SpaCE-VAE: Sparse and Confident Explanations using Variational Autoencoders

Alexander Liu
Eindhoven University of Technology
Groene Loper 5
Eindhoven, The Netherlands
a.liu1@student.tue.nl

Sibylle Hess
Eindhoven University of Technology
Groene Loper 5
Eindhoven, The Netherlands
s.c.hess@tue.nl

ABSTRACT

In the field of explainable AI (XAI), a significant challenge lies in the evaluation of explanation methods. State-of-the-art techniques identify the importance of input features (e.g., pixels) for the classification, based on untestable assumptions about the model. Lowering the intensity of pixels, identified as irrelevant, typically leads to a change in the predicted class. To address this, we propose SpaCE-VAE, a Variational Autoencoder (VAE) designed to generate sparse, confident explanations that are inherently evaluable. SpaCE-VAE produces a sparse representation of an image, putting as many pixels as possible to black, such that the resulting image is still assigned to the same class as the original image (with high confidence).

1. INTRODUCTION

Pairing the empirical success of Deep Neural Networks (DNNs) with a trustworthy explanation method can be considered a holy grail in supervised learning. XAI [6] is supposed to provide trust in DNN models by indicating the mechanisms behind the black box classifiers. A popular XAI approach is to create attribution maps that identify the most important features [13]. Focusing on images, attribution maps indicate sets of pixels that are most relevant to the model's output [12].

Although current state-of-the-art methods are capable of highlighting regions that look meaningful to a human eye, they rely on untestable assumptions. Surrogate evaluation techniques use pixel perturbations, for example varying the intensities of the top k attributions and then measuring the drop in a classification metric [16]. This way, the evaluation is only meaningful in comparison to other (untestable) explanation techniques, and is hence subjective.

We argue that a good explanation should at least be verifiable by the classifier. Figure 1 illustrates the challenges that we face when we want to generate evaluable classifications. The figure shows how an image of a dog is classified as a cat when pixels with a negative attribution to the class *dog* are removed. Setting here the negative attribution pixels to zero alters the shape of the dog, that is also having dark fur. Our proposed method SpaCE-VAE maintains the important information in the contrast of the dog and the background. Details, such as the eyes, snout, and paws of the dog are not visible anymore, while the classifier still predicts the class *dog* with a high confidence.

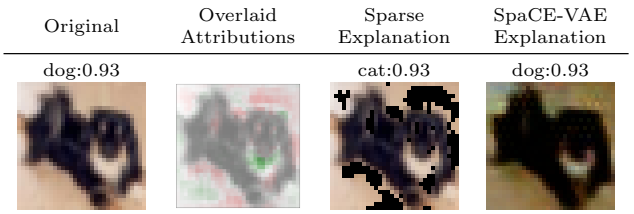

Figure 1: Illustrating the problem of evaluating attribution methods. Positive and negative attributions are indicated in green and red, respectively. The sparse explanation is generated by dropping the pixels with negative attributions. Removing *unimportant* pixels easily results in a misclassification.

In summary, our contributions are:

1. We propose SpaCE-VAE (Sparse Confident Explanations using Variational Autoencoders), a novel approach to generate sparse confident local explanations of a DNN that remain on the manifold of correctly classified images.
2. Our empirical analysis indicates that SpaCE-VAE is able to generate explanations that generalize over the test set.
3. Visual evaluation shows that SpaCE-VAE is able to highlight parts that give insight into the model, making sense from a human perspective.

1.1 Notation

We assume we are given a dataset

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

of n images $\mathbf{x}_i \in \mathbb{R}^{w \times h \times p}$ (width $\times$ height $\times$ channels) and their corresponding labels $y_i \in \{1, \ldots, c\}$. Let $C(x)$ be the classifier that we want to explain. We assume that $C : \mathbb{R}^d \to [0, 1]^c$ maps data points to a vector of class-confidences between zero and one.

2. RELATED WORK

In the taxonomy of XAI methods, we propose a local interpretation model, focusing on explaining individual predictions. Of those local interpretation models, we are closest to attribution methods.

