Incorporating Feature Attribution Priors into Unsupervised Models

Harry Amad
harryamad@gmail.com

Abstract

Beyond the transparency that post-hoc explainability methods provide to black-box models, feature attributions can be directly incorporated into model training to encourage models to conform to certain properties. This application has been investigated to incorporate prior preferences of feature attribution in supervised learning tasks, however it is currently underexplored in the unsupervised learning domain. In this short paper, I extend the use of feature attribution priors to unsupervised models, showing that they can effectively be used to tailor the latent representations learned by variational autoencoders (VAE). Furthermore, I use this framework to propose a new approach to increasing the interpretability of VAEs beyond disentanglement, by encouraging latent units to be sensitive to separate features, rather than generative factors which can be somewhat nebulous. I implement a few example feature attribution priors, and show some preliminary results of how they affect representation learning on the MNIST dataset.[1]

1 Introduction

Deep learning has had a profound impact on the capacity of machine learning models to solve complex tasks, both in supervised [1] and unsupervised [2, 3] settings. To achieve this increase in performance however, deep models often sacrifice interpretability. Understanding and explaining how models work is critical to establish trust, which is necessary for widespread adoption, particularly in high stakes fields such as healthcare [4, 5].

Post-hoc explainability To address this, post-hoc explainability methods have received a lot of interest, as they help to relate the importance of features and training instances to the outputs of deep models. In this paper I will focus on feature attribution methods which highlight the importance of each feature for a given model output. Many methods have been proposed for this purpose [6, 7, 8], althought they largely focused on supervised learning. [9] recently showed that these methods can also be used for unsupervised tasks. This paper builds off the discovery of label-free post-hoc explainability methods by investigating how informing unsupervised models of their own feature attributions can influence their training.

Feature importance priors This use of feature attribution informed training has been explored in the supervised setting [10, 11, 12, 13], however to the best of my knowledge it has yet to be used in unsupervised learning. This is perhaps because of the recency of [9], which was the first paper to demonstrate comprehensive feature attribution in the label-free setting. Given this new work, incorporating previous feature attribution priors proposed for supervised tasks into the learning of unsupervised models is quite straightforward, as I will demonstrate in the following sections. Furthermore, I focus on a particular use case for label-free feature attribution priors - increasing the interpretability of representations in VAEs. I propose novel priors to encourage the latent units in the

[1] Code available at `https://github.com/harrya32/encoder_attribution_priors`

encoder of a VAE to be sensitive to separate input features, and demonstrate empirically how this both changes the representations learned by VAEs and can increase their performance

2 Related Work

Increasing how explainable and interpretable machine learning models are has been a major area of recent research, and without continued work useful models will struggle to gain meaningful adoption due to how poorly understood they are [14]. Using post-hoc explainability metrics to directly influence training and build tailored, interpretable models has been explored a fair amount in supervised settings. Priors have been explored with applications in image classification [12], text classification [15], drug response prediction [13] and beyond. There has not been similar development for unsupervised tasks.

Disentangled VAEs Encouraging VAEs to learn lower dimensional representations of data that are disentangled—that is, latent units focus on independent generative factors—has been explored to improve model interpretability. Disentangled representations are more interpretable and better approximate how humans learn [16], and disentanglement has been substantially explored in VAEs [17, 18, 19, 20]. However, with the help of label-free feature attribution methods, [9] highlights that increasing the disentanglement in two popular disentangled VAE frameworks (β-VAE [17] and TC-VAE [20]) does not, in fact, increase the separation of their feature attribution maps. This is because these frameworks encourage the latent units to identify distinct generative factors from the data, which does not guarantee that distinct features will be learnt. There have not been similar works that seek to directly separate the sensitivity of latent units to input features through feature attribution priors. In this work I show that this avenue encourages latent units to better learn independent feature attribution maps than previous disentanglement works.

