the second image; nevertheless, both images represent the same digit. Moreover, feature activation does not appear to be region bound for each

latent unit. The second test image in the dSprites case shows similar feature activation for different latent units. Therefore, we cannot reach a conclusion about the activation of a latent unit with regard to certain features being present in an image based on these saliency maps, which supports the IDV claim.
The boxplots in Figure 5 display an increase of correlation between the saliency maps of latent units as β grows. Only the TC-VAE for the MNIST dataset does not exhibit this trend. This refutes the concept that increasing disentanglement leads to less correlation between the latent units of a disentangled VAE. However, it is significant to mention that the boxplots in Figure 5 are fairly different than the boxplots produced by Crabbé and Schaar. Combined with the large margins of uncertainty of each box, it can be argued that drawing a specific conclusion from these plots might be unproductive.



(a) MNIST

(b) dSprites

Figure 5. Pearson Correlation for the saliency maps of differing values of β.

4.2 Results beyond original paper

Integrated Gradients – Our results for feature importance using Integrated Gradients are compared to the results of the original paper, which uses Gradient Shap. These findings are presented in Table 1, which shows an average Pearson correlation for Integrated Gradients and Gradient Shap of 0.406 and 0.379 respectively. Combining this with the lower computational cost for Integrated Gradients (see Table 3 in Appendix E) would suggest that Integrated Gradients is a better feature importance method.

Addition of SCAE – It becomes apparent from Table 1 that the important features denoted by the SCAE correlate relatively weakly to the important features produced by reconstruction and denoising. This might possibly be due to the fact that the pretext tasks of reconstruction and denoising are significantly different from the task of the SCAE. We can observe that the SCAE correlates better with inpainting. Moreover, we can notice that Table 2 displays a weak correlation between the chosen training examples by the SCAE and the training examples of all other models. Lastly, the saliency maps displayed in Figure 3b show that the SCAE's highlighted features significantly vary from the important features of the other pretext tasks.

5 Discussion

Firstly, the reproduced results in Figure 1 and Table 1 support the LFI claim. Using the label-free feature importance framework proposed by Crabbé and Schaar, the three different methods significantly outperform the random baseline in the consistency checks. However, some deviations from the results of the original study were found for the CIFAR-10 dataset with the SimCLR model. This different curve shape is most likely caused by small differences in the SimCLR model, according to the original authors. Nonetheless, the conclusions that can be drawn from either figure remain the same.

The consistency checks indicated that the label-free feature framework performs best on the Integrated Gradients method, for both the reproduced and original results.
Secondly, the experimental results of this paper support the LEI claim; the results for the consistency checks for the label-free example importance methods align with the results of the original paper.
Moreover, the results in this paper also substantiate the IDV claim since the Pearson correlation between latent units does not decrease as the disentanglement parameter β is increased. Thus we can conclude that more disentanglement does not cause deterioration of the similarity between the saliency maps of varying latent units.
Lastly, the addition of the SCAE autoencoder in this paper demonstrates that the provided label-free framework also works on other autoencoders, which further strengthens the LFI and LEI claims. Further dissection of the results specific to the SCAE model can be found in Appendix D.2.

5.1 What was easy

The codebase provided by the authors was extensive and easy to execute with a command line interface. Moreover, the paper itself had a large appendix in which the authors explained the mathematics behind their label-free feature and example importance methods in considerable detail. This appendix also contained additional results of the experiments which made for a seamless comparison with our reproduction of the research.

5.2 What was difficult

Even though the source code was vast and thoroughly documented, we encountered several software-breaking bugs which needed to be resolved before reproduction was possible. The errors ranged from simply recovering a missing import to improper function arguments for example importance methods such as DKNN. Furthermore, the mathematics behind both Gradient Shap and TracIn required a solid understanding of fairly complicated multivariate calculus and statistics.