2.1 Attribution methods

Feature attributions assign a score to each input feature based on the perceived importance to the output [12]. Attribution maps $A(\mathbf{x}, \tilde{y}; C) \in \mathbb{R}^d$ indicate the importance of each pixel (or feature) in the input image $\mathbf{x}$ towards the class $\tilde{y}$ predicted by the classifier C. The idea is that the attribution values highlight parts of the input image that are crucial for the classifier's decision-making process. For positive attributions, increasing the pixel intensity is expected to yield a higher confidence. For negative attributions, decreasing the intensity is expected to yield a higher confidence. Attribution models largely rely on an interpretation of *importance* based on the sensitivity to perturbations. The perturbation effects are, for example, measured by a change in the prediction confidence. Let $C_{\tilde{y}}(\mathbf{x}) \in [0, 1]$ be the confidence of predicting class $\tilde{y}$ with classifier C for input image $\mathbf{x}$. We approximate the prediction confidence by the first-order Taylor expansion of input image $\mathbf{x}$

$$C_{\tilde{y}}(\mathbf{x} + \Delta) \approx C_{\tilde{y}}(\mathbf{x}) + \nabla C_{\tilde{y}}(\mathbf{x})^\top \Delta, \tag{1}$$

$$\Leftrightarrow C_{\tilde{y}}(\mathbf{x} + \Delta) - C_{\tilde{y}}(\mathbf{x}) \approx \sum_{i,j,s} \frac{\partial C_{\tilde{y}}(\mathbf{x})}{\partial x_{ijs}} \Delta_{ijs}. \tag{2}$$

Equation (2) states that the impact of perturbation Δ to image $\mathbf{x}$ on the prediction confidence depends on the partial gradients. This observation motivates *Saliency Maps*, which are a very basic attribution technique, identifying the most important pixels as those that have the largest absolute value of the gradient over the channels [18]

$$A(\mathbf{x}, \tilde{y}; C)_{ij} = \sum_{s=1}^{p} \left| \frac{\partial C_{\tilde{y}}(\mathbf{x})}{\partial x_{ijs}} \right|.$$

Other aggregation methods over the channels, such as taking the maximum absolute value, are also possible. While saliency maps identify pixels to which the predictions are sensitive, the question remains whether those pixels provide good explanations. Adversarial examples also use local sensitivities identified by the gradient to fool the classifier. However, those examples are considered as rather artificial, providing examples outside of the manifold of reasonable images [20]. The question arises whether saliency maps highlight pixels that are crucial for the predicted class, or only perturbation-sensitive pixels. Those things are not necessarily the same.

DeepLIFT [17] generalizes the indication of important features by the Taylor expansion to contribution scores. This way, not only small differences in the input can be evaluated towards the prediction outcome, but any differences to a specified reference input. The reference input typically represents a default or neutral input chosen for the problem at hand (for example, a black image). Similarly to the structure of the Taylor expansion in Equation (2), DeepLIFT assigns layer-wise contribution scores $C_{\Delta_{z\,i}\Delta_h}$ to the latent representation $\mathbf{z} \in \mathbb{R}^l$ of $\mathbf{x}$ and their consecutive hidden layer output $h(\mathbf{z})$, such that it satisfies the property

$$\Delta_h = h(z + \Delta_z) - h(z) = \sum_{i=1}^{l} C_{\Delta_{z\,i}\Delta_h}. \tag{3}$$

Similar to layer-wise relevance propagation methods [1], the scores of the final layer are backpropagated to the input image by specified rules, similar to the chain rule backpropagation. The quality of the provided explanation largely hinges on the reasonability of the established rules, which are again not easy to assess empirically.

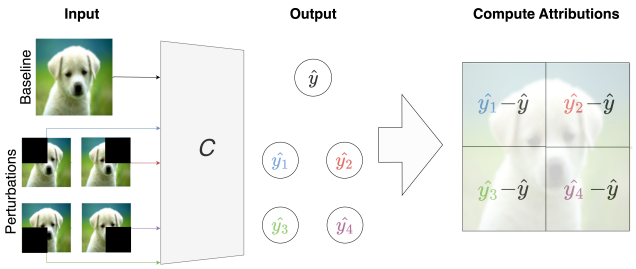

Figure 2: An illustrative example of Feature Ablation. The baseline image and perturbations are passed through the classifier C to obtain the reference score $\hat{y}$ and perturbed confidence scores $\hat{y_1}, \hat{y_2}, \hat{y_3}$, and $\hat{y_4}$. The attributions are computed by subtracting the reference score from each of the perturbed confidence scores.

