



# Explaining classifiers by constructing familiar concepts

Johannes Schneider<sup>1</sup> · Michalis Vlachos<sup>2</sup>

Received: 26 February 2021 / Revised: 6 January 2022 / Accepted: 19 February 2022 /  
Published online: 25 March 2022  
© The Author(s) 2022

## Abstract

Interpreting a large number of neurons in deep learning is difficult. Our proposed ‘CLAssifier-DECoder’ architecture (*ClaDec*) facilitates the understanding of the output of an arbitrary layer of neurons or subsets thereof. It uses a decoder that transforms the incomprehensible representation of the given neurons to a representation that is more similar to the domain a human is familiar with. In an image recognition problem, one can recognize what information (or concepts) a layer maintains by contrasting reconstructed images of *ClaDec* with those of a conventional auto-encoder(AE) serving as reference. An extension of *ClaDec* allows trading comprehensibility and fidelity. We evaluate our approach for image classification using convolutional neural networks. We show that reconstructed visualizations using encodings from a classifier capture more relevant classification information than conventional AEs. This holds although AEs contain more information on the original input. Our user study highlights that even non-experts can identify a diverse set of concepts contained in images that are relevant (or irrelevant) for the classifier. We also compare against saliency based methods that focus on pixel relevance rather than concepts. We show that *ClaDec* tends to highlight more relevant input areas to classification though outcomes depend on classifier architecture. Code is at <https://github.com/JohnTailor/ClaDec>

**Keywords** Deep learning · Explainability · XAI · Computer vision · Concept-based explanations

---

Editors: Annalisa Appice, Grigorios Tsoumakas.

This is an extended journal version of the conference publication “Explaining Neural Networks by Decoding Layer Activations”, accepted at the Intelligent Data Analysis Symposium(IDA), 2021.

---

✉ Johannes Schneider  
johannes.schneider@uni.li  
Michalis Vlachos  
michalis.vlachos@unil.ch

<sup>1</sup> Institute of Information Systems, University of Liechtenstein, Vaduz, Liechtenstein

<sup>2</sup> Department of Information Systems, HEC, University of Lausanne, Lausanne, Switzerland

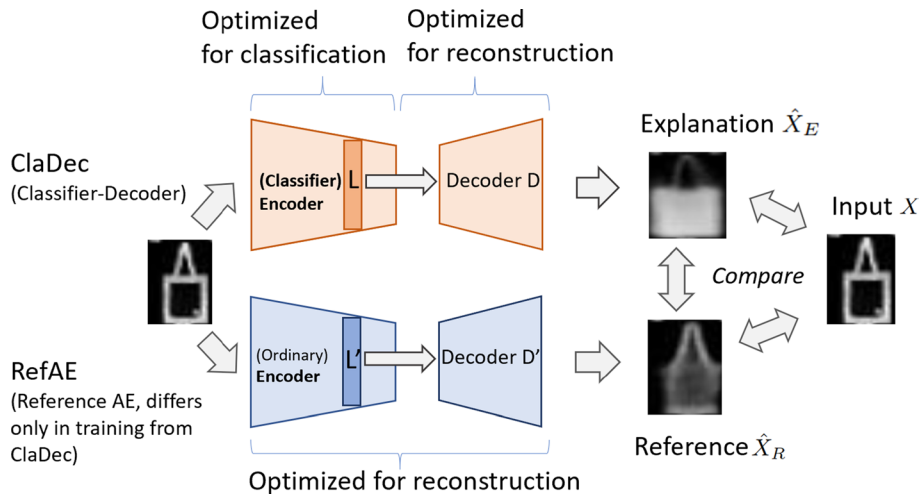


Fig. 1 Basic architecture of *ClaDec* and *RefAE* and explanation process

## 1 Introduction

Explaining predictive models is important for many reasons, including: (a) debugging or improving models, (b) fulfilling legal obligations such as the “right to explanation” as crystallized in the European GDPR data privacy law, and (c) increasing trust in models. Thus, explaining neural networks has received a lot of attention (Adadi and Berrada 2018; Schneider and Handali 2019; Confalonieri et al. 2021; Schneider et al. 2022). Understanding a neural network is a multi-faceted problem, ranging from understanding single decisions, single neurons and single layers, up to explaining complete models. Often, explainability methods touch on multiple of these aspects. In this work, we are primarily interested in better understanding a decision with respect to a user-defined layer (or a subset thereof) that originates from a complex feature hierarchy, as commonly found in deep learning models. In a layered model, each layer corresponds to a transformed representation of the original input. Thus, the neural network succinctly transforms the input into more useful representations for the task at hand, such as classification. From this point of view, we seek to answer the question: “Given an input  $X$ , what does the representation  $L(X)$  (or subsets thereof) produced in a layer  $L$  tell us about the decision and about the network?”. Our local, model-agnostic post-hoc explanation method requires comparing three images  $X$ ,  $\hat{X}_E$  and  $\hat{X}_R$  that together constitute the explanation as illustrated in the right of Fig. 1. We compare the original input image  $X$  with a transformation  $\hat{X}_E$  of the latent representation  $L(X)$  of the classifier to image space. This allows to visually identify what information is maintained in the layer and what information is discarded. We also compare against a reference  $\hat{X}_R$  consisting of an autoencoded image of  $X$ . This allows identifying what of the discarded information in  $\hat{X}_E$  is a result of inaccuracies of the explanation process.

Technically, we propose a classifier-decoder architecture called *ClaDec*. It uses a decoder to transform the representation  $L(X)$  produced by a layer  $L$  of the classifier, with the goal to explain the layer representation via a humanly understandable representation, i.e., one that is similar on the input domain. The layer in question provides the “code” that is fed into a decoder. The motivation for using an AE is that AEs are good at (re)

constructing high-dimensional data from a low-dimensional representation. Intuitively, a classifier to be explained should encode aspects relevant to the classification faithfully and ignore input information that does not impact decisions. Therefore, using a decoder can lead to accurate reconstruction of parts and attributes of the input that are essential for classification. In contrast, inputs with little or no influence on the classification will be reconstructed at lower fidelity. Attributes of an input might refer to basic properties such as color, shape, sharpness but also more abstract, higher-level concepts. That is, reconstructions of higher-level constructs might be altered to be more similar to prototypical, average-like instances.

Explanations should fulfill many partially conflicting objectives. We are interested in the trade-off between fidelity (How accurately does the explanation express the model behavior?) and comprehensibility (How easy is it to make sense of the explanation?). While these properties of explanations are well-known, existing methods typically do not accommodate adjusting this trade-off. In contrast, we propose an extension of our base architecture *Cladec* by adding a classification loss. It allows for balancing between producing reconstructions that are similar to the inputs, i.e., training data that a user is probably more familiar with (easier interpretation), and reconstructions that are strongly influenced by the model to explain (higher fidelity) but may deviate more from what the user knows or has seen.

Our approach relies on an auxiliary model, a decoder, to provide explanations. Similar to other methods that use auxiliary or proxy models, e.g., to synthesize inputs (Nguyen et al. 2016) or approximate model behavior (Ribeiro et al. 2016), we face the problem that a poor auxiliary model may negatively impact explanation fidelity. That is, reconstructions produced by AEs (or GANs) might suffer from artifacts. For example, AEs are known to produce images that might appear more blurry than real images. People have noticed that GANs can produce clearer images but they may suffer from other artifacts as shown in Nguyen et al. (2016). Neglecting that the explainability method might introduce artifacts can adversely impact understandability and even lead to wrong conclusions on model behavior. When looking at the reconstruction, a person not familiar with such artifacts might not attribute the distortion to the auxiliary model being used, but she might believe that it is due to the model to be explained. At the same time, evaluation of explainability methods has many known open questions (Yang et al. 2019), the community has not been aware of this one. To avoid any wrongful perceptions regarding artifacts in reconstruction, we suggest comparing outcomes of auxiliary models to a reference architecture. We employ an auto-encoder *RefAE* with the exact same architecture as *Cladec* to generate outputs for comparison as shown in Figure 1. The encoder of *RefAE* is not trained for classification, but the *RefAE* model optimizes the reconstruction loss of the original inputs as any conventional AE. Therefore, only the differences visible in the reconstructions of *RefAE* and *Cladec* can be attributed to the model to be explained. The proposed comparison to a reference model can also be perceived as a rudimentary sanity check, i.e., if there are no differences then either the explainability method is of little value or the features to be learned of the model to be explained are similar to that of the reference AE. This might be a consequence of a similar objective of the model to explain and the reference model, or the fact that there are universal features suitable for many tasks. We shall elaborate more in our theoretical motivation. We believe that such sanity checks are urgently needed since multiple explanation methods have been scrutinized for failing “sanity” checks and simple robustness properties (Adebayo et al. 2018; Ghorbani et al. 2019; Kindermans et al. 2019). For that reason, we also introduce a sanity check that formalizes the idea that explanations should be beneficial for downstream tasks. In our context, we even show that auxiliary

classifiers trained on either reconstructions from *RefAE* or *ClaDec* perform better on the latter, although the reference AE leads to reconstructions that are closer to the original inputs. Thus, the reconstructions of *ClaDec* are more amendable for the task to be solved.

While our reconstruction-based method conceptually differs strongly from a saliency-based method, we also quantitatively compare against one of the most prominent attribution techniques, i.e., GradCAM. To this end, we occlude relevant areas of inputs. The relevance score of input areas is well-suited to the idea of attribution-based techniques that assume that classification outcome depends on a subset of all input pixels. To get the relevance of an area, the individual scores of input pixels simply have to be summed. In turn, we argue more about concepts and how the classifier perceives them. Thus, our explanations do not directly yield a relevance score per pixel. Still, by introducing a measure based on reconstruction loss between reconstructed images with and without occlusion, we obtain a suitable relevance score allowing us to compare GradCAM and *ClaDec*. Our findings confirm the strengths of *ClaDec* showing that it better allows finding areas of minimum and maximum relevance, although the evaluation metrics are more tailored to GradCAM. Advantages of *ClaDec* are most profound when layers with limited spatial extent covering semantically meaningful concepts are used. Overall, the methods are less of a substitute but should be used together to better understand classifiers.

Our experiment with humans confirms that even non-experts can identify meaningful concepts discarded (and maintained) by a classifier. Overall, we make the following **contributions**:

- We present a novel method to understand layers of neural networks. It uses a decoder to translate incomprehensible outputs of an entire layer or a subset thereof into a humanly understandable representation. It allows to trade comprehensibility and fidelity. It eases the interpretation of explanations that also allow deriving general statements upon model behavior.
- We introduce a method dealing with artifacts created by auxiliary models (or proxies) through comparisons with adequate references. This includes evaluation of methods using novel sanity checks.
- This journal article extends the conference version by enhancing all manuscript sections with new material. It newly adds a user study, showing that our method enables even non-experts to identify a rich set of concepts shown on images that are maintained (or not maintained) in the classification process and a more detailed presentation of results and related work. It contains new use-cases, such as assessing subsets of layer activations down to individual neurons. It also adds an extensive evaluation using occlusion of inputs introducing novel measures to compare concept-based methods using reconstructions with saliency maps from GradCAM. The journal version also includes an extensive reflection (Discussion and Future Work section), an assessment of the impact of classifier performance on reconstructions, and a comparison against prototype-based methods.

## 2 Related work

We first discuss approaches that allow us to visualize single features and understand model decisions summarized in Fig. 2.

Input	Architecture	Class	Explanation	Visualization	What is explained?	Technique	Accounting for distortions due to auxiliary models
		Two		Saliency map using gradients	Decision	Analyze input sensitivity using gradients	(not applicable)
		Dog Guitar		Saliency map using activations	Decision	Analyze sensitivity using input perturbations	(not applicable)
		Bird		Prototypical patches	Decision	New classifier architecture	(not applicable)
Pool table		Pool table		Synthesized input (each instance needs optimization)	Single neuron	Activation Maximization	No
		Hand-bag		Synthesized input (no optimization needed per instance)	Decision	Reconstruction from feature space using decoder	Yes

**Fig. 2** Method overview. Figures are from cited papers

We categorize explainability methods (Schneider and Handali 2019) into methods that synthesize inputs (Agarwal and Nguyen 2020; Guidotti et al. 2019; like ours and Nguyen et al. 2016; Yosinski et al. 2015) and methods that rely on saliency maps (Simonyan et al. 2014) based on perturbation (Ribeiro et al. 2016; Zeiler and Fergus 2014) or gradients (Selvaraju et al. 2017; Bach et al. 2015). Saliency maps show the feature importance of inputs, whereas synthesized inputs often show higher level representations encoded in the network. Perturbation-based methods include occlusion of parts of the inputs (Wu et al. 2020; Zeiler and Fergus 2014) and investigating the impact on output probabilities of specific classes. Linear proxy models such as LIME (Ribeiro et al. 2016) perform local approximations of a black-box model using simple linear models by also assessing modified inputs. Saliency maps (Simonyan et al. 2014) highlight parts of the inputs that contributed to the decision. Many explainability methods have been under scrutiny for failing sanity checks (Adebayo et al. 2018) and being sensitive to factors not contributing to model predictions (Kindermans et al. 2019) or adversarial perturbations (Ghorbani et al. 2019). Even if many of them might nevertheless be considered helpful and there are attempts to remedy these issues (Yeh et al. 2019), explanations are still fairly trivial. That is, mere highlighting does not provide any insights into how (input) information is processed and how it is encoded in the network (Rudin 2019). For those methods that show gradients (or a function of the gradients), one primarily sees how (very) small changes would impact the output, e.g., red might improve confidence in a class and blue reduce. Still, it is generally unclear whether large changes would still yield the behavior as suggested in the explanation, i.e., either an increase or a decrease. This is because gradients are only valid locally and might give no information on function behavior far from the point they are computed.