3 Method

I will now describe how feature attribution priors can be implemented into unsupervised models, and VAEs specifically. Considering an input dataset $X \in \mathbb{R}^{d_X}$, where $d_X \in \mathbb{N}^*$ is the dimension of the input space, typical unsupervised deep learning models seek to find parameters θ such that

$$\theta = \text{argmin}_\theta \mathcal{L}(x, \theta) + \lambda\Omega(\theta) \tag{1}$$

where $\mathcal{L}(x, \theta)$ is some loss function and $\Omega(\theta)$ is a regularisation term weighted by λ. Considering a feature attribution method $A(\theta, x)$, which is a matrix where each entry $a_{i,j}$ is the importance of feature i for the output of sample j, we can view feature attribution priors as a regularisation term $\Omega(A(\theta, x))$ that encourages a models parameters to conform to a given condition. Ω can be defined as any differentiable function and it depends on the context of the problem which the model is trying to solve.

There are already some proposed priors $\Omega(A(\theta, x))$ for supervised learning tasks with various goals. One such example, proposed in [12], is a prior which encourages feature attribution maps in image classification to have low total variation, and therefore high smoothness, and this is defined as

$$\Omega_{\text{pixel}}(A(\theta, x)) = \sum_l \sum_{i,j} |a^l_{i+1,j} - a^l_{i,j}| + |a^l_{i,j+1} - a^l_{i,j}| \tag{2}$$

where $a^l_{i,j}$ is the attribution for the i, jth pixel on the lth training image. [12] propose this prior to encourage nearby pixels to have similar attributions, and they show that classification models are more robust to noisy test images when trained with this prior.

Shifting now to the context of VAEs, consider the input dataset X as a collection of random variables from a generative process $p(x)$. A VAE seeks to learn a latent variable model $p_\theta(x, z)$ that captures the generative process, where $z \in \mathbb{R}^{d_H}$ and $d_H \in \mathbb{N}^*$ is the dimension of the latent space. This is typically done with a parametric encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$, where θ and ϕ are parameters of a neural network which minimise the objective

$$\mathcal{L}(x;\theta,\phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)||p(z)) \tag{3}$$

With this task and loss function in mind, we can propose a VAE framework that encorporates feature attribution priors. In this paper, I will focus on feature attribution priors that only look to enforce behaviours on the encoder parameters θ. Such priors can be used to impart expert knowledge on a VAE, guiding its latent units z to be sensitive to particular features. A VAE with a feature attribution prior $\Omega(A(\theta, x))$ will look to optimise its parameters with the objective function

$$\mathcal{L}(x;\theta,\phi) + \lambda\Omega(A(\theta, x)) \tag{4}$$

As discussed above, it is important that we look to design models that are interpretable, and we can improve the interpretability of VAEs by separating which features each latent unit is sensitive to. In [9] two metrics are calculated from the saliency maps of each latent unit to measure how separate they are - entropy and Pearson correlation. It is desirable for a VAE to have latent saliency maps with low entropy, meaning that the importance of a given feature is not spread evenly among the latent units, and to have low correlation, as this implies that latent units are focusing on different features. With this in mind, I propose to incorporate these metrics directly into the training of VAEs with the following priors.

Formally, I follow the definitions in [9] for the entropy and Pearson correlation between saliency maps. Consider now the feature attribution notation $a_i(\mu_j, x)$ which is the importance of feature i on the average value of the latent unit j for input x. Define the proportion of the attribution for a feature $i \in d_X$ to a latent unit $j \in d_H$ for instance x as

$$p_j(i, x) = \frac{|a_i(\mu_j, x)|}{\sum_{k=1}^{d_H} |a_i(\mu_k, x)|} \tag{5}$$

Since $\sum_{j=1}^{d_H} p_j(i, x) = 1$ by construction, this can be interpreted as the probability of saliency, allowing the calculation of entropy over the latent units

$$S(i, x) = -\sum_{j=1}^{d_H} p_j(i, x) \log p_j(i, x) \tag{6}$$

This entropy is minimised as 0 when only one latent unit j is sensitive to feature i, and maximised as $\log d_H$ when the saliency is distributed uniformly over the latents. Ultimately, I define a prior using the average entropy over a dataset X and its features I as