Integrated Gradients [19] evaluates the importance of an input feature with respect to a reference image by means of fluctuations in the gradient on the path between both images. Features that exhibit significant changes along the path are attributed higher scores. That is, the importance value for the image tensor element at index (ijs) is defined as

$$IG_{ijs}(\mathbf{x} + \Delta, \mathbf{x}) = \Delta_{ijs} \cdot \int_0^1 \frac{\partial C_{\tilde{y}}(\mathbf{x} + \alpha\Delta)}{\partial x_{ijs}} d\alpha. \tag{4}$$

The integral is efficiently computed with the Riemann approximation. The corresponding attribution map aggregates the integrated gradients over the channels, for example, computing a norm or the sum. Integrated Gradients rely again on the assumption that local sensitivity to features, indicated by the gradients, indicates the importance of that feature for classification.

Feature Ablation [9] computes the attributions of the input by an empirical evaluation of the effect of perturbations of the original image. It removes regions or pixels by setting pixel intensities to zero (black). The attributions are computed by subtracting the confidence score of the perturbed image from the reference confidence score of the original image.

Figure 2 illustrates this approach by means of four perturbations. The confidence scores of the original image for the predicted class is $C_{\tilde{y}}(\mathbf{x}) = \hat{y}$ and $\hat{y}_1, \ldots, \hat{y}_4$ indicate the confidences of class $\tilde{y}$ for each of the perturbations. Afterwards, the attributions for each blackened region are given by the drop in confidence if the pixels of each region are removed. Hence, Feature Ablation provides a baseline for evaluable explanations. The clear drawback of Feature Ablation is that this is computationally very expensive.

The attribution methods discussed above (with the exception of feature ablation) are difficult to evaluate, as no ground truth is available [14]. Consequently, perturbation strategies are often employed [4]. These approaches assess model performance by systematically altering the input data or model parameters according to the attribution maps. However, they suffer from the drawback that sensitivity to feature changes does not necessarily imply that

those features are most important for the decision-making process. Or, in other words, the pixels that are close to the decision boundary are not necessarily the ones that have the most impact on the classification outcome.

2.2 Generative Model Explanations

An approach related to ours is GasTeN, generating example-based global explanations using a generative model [3]. GaSTeN uses GANs to generate images that are close to the decision boundary of two classes. These images can be used to gain insight into the classifier decision-making process. Like in our method, a regularization term is added to the reconstruction loss to generate images that are expected to provide good explanations (here, those are images close to the decision boundary).

Dabkowski and Gal [5] train a model that generates a binary mask to select pixels, or preferably areas, that are important for the prediction. The authors acknowledge that the usage of *pixel removal* strategies, such as applying the binary mask to the input image, introduces artifacts to the image that affect the performance of DNNs [7]. They propose evaluating images cropped to the smallest rectangle enclosing the salient region. This approach enables the assessment of the generated saliency maps, but the evaluation is restricted to rectangular crops, which may overlook important cases such as when the model relies on background features for its prediction.

Another class of related models, exemplified by Deep Dream [22], highlights structures and patterns that increase neuron activations. Unlike our method, which aims to remove important information, generative approaches such as Deep Dream add patterns to the explanation image. Since our approach also relies on a generative model, it occasionally introduces patterns into the image as well. As we will see in the evaluation, this can be useful for explaining misclassifications by indicating what the model *perceives*.

3. SPACE-VAE

We aim to generate explanations that lower the pixel intensity (increasing sparsity) of an image such that it is still classified confidently by the target classifier like the original image. We list the following desired properties for our explanations:

- **Accurate** The explanations should at least be classified coherently: the prediction of the original image should be equal to the prediction of the explanation.
- **Sparse** The explanation should reduce the number of activated pixels or the pixel intensities.
- **Confident** Ideally, the explanation should have a confidence at least as high as the model's confidence on the original image, so that it captures the features that actually drive the model's high-confidence prediction.
- **Interpretable** Although we encourage sparseness, the explanations should still be interpretable and the features or regions with high importance should be clear and visible.

We aim to develop a framework capable of producing explanations that are both sparse and confident. To achieve this objective, we introduce SpaCE-VAE, an extension of the Vector Quantized Variational Autoencoder (VQ-VAE) model [21]. The VQ-VAE uses in contrast to the vanilla VAE discrete distributions to model the latent space. It has demonstrated remarkable versatility in the generation of high-quality images [15], and we believe that the discrete representations fit the goal of SpaCE-VAE. Discrete variables are likely more suitable to push the learned representations towards identifying concepts of images that explain the classifiers behavior. Furthermore, the VQ-VAE architecture avoids common problems in VAEs, such as the posterior collapse problem, where a decoder simply ignores learned latent representations and variance issues.

