Extensions the Label-free Explainability Framework for Unsupervised Models

Ying Huang [1]

Abstract

Interpretability remains a significant challenge in the field of machine learning, particularly within the unsupervised learning setting. This report proposes a set of extensions building upon the foundational work of the Label-free Explainability Framework for Unsupervised Models (Crabbé & van der Schaar, 2022). We introduce four novel directions accompanied by corresponding experimental designs, aiming to enhance the framework's capacity to elucidate the underlying mechanisms of unsupervised algorithms. Our approach seeks to bridge the gap between complex model behaviours and human-understandable explanations, thereby advancing the interpretability of machine learning models without the reliance on labelled data.

1. Introduction

In recent years, the field of machine learning has experienced remarkable growth, leading to significant advancements in various domains such as image processing, natural language understanding, and medical diagnostics. One of the fundamental challenges in this field, however, is the interpretability of machine learning models, particularly in unsupervised learning scenarios where the lack of labelled data adds complexity to understanding model behaviours and decisions.

This report presents two novel extensions to the Label-free Explainability Framework for Unsupervised Models(Crabbé & van der Schaar, 2022). Our contributions aim to enrich this framework by introducing innovative approaches that enhance the interpretability of unsupervised models. We concentrate on their application in the realm of medical imaging and propose a prospective variant of the Variational Autoencoder (VAE). Additionally, we explore the fusion of diverse VAE architectures, aiming to significantly elevate the explainability of these models.

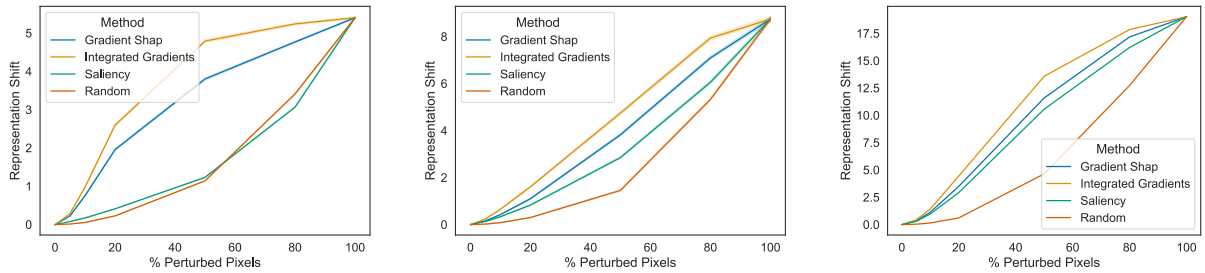
Structured in a coherent manner, this report first addresses the proposed extensions in Section 2, focusing on their applicability to medical image analysis and interpretation. Section 3 introduces our innovative VAE proposal, complemented by both qualitative and quantitative analyses, and delineates prospective research trajectories. The report culminates in Section 4 with a concise summary of our findings and contributions, reflecting the advancements made in unsupervised model interpretability.

2. Application to Medical Image

Machine learning and deep learning techniques have become increasingly prevalent in medical image analysis, offering novel methods to enhance diagnostic accuracy and efficiency (Chen et al., 2022). However, the integration of unsupervised learning techniques into clinical practice faces resistance due to the inherent opacity of such 'black-box' models. This reluctance stems from the challenges healthcare professionals encounter when attempting to interpret and trust the decision-making processes of these models(Raza & Singh, 2021). The proposed framework is poised to demystify the workings of these black-box algorithms, thereby fostering a deeper understanding and acceptance of unsupervised models in healthcare settings, and potentially widening their application in the industry.

We conducted an analysis on three distinct subsets derived from the MedMNIST dataset (Yang et al., 2023): ChestMNIST, which consists of chest radiographs; OrganSMNIST, encompassing abdominal CT images of human organs from various angles; and BreastMNIST, containing breast ultrasound scans. These datasets are characterized by their unique imaging modalities. Our methodology involved obscuring the top M salient features to compute the feature importances as per the protocol described in (Crabbé & van der Schaar, 2022), with the outcomes depicted in Figure 1.