We anticipate that our work is less sensitive to targeted, hard to notice perturbations (Ghorbani et al. 2019) as well as translations or factors not impacting decisions (Kindermans et al. 2019) since we rely on encodings of the classifier. Thus, explanations only change if these encodings change, i.e., if the classifier is sensitive to the perturbations. The idea to evaluate explanations on downstream tasks is not new, however a comparison to

a “close” baseline like our *RefAE* is. Our “evaluation classifier” using only explanations (without inputs) is more suitable than methods like (Schneider and Vlachos 2020) that use explanations together with inputs in a more complex, non-standard classification process. Using inputs and explanations for the evaluation classifier is diminishing differences in evaluation outcomes since a network might extract missing information in the explanation from the input. So far, inputs have only been synthesized to understand individual neurons (Barbalau et al. 2020; Nguyen et al. 2016), where the pioneering work (Nguyen et al. 2016) used activation maximization in an optimization procedure. The idea is to identify inputs that maximize the activation of a given neuron. This is similar to the idea to identify samples in the input that maximize neuron activation. (Nguyen et al. 2016) uses a (pre-trained) GAN on natural images relevant to the classification problem. It identifies through optimization the latent code that when fed into the GAN results in a more or less realistic looking image that maximally activates a neuron. (Yosinski et al. 2015) uses regularized optimization as well, yielding artistically more interesting but less recognizable images. Regularized optimization has also been employed in other forms of explanations of images, e.g., to make humans understand how they can alter visual inputs such as handwriting for better recognizability by a CNN (Schneider 2020). Agarwal and Nguyen (2020) uses a GAN to replace removed input features, e.g., through cropping of parts of the input image, using realistic in-painting in the context of explainability. Guidotti et al. (2019) trains an adversarial auto-encoder (AAE) on the training data. The idea is to generate similar samples to an input  $X$  using the AAE by distorting the latent encoding of  $X$ . The generated samples are labeled using the original classifier. Then, an approximate model using a decision tree is trained using the labeled data. This also allows to obtain contrastive explanations. Rather than using an AAE one might use *ClaDec* to generate similar samples. This might be even more appropriate since these reconstructions would be based on the latent space of the classifier instead of the latent space of the AAE. That is, *ClaDec* focuses more strongly on differences in generated samples that are also relevant to classification. In contrast to our work, (Guidotti et al. 2019) does not propose to compare to a reference, i.e., the generated samples might exhibit distortions stemming from the AAE that are misleading. van Doorenmalen and Menkovski (2020) uses a variational AE for contrastive explanations. They use distances in latent space to identify samples that are closest to a sample  $X$  of class  $Y$  but actually classified as  $Y'$ . *ClaDec* might also be used to this end to work directly using the latent space of the classifier rather than the one of the encoder from a separate AE.

Kim et al. (2018), Ghorbani et al. (2019) investigate high level concepts that are relevant to a specific decision. DeepLift (Shrikumar et al. 2017) compares activations to a reference and propagates them backward. Defining the reference is non-trivial and domain specific. Koh and Liang (2017) estimates the impact of individual training samples. Liu et al. (2020) discusses how to explain variational AEs using gradient-based methods. *ClaDec* could also be used to explain AEs. Chen et al. (2019), Li et al. (2018) propose new network architectures that are based on encoding prototypes. In contrast to other methods that allow for post-hoc explanations, explainability is built into the model. The reasoning process is based on using dedicated layers/convolutions that encode prototypical patches/samples. While (Chen et al. 2019) achieves good classification performance, it imposes constraints on the classifier design that can lead to inferior classifier performance. In contrast, our method is not imposing any constraints and is universally applicable. We also do not explicitly learn parts or patches like (Chen et al. 2019). Wu et al. (2020) aims at global explanations. They use a two stage process using occlusion of inputs and ad-hoc semantic analysis. The visualization of class concepts (Fig. 3 in Wu et al. (2020)) allows to draw some conclusions on network behavior, but

appears highly distorted since it is purely based on activation maximization. Rafegas et al. (2020) used pre-defined concepts such as color and class association to classify neurons. The idea to focus on individual neurons has been criticized (Fong and Vedaldi 2018) since concepts are often not encoded by a single neuron but by groups. Our work aims more at the question “What concepts are relevant given layer activations?”. Other works investigated the usage of domain knowledge rather than high-level concepts (Confalonieri et al. 2020).

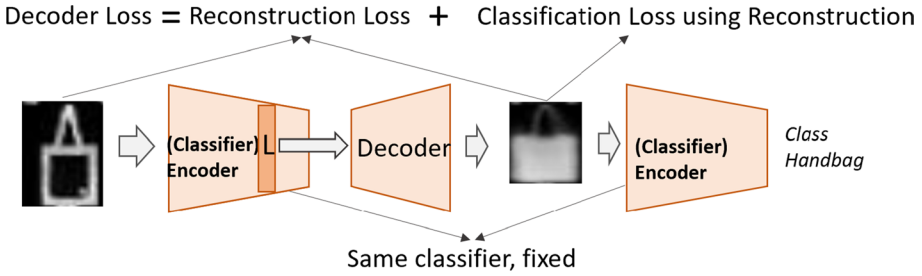
Denoising AEs are well established (Du et al. 2016; Vincent et al. 2010). They can be used to remove noise from images, reconstruct images and leverage data in an unsupervised manner (Du et al. 2016; Vincent et al. 2010). Ideas to combine unsupervised learning approaches to remove noise and supervised learning by extending loss functions have been presented since the early 90’s (Deco et al. 1993). In the context of explanations, AEs are also common (Guidotti et al. 2019; Qi et al. 2019; van Doorenmalen and Menkovski 2020), e.g., (Qi et al. 2019) used an AE with skip-connections for saliency map predictions. However, the encoder is fixed in our work, i.e., only we use the classifier directly as encoder. Furthermore, only in our work an AE is trained to identify distortion within the explanation process.

### 3 Method and architecture

In our local, model-agnostic post-hoc explanation method, an explanation consists of three images  $X$ ,  $\hat{X}_E$  and  $\hat{X}_R$ , which are compared among each other as illustrated in the right of Fig. 1. The *ClaDec* architecture is shown on the top portion of Fig. 1. The entire classifier has been trained beforehand to optimize classification loss. Its parameters remain unchanged during the explanation process. To explain layer  $L$  of the classifier (= encoder) for an input  $X$ , we use the activations of the entire layer  $L(X)$  (or a subset  $S(X) \subseteq L(X)$ ). The decoder is trained to optimize the reconstruction loss given the activations  $L(X)$  with respect to the original inputs  $X$ . The *RefAE* architecture is identical to *ClaDec*. It differs only in the training process and the objective. For the reference AE, the encoder and decoder are trained jointly to optimize the reconstruction loss of inputs  $X$ . In contrast, the encoder is treated as fixed in *ClaDec*. Once the training of all components is completed, explanations can be generated without further need for optimization. That is, for an input  $X$ , *ClaDec* computes the reconstruction  $\hat{X}_E$  serving together with the original input and the reconstruction from *RefAE* as the explanation.

However, comparing the reconstruction  $\hat{X}_E$  to the input  $X$  may be difficult and even misleading since the decoder can introduce distortions. Therefore, it is unclear, whether the differences between the input and the reconstruction originate from the encoding of the classifier or the inherent limitations of the decoder. Thus, we propose to use both the *RefAE* (capturing unavoidable limitations of the model or data) and *ClaDec* (capturing model behavior). The evaluation proceeds by comparing the reconstructed “reference” from *RefAE*, the explanation from *ClaDec* and the input. Only differences between the input and the reconstruction of *ClaDec* that do not occur in the reconstruction *RefAE* can be attributed to the classifier. In case, reconstructions by the *RefAE* are (almost) identical to the original images, it suffices to compare only the reconstruction  $\hat{X}_E$  by *ClaDec* to the input  $X$ .

The resulting reconstructions might be easy to interpret, but in some cases it might be preferable to allow for explanations that are more fidel, i.e., capturing more aspects of the model that should be explained. Figure 3 shows an extension of the base architecture



**Fig. 3** Extension of the *ClaDec* architecture. The decoder is optimized for reconstruction and classification loss

of *ClaDec* (Fig. 1) using a second loss term for the decoder training. It is motivated by the fact that *ClaDec* seems to yield reconstructions that capture more aspects of the input domain than of the classifier.

More formally, for an input  $X$ , a classifier  $C$  (to be explained) and a layer  $L$  to explain, let  $L(X)$  be the activations of layer  $L$  for input  $X$ , and  $Loss(C(X), Y)$  the classification loss of  $X$  depending on the true classes  $Y$ . Let  $S(X) \subseteq L(X)$  be the subset of neurons to explain, i.e.,  $S(X) = L(X)$  implies neurons of the entire layer should be explained. The decoder  $D$  transforms the representation  $S(X)$  into the reconstruction  $\hat{X}$ . For *ClaDec* the decoder loss is:

$$Loss_{ClaDec}(X) := (1 - \alpha) \cdot \sum_i (X_i - \hat{X}_{E,i})^2 + \alpha \cdot Loss(C(\hat{X}_E), Y) \tag{1}$$

with  $\hat{X}_E := D(S(X))$  and  $\alpha \in [0, 1]$

The trade-off parameter  $\alpha$  allows controlling whether reconstructions  $\hat{X}_E$  are more similar to inputs with which the domain expert is more familiar, or reconstructions that are more shaped by the classifier and, thus, they might look more different than training data a domain expert is familiar with. For reconstructions  $X_R$  of *RefAE* the loss is only the reconstruction loss:

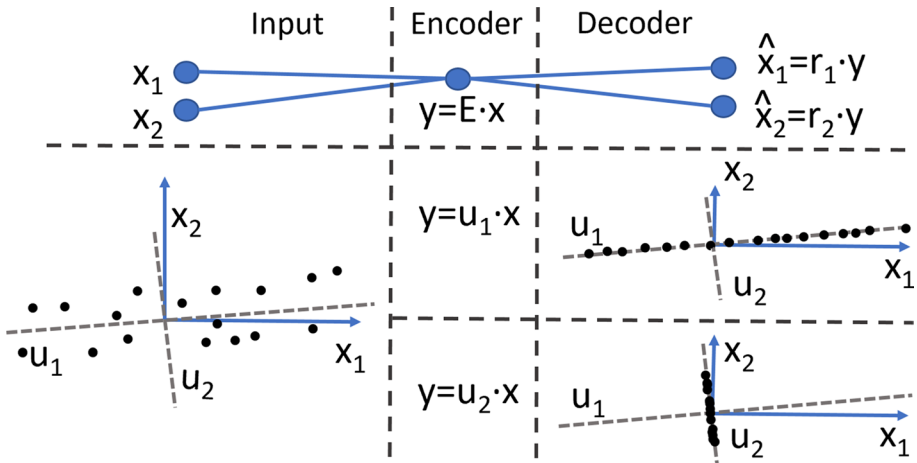
$$Loss_{RefAE}(X) := \sum_i (X_i - \hat{X}_{R,i})^2$$

### 4 Theoretical motivation of *ClaDec*

We provide rationale for reconstructing explanations using a decoder from a layer of a classifier that should be explained, and comparing it to the output of a conventional AE, i.e., *RefAE* (see Fig. 1). AEs perform a transformation of inputs to a latent space and then back to the original space. This comes with information loss on the original inputs because reconstructions are typically not identical to inputs.<sup>1</sup> To provide intuition, we focus on

<sup>1</sup> It may appear that this information loss is due to forcing high-dimensional data to be represented in a low dimensional space. However, as claimed in Goodfellow et al. (2016)(p.505), a non-linear encoder and decoder (theoretically) only require a single dimension to encode arbitrary information without any loss. The deeper mathematical reason is that a dimension  $d$  is a real number, i.e.,  $d \in \mathbb{R}$  and real numbers are uncountable infinite. Thus, there are (more than) enough options to encode an infinite amount of inputs. Therefore, information loss is more a failure of the model to encode and decode the input error free using just one dimension.





**Fig. 4** An AE with optimal encoder  $y = u_1 \cdot x$  (and decoder) captures more information than any other encoder. But a regression/classification model serving as encoder, e.g.,  $y = u_2 \cdot x$ , combined with an optimized decoder, might capture some input attributes more accurately, e.g.,  $x_2$

a simple architecture with a linear encoder (consisting of a linear model that should be explained), a single hidden unit and a linear decoder as depicted in Fig. 4.

An AE, i.e., the reference AE *RefAE*, aims to find an encoding vector  $E$  and a reconstruction vector  $R$ , so that the reconstruction  $\hat{x} = R \cdot y$  of the encoding  $y = E \cdot x$  is minimal using the L2-loss:

$$\min_{R,E} ||x - R \cdot E \cdot x||^2$$

The optimal solution which minimizes the reconstruction loss stems from projecting onto the eigenvector space (as given by a Principal Component Analysis) (Baldi and Hornik 1989). Given that there is just a single latent variable, the optimal solution for  $W = R \cdot E$  is the first eigenvector  $u_1$ . This is illustrated in Fig. 4 in the upper part with  $y = u_1 \cdot x$ . For *ClaDec* the goal is to explain a linear regression model  $y = E \cdot x$ . The vector  $E$  is found by solving a regression problem. We fit the decoder  $R$  to minimize the reconstruction loss on the original inputs given the encoding:

$$\min_R ||x - R \cdot y||^2 \text{ with } y = E \cdot x$$

The more similar the regression problem is to the encoding problem of an AE, the more similar are the reconstructions. Put differently, the closer  $E$  is to  $u_1$ , the lower the reconstruction loss and the more similar are the optimal reconstructions for the reference AE and *ClaDec*. Assume that  $E$  differs strongly from  $u_1$ , e.g., the optimal solution to the regression problem is the second eigenvector  $y = u_2 \cdot x$ . This is shown in the lower part of Fig. 4. When comparing the optimal reconstruction of the *RefAE*, i.e., using  $y = u_1 x$ , and the illustrated reconstruction of *ClaDec*, i.e., using  $y = u_2 x$ , it becomes apparent that for the optimal encoding  $y = u_1 x$  the reconstructions of both coordinates  $x_1$  and  $x_2$  are fairly accurate on average. In contrast, using  $y = u_2 x$ , coordinate  $x_2$  is reconstructed more accurately (on average), whereas the reconstruction of  $x_1$  is mostly poor.