Figure 3 illustrates the overall idea of SpaCE-VAE. A VQ-VAE model, consisting of encoder E, decoder D, and embedding space e is trained to explain predictions of classifier C. The reconstruction generated by the VAE lowers pixel intensities such that the classifications of the reconstructed and the original image remain the same.

3.1 Loss Function

VQ-VAE models pass an image $\mathbf{x}$ to the encoder E, which generates an output $z_e(\mathbf{x}) \in \mathbb{R}^{d_q}$. Then, $z_e(\mathbf{x})$ is passed to the quantizer Q. Here, the discrete latent variables are computed using a nearest centroid lookup using the shared embedding space $e = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\} \subseteq \mathbb{R}^{d_q}$

$$z_q(\mathbf{x}) = \mathbf{e}_l,\ l = \arg\ \min_j \|z_e(\mathbf{x}) - e_j\|. \tag{5}$$

The quantization process returns $z_q(\mathbf{x}) \in \mathbb{R}^{d_q}$. Afterwards, the decoder D maps the quantized representation $z_q(\mathbf{x})$ back to a reconstructed image $\hat{\mathbf{x}}$. Here, the gradient $\nabla_z L$ of the VQ-VAE loss is passed unaltered to the encoder to circumvent issues in the optimization of discrete variables [21].

We employ the method to update the embedding space by exponential moving averages (EMA) [21]. The updating process follows a similar procedure as k-Means [11] using an online version to update the embedding space. Using this approach, the model typically yields better performance and quicker convergence. Denoting the collective parameters of the encoder, the decoder and the embedding by the vector θ, the loss function of VQ-VAE-EMA is

$$\mathcal{L}_{\text{VQ-EMA}}(\theta, \mathbf{x}) = \log p(\mathbf{x}|z_q(\mathbf{x})) + \delta\|z_e(\mathbf{x}) - \text{sg}[e]\|^2. \tag{6}$$

Here, the first term $\log p(\mathbf{x}|z_q(\mathbf{x}))$ models the reconstruction loss. The function $\text{sg}[\cdot]$ denotes the stopgradient operator, which is defined as the identity at forward computation time, but it *stops the gradient* because it returns a gradient of zero at backward passes. As a result, the embedding space will not be optimized by the reconstruction loss. The last term $\delta\|z_e(\mathbf{x}) - \text{sg}[e]\|^2$ incentivizes the encoder outputs to commit to the embedding and it ensures that the output does not grow arbitrarily.

We adapt the VAE-VQ loss function to generate explanations of a given classifier C. We add as a sparseness term the L_1-norm of the reconstructed image $\hat{\mathbf{x}}$. The L_1-norm provides a good trade-off between the desired sparse representations (having a low L_0-norm) and the ability to optimize this term via numerical methods. To ensure that the explanations are coherently classified, we also add a cross-entropy term to the loss. Here, the parameters of the classifier are not optimized, only the parameters of the VAE, generating the explanations $\hat{\mathbf{x}} = vae(\mathbf{x})$. Denoting the predicted class