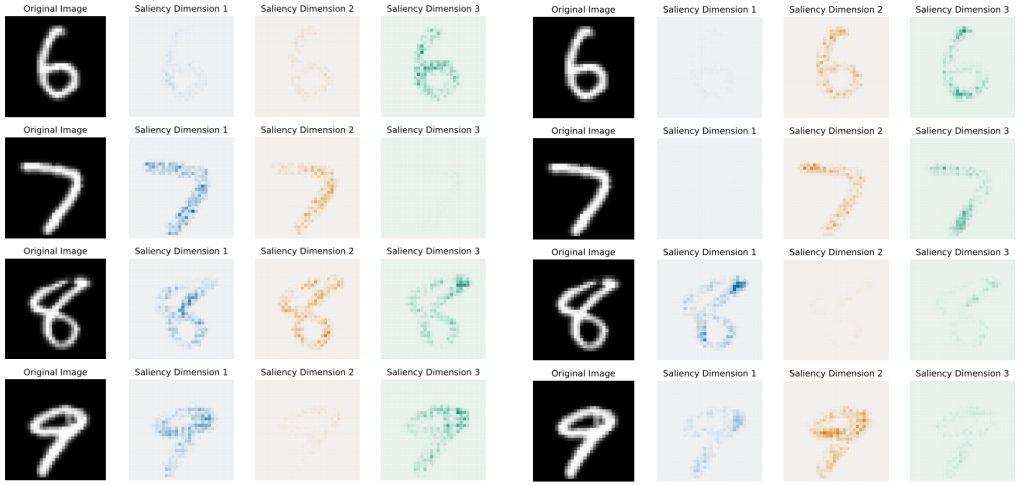
$$\Omega_{\text{entropy}}(A(\theta, x)) = \mathbb{E}_{X,I}[S(I, X)] \tag{7}$$

As for the Pearson correlation, given two latent units i, j, their correlation can be calculated as

$$r_{i,j} = \frac{\text{cov}_{X,I}[a_I(\mu_i, X), a_I(\mu_j, X)]}{\sigma_{X,I}[a_I(\mu_i, X)]\sigma_{X,I}[a_I(\mu_j, X)]} \tag{8}$$

where X and I are chosen randomly from the data distribution and the feature set respectively. To get the average correlation for a VAE's latent units, we sum this value across all pairs of latents, which I define as the Pearson correlation prior

$$\Omega_{\text{pearson correlation}}(A(\theta, x)) = \frac{1}{d_H(d_H - 1)} \sum_{i=1, i\neq j}^{d_H} \sum_{j=1}^{d_H} r_{i,j} \tag{9}$$



(a) Ordinary VAE.

(b) VAE with total variation prior with $\lambda = 0.01$.

Figure 1: Saliency maps for each latent unit of VAEs with and without a total variation prior.

4 Results

I will now conduct a few experiments to show the utility of these feature attribution priors on VAEs. While I only focus on one kind of unsupervised learning task here, representation learning, there is no reason that the methods discussed above cannot be applied to a wider array of applications beyond the VAE framework. These experiments are quite preliminary and non-exhaustive because of limited computing power and time, however they do highlight the immediate effects that feature attribution priors can have, and scaling this to more complicated datasets is a task for future work. The following experiments are all conducted on the MNIST dataset [21] using the same architecture as [9] for the VAEs, and using Gradient Shap to measure feature importance.