Generally, this suggests that a representation obtained from a model (trained for some task) captures less information of the input than an encoder optimized for reconstructing inputs. But aspects of inputs relevant to the task should be captured relatively in more detail than those irrelevant. Reconstructions from *ClaDec* should show more similarity to original inputs for attributes relevant to classification and less similarity for irrelevant attributes. But, overall reconstructions from the classifier will show less similarity to inputs than those of an AE. Our fidelity assessment builds on this idea by proclaiming that reconstructions from an AE, i.e., *RefAE*, capture more information on inputs (measured using reconstruction loss) than those from *ClaDec*, but are less suitable for classification (measured using accuracy on a classifier trained on explanations).

## 5 Assessing comprehensibility and fidelity

*Fidelity* is the degree to which an explanation captures model behavior. That is, a “fidel” explanation captures the decision process of the model accurately.

The proposed evaluation (also serving as a sanity check) uses the rationale that fidel explanations for decisions of a well-performing model should help performing the task the model addresses. That is, learning from explanations of a well-performing model should also lead to a well-performing model for the same task. Concretely, training a new classifier  $C_{eval}^{ClaDec}$  on explanations should yield a better performing classifier than relying on reconstructed inputs only. That is, we train a baseline classifier  $C_{eval}^{RefAE}$  on the reconstructions  $\hat{X}_R$  of the *RefAE* and a second classifier with identical architecture  $C_{eval}^{ClaDec}$  on reconstructions  $\hat{X}_E$  from *ClaDec*. The latter classifier should achieve higher accuracy. This is a much stronger requirement than the common sanity check demanding that explanations must be valuable to perform a task better than a “guessing” baseline. More formally, we compare the accuracy  $Acc(C_{eval}^{ClaDec})$  of the model  $C_{eval}^{ClaDec}$  trained on reconstructions  $\hat{X}_E$  and that of a model  $C_{eval}^{RefAE}$  trained on reconstructions  $\hat{X}_R$  of the *RefAE*. Thus, as a (proxy) measure for fidelity  $\Delta Acc$ , we use

$$\Delta Acc := Acc(C_{eval}^{ClaDec}) - Acc(C_{eval}^{RefAE})$$

One must be careful that explanations do not contain additional external knowledge (not present in the inputs or training data) that helps solving the task. For most methods, including ours, this holds. It is not obvious that training on explanations improves classification performance compared to training on inputs that are more accurate reconstructions of the original inputs. Improvements seem only possible if an explanation is a more adequate representation to solve the problem. Formally, we measure the similarity between the reconstructions  $\hat{X}_R$  (using *RefAE*) and  $\hat{X}_E$  (of *ClaDec*) with the original inputs  $X$ . We show that explanations (from *ClaDec*) bear less similarity with original inputs than reconstructions from *RefAE*. Still, training on explanations  $\hat{X}_E$  only yields classifiers with better performance than on the more informative outputs  $\hat{X}_R$  from *RefAE*.

*Comprehensibility* is the degree to which the explanation is human-understandable. In our case, this means whether a person can make sense of the concepts depicted in the explanation. We build upon the intuitive assumption that a human can better and more easily comprehend explanations made of concepts that she is more familiar with. We argue that a user is more familiar with real-world phenomena and concepts as captured in the training data than possibly unknown concepts captured in representations of a neural network. This implies that more similar explanations to the training data are more comprehensible than

**Table 1** Encoder/Decoder, where “C” is a convolution, “DC” a deconv; a BatchNorm and a ReLu layer follow each “C” layer; a ReLu layer follows each “DC” layer

VGG-style Encoder		Decoder	
Type/Stride	filter shape	Type/stride	Filter shape
C/s2	3×3×1×32	FC	nClasses
C/s2	3×3×32×64	DC/s2	3×5×5×256
C/s1	3×3×64×128	DC/s2	3×5×5×128
C/s2	3×3×128×128	DC/s2	3×5×5×64
C/s1	3×3×128×256	DC/s2	3×5×5×32
C/s2	3×3×256×256	DC/s2	3×5×5×1
C/s1	3×3×256×512		
C/s2	3×3×512×512		
FC	256×nClasses		
Softmax/s1	Classifier		

those with strong deviations from the training data. Therefore, we quantify comprehensibility of a reconstruction  $\hat{X}_E$  by measuring the distance to the original input  $X$ , i.e., the reconstruction loss  $\|X - \hat{X}_E\|$ . If reconstructions show fidelitous but non-intuitive concepts (high reconstruction loss) then a user can experience difficulties in making sense of the explanation. In contrast, a trivial explanation (showing the unmodified input) is easy to understand but it will not reveal any insights into the model behavior, i.e., it lacks fidelity. We consider the best “trivial” explanation that can be obtained through a reconstruction process that of the *RefAE*. We discuss the reconstruction loss of the reconstructions  $\hat{X}_E$  from *ClaDec* with respect to that of the *RefAE*. That is, for a test dataset  $D = \{(X, Y)\}$ , we report separately the reconstruction loss for *RefAE* and *ClaDec* and the difference of these losses:

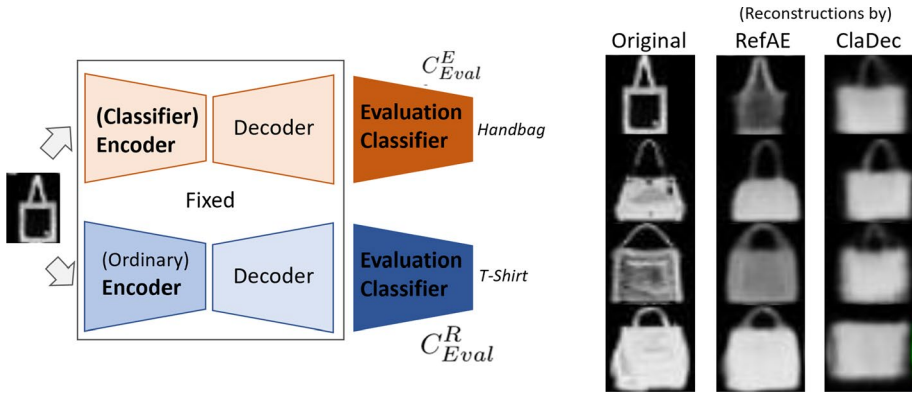
$$\Delta Rec := \frac{1}{|D|} \sum_{X, Y \in D} \|X - \hat{X}_E\| - \|X - X_R\|$$

## 6 Evaluation

We perform a qualitative and quantitative evaluation including a user study focusing on image classification using CNNs. We perform the following experiments:

- Explaining multiple layers for correct and incorrect classifications (Sect. 6.1)
- Varying the fidelity and comprehensibility trade-off (Sect. 6.2)
- Impact of encoder training and accuracy on explanations (Sect. 6.3)
- Using subsets of layer activations down to individual neurons (Sect. 6.4)
- Comparing *ClaDec* explanations to class prototypes learnt in Li et al. (2018) (Sect. 6.5)
- Investigating impact of occluding parts of the input (Section 6.6)
- A user study with non-experts asking them to make sense of explanations (Sect. 6.7)

*Setup* The decoder follows a standard design, i.e., using 5x5 deconvolutional layers. For the classifier and encoder we used the same architecture, i.e., a VGG-11 and ResNet-10. Architectures are shown in Fig. 1. For ResNet-10 we reconstructed after each block. For



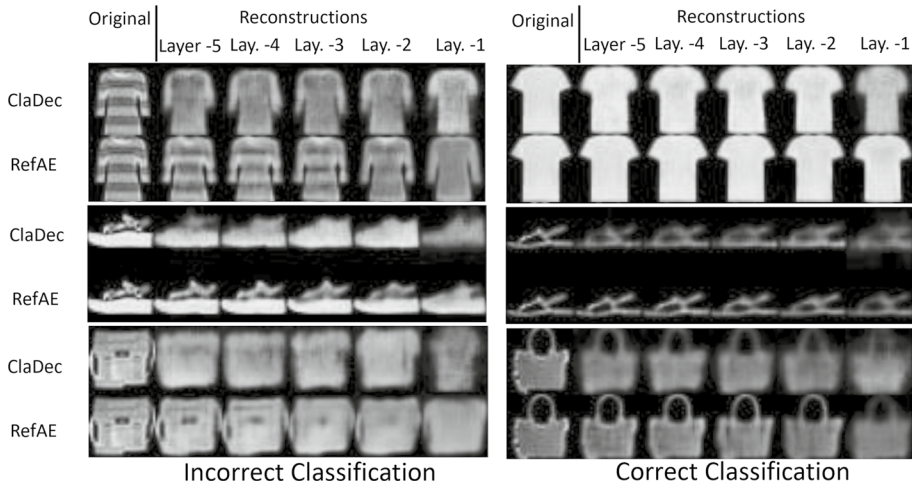
**Fig. 5** Left panel: Evaluation setup using a dedicated evaluation classifier for *ClaDec* and *RefAE*. Right panel: Comparison of original inputs and reconstructions using the FC layer of the encoder for handbags. *RefAE* and *ClaDec* both do not reconstruct detailed textures. The classifier does not rely on gray tones, which are captured by *RefAE*. It uses prototypical shapes

VGG-11 after a ReLU unit associated with a conv layer. The same classifier architecture (but trained with different input data) serves as encoder in *RefAE*, classifier in *ClaDec* and for classifiers used for evaluation of reconstructions, i.e., classifier  $C^{ClaDec}_{Eval}$  (for assessing *ClaDec*) and  $C^{RefAE}_{Eval}$  (for *RefAE*) shown in the left panel of Figure 5. We report the validation accuracy “Acc  $C^{ClaDec}_{Eval}$ ” and “Acc  $C^{RefAE}_{Eval}$ ” of these classifiers. Code is available at <https://github.com/JohnTailor/ClaDec>.

Note that the decoder architecture varies depending on which layer is to be explained. The original architecture allows to either obtain reconstructions from the last convolutional layer or the fully connected layer. For a lower layer, the highest deconvolutional layers from the decoder have to be removed, so that the reconstructed image  $\hat{X}$  has the same width and height as the original input  $X$ . We employed three datasets namely Fashion-MNIST, MNIST and CIFAR-100. Fashion-MNIST consists of 70000 28x28 images of clothing stemming from 10 classes that we scaled to 32x32. MNIST of 60000 digits and CIFAR-100 of 60000 objects in color. 10000 samples are used for testing. We train all models for reconstruction using the Adam optimizer for 64 epochs, i.e., *RefAE* and the decoder of *ClaDec*. The classifier serving as encoder in *ClaDec* as well as the classifiers used for evaluation for SGD were trained using SGD for 64 epochs starting from a learning rate of 0.1 that was decayed twice by 0.1. We conducted 5 runs for each reported number. We show both averages and standard deviations. The classifier performance for each of the dataset and architecture is comparable to those in other papers without data augmentation, i.e., for MNIST we achieved mean accuracy above 99%, for FashionMNIST above 92%, and for CIFAR-100 above 45% for both architectures.