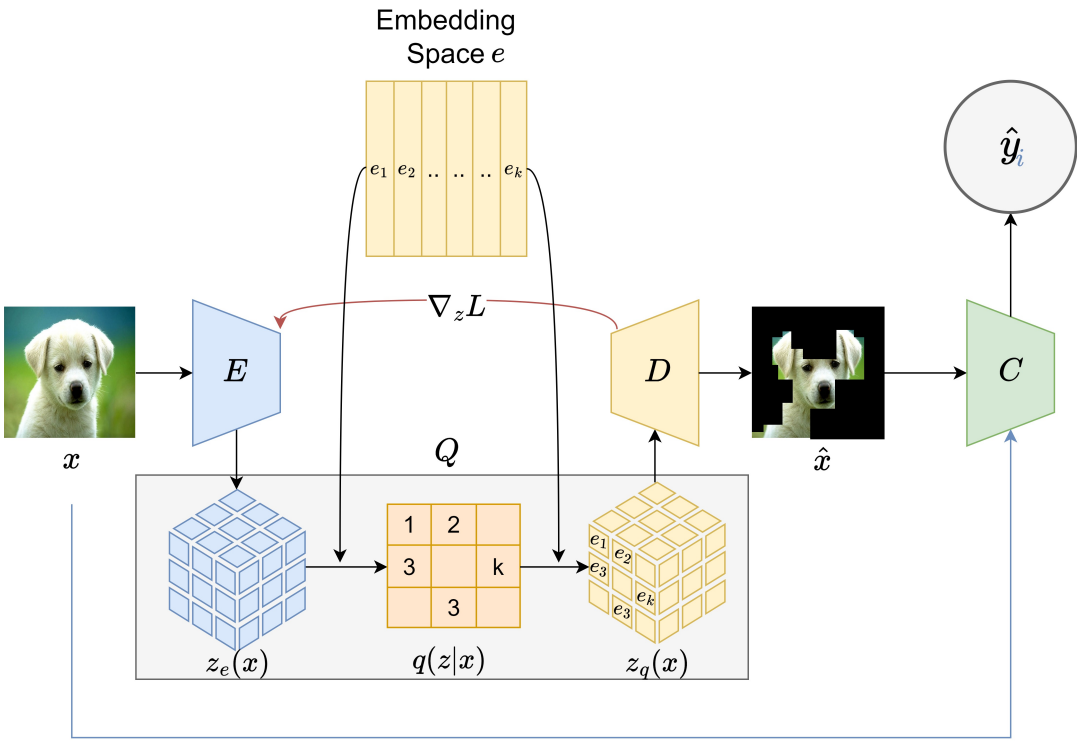

Figure 3: The SpaCE-VAE framework. A VQ-VAE (consisting of encoder E, embedding basis vectors $\mathbf{e}$ and decoder D) is trained to generate *sparse* images $\hat{\mathbf{x}}$ that approximates the input image $\mathbf{x}$ while decreasing the intensity of nonrelevant pixels such that the resulting image is still classified like the original image.

by $\tilde{y} = \arg\max_j C_j(\mathbf{x})$ and the confidence of the predicted class as $C_{\tilde{y}}(\hat{\mathbf{x}})$, we define our loss term as

$$\mathcal{L}_{\text{SpaCE}}(\theta, \mathbf{x}) = \log p(\mathbf{x}|z_q(\mathbf{x})) + \delta\|z_e(\mathbf{x}) - \text{sg}[e]\|^2 + \alpha|\hat{\mathbf{x}}| - \beta \log C_{\tilde{y}}(\hat{\mathbf{x}}). \quad (7)$$

The term $\alpha|\hat{\mathbf{x}}|$ incentivizes sparsity, and the term $-\beta \log C_{\tilde{y}}(\hat{\mathbf{x}})$ increases the confidence of the predicted class for the reconstruction. In the remainder of this paper, we choose a value of $\alpha = 0.05$, since it provides a notable drop in pixel intensities while not making the explaining images too dark.

4. EXPERIMENTS

We evaluate our method against the attribution methods DeepLIFT, Feature Ablation, and Integrated Gradients. Those methods are designed to provide explanations against a reference picture, such as a black image, and thus have a similar goal as SpaCE, to provide sparse explanations. We use the existing implementation of attribution methods by Captum [9].

We normalize the attribution methods such that competitors generate attributions $A \in [-1, 1]^d$. We try to find a good threshold value such that putting all pixels to zero (black), if they do not exceed the threshold, does not decrease the accuracy too much. For that reason, we try thresholds $\sigma \in \{-0.9, \ldots, -0.1, 0\}$ to generate explanations

$$\mathbf{x}_\sigma = \mathbb{1}[A \geq \sigma] \circ \mathbf{x},$$

where $\circ$ denotes the Hadamard product. The best threshold value is determined on the test set.

We evaluate our method and the competitors on the CIFAR10 dataset [10]. Training DNNs on CIFAR10 is reasonable, and it allows evaluating the computationally expensive Feature Ablation method. We sample a validation dataset containing 15% of the training data points by means of stratified sampling, ensuring that the relative proportions of each class remain consistent. The validation set is used to select the SpaCE model that achieves the lowest validation loss during training.

The classifier that we aim to provide explanations for is a VGG11 model. The training procedure uses a batch size of 256, and the network was optimized using stochastic gradient descent with a learning rate set to 0.01, weight decay of 0.01, and momentum of 0.9. After training, the model achieved 99.99% and 92.39% on the train and test sets, respectively.

4.1 Reproducibility details

To train the VQ-VAE for our SpaCE model, we use the ADAM optimizer with a learning rate of 0.0003. Following the commitment loss recommendations [21], we set $\delta = 0.25$, and we use an embedding space of 512 elements. We train the model for maximum 200 epochs, as long as the validation loss has not decreased for 20 epochs. We save the model with the highest validation loss. We provide the implementation of our method together with evaluation scripts[1]. Our code

[1]https://github.com/AlexanderLLLiu/SpaCE-VAE

builds upon existing implementations of the VGG11[2] and VQ-VAE[3].
All experiments are performed using an NVIDIA GeForce RTX 4070 Ti GPU and an Intel Core i9-13900K CPU. The complete training of one configuration takes approximately one hour.