4.1 Pixel attribution prior

I will begin by implementing the pixel attribution prior in (2), as a short exploration to show that label-free feature attributions enable the porting of previous works, such as those in [12], into the unsupervised realm. In Figure 1 I compare the saliency map for the three latent units in an ordinary VAE, with no pixel attribution prior, with those from a VAE with the prior enabled, with $\lambda = 0.01$. This prior does indeed have an effect, as the saliency maps in Figure 1b are slightly visually smoother, as neighbouring pixels tend to have more similar importance levels. Furthermore, the average total variation for the saliency maps of the VAE with the pixel attribution prior is $50\times$ lower than the base model, with less than 1% increase in reconstruction error across the MNIST test dataset. These results were achieved without any significant hyperparameter optimisation, and so it is reasonable to assume that even greater increases in saliency map smoothness can be achieved without sacrificing performance.

4.2 Entropy attribution prior

Having established that feature attribution priors can impact how a VAE learns in practice, I now move to the novel investigation of separating VAE latent feature attributions to increase interpretability. I implement the prior for the entropy of feature attributions as described in (7) and investigate how latent representations change with λ. To determine the extent to which latent units are paying attention to separate features, I primarily use the Pearson correlation and the entropy between their saliency maps, where low entropy and low correlation indicate good separation and therefore better interpretability.

Figure 2 show boxplots of these metrics at different λ, with five models trained at each level. As λ increases, the entropy does indeed decrease as desired, however this also results in an increased Pearson correlation, which is detrimental to the interpretability of the learned representations.

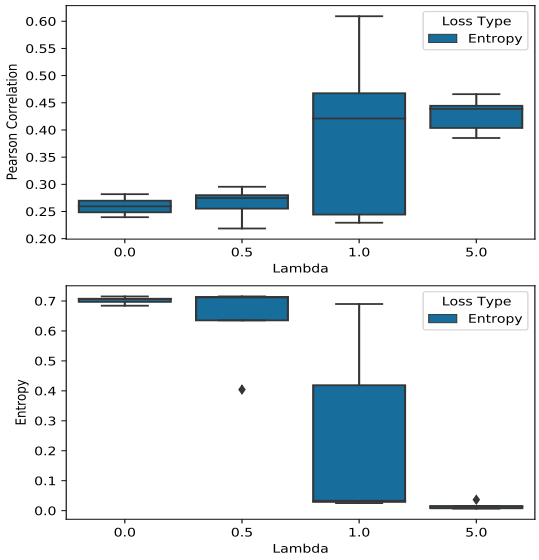


Figure 2: Pearson correlation and entropy of saliency maps for VAE with entropy prior at different λ.

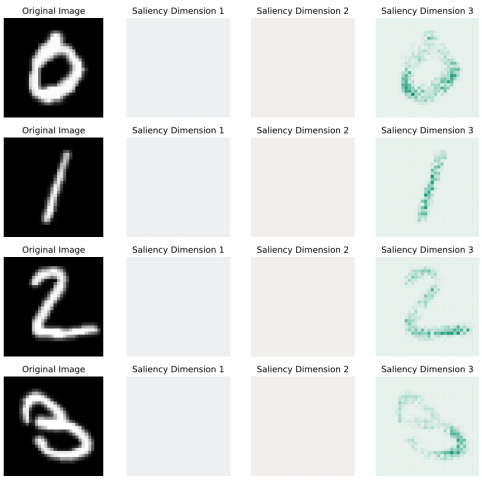


Figure 3: Saliency map for each latent unit of VAE with entropy prior with $\lambda = 1$.

It should be noted that I found training unstable using this prior, particularly with larger λ, as this pushed the model towards a failure state of having almost all of the feature importance in a single latent unit, in search of low entropy. This case is shown in Figure 3, where with $\lambda = 1$ only the third latent unit is receptive to the images' pixels, while the other two latents are dormant.

This is clearly undesirable, as this artificially decreases the model capacity, resulting in a decrease in performance. Figure 4 shows that as λ increases, so too does the reconstruction loss on the MNIST test dataset.