### 6.1 Explaining layers for correct and incorrect classifications

Reconstructions based on *RefAE* and *ClaDec* for FashionMNIST are shown in Figs. 5, 6 and 7. Overall reconstructions from *ClaDec* resemble more prototypical, abstract features and they allow to identify relevance of input details due to imprecise reconstruction (blurriness, change of shape) or complete absence of concepts such as textures or gray tones. That is, omitted concepts are not relevant, while for poor reconstructions, the (blurry) input

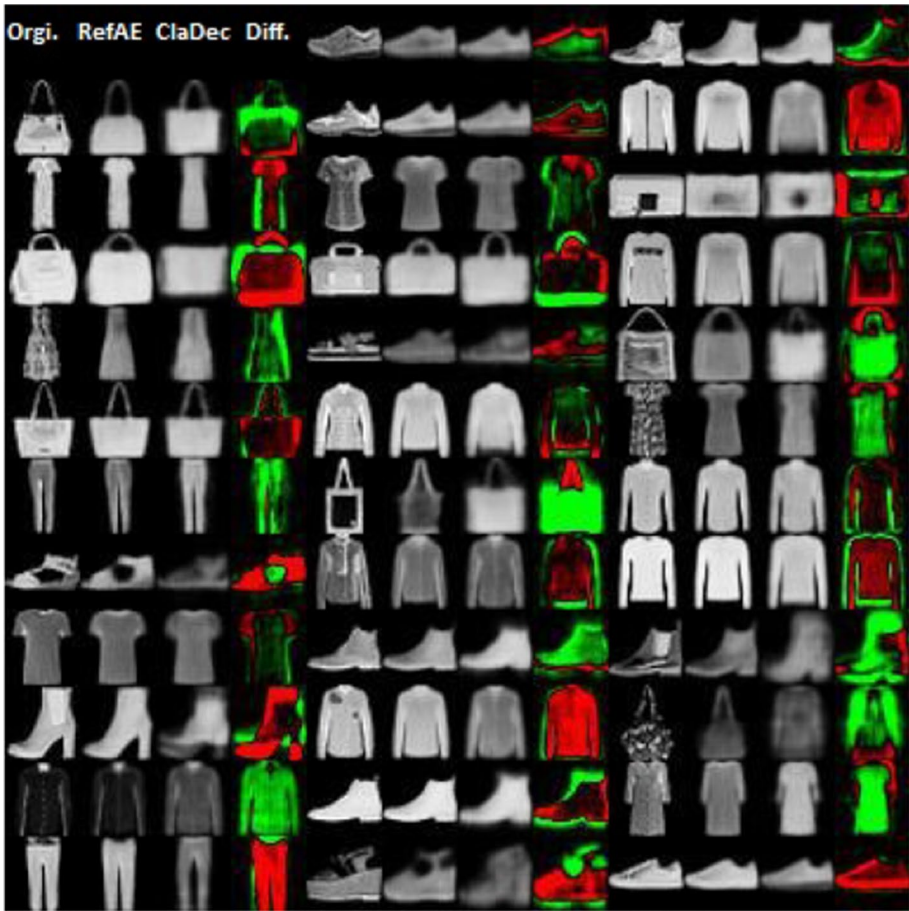


**Fig. 6** Comparison of original inputs and reconstructions using multiple layers of the encoder. For incorrect samples it shows a gradual transformation into another class. Differences between *RefAE* and *ClaDec* increase with each layer

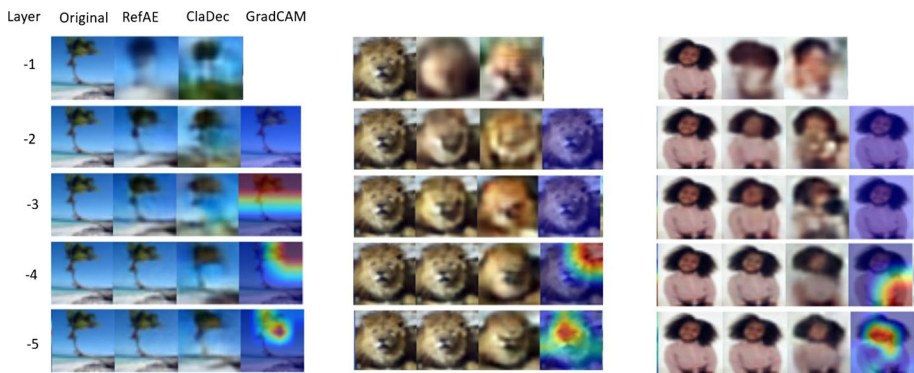
parts are not relevant at a high level of detail to discriminate between classes. In Fig. 5 the difference between reconstructions of *RefAE* and *ClaDec* are not the same as in a saliency map, i.e., they do not indicate that a red part is relevant according to *ClaDec* and a green part is not or vice versa. It only shows differences in pixel values. This is not directly translatable to relevance. The classifier (and thus the reconstruction from *ClaDec*) can turn a fairly dark area in the input into a gray area or a fairly white area into a gray area. This primarily indicates that the color or gray tone is not important, but it does not mean that in one case importance of the pixel is low and in the other it is high.

MNIST and CIFAR-100 exhibited similar behavior as Fashion-MNIST as can be seen in Figs. 8 and 9. For CIFAR-100 analyzing explanations is more cognitively demanding since there are more classes, classes exhibit more diversity and reconstructions of both *RefAE* and *ClaDec* are of worse quality compared to the other two datasets. First of all, it should be noted that GradCAM does not give favorable results for layers close to the output since these layers lack spatial extent. For example, for the third last layer the spatial dimensions of activation layers is  $2 \times 2$  requiring an upsampling by a factor of 16 for each dimension to the final output, which yields essentially a uniform attribution map of GradCAM. For the very last layer GradCAM yields meaningful results, i.e., for CIFAR-100 in Fig. 8 they show that the classifier focuses on the treetop as well as the girl's and lion's face. In fact, the situation is more delicate since GradCAM explanations do not clarify how the classifier perceives the highlighted pixels, i.e., does the classifier rely either on skin color or facial features such as nose or eyes or both?

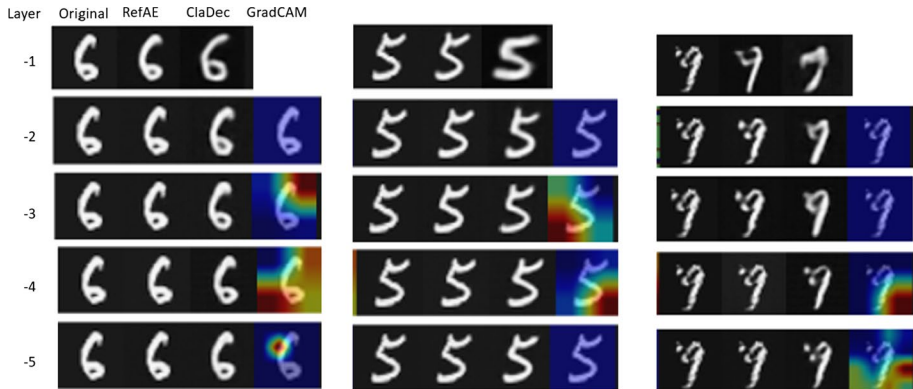
For *RefAE* reconstructions tend to lack details for CIFAR-100 (Fig. 8), in particular for the last layer, where the number of dimensions is just 100 compared to 512 of the prior layer. However, the original input and *RefAE* are overall still fairly similar. Comparing *RefAE* and *ClaDec* shows that both exhibit similar reconstructions for the two layers closest to the input with *ClaDec* appearing slightly more blurry. For upper layers, reconstructions show partially semantic differences. For the tree, the shape is changed to be more prototypical for a palm tree. While the blue sky remains well visible, the ground does not appear



**Fig. 7** Comparison of original inputs and reconstructions using the last layer, i.e., FC, of the VGG encoder in Table 1. Differences between reconstructions are shown in the last column. Red areas show where reconstructions from *RefAE* are brighter than those from *ClaDec*. Green shows the opposite



**Fig. 8** Comparison of original and reconstructions and GradCAM for different layers using ResNet on CIFAR-100

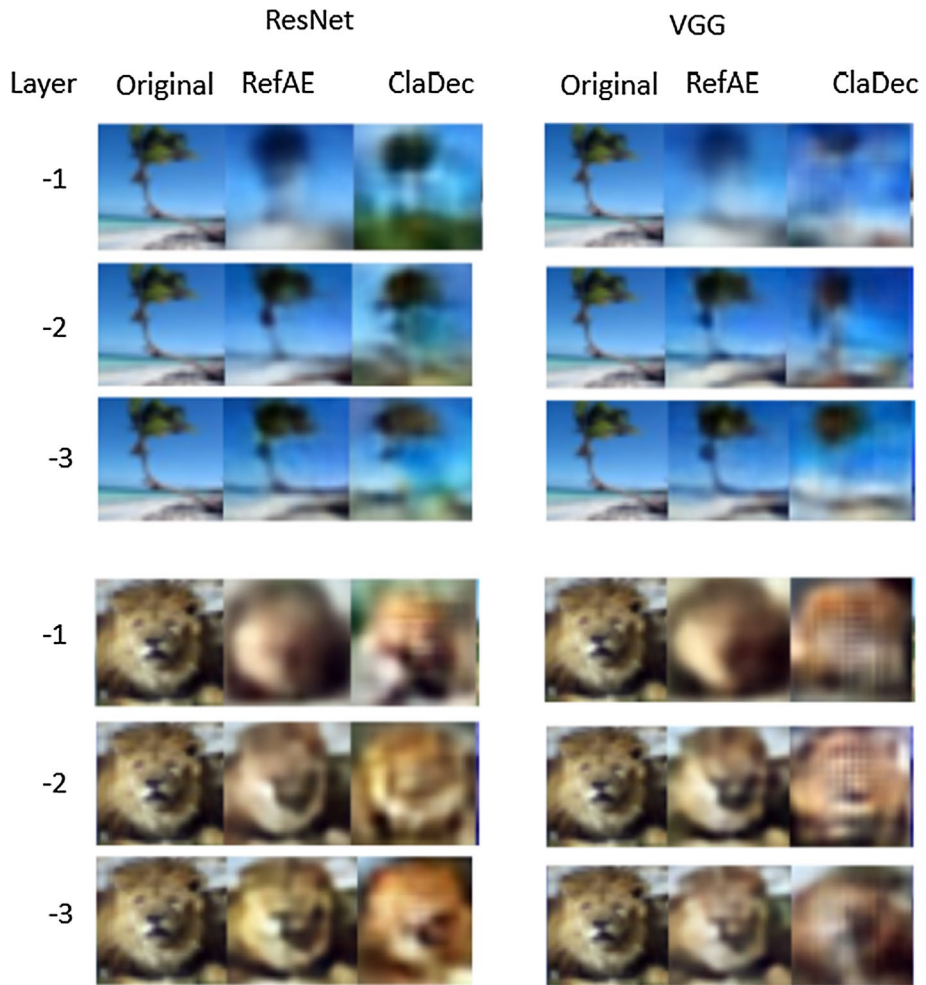


**Fig. 9** Comparison of original and reconstructions and GradCAM for different layers using ResNet on MNIST

as a white sandy beach including parts of the sea but it is altered from a more yellowish (sandy) ground to greenery. Compared to GradCAM, *ClaDec* provides a more diversified understanding of the layers since the reconstructions give more information for each layer. There is agreement between the two methods that the treetop is highly relevant. *ClaDec* shows that the tree shape is somewhat different for the input than for a typical sample, i.e., the collection of branches face more upwards in the original but more downward on the reconstruction (layers  $-2$  and  $-3$ ), which seems to be more characteristic (average) representation. Interestingly, the sandy beach is not well-reconstructed at the very top layer, whereas blue sky remains. This hints that the sandy ground and the ocean in the original do not lead the classifier to move towards a setting at a beach with sand and ocean. For the girl on the right the behavior is also interesting since starting from layer  $-3$  upwards the reconstructions from *ClaDec* appear to contain significant distortion in the center ultimately leading to wrong classification. That is, the very top does not show a black girl but a (white) woman. The girl's hair is changed from layer  $-4$  to  $-3$  from its fluffy appearance to a more common appearance (found more often in the training data). The change of skin color occurs from layer  $-3$  to  $-2$ . This shows that the classifier is not picking up less common looks such as fluffy hair as a criterion for identification of the person and it highlights how the hair is transformed. For the lion, colors appear more vivid, i.e., brighter with larger hue, for *ClaDec* showing how the more pale input deviates from the typical representation of colors. Also on the topmost layer the face of the lion is shrunk and more background is added, highlighting that such a setting is more common.

For MNIST (Fig. 9) the reconstructions of *RefAE* are highly accurate. There are clearly visible differences in the top layer only for digit “9”. For *ClaDec* reconstructions for digits “5” and “6” are highly accurate except for the last layer, where more typical samples are reconstructed and sharpness also decreases for “5”. The digit “9” is most interesting - it is also classified wrongly as “7” as can be seen well from the top layer reconstruction with the change already emerging at lower layers. At layer  $-3$  the reconstruction still seems to point to a “9”, with some uncommon remains of the “big dot” and the loop not being fully closed. At layer  $-2$  the opening of the loop increases somewhat and now the distinction between “7” and “9” is less clear. GradCAM explanations are not so insightful.

We also compared reconstruction of ResNet and VGG shown in Fig. 10. The reconstructions of the *RefAE* are similar for both architectures, hinting that there is consistency



**Fig. 10** Comparison of VGG and ResNet reconstructions on CIFAR-100

– at least for the reference. This is not obvious since the encoder architectures differ. However, for *ClaDec* the reconstructions of the two classifiers show more variation. Quality of both, i.e., sharpness and shape characteristics, are similar with ResNet having arguably a slight advantage. But reconstructed concepts show differences. The ground for the beach differs strongly between the two as well as the shape of the treetop, while both maintain the blue sky. For the lion color tones differ also, with ResNet showing more saturated colors.