4.2 Quantitative Analysis

We quantify our results by means of the average accuracy on the test set and the average sparsity measured in L_0-norm and L_1-norm. The sparsity is calculated for the input image $\mathbf{x}$ and sparse explanation $\hat{\mathbf{x}}$ as

$$L_p\text{-Sparsity} = \frac{\text{avg}(\{\|\mathbf{x}\|_p - \|\hat{\mathbf{x}}\|_p \mid \mathbf{x} \in \mathcal{D}\})}{\text{avg}(\{\|\mathbf{x}\|_p \mid \mathbf{x} \in \mathcal{D}\})} \cdot 100$$

where $\text{avg}(\cdot)$ computes the average of a set. Figure 4 illustrates the drop in accuracy that arises when generating explanations that put an increasing amount of pixels to zero, depending on the threshold of negative attributions. We observe that in particular DeepLIFT's and Integrated Gradients' masked explanations drop to an average accuracy of approximately 0.3 when removing only 10% of the pixels (equating a sparsity in L_0-norm of 10%). The computationally much more expensive Feature Ablation method yields better results, but the accuracy still drops steeply with an increase in sparsity.
For SpaCE-VAE, we plot the results for the three values of the hyperparameter β. We observe that, particularly for the L_1-norm, subject to which the model has been trained, SpaCE-VAE achieves a comparatively high accuracy for high L_1-norm sparsity. Yet, also with regard to the L_0-norm, SpaCE-VAE provides explanations with an accuracy above 70% for 20% of the pixels being black in average. The results indicate that our proposed method is indeed able to find explanations that still lie on the manifold of the correctly classified images.
Table 1 summarizes our experimental results. For our competitors, we display the results for the largest threshold σ that yields a L_0-sparsity smaller than 10%. Since the accuracy of competitors is dropping so rapidly, we choose the threshold this way to obtain informative explanations, which *remove* at least around 10% of the pixels. The maximum achievable accuracy is given this way by Feature Ablation with 0.65. The accuracy of SpaCE-VAE is around 0.73, with an L_1-sparsity around 50%. Considering that SpaCE-VAE achieves an accuracy of around 0.9 on the training set, it demonstrates a solid degree of generalization, though there is still room for improvement. This is a nontrivial result. While it is not surprising that many *explanations* can be constructed for an image—since numerous perturbations can increase a model's confidence in its prediction — the ability to transform images into sparse representations that still support high confidence in the predicted class indicates that the autoencoder is effectively learning to emphasize features relevant for classification while suppressing irrelevant ones.

4.3 Qualitative Evaluation

In Figure 5, we provide example explanations given by the binary mask of competitors and our method SpaCE-VAE. Above each image, you find the predicted class of the image and its confidence. We observe that the sparse explanations of DeepLIFT and Integrated Gradients are not very insightful, and they also lead to frequent misclassifications of the images where *nonimportant* pixels are set to black. Feature Ablation suffers (as expected) less from misclassification and provides, as well, some more coherent regions that are deemed unimportant. Still, we argue that the provided explanations are not suitable to identify what is important information of that image, but rather what is really unimportant (such as the small background area on the top left of the dog).
Regarding SpaCE-VAE, we observe that a parameter value of $\beta = 1$ leads to rather smooth, darker images that seem to be composed of a smaller color palette. Particularly interesting is the example of explaining a misclassified bird (as a cat) in row three. The explanation created by SpaCE-VAE with $\beta = 1$ emphasizes a cat-like form: in particular, there are three regions that could be interpreted as legs, smoothing out the background that does not contribute to the cat-like figure. Regarding the explanation of a *truck* in the last row, SpaCE-VAE ($\beta = 1$) suggests that the cabin is important for classification. SpaCE-VAE with a value of $\beta = 100$ reconstructs the texture of objects and animals more than SpaCE-VAE with $\beta = 1$. In the examples of a dog and a bird in the first two rows, we also observe that the boundary between the animal and the background is highlighted. This is insofar interesting as DNNs are usually identified to be reliant on texture in the first place [2; 8]. The reconstructions suggest that the VAE does indeed emphasize the shape information paired with some general color patterns (as seen in the intensified colors of the bird and the intensified cat-fur colors).