5.3 Communication with original authors

Contact with the original authors of the paper was established through email. We have posed a series of questions regarding the reproducibility of the original results. First of all, the LFI results reproduced for CIFAR-10 in Figure 1c differed from the paper to an extent where only Integrated Gradients truly matched. The response was that SimCLR's weight initialisation might have differed in both studies and therefore created a different baseline for the results. Furthermore, we noticed the authors applied Gradient Shap to the pretext use case, regardless of the fact that Integrated Gradients displayed better performance in the consistency checks. The authors utilised Gradient Shap because they stated the method is less computationally expensive than Integrated Gradients. In our reproduction however, we observed the converse to be the case (see Appendix E Table 3). Finally, the obtained boxplots for the IDV claim differed significantly. The authors mentioned that VAEs are extremely unstable and difficult to seed, thus giving a high likelihood of dissimilarity between runs.

References

1. P. W. Koh and P. Liang. "Understanding Black-box Predictions via Influence Functions." In: **Proceedings of the 34th International Conference on Machine Learning**. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1885–1894. URL: https://proceedings.mlr.press/v70/koh17a.html.
2. A. I. Poon and J. J. Sung. "Opening the black box of AI-Medicine." In: **Journal of Gastroenterology and Hepatology** 36.3 (2021), pp. 581–584.
3. J. Petch, S. Di, and W. Nelson. "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology." In: **Canadian Journal of Cardiology** 38.2 (2022). Focus Issue: New Digital Technologies in Cardiology, pp. 204–213. DOI: https://doi.org/10.1016/j.cjca.2021.09.004. URL: https://www.sciencedirect.com/science/article/pii/S0828282X21007030.
4. G. Varoquaux and V. Cheplygina. "Machine learning for medical imaging: methodological failures and recommendations for the future." In: **npj Digital Medicine** 5 (Apr. 2022), p. 48. DOI: 10.1038/s41746-022-00592-y.
5. A. Gosiewska and P. Biecek. "iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models." In: **CoRR** abs/1903.11420 (2019). arXiv:1903.11420. URL: http://arxiv.org/abs/1903.11420.
6. P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, and S. Melacci. "Entropy-based Logic Explanations of Neural Networks." In: **CoRR** abs/2106.06804 (2021). arXiv:2106.06804. URL: https://arxiv.org/abs/2106.06804.
7. P. Hall, N. Gill, and N. Schmidt. **Proposed Guidelines for the Responsible Use of Explainable Machine Learning**. 2019. DOI: 10.48550/ARXIV.1906.03533. URL: https://arxiv.org/abs/1906.03533.
8. Z. C. Lipton. "The Mythos of Model Interpretability." In: **CoRR** abs/1606.03490 (2016). arXiv:1606.03490. URL: http://arxiv.org/abs/1606.03490.
9. B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." In: **Proceedings of the 35th International Conference on Machine Learning**. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 2668–2677. URL: https://proceedings.mlr.press/v80/kim18d.html.
10. T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, et al. "Opportunities and obstacles for deep learning in biology and medicine." In: **Journal of The Royal Society Interface** 15.141 (2018), p. 20170387.
11. E. Tjoa and C. Guan. "A survey on explainable artificial intelligence (xai): Toward medical xai." In: **IEEE transactions on neural networks and learning systems** 32.11 (2020), pp. 4793–4813.
12. J. Crabbé and M. van der Schaar. "Label-Free Explainability for Unsupervised Models." In: **Proceedings of the 39th International Conference on Machine Learning**. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 4391–4420. URL: https://proceedings.mlr.press/v162/crabbe22a.html.
13. L. Deng. "The mnist database of handwritten digit images for machine learning research." In: **IEEE Signal Processing Magazine** 29.6 (2012), pp. 141–142.
14. A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." In: **circulation** 101.23 (2000), e215–e220.
15. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations." In: **International conference on machine learning**. PMLR. 2020, pp. 1597–1607.
16. K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016, pp. 770–778.
17. A. Krizhevsky, G. Hinton, et al. "Learning multiple layers of features from tiny images." In: (2009).
18. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. "beta-vae: Learning basic visual concepts with a constrained variational framework." In: (2016).
19. R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. "Isolating sources of disentanglement in variational autoencoders." In: **Advances in neural information processing systems** 31 (2018).
20. L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. **dSprites: Disentanglement testing Sprites dataset**. https://github.com/deepmind/dsprites-dataset/. 2017.
21. S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions." In: **Advances in neural information processing systems** 30 (2017).
22. M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic attribution for deep networks." In: **International conference on machine learning**. PMLR. 2017, pp. 3319–3328.
23. K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." In: **arXiv preprint arXiv:1312.6034** (2013).
24. P. W. Koh and P. Liang. "Understanding black-box predictions via influence functions." In: **International conference on machine learning**. PMLR. 2017, pp. 1885–1894.