Quantitative Results in Tables 2 and 3 contain two key messages: First, the reconstruction loss is lower for *RefAE* than for *ClaDec*. This follows since *RefAE* is optimized entirely towards minimal reconstruction loss of the original inputs. Second, the classification (evaluation) accuracy is (almost always) higher, when training the evaluation classifier  $C_{Eval}$  using reconstructions from *ClaDec* rather than from *RefAE*. This behavior is not obvious and it might be seen as surprising since the reconstructions from *ClaDec* are poorer according to the reconstruction loss. That is, they contain less information about



**Table 2** Explaining layers for VGG: *ClaDec* has larger reconstruction loss but the evaluation classifier has higher accuracy on *ClaDec*'s reconstructions

Layer	Rec Loss <i>ClaDec</i>	Rec Loss <i>RefAE</i>	$\Delta$ Rec	Acc $C_{Eval}^{ClaDec}$	Acc $C_{Eval}^{RefAE}$	$\Delta$ Acc
<i>FashionMNIST</i>						
-1	0.24±0.0183	0.038±0.0003	0.203	0.905±0.0003	0.917±0.0019	-0.011
-2	0.109±0.0047	0.017±0.0004	0.092	0.918±0.0021	0.921±0.0011	-0.003
-3	0.062±0.0015	0.013±0.0005	0.049	0.92±0.0022	0.919±0.0023	0.002
-4	0.042±0.0007	0.003±0.0004	0.039	0.925±0.0016	0.917±0.0029	0.008
-5	0.038±1e-04	0.002±0.0002	0.036	0.927±0.0013	0.915±0.0005	0.012
<i>MNIST</i>						
-1	0.374±0.0042	0.028±0.0005	0.346	0.993±0.0004	0.989±0.0009	0.004
-2	0.139±0.0022	0.012±0.0005	0.127	0.995±0.0005	0.99±0.0003	0.005
-3	0.074±0.0013	0.01±0.0004	0.064	0.995±0.0002	0.992±0.0006	0.003
-4	0.038±0.0013	0.003±0.0001	0.035	0.995±0.0006	0.992±0.0006	0.003
-5	0.029±0.0004	0.002±0.0002	0.027	0.995±0.0003	0.994±0.0004	0.001
<i>CIFAR-100</i>						
-1	0.16±0.0	0.12±0.001	0.04	0.31±0.003	0.37±0.002	-0.05
-2	0.11±0.001	0.06±0.002	0.05	0.41±0.004	0.37±0.007	0.05
-3	0.13±0.002	0.05±0.001	0.08	0.43±0.002	0.37±0.009	0.06
-4	0.1±0.001	0.02±0.002	0.09	0.48±0.002	0.41±0.005	0.07
-5	0.08±0.001	0.01±0.0	0.07	0.49±0.004	0.43±0.003	0.06

**Table 3** Explaining layers for ResNet: *ClaDec* has larger reconstruction loss but the evaluation classifier on reconstructions from *ClaDec* achieves higher accuracy

Layer	Rec Loss <i>ClaDec</i>	Rec Loss <i>RefAE</i>	$\Delta$ Rec	Acc $C_{Eval}^{ClaDec}$	Acc $C_{Eval}^{RefAE}$	$\Delta$ Acc
<i>FashionMNIST</i>						
-1	0.1444±0.00338	0.0287±0.00048	0.1158	0.916±0.00248	0.922±0.00107	-0.006
-2	0.0222±0.00075	0.0006±0.00011	0.0216	0.9345±0.00115	0.9256±0.00212	0.0089
-3	0.0069±0.00056	0.0004±8e-05	0.0065	0.9353±0.00047	0.931±0.00035	0.0043
-4	0.0171±0.00487	0.0006±0.00012	0.0165	0.936±0.00122	0.9308±0.00111	0.0052
-5	0.0023±0.00069	0.0015±0.0005	0.0008	0.9362±0.0015	0.935±0.00096	0.0012
<i>MNIST</i>						
-1	0.1722±0.00184	0.0194±0.00079	0.1528	0.9952±0.00098	0.9833±0.00226	0.0119
-2	0.0148±0.00025	0.0004±3e-05	0.0143	0.9964±0.00024	0.9955±0.00032	0.0009
-3	0.0045±0.00016	0.0002±0.00016	0.0043	0.9964±0.00019	0.9958±0.00022	0.0006
-4	0.0026±0.00039	0.0001±3e-05	0.0025	0.9963±0.00022	0.9962±0.00039	0.0001
-5	0.001±0.00015	0.0006±8e-05	0.0004	0.996±0.00036	0.9962±0.00044	-0.0002
<i>CIFAR-100</i>						
-1	0.185±0.002	0.071±0.0017	0.114	0.389±0.0038	0.508±0.0027	-0.12
-2	0.057±0.0015	0.002±0.0003	0.055	0.602±0.0022	0.532±0.0025	0.07
-3	0.014±0.0008	0.001±0.0002	0.013	0.601±0.0033	0.583±0.0036	0.018
-4	0.015±0.0025	0.002±0.0003	0.013	0.607±0.0027	0.592±0.0031	0.015
-5	0.003±0.0005	0.005±0.0006	-0.002	0.604±0.0032	0.605±0.0029	-0.001

the original input than those from *RefAE*. However, it seems that the “right” information is encoded using a better suited representation. Only for the very last layer, i.e., the linear layer, the accuracy for the evaluation classifier tends to be higher if trained on reconstructions from *RefAE*. The last layer acts most discriminatively and has lower dimensions, i.e., for the classifier often only relatively few neurons are active in the last layer. The reconstructions from *ClaDec* are often very different from the original and of poor quality, i.e., blurry, or resembling fairly different looking, prototypical objects (though of the same category). For CIFAR-100 also once the reconstruction loss is lower for *ClaDec* though not by a large margin and a visual assessment generally attests lower quality to reconstructions. Aiming for a smooth, average-like reconstruction seems to be beneficial.

Aside from these two key observations there are a set of other noteworthy behaviors: The reconstruction loss increases the more encoder layers, i.e., the more the input is transformed. In absolute numbers, the impact is significantly stronger for *ClaDec*. The difference between *RefAE* and *ClaDec* increases the closer the layer to explain is to the output. For Resnet (Table 3) the accuracy of the evaluation classifier is fairly constant across layers for *ClaDec* (except for the very last layer, where it drops) and the reconstruction loss is significantly lower for both *RefAE* and *ClaDec* compared to VGG. This is not unexpected, since Resnets do not downsample that much (up to the last layer the spatial dimension is still 4x4) and it contains residual connections that facilitate information flow. However, also for ResNets the reconstruction error clearly increases for *ClaDec* with more layers.

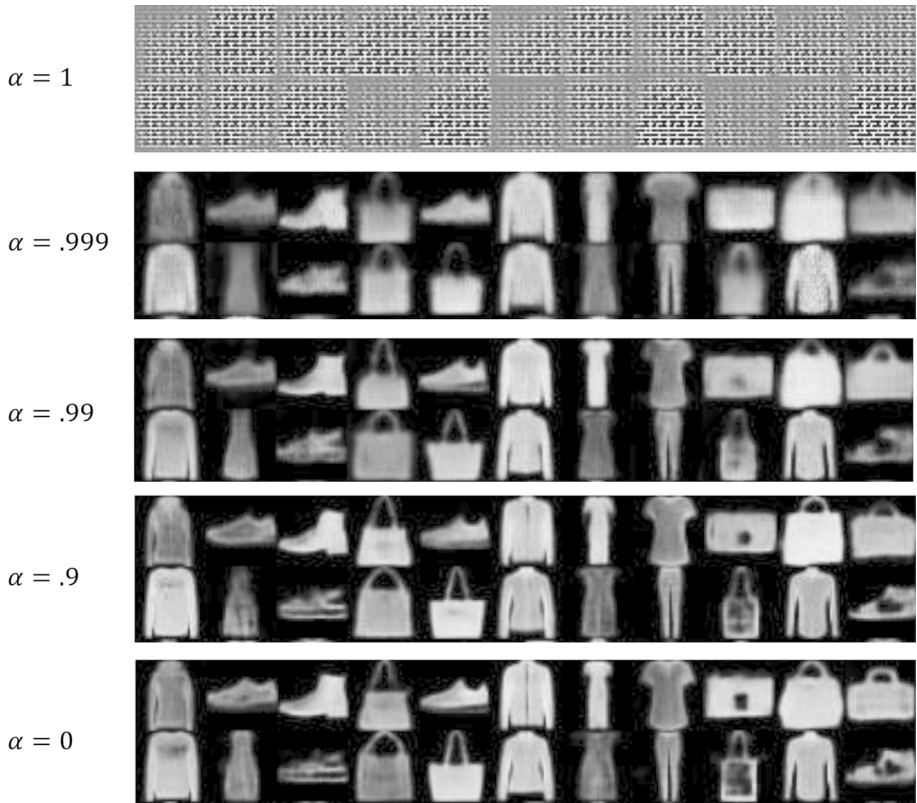
## 6.2 Fidelity and comprehensibility trade-off

We assess the impact of adding a classification loss (Fig. 3) to modulate using a parameter  $\alpha \in [0, 1]$  how much the classifier model impacts reconstructions. Neglecting reconstruction loss, i.e.,  $\alpha = 1$ , yields non-interpretable reconstructions as shown in the first row in Fig. 11. Already modest reconstruction loss leads to well-human recognizable shapes. The quality of reconstructions in terms of sharpness and amount of captured detail constantly improves the more emphasis is put on reconstruction loss. It also becomes evident that the neural network learns “prototypical” samples (or features) towards which reconstructed samples are being optimized. For example, the shape of handbag handles shows much more diversity for values of  $\alpha$  close to 0, it is fairly uniform for relatively large values of  $\alpha$ . Thus, the parameter  $\alpha$  provides a means to reconstruct a compromise between the sample that yields minimal classification loss and a sample that is true to the input. It suggests that areas of the reconstruction of *ClaDec* that are similar to the original input are also similar to a “prototype” that minimizes classification loss. The network can recognize them well, whereas areas that are strongly modified, resemble parts that seem non-aligned with “the prototype” encoded in the network.

Quantitative results in Tables 4 and 5 shows that evaluation accuracy increases when adding a classification loss, i.e.,  $\alpha > 0$  yields an accuracy above 90% whereas  $\alpha = 0$  gives about 88%. Reconstructions that are stronger influenced by the model to explain (larger  $\alpha$ ) are more truthful to the model, but they exhibit larger differences from the original inputs. Choosing  $\alpha$  slightly above the minimum, i.e., larger than 0, already has a strong impact.

## 6.3 Impact of encoder training

Comparing reconstructions of the *RefAE* with those of *ClaDec* for an untrained classifier might be used to assess the relevance of training the encoder of the classifier or an AE



**Fig. 11** Adding classification loss ( $\alpha > 0$ ) yields worse reconstructions for the last conv. layer. Using classification loss only ( $\alpha = 1$ ), reconstructions are not human recognizable

**Table 4** For VGG, adding classification loss  $\alpha > 0$  (Equation 1) yields worse reconstructions, but higher evaluation accuracy

$\alpha$	Total Loss <i>ClaDec</i>	Rec Loss	Classifier Loss	Acc $C_{Eval}^{ClaDec}$
1.0	0.01±0.003	285.5±52.01	0.0±0.0	0.903±0.003
0.999	0.03±0.002	25.4±0.929	0.03±0.0	0.903±0.002
0.9	0.84±0.013	8.35±0.12	0.75±0.011	0.901±0.002
0	7.49±0.112	7.49±0.112	4.4±0.254	0.882±0.004

itself, i.e., “How does training an encoder impact reconstructions compared to a non-trained encoder?”. Figure 12 shows reconstructions if the classifier is not trained at all. The figure suggests that a trained encoder does lead to encodings that allow to better reconstruct original inputs than an untrained encoder utilizing randomly initialized layers. While sharpness is generally comparable for both reconstructions, there are several samples where shapes of objects are altered, or details differ. Overall, this behavior can be understood using theory on random projections as we explain next for our quantitative results in Table 6. The table

**Table 5** For ResNet, adding classification loss  $\alpha > 0$  yields worse reconstructions, but higher evaluation accuracy

$\alpha$	Total Loss <i>ClaDec</i>	Rec Loss	Classifier Loss	Acc $C_{Eval}^{ClaDec}$
1.0	0.009±0.001	416.5±76.69	0.0±0.0	0.912±0.003
0.999	0.043±0.003	34.5±1.6	0.035±0.002	0.911±0.002
0.99	0.195±0.009	19.0±0.704	0.188±0.007	0.911±0.003
0.9	1.33±0.017	13.3±0.164	1.193±0.015	0.909±0.002
0	12.2±0.196	12.2±0.196	4.936±0.057	0.898±0.005

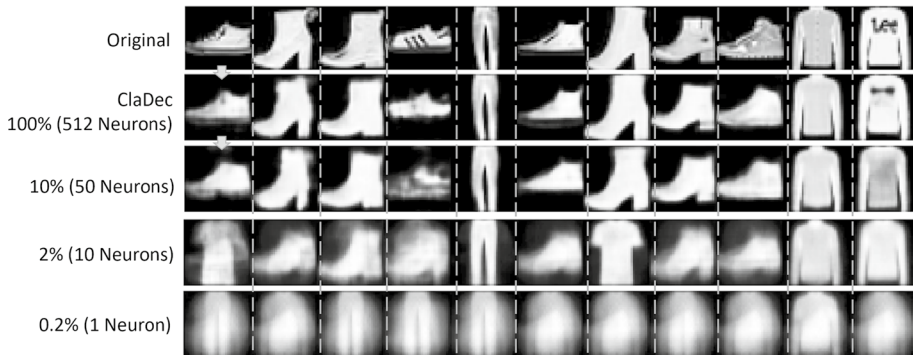


**Fig. 12** Comparison of original inputs and reconstructions using the last conv. layer of the encoder in Table 1 without any training of the classifier in *ClaDec*. Green indicating brighter values of *ClaDec* than *RefAE* and red the opposite

**Table 6** Impact of Classifier Accuracy (modulated through training epochs): Evaluation Accuracy increases with higher classifier accuracy, behavior of rec.loss follows an inverted U shape