4.3 Pearson correlation attribution prior

Finally, I implement the Pearson correlation prior as described in (9). Figure 5 shows how the Pearson correlation and entropy between the latent units' saliency maps change with λ, again with

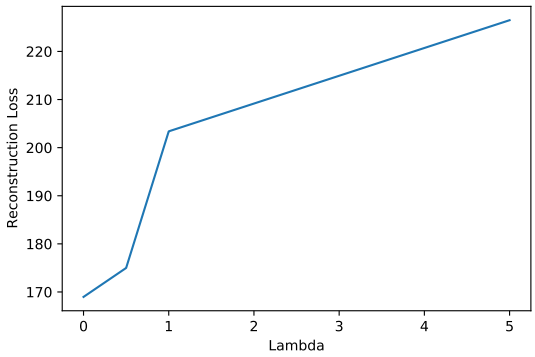


Figure 4: Reconstruction loss on MNIST test dataset for VAE with entropy prior at different λ.

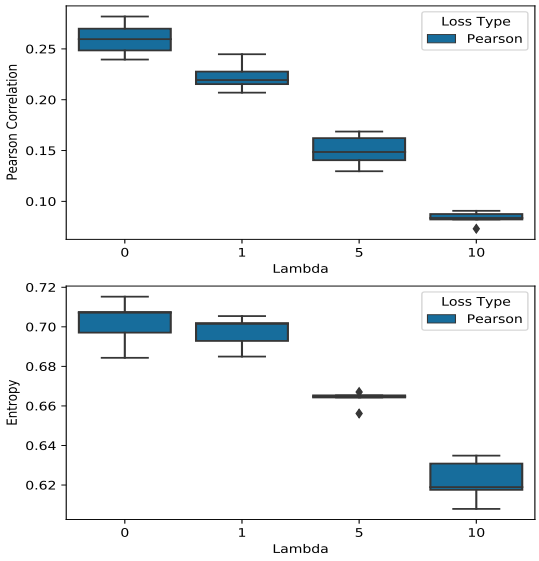


Figure 5: Pearson correlation and entropy of saliency maps for VAE with Pearson correlation prior at different λ.

five models trained at each level. As λ increases, both the Pearson correlation and the entropy decrease substantially, indicating that the latent units are becoming more separate, and therefore more interpretable. At the highest level of $\lambda = 10$, the Pearson correlation is close to 0, meaning that the latent units are almost uncorrelated. This is well below the correlation found for β-VAEs and TC-VAEs in [9], demonstrating that this feature attribution prior better separates features of interest between the latent units than these disentanglement methods.

This prior was much more stable than the entropy prior, and no obvious failure cases where reached during training. Furthermore, not only does this prior encourage the latent units to be more separate, it also resulted in increased performance. Figure 6 shows that as λ increases, the reconstruction loss on the MNIST test dataset decreases, in stark contrast to what resulted from the entropy prior. This improvement does not plateau for the λ values explored here, suggesting that exploration of larger λ values could result in further performance increases.

While this quantitative improvement in both reconstruction performance and separation is encouraging, it is worth examining the saliency maps that this prior induces, to further investigate how these