25. G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. “Estimating training data influence by tracing gradient descent.” In: **Advances in Neural Information Processing Systems** 33 (2020), pp. 19920–19930.
26. N. Papernot and P. McDaniel. “Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning.” In: **arXiv preprint arXiv:1803.04765** (2018).
27. J. Crabbé, Z. Qian, F. Imrie, and M. van der Schaar. “Explaining Latent Representations with a Corpus of Examples.” In: **Advances in Neural Information Processing Systems** 34 (2021), pp. 12154–12166.
28. P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher. “Inference with deep generative priors in high dimensions.” In: **IEEE Journal on Selected Areas in Information Theory** 1.1 (2020), pp. 336–347.
29. O. Le Meur and T. Baccino. “Methods for comparing scanpaths and saliency maps: strengths and weaknesses.” In: **Behavior research methods** 45.1 (2013), pp. 251–266.
30. A. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton. “Stacked Capsule Autoencoders.” In: **Advances in Neural Information Processing Systems**. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/2e0d41e02c5be4668ec1b0730b3346a8-Paper.pdf.

A Mathematical definitions

The following definitions are directly quoted from the original paper [12].

A.1 Definition 1 (Label-Free Feature Importance)

Let $f : \mathcal{X} \to \mathcal{H}$ be a black-box latent map and for all $i \in [d_X]$ let $a_i(\cdot, \cdot) : \mathcal{A}(\mathbb{R}^{\mathcal{X}}) \times \mathcal{X} \to \mathbb{R}$ be a feature importance score linear w.r.t. its first argument. We define the label-free feature importance as a score $b_i(\cdot, \cdot) : \mathcal{A}(\mathcal{H}^{\mathcal{X}}) \times \mathcal{X} \to \mathbb{R}$:

$$b_i(f, x) \equiv a_i(g_x, x)$$

$g_x : \mathcal{X} \to \mathbb{R}$ such that for all $\tilde{x} \in \mathcal{X}$:
$g_x(\tilde{x}) = \langle f(x), f(\tilde{x}) \rangle_{\mathcal{H}}$,
where$\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes an inner product for the space $\mathcal{H}$.

A.2 Definition 2 (Label-Free loss-based example importance)

Let $f_{\theta_r} : \mathcal{X} \to \mathcal{H}$ be a black-box latent map trained to minimize a loss $L : \mathcal{X} \times \Theta \to \mathbb{R}$ on a training set $\mathcal{D}_{train} = \{x^n | n \in [N]\}$ (N is the length of the training set). To measure the impact of removing example x^n from $\mathcal{D}_{train}$ with $n \in [N]$, we define the *Label-Free loss-based example importance* as a score $c^n(\cdot, \cdot) : \mathcal{A}(\mathcal{H}^{\mathcal{X}}) \times \mathcal{X} \to \mathbb{R}$ such that:

$$c^n(f_{\theta_r}, x) = \delta^n_{\theta_r} L(x, \theta_*). \tag{1}$$

Where $\delta^n_{\theta_r} L(x, \theta_*)$ is an estimation of the loss shift that can be evaluated by using the Influence Functions method[24] or the TracIn method [25].

A.3 Definition 3 (Label-Free representation-based example importance)

Besides the loss based-example importance methods, representation-based example importance methods are used to attribute an importance score to examples.