Train. Epochs of Classifier to be expl.	Acc. of Classifier to be expl.	Rec Loss		$\Delta$ Rec	Acc Eval		$\Delta$ Acc
		<i>ClaDec</i>	<i>RefAE</i>		<i>ClaDec</i>	<i>RefAE</i>	
0	0.1±0.0	5.445±0.164	3.333±0.04	2.112	0.85±0.003	0.886±0.004	- 0.036
1	0.506±0.012	6.417±0.152	3.3±0.038	3.118	0.88±0.002	0.888±0.003	- 0.007
4	0.885±0.003	6.608±0.079	3.299±0.049	3.309	0.893±0.004	0.893±0.004	0.0
16	0.902±0.003	6.233±0.145	3.334±0.062	2.898	0.896±0.005	0.891±0.003	0.005
64	0.904±0.003	6.081±0.069	3.341±0.062	2.74	0.895±0.002	0.889±0.004	0.006

shows for *ClaDec* that classifiers that are trained longer and, therefore, achieve higher validation accuracy also lead to better accuracy for the evaluation classifier. More surprising is the dependence of the reconstruction loss on the number of training epochs. For *ClaDec*, it is lowest without any training, increases quickly and then steadily decreases again. This pattern is highly statistically significant. We conducted *t*-tests to verify that means between subsequent rows are different, yielding *p*-values below 0.01. The fact that an untrained



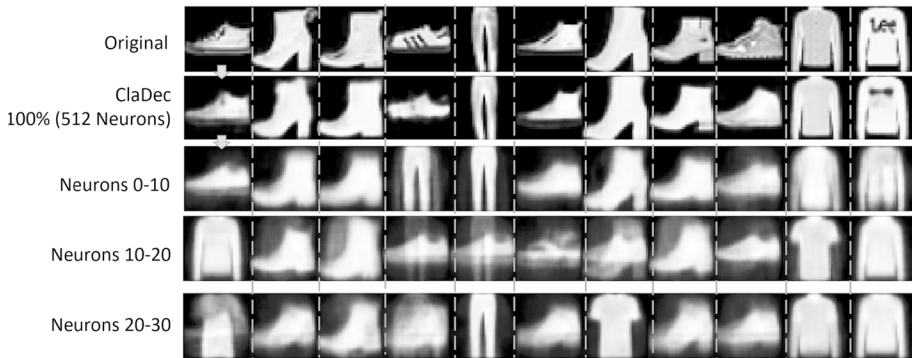
**Fig. 13** Reconstructions for *ClaDec* when using subsets of all activations of varying size for the last conv layer of VGG. Reconstructions quality decreases with fewer neurons

network, i.e., using random weights, achieves lower reconstruction loss than the trained classifier can be explained as follows: First, it should be noted that the reconstruction loss using random weights is significantly higher than for the reference architecture, where the encoder is optimized. Second, it is known from extreme learning, e.g., (Sun et al. 2017), that encoders with randomly chosen weights can yield good results, if just the decoder is optimized. More generally, good encoding properties might be deduced from the behavior of random projections formulated in the Johnson–Lindenstrauss lemma. It says that distances are well-preserved in a low-dimensional space originating from random projections. The theorem is commonly used in machine learning, e.g., (Schneider and Vlachos 2013, 2014). Training the classifier seems to destroy some of the desirable properties of random initialization by focusing on information needed for classification (but not for reconstruction) – as motivated theoretically. The reconstruction improves with more training, indicating that the initial encodings are noisy.

## 6.4 Explaining subsets of layer activations

We aim to assess reconstructions that are only based on some neurons of a layer. While neurons in lower layers often encode generic, class-independent information, neurons in higher layers are more strongly associated with specific concepts found in a dataset or a specific class. That is, neurons from upper layers often show large activations for one class or few classes only. This leads to the hypothesis that a subset of neurons might describe some classes well and others not so well. Since activations of different neurons often correlate, the idea that a neuron encodes one specific pattern that can only be reconstructed using this specific neuron is generally not valid (Fong and Vedaldi 2018). However, more information on a concept certainly helps in obtaining a better reconstruction as we shall show.

We train *ClaDec* using only a subset of neurons, i.e., their activations. We pick a subset  $S \subseteq L$  where  $L$  are all neurons of a layer. We reconstruct the input only based on the activations of this subset  $S$  of neurons. In our evaluation, we decode subsets of activations of the last convolutional layer. In Fig. 13 we vary the subset size so that subsets with fewer neurons are subsets of those with more neurons. Formally, let  $L'$  and  $L''$  be two subsets of  $L$ , i.e.,  $L', L'' \subset L$ . If  $|L'| < |L''|$  then  $L' \subset L''$ . In Figure 14 we maintain the same subset size but use disjoint subsets, that is  $L' \cap L'' = \{\}$ . Neurons can be chosen arbitrarily. We chose



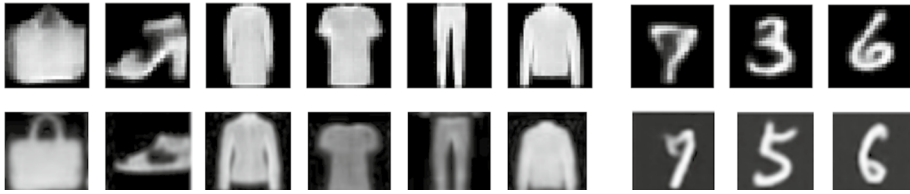
**Fig. 14** Reconstructions for *ClaDec* when using disjoint subsets of all activations of the last conv layer of VGG. Class associations of subsets are well-recognizable

subsets consisting of a sequence of neurons, which is as good as choosing them randomly since indexes of neurons and their parameters and semantics bear no relationship.

Overall, it is well visible that using fewer activations to reconstruct leads to less detailed and more blurry reconstructions - see Figs. 14 and 13. However, it is well-known that deep learning networks contain significant redundancy and inter-dependencies (Fong and Vedaldi 2018). Therefore, using only a fraction of all neurons is likely to lead to samples that are still resembling the input to a significant degree, i.e., neuron activations tend to correlate.

When using disjoint sets (Fig. 14) some samples are very poorly reconstructed compared to those based on all neurons. Poor reconstructions happen when we use only a small number of neurons of a layer. These neurons relate to concepts that are mostly associated with one class, i.e., we use 2% of all neurons of the last convolutional layer. Thus, for some classes  $A$ , there are no neurons that contain any strongly related concept. That is, the decoder cannot distinguish inputs, i.e., subsets of activations, belonging to such classes  $A$  well. That is, to the decoder samples from classes  $A$  look alike. In this case, reconstructions resemble overlays corresponding to prototypes (or averages) of multiple classes. They appear in low intensity and blurry. More precisely, the (weak) overlays show the classes the decoder is uncertain about. For example, the figure shows overlays of sneakers and pants in row 4 (Neurons 10–20). Thus, blurriness indicates a lack of information (in the activations) to reconstruct input details or even just the class. It also becomes apparent that in case there are neurons associated with a class, reconstructions are fairly decent, in the sense that they resemble a well-recognizable class prototype. However, a closer look reveals that they lack many details compared to reconstructions based on all neurons.

When using smaller subsets of neurons (Fig. 13) one can see a steady decrease in quality. For a single neuron classes can only be guessed. For a particular class and a subset of neurons  $S$  reconstructions could be good because most neurons in  $S$  contribute some information to the reconstruction or a few neurons in  $S$  are very strongly related to the class and others are not at all.



**Fig. 15** Comparison of explanations from *ClaDec* from last conv layer (bottom row) with class prototypes from Li et al. (2018) (top row) for MNIST and FashionMNIST

## 6.5 Comparing *ClaDec* explanations to learnt class prototypes

Figure 15 shows a subset of the 15 class prototypes learnt in Li et al. (2018). Li et al. (2018) proposes a special architecture to learn prototypes that are also used during the classification process. The number of prototypes to be learnt is a hyperparameter set to 15 for the datasets in Li et al. (2018). *ClaDec* can be said to compute a distinct reconstruction for each of the (thousands of) samples that might appear prototypical. Figure 15 indicates that explanations from *ClaDec* and prototypes from Li et al. (2018) both focus on information relevant for classification, e.g., they do not show texture and stripes. The class prototypes from Li et al. (2018) do not appear to be a “typical” (or average) representation but bear similarity to an explanation of *ClaDec* for a seemingly randomly chosen sample. That is, class prototypes contain concepts that are not widely present across samples. For instance, the “7” contains several small uncommon artifacts, and the handbag is also a specific type of bag with a handle that is clearly shorter than those of the average.<sup>2</sup> In the proposed architecture (Li et al. 2018) these class prototypes reside in the “prototype” layer that is followed by a fully connected layer as the last layer before the output is computed (using a Softmax). This implies that the class prototypes are also encoding concepts that occur at this abstraction level. As the prototypes are not reconstructed based on the very last layer but are followed by a fully connected layer, classification depends on the combination of multiple prototypes. For instance, there might be a prototype showing a red T-shirt and a black pullover. When a red pullover is classified as a pullover, the match with both prototypes might contribute to the decision of the pullover (though the red T-shirt prototype might be stronger associated with the T-shirt class). As such a prototype showing a sample of a particular class might lack concepts that are also common for this class.

To summarize, the explanations of both methods lead to similar insights about the general behavior of the classifier. While (Li et al. 2018) has explainability built into the model, *ClaDec* also comes with a set of advantages: It is model-agnostic (i.e., to all state-of-the-art models), it explains individual decisions (i.e., it highlights for each sample what is relevant and what not), it allows to explain different layers and subsets thereof.

## 6.6 Impact of occlusions of inputs

We qualitatively and quantitatively assess the impact of input occlusions. The idea is to assess the change in appearance of reconstructions in our qualitative evaluation and the

<sup>2</sup> We inspected about 100 handbags from the dataset to derive this conclusion.

change in prediction accuracy when altering more or less relevant parts of the input. The idea to obstruct parts of the input to understand their relevance is common in XAI, e.g., (Petsiuk et al. 2018). For saliency maps obtaining relevance is trivial since saliency maps output a relevance score for each pixel. Thus, to get the relevance of an arbitrary area, relevance score of pixels can be summed. For reconstructions from *ClaDec* determining the relevance of a pixel and area is more tricky since it is based on reconstructing concepts and not on computing scores for (input) pixels. To obtain a pixel relevance score, we use the idea that if parts relevant to the classification in the input are distorted, the classifier's layer activation also gets heavily distorted. Thus, reconstructions by *ClaDec* get strongly distorted. If parts of the input get occluded that are not impacting the classifier, then the activations and, consequently, the reconstructions  $\hat{X}_E$  by *ClaDec* should not change much. Therefore, we can assess how much the overall reconstructed image gets altered due to occluding parts of the input. Thus, assume we occlude a set of pixels  $O$ , then the relevance  $R(O)$  of these pixels is given by the sum of differences in reconstructions between the reconstruction  $\hat{X}_E$  for the original input  $X$  by *ClaDec* and the reconstruction  $\hat{X}_E^{occ}$  of the partially occluded input  $X^{occ}$  by *ClaDec*. Formally, we define

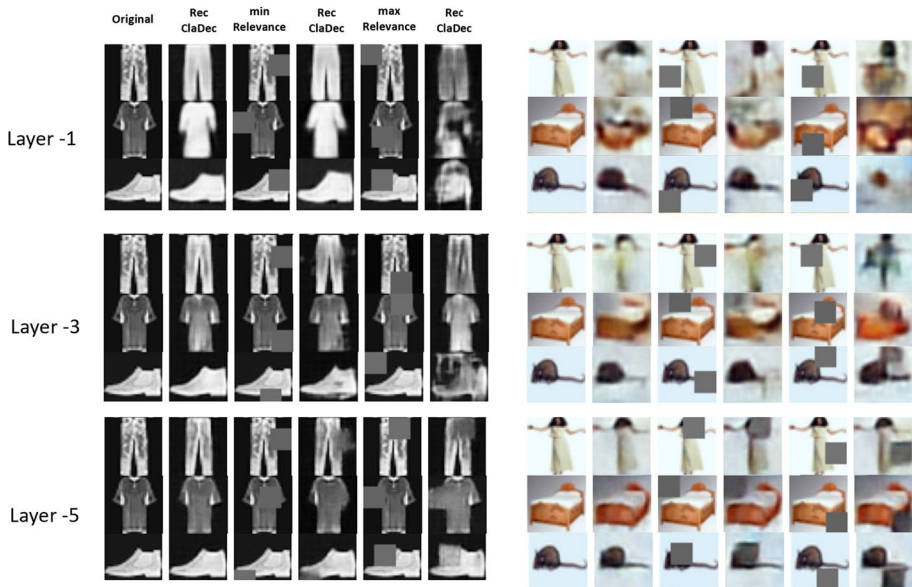
$$R^{ClaDec}(O) := \sum_{i \in O} \|\hat{X}_{E,i}^{occ} - \hat{X}_{E,i}\|^2$$

In our evaluation, we consider for each input 16 options for occlusion. Each occlusion having the shape of a square of size  $12 \times 12$  pixels with the upper left corner  $(x_u, y_u)$  given by  $x_u, y_u \in [0, 6, 12, 20]$ . We replace the occluded pixels with the mean value of all pixels in the training data, i.e., a gray tune.

In our quantitative evaluation, we compute the accuracy of the classifier  $C$  to explain occluded inputs. We compare the accuracy when occluding least and most relevant parts of the input according to *ClaDec*, i.e., the measure  $R^{ClaDec}(O, X)$ , and according to GradCAM, where we defined the relevance  $R^{GradCAM}(O, X)$  as being the sum of the value of the saliency map  $S(X)$  normalized to  $[0, 1]$  of pixel  $i \in O$ , i.e.,  $R^{GradCAM}(O, X) := \sum_{i \in O} S_i(X)$ . In particular, we are interested in each image in the occlusion out of the 16 possibilities yielding minimum (or maximum) relevance according to an explainability method, i.e., either GradCAM or *ClaDec*. For a relevance score  $R \in \{R^{ClaDec}, R^{GradCAM}\}$ , we define  $X_{O,min}$  with an occlusion  $O_{min}$  of areas of minimum relevance. Since there can be many occlusions with the same minimum relevance,  $O_{min}$  is chosen uniformly at random out of all of the occlusions  $\{O' | R(O', X) = \min_O R(O, X)\}$  yielding minimum relevance.  $X_{O,max}$  and  $O_{max}$  are defined analogously. We define  $Acc(C(X_{O,min}))$  as the accuracy of the classifier  $C$  to explain on samples  $X_{O,min}$ , i.e. samples  $X$  where an area of minimum relevance is occluded.