Upon examination, it was observed that the Integrated Gradients and Gradient Shap methods consistently surpassed the performance of alternative techniques. Conversely, the method employing random feature occlusion was invariably identified as the least effective strategy. Notably, the degree of representation shift displayed dataset-specific sensitivity. For example, within the OrganSMNIST dataset, the shift maintains a linear trend across a broad range, implying that



(a) ChestMNIST (b) OrganSMNIST (c) BreastMNIST

Figure 1. Comparative Analysis of MedMNIST Subsets: ChestMNIST, OrganSMNIST, and BreastMNIST.

occluding between 20% to 80% of the most critical features results in a uniform impact on the representation shift. This suggests that there may not be any pivotal features that initially cause significant fluctuations in the shift.

Consequently, the variability in performance across different datasets necessitates further investigation and resolution. Moving forward, this framework can be leveraged to discern the true significance of the most important features identified. For instance, it has the potential to aid in isolating specific regions within medical images that could distinguish between two similar diseases.

3. New VAE proposal

The β-VAE and TC-VAE models have been scrutinized for their disentanglement efficacy. It has been observed that they do not consistently demonstrate a uniform response to a single factor across the same latent dimensions. Moreover, enhancing the model's disentanglement does not necessarily ensure an increase in the independence of latent dimensions. Consequently, we propose the adoption of a new disentanglement VAE approach to delineate the role of each latent factor. The FactorVAE, as introduced in (Kim & Mnih, 2019), fosters factorial distribution of representations, thereby promoting independence across dimensions.

3.1. Setup

The VAE involves a variational encoder computing the expected representation $\mu : X \to H$ as well as its standard deviation $\sigma : X \to H$, and a decoder $f_d : H \to X$. To evaluate the performance of our VAE model, we conducted training on the dSprites dataset, adhering to a 90% training and 10% testing data split. The primary goal was to minimise the model's objective function. During this process, we set the dimensionality of the latent space to $d_H = 6$ and experimented with varying levels of the hyperparameter γ, specifically testing the values $\gamma = [1, 5, 10]$. This approach allowed us to assess the model's behaviour under different degrees of regularization and disentanglement within the latent space.

3.2. Results

Qualitative Analysis We expect that the saliency maps for each latent dimension will distinctly highlight different sections of the original images, exhibiting consistency within the same latent dimension for similar images (such as the same shape at different positions or sizes). The saliency maps presented in Figure 2 suggest that the latent dimensions of the FactorVAE exhibit greater consistency in their response to specific factors (for instance, variations in the images affect primarily latent dimensions 1, 4, and 5). Moreover, the features focused on by different latent dimensions vary (as exemplified by image 1's latent dimensions 1 and 4, which concentrate on the left and right sides, respectively), demonstrating a significant degree of complementarity. This indicates a heightened level of independence within the latent space.

Quantative Analysis We also want to investigate the relationship between the disentanglement of a VAE and the independence of latent units. We can see that as the disentanglement parameter γ increases, the Pearson correlations decrease correspondingly, which implies that FactorVAE exactly has better performance on the independence of latent units.

3.3. Future Work

Although FactorVAE represents an advancement over β-VAE and TC-VAE, delineating each latent unit with distinct and specific feature factors remains a challenge. VQ-VAE may offer some insights, as it employs vector quantisation techniques to discretise the latent space (van den Oord et al., 2017). Ideally, this leads to each discrete unit capturing a different aspect of the input data, thereby minimising overlap and correlation between latent units. The discrete nature of the latent units ensures that each code vector serves as a definitive representation of the data, which is typically more