To quantify the affinity between x and the training set examples, we attempt a reconstruction of $f_e(x)$ with training representations from $f_e(D_{train})$: $f_e(x) \approx \sum_{n=1}^{N} w_n(x) \cdot f_e(x_n)$. The first approach [26] to define weights $w^n(x)$ is to identify the indices $KNN(x) \subset [N]$ of the K nearest neighbours (DKNN) of $f_e(x)$ in $f_e(D_{train})$ and weigh them according to a Kernel function $\kappa : \mathcal{H}^2 \to \mathbb{R}^+$:

$$w^n(x) = \mathbf{1}[n \in KNN(x)] \cdot \kappa[f_e(x_n), f_e(x)], \tag{2}$$

Where $\mathbf{1}$ denotes the indicator function.
A second method (SimplEx) to learn the weights [27] is by solving:

$$w(x) = argmin_{\lambda \in [0,1]^N} \left\| f_e(x) - \sum_{n=1}^{N} \lambda^n f_e(x^n) \right\|_{\mathcal{H}}, \tag{3}$$

where $\sum_{n=1}^{N} \lambda^n = 1$. For the Label-Free setting, we can take $c^n = w^n$ without any additional work and thus compute the Label-Free example importance score this way.

B Datasets

Three different datasets are used for the quantitative and qualitative experiments regarding label-free feature and example importance. An additional fourth dataset is used for the study on disentangled VAEs. The datasets will be briefly discussed below.

MNIST — The MNIST dataset is a commonly used dataset for image classification tasks. It comprises 60,000 training and 10,000 test images of handwritten digits in the range 0-9, each with corresponding labels. All images are in grayscale and have a size of 28x28 pixels. We corrupt each training image with random noise $\epsilon \sim \mathcal{N}(0, \frac{1}{3}\mathbf{I})$ where $\mathbf{I}$ is the identity matrix.

ECG5000 — This dataset contains 5000 univariate time series, each describing the heartbeat of a patient. Each time series comes with a binary label indicating whether the heartbeat is normal or not.

CIFAR-10 — A dataset consisting of 60,000 32x32 color images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images. The 10 classes are: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

dSprites — The dSprites dataset, which is commonly used to assess the disentanglement properties of unsupervised models, consists of 2D shapes that are procedurally generated from 6 ground truth independent latent factors: color, shape, scale, rotation and x and y positions of a sprite. All possible combinations of the latent variable values add to a total of 737.280 64x64 black and white images.

C Baselines latent shift

To compute the latent shift, several baselines $\tilde{x}$ are used for different models. As a baseline for determining feature importance in the MNIST dataset, a black image is used represented by $\tilde{\boldsymbol{x}} = 0$. In the ECG5000 dataset, the average normal heartbeat is used as the baseline represented by $\tilde{\boldsymbol{x}} = \sum_{x \in \mathcal{D}_{train}} \frac{x}{|\mathcal{D}_{train}|}$. In the CIFAR-10 dataset, a blurred version of the image being explained is used as the baseline, represented by $\tilde{\boldsymbol{x}} = G_\sigma \otimes \boldsymbol{x}$, where G_σ is a Gaussian blur of kernel size 21 with width $\sigma = 5$ and $\otimes$ represents the convolution operator.

D Stacked Capsule Autoencoder (SCAE)

D.1 SCAE setup

The Stacked Capsule Autoencoder [30] (SCAE) employs geometric relationships between parts of objects to make a reasonable reconstruction while also estimating which object parts are present within a given image. The SCAE comprises of two separate models, the Part Capsule Autoencoder (PCAE - regarded as the encoder) and the Object Capsule Autoencoder (OCAE - regarded as the decoder). The PCAE receives an input image $\mathbf{y}$ and transforms it into an M number of capsules. Each capsule m contains a pose vector $\mathbf{x}_m$ which denotes spatial information about a part of the image, a presence probability $d_m \in [0, 1]$ and a vector of special features $\mathbf{z}_m$. The special features are a distinguishable attribute of that specific part of the image; in the paper, the authors simply use colour as the special feature. Alongside colour, the special features z_m for a particular capsule also track the level of transparency of a given object to simulate occlusions. Each of these capsules is then applied to a template T_m to create a transformed template $\hat{T}_m$. These templates are learned sets of parameters. The core concept is that the received geometric information from the part capsules and the transformed templates are sufficient for the reconstruction of the image by the OCAE. Each capsule is only allowed for a single usage, meaning that an image can be reconstructed as a weighted combination of M capsules containing geometrical information and a single template. The output of the PCAE is therefore a capsule plus a flattened transformed template for each capsule in M.