**Qualitative evaluation** Figure 16 shows the most and least relevant occlusions. It can be seen that occluding areas in inputs deemed more relevant lead to larger distortion in the reconstructions. Relevance is not stable across layers, i.e., the areas with minimum (and maximum) relevance scores depend on the layer. Occluding an area can lead to specific features at a layer to activate or fail to activate. Convolutional kernels have a larger receptive field and more semantic features for higher-level layers. This implies that activating a feature at a higher level might still be possible, although parts of the input that activate it are occluded. Furthermore, the impact on the reconstruction is expected to cover a larger area and yield more extreme distortions in the reconstruction when a higher-level feature is not activating. This can be well-seen in Fig. 16 when comparing the reconstruction of occluding the maximum relevance area for the shoe (left-hand side) or bed (right-hand side) layer  $-1$  and layer  $-5$ . For layer  $-5$  the reconstructed image is almost identical





**Fig. 16** Reconstructions for *ClaDec* when occluding parts of the image. Columns are original image  $X$ , its reconstruction, occluded image  $X_{O,min}$ , its reconstruction,  $X_{O,max}$  and its reconstruction for FashionMNIST and CIFAR-100 for VGG

to the original except for the occluded square, whereas for layer  $-1$  it looks completely different. Aside from that, lower layers allow reconstructing the image better than higher layers since fewer transformations (and possibly loss due to downsampling) of input information has occurred. The gray area indicating an occlusion is well-visible in the reconstruction in lower layers in Fig. 16, and the rest of the object is also commonly undistorted. Since our relevance measure is based on the difference between the reconstructed image without occlusion and with occlusion, the occlusion that maximizes relevance is related primarily to the difference between the pixels of the occluded area in the two compared reconstructions since the rest is almost identical, i.e., it has low relevance. This can lead to the situation where the maximum relevance area is part of an (uninformative) background that seems irrelevant for classification and the minimum relevance area contains the object to be classified. This is illustrated well in Fig. 16 for the mouse and the woman on layer  $-5$ . The area of maximum relevance is the white background in both cases since the difference between white and gray pixels is very large. In contrast, it is the object itself for minimum relevance since it is darker and therefore, the difference (and relevance) is smaller. Lower layers of the classifier do not alter the information flow very much, i.e., they do not filter much information of the input, no matter where the occlusions occur. Thus, if only a few layers of the classifier processed the occluded input, it can be well-reconstructed independent of the occlusion location. However, if more layers are involved, the reconstruction depends heavily on the activation of the input features that are semantically meaningful and tight to specific input areas.

Reconstructions also differ strongly depending on whether an area of minimum or maximum relevance is distorted. As expected, the visual appearance does not vary much across layers when areas of minimum relevance are occluded since all images contain areas that do not impact layer representations of upper layers, i.e., they do not help discriminate

**Table 7** Accuracy when occluding least and most relevant parts of input for *ClaDec* and *GradCAM* for VGG

Layer	<i>ClaDec</i>			<i>GradCAM</i>		
	Accuracy		$\Delta Acc$	Accuracy		$\Delta Acc$
	$C(X_{O,max})$	$C(X_{O,min})$		$C(X_{O,max})$	$C(X_{O,min})$	
<i>CIFAR-100</i>						
-1	0.277±0.006	0.399±0.005	0.122	0.334±0.004	0.334±0.004	0.0
-3	0.298±0.006	0.386±0.003	0.088	0.355±0.003	0.377±0.002	0.023
-5	0.325±0.008	0.367±0.002	0.042	0.321±0.002	0.371±0.003	0.05
<i>FashionMNIST</i>						
-1	0.783±0.017	0.907±0.001	0.124	0.859±0.006	0.859±0.006	0.0
-3	0.771±0.018	0.906±0.002	0.135	0.878±0.002	0.882±0.003	0.004
-5	0.824±0.013	0.888±0.004	0.064	0.843±0.011	0.879±0.005	0.036
<i>MNIST</i>						
-1	0.772±0.028	0.993±0.0	0.221	0.914±0.006	0.914±0.006	0.0
-3	0.76±0.011	0.994±0.001	0.234	0.989±0.0	0.988±0.001	0.0
-5	0.766±0.011	0.994±0.001	0.227	0.858±0.008	0.947±0.007	0.089

**Table 8** Accuracy when occluding least and most relevant parts of input for *ClaDec* and *GradCAM* for Resnet

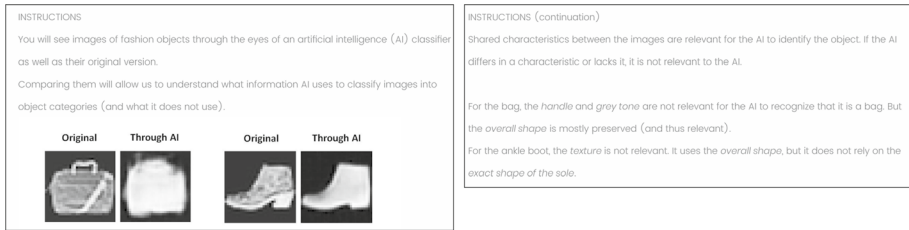
Layer	<i>ClaDec</i>			<i>GradCAM</i>		
	Accuracy		$\Delta Acc$	Accuracy		$\Delta Acc$
	$C(X_{O,max})$	$C(X_{O,min})$		$C(X_{O,max})$	$C(X_{O,min})$	
<i>CIFAR-100</i>						
-1	0.377±0.007	0.52±0.004	0.143	0.381±0.006	0.533±0.004	0.152
-3	0.458±0.005	0.462±0.005	0.004	0.398±0.003	0.502±0.003	0.105
-5	0.466±0.005	0.453±0.006	-0.013	0.441±0.002	0.456±0.005	0.015
<i>FashionMNIST</i>						
-1	0.814±0.023	0.914±0.003	0.1	0.875±0.006	0.912±0.007	0.037
-3	0.876±0.012	0.887±0.005	0.011	0.874±0.004	0.893±0.01	0.019
-5	0.878±0.006	0.885±0.008	0.006	0.888±0.004	0.889±0.004	0.001
<i>MNIST</i>						
-1	0.548±0.147	0.877±0.075	0.329	0.757±0.093	0.774±0.165	0.023
-3	0.573±0.043	0.824±0.075	0.251	0.523±0.066	0.809±0.046	0.286
-5	0.509±0.064	0.773±0.072	0.264	0.564±0.109	0.566±0.051	0.001

between classes. The object to be recognized remains mostly intact in the reconstruction. For maximum relevance, this is not the case when reconstructions from activations of upper layers are computed. The reconstructed image can appear differently because the occluded area and semantics can change. This becomes most apparent when considering the outcome of the quantitative evaluation.

*Quantitative evaluation* Tables 7 and 8 show the impact on accuracy on the classifier to explain when areas of minimum and maximum relevance are occluded based on GradCAM

and *ClaDec* for various networks and datasets. Aligned with our qualitative evaluation, for *ClaDec* we can observe that occluding areas of maximum relevance have a larger impact on accuracy than areas of lower relevance. Furthermore, the impact tends to be larger for upper layers than for lower layers for areas of maximum relevance. For ResNet (Table 8) the differences in accuracy ( $\Delta$  Acc) can be small for lower layers or even negative. This is aligned with our qualitative analysis, indicating that reconstructions are not depending strongly on the occluded area as well as the observation that the relevance measure is not so much tied to the object to be classified, e.g., the background rather than the object to be classified might yield maximum relevance. While the absolute numbers vary across networks and datasets, the general behavior tends to be very similar. Most notably, the difference in ( $\Delta$  Acc) varies less across layers for MNIST than CIFAR-100, and the impact is also largest for MNIST. MNIST is a fairly simple dataset, where only a relatively small area of the input decides on the class, e.g., adding a small line can change the class from 6 to 8 or from 3 to 9. Thus, occlusions can be chosen to have no impact or to change the class with a high likelihood. Interestingly, while Resnets perform better in general for MNIST they are more sensitive to occlusions than VGG networks, e.g., accuracy is lower for an occluded area. We attribute this to the fact that Resnets are downsampling differently. They maintain more spatial information in early layers and also up to the last dense layer and use average-pooling, whereas VGG networks downsample more aggressively already in lower layers by max-pooling. This allows occluded areas that might resemble features similar to actual classes to directly influence the output. For example a gray square indicating occlusions in a background area might appear as a 0 or parts of an 8 or 9.

*Comparing GradCAM and ClaDec* For GradCAM, only lower layers yield a significant difference in accuracy when comparing occluded areas of minimum and maximum relevance for VGG but not so for Resnet. This is a consequence of the spatial extent of the layer. If GradCAM is applied to layers with little spatial extent, relevance scores are imprecise. In the most extreme case, they are identical for all pixels, i.e., if the layer has a spatial extent of  $1 \times 1$ , the upsampling of relevance scores to the full  $32 \times 32$  image yields equal scores across pixels. As already described, Resnet preserves spatial dimensions up to higher layers than VGG. This explains the different behavior for GradCAM shown in Tables 7 and 8. For Resnet, removing maximum relevant areas yields larger changes in classifier performance except for MNIST. However, for MNIST and layer -1 we observed very large standard deviations when removing the minimum relevant areas using occlusions. Most interesting is arguably the question of which of the two explainability methods yields better outcomes. This might be judged based on whether occluding more or less relevant parts strongly impacts accuracy, i.e.,  $\Delta$  Acc. Choosing areas randomly as minimum and maximum relevant would yield a  $\Delta$  Acc of zero. Thus, larger  $\Delta$  Acc means that a method can better anticipate what is relevant and irrelevant for a classifier. Overall,  $\Delta$  Acc is larger for *ClaDec*, in particular for VGG. This is remarkable, in particular, since the relevance measure used in the computation is well-suited for GradCAM but rather poor for *ClaDec*. GradCAM is developed to output relevance score for pixels which can be aggregated to get the relevance of an (occluded) area. *ClaDec* requires a rather complex computation of relevance score using differences between reconstructions which introduces additional noise, in particular for datasets such as CIFAR-100, where reconstructions are generally not of very high quality. Noise, i.e., randomly choosing areas as minimum and maximum relevant, reduces  $\Delta$  Acc. Therefore, it is somewhat surprising that overall our evaluation indicates that *ClaDec* can better identify relevant areas than *GradCAM* that is developed for this purpose. However, we believe that both methods are of high value, particularly for datasets where reconstructions of *ClaDec* are not of very high quality.



**Fig. 17** Introduction for participants of human study

**Table 9** Participants’ demographics and prior knowledge

Question	Response distribution			
Gender	30 males	29 females		
Age	41 (age 18-24)	12 (age 35-44)	6 (age 35-44)	
Education	26 (some college)	13 (high school)	14 (4 years degree)	6 (Other)
Do you know what .				
... deep learning is?	49 (no)	10 (yes)		
... a CNN is?	54 (no)	5 Yes		

### 6.7 User study

We conducted an experiment asking humans for their judgment. Our goals were to (i) assess whether non-experts can make sense of our explanations, (ii) assess whether humans can identify a diverse set of concepts in the explanations, i.e., if they can determine many concepts abstracted by the AI such as texture, shape, color, etc., and (iii) conduct an exploratory analysis of free-text responses in a qualitative manner to highlight other interesting findings. We recruited people from the general public who were mostly unfamiliar with convolutional neural networks and deep learning. That is, they were neither aware of the concept of layers nor of the workings of an autoencoder. We decided to focus on a simple scenario where the reconstructions from *RefAE* are similar to the original image. We neglected differences that could be attributed due to distortion of the decoder since they were mostly relatively small compared to differences between the reconstructions from *ClaDec* and we did not want to overload inexperienced participants in a single experiment.


We used the second-to-last layer of the VGG-11 network and 100 samples of the Fashion-MNIST dataset. That is, a sample consists of the original image as well as the reconstruction from *ClaDec*. We recruited 59 participants through the platform “Prolific”. Their demographics and prior knowledge are summarized in Table 9.

Participants were given a short general introduction on AI and image recognition and a more detailed one for the task shown in Fig. 17. Participants were asked to compare the original image and the one from *ClaDec*, i.e., the image as seen through an AI classifier as shown in the example in Figure 18. Each participant analyzed five randomly selected image pairs qualitatively using textual analysis so that each pair was analyzed equivalently often.

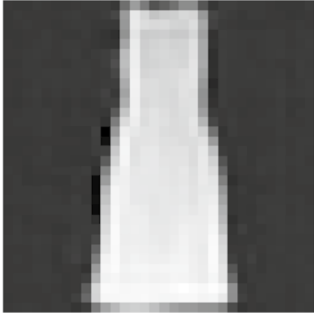
*Analysis and findings* We obtained 295 free-text answers (5 from each of the 59 participants), e.g., most images had three responses and a few just two. We did not exclude any answers. To understand whether participants reached valid conclusions, we checked

Compare the original image (left) with the one seen through an AI classifier (right).  
What is different or lacking for the AI? What does it maintain?