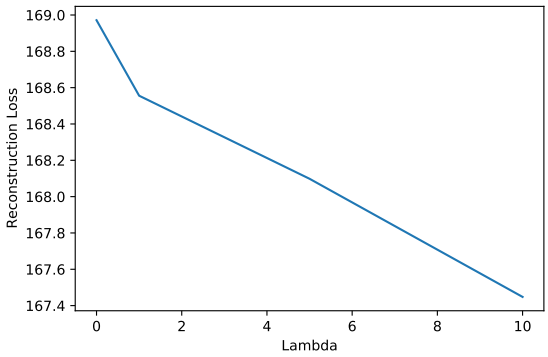
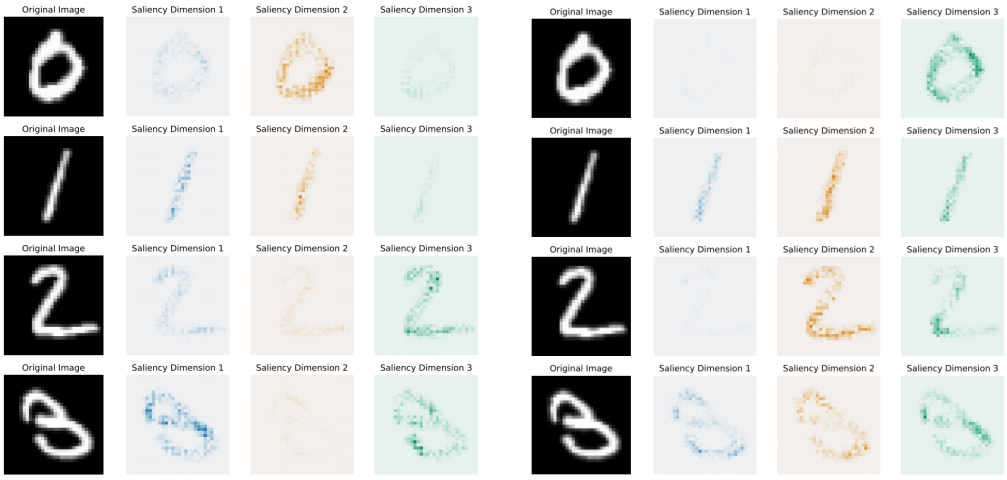


Figure 6: Reconstruction loss on MNIST test dataset for VAE with Pearson correlation prior at different λ.



(a) Ordinary VAE.

(b) VAE with Pearson correlation prior with $\lambda = 10$.

Figure 7: Saliency maps for each latent unit of VAEs with and without Pearson correlation prior.

results are achieved. Figure 7 compares the saliency maps of a base VAE with those of a VAE with the Pearson correlation prior with $\lambda = 10$. Despite the large difference in the Pearson correlation as shown in Figure 5, it is difficult to visually see this, as there is no immediately obvious characteristic differences between these saliency maps. This is worth exploring further, on more datasets with different applications, as ideally this increase in the separation of the focus each latent unit would be more visible.

5 Discussion

In this paper, I showed that label-free feature attributions can be used to imbue prior knowledge of a desried feature attribution into the learning of unsupervised models. Beyond the examples displayed here, this enables the use of a variety of priors that can be tailored to the individual contexts in which the models are being used. This allows expert knowledge to be better incorporated into unsupervised models, which could be used to help increase their performance, explainability, generalisation, and more

The particular use case of feature attribution priors that I demonstrated here increased both the interpretability and performance of VAEs by informing the model of its own saliency maps during training. Encouraging VAE latent units to focus on separate features can lead to a better understanding of the representation space. This increases the interpretability of VAEs beyond previous works, as

I showed my method pushes latent units to focus on separate features more so than β-VAEs and TC-VAEs. Increasing the interpretability of VAEs will help to increase trust that practitioners and users can have in them, which is important in a number of fields, not least of which is medical care where VAEs have multiple potential use cases [22, 23].

There is ample room to explore this space further in future work. Improvements on the priors proposed for latent unit separation should be investigated, such as combining the entropy and Pearson correlation priors into one regularisation function. Furthermore, experimentation on more complicated and varied datasets is necessary.

References

[1] Yanzheng Yu. "Deep Learning Approaches for Image Classification". In: *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*. EITCE '22. Xiamen, China: Association for Computing Machinery, 2023, pp. 1494–1498. ISBN: 9781450397148. DOI: 10.1145/3573428.3573691. URL: https://doi.org/10.1145/3573428.3573691.

[2] Absalom E. Ezugwu et al. "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects". In: *Engineering Applications of Artificial Intelligence* 110 (2022), p. 104743. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2022.104743. URL: https://www.sciencedirect.com/science/article/pii/S095219762200046X.