For this research, we flatten the capsules to create a Z dimensional latent representation when computing the feature and example importance scores (not during initial training) because the methods of importance attribution enforce this form of latent representation. This essentially does not harm the information stored in the latent space, since the Z dimensional latent space now just comprises of $M \times (\mathbf{x}_m, d_m, \mathbf{z}_m, \hat{T}_m)$ vectors stacked on top of each other and we can still evaluate the latent shift accordingly.

D.2 SCAE discussion

Table 1 displayed that the SCAE has the strongest measure of correlation with an inpainting autoencoder. As mentioned in the previous subsection, the part capsules track the level of transparency of the corresponding parts to inform the decoder about possible occlusions. The main concept of inpainting is reconstructing parts of an image hidden by specific occlusions. This might cause the models to more deeply correlate in their hidden representation.

Furthermore, there is an overall weak correlation with the other pretext tasks for the example importance case (shown in Table 2). The latent space of SCAE's encoder is specifically modelled towards a given structure. The hidden representation is dissected in part capsules which all contain information about separate locations of the image plus an additional template which it deforms with that same information. This strongly distinct manner of organising the latent space might explain the low correlation of selected top examples to top examples of other models.

Lastly, the saliency maps of the SCAE (Figure 3b) vary substantially from the other models' saliency maps. This is more evidence that the task of the SCAE greatly differs from the tasks presented in the original paper. The SCAE pays attention to alternative pixels within the MNIST digits, presumably because it is attempting to obtain geometric objects from the images rather than performing a more standard task such as reconstruction.

E Computational requirements

All our experiments are run on a cluster whose nodes are equipped with Nvidia Titan RTX GPUs. Running the reproducibility experiments comes at a total computational cost of 68 GPU hours. Our own experiments, which make use of the model checkpoints, come at a total computational cost of 4.5 GPU hours. A further breakdown of the computational costs is presented in Table 3.

Experiment type	Experiment name	Section	Dataset	Dataset specific GPU hours	Total GPU hours
Reproducibility	Consistency features	4.1	MNIST	0.376	0.765
			ECG5000	0.033	
			CIFAR10	0.356	
	Consistency examples	4.1	MNIST	3.212	23.791
			ECG5000	20.567	
			CIFAR10	0.012	
	Use Case: representations learned with different pretext tasks	4.2	MNIST	3.550	3.550
	Challenging our assumptions with disentangled VAEs	4.3	MNIST	7.725	39.941
			dSprites	32.216	
Additional experiments	Extension 1 of the pretext use case: The Stacked Capsule Autoencoder (SCAE)	-	MNIST	1.106	1.106
	Extension 2 of the pretext use case: Integrated Gradients instead of Gradient Shap for feature attribution	-	MNIST	3.386	3.386

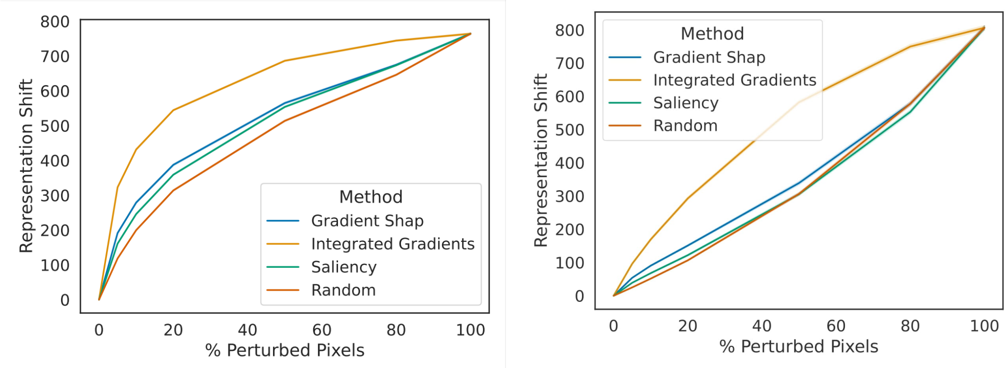
Table 3. Overview of the computational cost, in terms of GPU hours, for each experiment.