4.4 Limitations

A key assumption underlying our approach is that uninformative or irrelevant pixels can be effectively suppressed by driving them toward zero, which corresponds to black in standard image representations. However, this intuition does not hold universally across all visual contexts. For example, in an image of a dark-colored dog standing in front of a white-tiled wall (like in Figure 1), the semantically important pixels may be dark, while the background (which is irrelevant) is predominantly bright. In such cases, a sparsity constraint that pushes pixel values toward zero may inadvertently retain irrelevant regions while suppressing important ones.
This reflects a more general limitation of using mathematical simplicity — here, via an L_1 penalty — as a proxy for interpretability in image-based tasks. While L_1 regularization promotes sparsity, it does not necessarily align with the semantics of visual data. Furthermore, applying an L_1 penalty often results in globally darker images, which can make the explanations visually less intuitive or harder to interpret.

5. CONCLUSIONS

We propose SpaCE-VAE, a new framework to generate local explanations that enable a quantitative evaluation. We propose a loss function to train a VQ-VAE that generates images that are close to the input image, while lowering the intensity of as many pixels as possible without changing the

[2]https://github.com/huyvnphan/PyTorch_CIFAR10
[3]https://github.com/swasun/VQ-VAE-Images

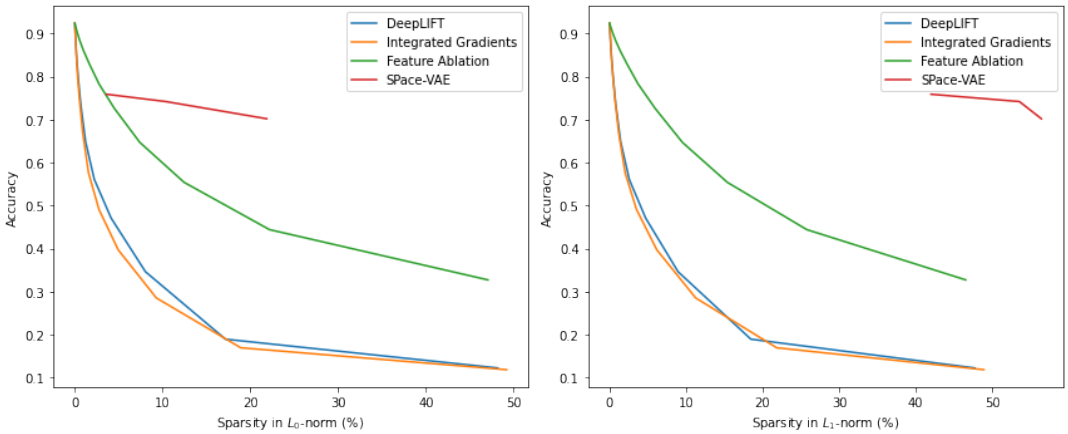

Figure 4: Plot of sparsity against the accuracy on competitors and our model SpaCE-VAE on the test set. The competitor plots connect the results given by varying the threshold value $\sigma \in \{-0.9, \ldots, -0.1\}$. The SpaCE plot visualizes the results for varying the hyperparameter $\beta \in \{1, 10, 100\}$.

Table 1: Comparative evaluation of SpaCE-VAE with baselines and state-of-the-art attribution methods on the test set. The best performing XAI models are highlighted in bold.

			Acc ↑	Conf ↑	L_0-Sparsity ↑	L_1-Sparsity ↑
Baselines	Original Images		0.92	0.94	0%	0%
	VQ-VAE		0.56	0.56	-0.18%	-0.88%
XAI	DeepLIFT	$\sigma = -0.2$	0.35	0.32	8.05%	8.95%
	Integrated Gradients	$\sigma = -0.2$	0.29	0.27	9.28%	11.250%
	Feature Ablation	$\sigma = -0.3$	0.65	0.64	7.40%	9.52%
	SpaCE-VAE	$\beta = 1$	0.70	0.66	**21.86%**	**56.42%**
		$\beta = 10$	0.74	0.69	10.37%	53.56%
		$\beta = 100$	**0.76**	**0.72**	3.62%	42.01%

class prediction.

We evaluate SpaCE-VAE against three competitors that identify pixels that contribute negatively to the predicted class. Two of those competitors (DeepLIFT and Integrated Gradients) have no sensible method to evaluate those explanations, since *removing* the negatively contributing pixels often changes the predicted class. The other competitor, Feature Ablation, is a computationally expensive trial-and-error approach that identifies negative attributions based on an actual increase in the prediction confidence when pixels are removed.