**Original**



**Through AI**



Your answer (10–100 words):

**Fig. 18** Question presented to participant

for erroneous judgments by participants. We investigated whether participants claimed images were (almost) identical although there were clear differences. We found that about 1%, i.e., three responses, claimed to see no differences though we (and other participants) could identify clear differences. We found that about 1% of responses contained objectively false claims, e.g., they claimed that the original and reconstructed objects differed (significantly) in size though they did not. Many responses contained imprecise language, i.e., respondents used generic terms like “design differs”, which could imply differences in shape, texture, color, etc. Furthermore, terms like texture, patterns and characteristics were often used interchangeably. Overall, 2% of responses are incorrect, i.e., they contradict responses of other participants and the ground truth (defined by the author team). The remaining 98% of responses are deemed valid, i.e., they describe well-visible differences. However, some of the responses contain linguistic imprecision, and most responses only contain a subset of all visible differences.

We identified generic concepts, i.e., comparison criteria that can be applied to multiple classes through a qualitative data analysis method, i.e., “open coding” (Corbin and Strauss 1990). We found that responses used as criteria for comparison: “outer” or “overall” shape (66% of replies contained this concept), a feature or specific characteristic (40% of replies), color or grey tone (29%), pattern or texture (24%), logo or brand name (15%), shading (10%), design (8%), sharpness or blurriness (5%), perspective/orientation of object (3%), artifacts introduced by AI (1%), contrast (of images) (1%),

image depth (1%), symmetry/skewness (1%). Overall, the responses of users show that our explainability process of comparing the original image and the reconstructed image by *ClaDec* allows even non-experts to identify a rich set of concepts. Note that some concepts were rare, e.g., only a few explanations contained obvious artifacts introduced by the reconstruction or a change in perspective. Also, our introduction mentioned the concepts “overall shape”, “texture” and “grey tone”, which might have primed users to use these more often.

Most concepts were used primarily to indicate differences except for (overall) shape, which was mostly mentioned to be preserved (more than 90% of responses). Color and pattern/textures were mentioned commonly in contexts where they differed and where they were identical. Thus, the comparison seems to be biased towards differences. We also found that about 15% of answers only focused on differences though we explicitly asked also what the AI maintains. Emphasizing differences rather than commonalities is not unexpected since our visual system is primed to focus attention on movement and change, i.e., differences (Franconeri and Simons 2003).

Other interesting findings were that a few responses (about 2%) stated that the reconstructions tend to look like another class due to the absence of specific features or other changes. For example, one response was “AI preserves the shape, but erases the patterns and also the zipper, so this image is no longer a sweater but a shirt”. Furthermore, while most replies tended to indicate that there were either little differences between the original and the reconstruction, or the reconstruction lost some information of the original, some replies also indicated the opposite, i.e., that the reconstructions are “better” or easier to recognize than the original. In the most extreme case, participants could recognize the object reconstructed by the AI, but not the original (1% of replies), e.g., one reply was “I can’t tell what the original image shows, but through AI I can easily tell that it’s a dress”. This is not unexpected since the AI tends to reconstruct prototypical images replacing uncommon features with more common ones. When it comes to responses, aligned with our perception, users mostly mentioned reconstructions to be more blurry but interestingly, a few responses also indicated the opposite (1%), i.e., for two images where the pattern contained many dark patches making the contour more difficult to recognize.

In summary, our user study shows that non-experts can analyze our explanations and identify a rich set of concepts within images that are relevant (or irrelevant) to the classification process. To interpret the explanations, another small step is needed: If the concepts in the input differ from concepts in the *ClaDec* reconstruction (but not from those of the RefAE), the input concept differs significantly from those encoded by the classifier, i.e., the sample differs significantly from a prototypical instance and details of the input concept that could not be reconstructed are not relevant to distinguish between classes.

If the input shape differs from the shape of the *ClaDec* reconstruction (but not from the shape of the RefAE), this means that the classifier does not need to be able to reconstruct it well to classify the object. The sample differs significantly from a prototypical instance.

## 7 Discussion and future work

To create explanations, *ClaDec* must be trained on a dataset that should be similar to the training dataset though no labels are needed. This is a disadvantage compared to other methods such as GradCAM that can be used in situ. If the dataset is too small, the reference AE and *ClaDec* both cannot produce high quality reconstructions limiting

the value of our method. This is often manifested in poor reconstructions of the *RefAE*, which correlates strongly with difficult to comprehend reconstructions  $\hat{X}_E$  by the *ClaDec*. In our case, the CIFAR-100 dataset is the most complex and the qualitative analysis suggests that only coarse differences between reconstructions from *RefAE* and *ClaDec* such as shapes and color tones are easy to identify by humans. The CIFAR-100 classifiers also perform poorest among all datasets. Thus, it can be concluded that explanations are better for well-performing classifiers that are typically also based on a large dataset. We might use similar tricks used for classifiers to improve our explainability method, i.e. the decoders. For example, data augmentation techniques could be used to enhance a dataset. That is, even if the classifier is trained on non-augmented data, we might train *RefAE* and *ClaDec* using data augmentation. Furthermore, pre-trained AE architectures could be leveraged using transfer learning. Our evaluation supports the idea that reconstruction-based techniques should employ a reference since generative models tend to introduce some distortion and artifacts. Non-reconstruction based explanation methods do not require comparing to a reference. However, for a sufficiently large dataset and an adequate architecture that yields reconstructions of the *RefAE* that are not distinguishable from the original, this step is also not necessary. For the MNIST dataset, which is relatively large compared to the number of classes and the complexity of samples, reconstructions are very good for both *RefAE* and they are essentially only needed to highlight differences of the topmost layer, where the number of dimensions is small, and thus distortion are largest. On the contrary, for a more complex dataset with many more classes but the same number of samples like CIFAR-100, the comparison to the *RefAE* is helpful. While our method still provides many interesting insights, interpretation is more tricky due to distortions in reconstruction. However, other methods like GradCAM are of little help in such cases, i.e., when the difference between the spatial extent of the feature map and the reconstructed input is very large. This is also a finding of our occlusion-based analysis. Overall, we believe that reconstruction-based methods and saliency-based methods are complementary.

Our user study confirmed that users could easily identify differences in actual input images and reconstructions by *ClaDec*. It might also be interesting to perform a quantitative study, e.g., asking users explicitly if they notice differences with respect to a specific concept. Furthermore, our user study focused on the case where *RefAE* and original inputs show little differences. This limitation could also be remedied with another user study.

Our reconstruction method is limited to explaining actual classifier behavior, but cannot as easily be used for contrastive explanations to answer questions such as “Why is  $X$  not classified as class  $Y'$ ?” or “How to change  $X$  to get class  $Y'$ ?” as other methods. One way to determine (small) changes to layer activations  $L(X)$  for a sample  $X$  to get class  $Y'$ , is to find a sample  $X'$  classified as  $Y'$  so that  $(L(X) - L(X'))^2$  is small and to linearly interpolate between them. Such ideas have been pursued in van Doorenmalen and Menkovski (2020), Guidotti et al. (2019). Related to the prior limitation, our method lacks a relevance score, meaning that it is unclear how important one (reconstructed) part of the input is. Techniques based on occluding parts of the input might be used to determine which parts of the reconstruction are crucial (Zeiler and Fergus 2014). Furthermore, we have focused on one network type and data type, i.e., CNNs and images. While images are very illustrative, future work might touch the validity of the approach on other datasets.

Our work also touches on fundamental questions in deep learning, i.e., the information bottleneck (Geiger 2021; Saxe et al. 2019). Our work contributes to this discussion since it indicates that information is discarded from layer to layer because reconstructions get

poorer for upper layers as shown qualitatively and quantitatively. Still, more work is needed in this direction.

Another aspect is computation time for an explanation. Our method incurs computational setup costs due to the initial training of *ClaDec* and *RefAE* as well as for computing each explanation. Roughly speaking, the costs for explaining are about the same as for model training. Some methods like GradCAM do not require any setup costs, while methods like LIME require training of many approximate models, which can incur much higher computational costs.

## 8 Conclusions

Our explanation method synthesizes human understandable inputs based on layer activations or subsets thereof. It takes into account distortions originating from the reconstruction process. Rather than pinpointing to individual neurons or parts of an input, we are interested in understanding what information of the input can be reconstructed. We believe that our method might form the basis for many more methods that further expand and contribute to the field of explainability.

**Acknowledgements** We thank Jeroen van Doorenmalen for valuable discussions.  
None

**Author Contributions** MV: Proof reading, GradCAM visualizations JS: all the rest

**Funding** Open access funding provided by University of Liechtenstein. Not Applicable

**Availability of data and material** Code is at <https://github.com/JohnTailor/ClaDec>. All data used is public.

## Declarations

**Competing interests** The authors declare that they have no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160.
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31.
- Agarwal, C., & Nguyen, A. (2020). Explaining image classifiers by removing input features using generative models. In: *Proceedings of the Asian Conference on Computer Vision*.



- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10(7), e0130140.
- Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1), 53–58.
- Barbalau, A., Cosma, A., Ionescu, R. T., & Popescu, M. (2020). A generic and model-agnostic exemplar synthetization framework for explainable ai. In: *ECML-PKDD*.
- Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32.
- Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11(1), e1391.
- Confalonieri, R., Weyde, T., Besold, T. R., & Moscoso del Prado Martín, F. (2020). Trepan reloaded: A knowledge-driven approach to explaining black-box models. In: *ECAI 2020*
- Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1), 3–21.
- Deco, G., Finnoff, W., & Zimmermann, H. (1993). Elimination of overtraining by a mutual information network. In: *Int. Conference on artificial neural networks*.
- van Doorenmalen, J., & Menkovski, V. (2020). Evaluation of cnn performance in semantically relevant latent spaces. In: *Int. Symposium on intelligent data analysis*.
- Du, B., Xiong, W., Wu, J., Zhang, L., Zhang, L., & Tao, D. (2016). Stacked convolutional denoising auto-encoders for feature representation. *IEEE Transactions on Cybernetics*, 47(4), 1017–1027.
- Fong, R., & Vedaldi, A. (2018). Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8730–8738.
- Franconeri, S. L., & Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception and Psychophysics*, 65(7), 999–1010.
- Geiger, B.C. (2021). On information plane analyses of neural network classifiers—a review. *IEEE Transactions on Neural Networks and Learning Systems*.
- Ghorbani, A., Abid, A., Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 3681–3688.
- Ghorbani, A., Wexler, J., Zou, J. Y., Kim, B. (2019). Towards automatic concept-based explanations. In: *Advances in Neural Information Processing Systems*.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Guidotti, R., Monreale, A., Matwin, S., Pedreschi, D. (2019). Black box explanation by learning image exemplars in the latent feature space. In: *Joint European conference on machine learning and knowledge discovery in databases*, pp. 189–205.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: *International conference on machine learning*.
- Kindermans, P. J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (un) reliability of saliency methods. In: *Explainable AI: Interpreting, explaining and visualizing deep learning*.
- Koh, P.W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In: *Proceedings of International conference on machine learning*.
- Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: *Proceedings of the AAAI conference on artificial intelligence*.
- Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R. J., Camps, O. (2020). Towards visually explaining variational autoencoders. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in Neural Information Processing Systems*, 29.
- Petsiuk, V., Das, A., Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. arXiv preprint [arXiv:1806.07421](https://arxiv.org/abs/1806.07421).
- Qi, F., Lin, C., Shi, G., & Li, H. (2019). A convolutional encoder-decoder network with skip connections for saliency prediction. *IEEE Access*, 7, 60428–60438.
- Rafegas, I., Vanrell, M., Alexandre, L. A., & Arias, G. (2020). Understanding trained cnns by indexing neuron selectivity. *Pattern Recognition Letters*, 136, 318–325.

- Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5), 206–215.
- Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., & Cox, D. D. (2019). On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 124020.
- Schneider, J. (2020). Human-to-ai coach: Improving human inputs to ai systems. In: *International Symposium on Intelligent Data Analysis*.
- Schneider, J., & Handali, J. P. (2019). Personalized explanation for machine learning: A conceptualization. In: *European Conference on Information Systems (ECIS)*.
- Schneider, J., Meske, C., & Vlachos, M. (2022). Deceptive AI explanations: Creation and detection. In: *International Conference on Agents and Artificial Intelligence (ICAART)*.
- Schneider, J., & Vlachos, M. (2013). Fast parameterless density-based clustering via random projections. In: *Proc. of the international conference on Information & Knowledge Management (CIKM)*.
- Schneider, J., & Vlachos, M. (2014). On randomly projected hierarchical clustering with guarantees. In: *Proceedings of the SIAM International Conference on Data Mining*, pp. 407–415.
- Schneider, J., & Vlachos, M. (2020). Reflective-net: Learning from explanations. In: [arxiv: 2011.13986](https://arxiv.org/abs/2011.13986).
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626.
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In: *Int. Conf. on Machine Learning*.
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In: *In Workshop at International Conference on Learning Representations*.
- Sun, K., Zhang, J., Zhang, C., & Hu, J. (2017). Generalized extreme learning machine autoencoder and a new deep neural network. *Neurocomputing*, 230, 374–381.
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec), 3371–3408.
- Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M. R., & Tai, Y. W. (2020). Towards global explanations of convolutional neural networks with concept attribution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661.
- Yang, F., Du, M., & Hu, X. (2019). Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint [arXiv:1907.06831](https://arxiv.org/abs/1907.06831).
- Yeh, C. K., Hsieh, C. Y., Suggala, A., Inouye, D. I., & Ravikumar, P. K. (2019). On the (in) fidelity and sensitivity of explanations. In: *Advances in Neural Information Processing Systems*, pp. 10965–10976.
- Yosinski, J., Clune, J., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. In: *In ICML Workshop on Deep Learning*.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In: *European conference on computer vision*.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.