F Reproducability differences

While reproducing the experiments from the original paper some differences were encountered with respect to the original figures in the study. Since our main report only details the results obtained when replicating the experiment, this section also highlights the original figures copied from the paper by Crabbé and Schaar.

F.1 Label-Free feature importance

While the three different feature importance methods and the random baseline showed the same effect in representation shift on the MNIST and ECG5000 datasets as the original study, our results differed quite substantially on the CIFAR10 dataset. This different curve shape is most likely caused by the difference in weight initialisation for the SimCLR model.



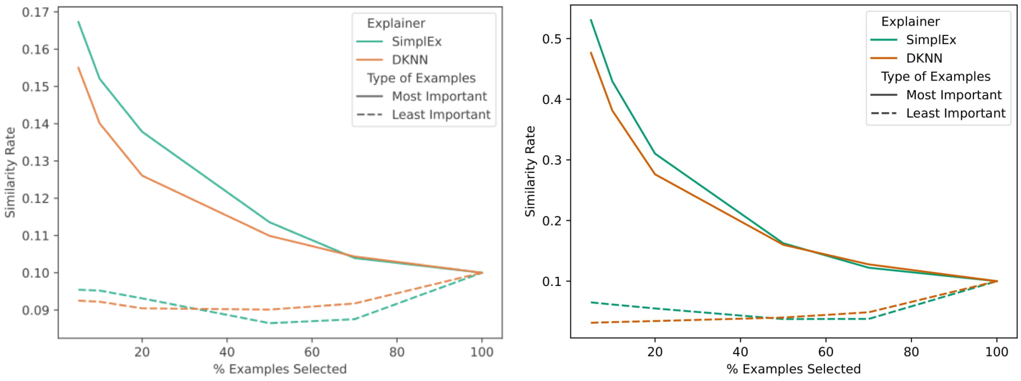
(a) Result original paper

(b) Reproducibility study result

Figure 6. Differences between the original paper and our reproducibility study for the consistency check for label-free feature importance on the CIFAR10 dataset (average and 95% confidence interval).

F.2 Label-Free example importance

While the shape of the original and reproduced sets of graphs in figure 7 match, the similarity rates of our results are scaled up compared to the original study, in a very consistent manner. Again, this deviation is most likely caused by a difference in weight initialisation for the SimCLR model.



(a) Result original paper

(b) Reproducibility study result

Figure 7. Differences between the original paper and our reproducibility study for the consistency check for label-free example importance on the CIFAR10 dataset (average and 95% confidence interval). Notice the difference in scale for the similarity rate on the y-axis.

F.3 Saliency correlations for different disentangled VAEs

In the final experiment of the paper, different combinations of two types of disentangled VAEs and corresponding β values are trained for multiple runs. For a qualitative analysis, the authors state they showcase saliency maps for the configurations with the lowest Pearson correlation coefficient, as that corresponds to latent units paying attention to distinct parts of the images, which is desirable for disentanglement. It is reported that this entails a β-VAE with $\beta = 10$ for MNIST and a TC-VAE with $\beta = 1$ for dSprites, but the actual Pearson scores this decision is based on are not reported. In our reproduction study however, the criterion of using the lowest Pearson score, averaged over 5 runs, leads to a different conclusion for the configuration to use for both datasets. Therefore selecting a TC-VAE with $\beta = 10$ for MNIST and a β-VAE with $\beta = 5$, as per table 4.

VAE Type	β	Average Pearson (n=5)	
		MNIST	dSprites
Beta	1	0.024426	0.303969
TC	1	0.023251	0.266775
Beta	5	0.026577	**0.237737**
TC	5	0.024480	0.283371
Beta	10	0.028552	0.369678
TC	10	**0.018717**	0.403007

Table 4. Averaged Pearson correlation coefficient for different types of VAEs and values of β.