Our evaluation based on the CIFAR10 dataset shows that SpaCE-VAE is vastly able to outperform the three competitors in the achieved accuracy when removing the pixel intensity of negative attributions. Our qualitative evaluation shows that SpaCE-VAE is able to provide interesting and insightful explanations.

Our proposed framework is flexible, and simple changes in the loss function enable the identification of various explanations. For example, we could incorporate a regularization term to increase the contrast of the explanations, to decrease the number of used colors, or to smooth the color variations as much as possible.

6. REFERENCES

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS ONE*, 2015.

[2] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 2018.

[3] L. Cunha, C. Soares, A. Restivo, and L. F. Teixeira. Gasten: Generative adversarial stress test networks. In *International Symposium on Intelligent Data Analysis (IDA)*, 2023.

[4] J. da Costa Feitosa, M. Roder, J. P. Papa, and J. R. F. Brega. Influence of pixel perturbation on explainable artificial intelligence methods. In *VISIGRAPP : VISAPP*, 2024.

[5] P. Dabkowski and Y. Gal. Real time image saliency for black box classifiers. In *Neural Information Processing Systems*, 2017.

[6] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 2023.

[7] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *IEEE In-*

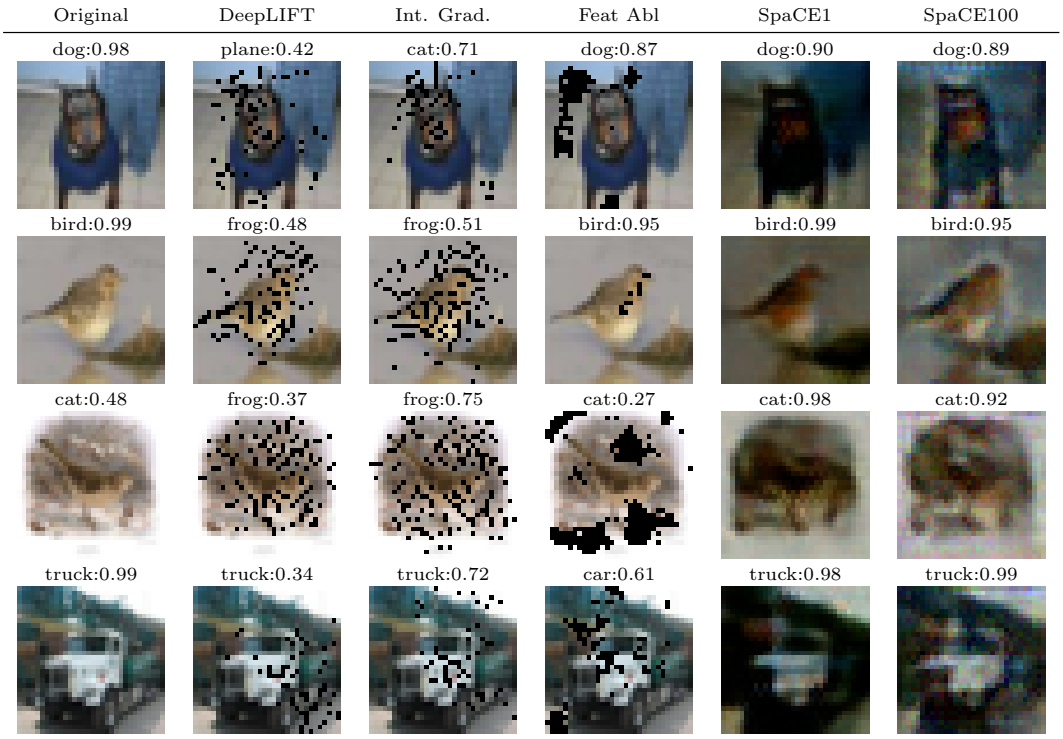

Figure 5: Original images, their explanation, and the predicted class with confidence. For competitors, we apply the binary mask for thresholds indicated in Table 1. SpaCE1 uses a value of $\beta = 1$ and SpaCE100 uses a $\beta = 100$.

ternational Conference on Computer Vision (ICCV), 2017.

[8] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[9] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[10] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[11] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967.

[12] C. Molnar. *Interpretable Machine Learning*. 3 edition, 2025.

[13] G. Nguyen, D. Kim, and A. Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 2021.

[14] S. Rao, M. Böhle, and B. Schiele. Towards better understanding attribution methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 2019.

[16] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, 2022.

[17] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, 2017.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[19] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, 2017.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

[21] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 2017.

[22] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV*, 2014.