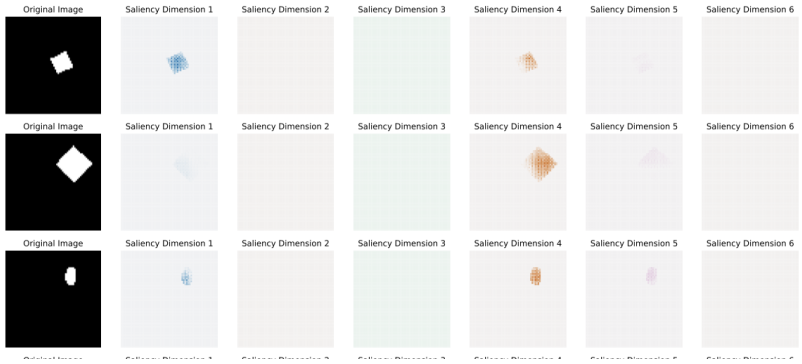


Figure 2. Saliency maps for each unit of the disentangled FactorVAE on dSprites Dataset.

interpretable than a vague, continuous latent space. Therefore, for future developments, we could envisage a new VAE framework that combines the strengths of VQ-VAE with those of disentangled VAEs by maintaining the FactorVAE's discriminator to encourage independence among the latent variables. Such a framework would facilitate a more intuitive understanding of the diverse meanings each latent unit represents.

However, a common issue arises when using autoencoders, such as VAE or VQ-VAE, for feature extraction: the models may lean towards employing convolutional computations and compressed representations to enhance efficiency. This approach effectively reduces the data's dimensionality and increases the information compression rate but can lead to the so-called 'factor mixing' problem. Therefore, in designing a new VAE framework, it's crucial to find solutions to address this issue. One potential approach is to integrate non-convolutional layers, such as fully connected or recurrent layers, which could help lower the feature compression rate, thereby mitigating the problem of feature mixing. Recent studies have corroborated the efficacy of this approach for feature extraction, exemplified by RNN-based VAEs' capability in efficiently extracting and classifying features from time-series data (Huang et al., 2019).

4. Conclusion

In this report, we have delved into two innovative extensions within the realm of machine learning, particularly focusing on image analysis and exploring the potential of a new VAE framework. These extensions represent significant strides in our ongoing efforts to enhance the interpretability and applicability of machine learning models in various domains.

A pivotal aspect of our research was the investigation of feature importance in medical images. Our findings indicate a certain sensitivity of feature importance to different parts or modalities of medical images. Specifically, in the case of the OrganSMNIST dataset, which includes images of human organs from various angles, the performance was notably subpar. This underscores the need for further enhancement in the generalization capabilities of our framework, especially when it's applied to feature explanation and analysis in medical imaging.

Additionally, we experimented with a new disentangled VAE, namely FactorVAE, and observed its superior performance in terms of the independence of latent units. A noteworthy discovery was the apparent negative correlation between the degree of disentanglement and the independence of these units. However, despite these advancements, the direct interpretation of significant feature factors represented by the latent units from the images remains elusive. To address this challenge and push the boundaries of our current understanding, we propose a potential new VAE framework. This framework aims to combine the discrete nature of VQ-VAE, which minimizes feature overlap, with the independence of latent units found in FactorVAE. Such a synthesis could pave the way for a more intuitive and interpretable understanding of the features represented by the latent units in VAE models.

In conclusion, our journey towards fully explainable and generalizable label-free models continues. The insights gained from this research highlight both the advancements and the challenges in the field. We believe that the future lies in further refining these methods, exploring their applicability in broader contexts, and continually striving for models that are not only powerful in their predictive capabilities but also transparent and interpretable in their decision-making processes.

References

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T. C., Moore, K., Mannel, R. S., Liu, H., Zheng, B., and Qiu, Y. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79(102444), 2022.

Crabbé, J. and van der Schaar, M. Label-free explainability for unsupervised models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4391–4420. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/crabbe22a.html.

Huang, Y., Chen, C.-H., and Huang, A. C.-J. Motor fault detection and feature extraction using rnn-based variational autoencoder. *IEEE Access*, 7, 2019. doi: 10.1109/access.2019.2940769.

Kim, H. and Mnih, A. Disentangling by factorising. In *Learning Disentangled Representations: from Perception to Control NIPS 2017 Workshop*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Raza, K. and Singh, N. K. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging Reviews*, 17(9), 2021.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1711.00937.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., and Ke, B. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), 2023.

A. You *can* have an appendix here.