[3] Sam Bond-Taylor et al. "Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (Nov. 2022), pp. 7327–7347. ISSN: 1939-3539. DOI: 10.1109/tpami.2021.3116668. URL: http://dx.doi.org/10.1109/TPAMI.2021.3116668.

[4] Riccardo Miotto et al. "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in Bioinformatics* 19.6 (May 2017), pp. 1236–1246. ISSN: 1477-4054. DOI: 10.1093/bib/bbx044. eprint: https://academic.oup.com/bib/article-pdf/19/6/1236/27119191/bbx044.pdf. URL: https://doi.org/10.1093/bib/bbx044.

[5] Travers Ching et al. "Opportunities and obstacles for deep learning in biology and medicine". In: *bioRxiv* (2018). DOI: 10.1101/142760. eprint: https://www.biorxiv.org/content/early/2018/01/19/142760.full.pdf. URL: https://www.biorxiv.org/content/early/2018/01/19/142760.

[6] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV].

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

[8] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI].

[9] Jonathan Crabbé and Mihaela van der Schaar. *Label-Free Explainability for Unsupervised Models*. 2022. arXiv: 2203.01928 [cs.LG].

[10] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. *Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations*. 2017. arXiv: 1703.03717 [cs.LG].

[11] Laura Rieger et al. *Interpretations are useful: penalizing explanations to align neural networks with prior knowledge*. 2020. arXiv: 1909.13584 [cs.LG].

[12] Gabriel Erion et al. *Improving performance of deep learning models with axiomatic attribution priors and expected gradients*. 2020. arXiv: 1906.10670 [cs.LG].

[13] Ethan Weinberger, Joseph Janizek, and Su-In Lee. *Learning Deep Attribution Priors Based On Prior Knowledge*. 2020. arXiv: 1912.10065 [cs.LG].

[14] Leander Weber et al. "Beyond explaining: Opportunities and challenges of XAI-based model improvement". In: *Information Fusion* 92 (2023), pp. 154–176. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2022.11.013. URL: https://www.sciencedirect.com/science/article/pii/S1566253522002238.

[15] Frederick Liu and Besim Avci. *Incorporating Priors with Feature Attribution on Text Classification*. 2019. arXiv: 1906.08286 [cs.CL].

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. *Representation Learning: A Review and New Perspectives*. 2014. arXiv: 1206.5538 [cs.LG].

[17] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[18] Mhd Hasan Sarhan et al. *Learning Interpretable Disentangled Representations using Adversarial VAEs*. 2019. arXiv: 1904.08491 [cs.LG].

[19] Emile Mathieu et al. *Disentangling Disentanglement in Variational Autoencoders*. 2019. arXiv: 1812.02833 [stat.ML].

[20] Ricky T. Q. Chen et al. *Isolating Sources of Disentanglement in Variational Autoencoders*. 2019. arXiv: 1802.04942 [cs.LG].

[21] Li Deng. "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142. DOI: 10.1109/MSP.2012.2211477.

[22] Dimitris Papadopoulos and Vangelis D. Karalis. "Variational Autoencoders for Data Augmentation in Clinical Studies". In: *Applied Sciences* 13.15 (2023). ISSN: 2076-3417. DOI: 10.3390/app13158793. URL: https://www.mdpi.com/2076-3417/13/15/8793.

[23] Jan Ehrhardt and Matthias Wilms. "Chapter 8 - Autoencoders and variational autoencoders in medical image analysis". In: *Biomedical Image Synthesis and Simulation*. Ed. by Ninon Burgos and David Svoboda. The MICCAI Society book Series. Academic Press, 2022, pp. 129–162. ISBN: 978-0-12-824349-7. DOI: https://doi.org/10.1016/B978-0-12-824349-7.00015-3. URL: https://www.sciencedirect.com/science/article/pii/B9780128243497000153.