Information Fusion 108 (2024) 102407

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

ExplainLFS: Explaining neural architectures for similarity learning from local perturbations in the latent feature space

Marilyn Bello [a,*], Pablo Costa [a], Gonzalo Nápoles [b], Pablo Mesejo [a], Óscar Cordón [a]

[a] Department of Computer Science and Artificial Intelligence (DECSAI) and Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Spain

[b] Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands

ARTICLE INFO

Keywords:
Similarity learning networks
Face recognition
Image retrieval
Latent feature space
Visual similarity explanations

ABSTRACT

Despite the increasing development in recent years of explainability techniques for deep neural networks, only some are dedicated to explaining the decisions made by neural networks for similarity learning. While existing approaches can explain classification models, their adaptation to generate visual similarity explanations is not trivial. Neural architectures devoted to this task learn an embedding that maps similar examples to nearby vectors and non-similar examples to distant vectors in the feature space. In this paper, we propose a *post-hoc* agnostic technique that explains the inference of such architectures on a pair of images. The proposed method establishes a relation between the most important features of the abstract feature space and the input feature space (pixels) of an image. For this purpose, we employ a relevance assignment and a perturbation process based on the most influential latent features in the inference. Then, a reconstruction process of the images of the pair is carried out from the perturbed embedding vectors. This process relates the latent features to the original input features. The results indicate that our method produces "continuous" and "selective" explanations. A sharp drop in the value of the function (summarized by a low value of the area under the curve) indicates its superiority over other explainability approaches when identifying features relevant to similarity learning. In addition, we demonstrate that our technique is agnostic to the specific type of similarity model, e.g., we show its applicability in two similarity learning tasks: face recognition and image retrieval.

1. Introduction

Similarity learning is a task within computer vision that has gained significant interest due to its multiple applications, such as face recognition [1,2], person re-identification [3,4], image retrieval [5,6], and one-shot learning [7,8]. This task addresses the question of whether a pair of images is similar or to what degree they exhibit similarity. The obtained result is subsequently applied to address a specific problem. For instance, in the context of face recognition for identity validation, the image of an individual seeking access is compared with all stored images of persons in the access database, leading to either confirmation or denial. In the scenario of image retrieval, when presented with a query image of an object, the goal is to locate, from a collection of reference images, the object image that bears the highest resemblance to the query image.

Measuring similarity requires learning an embedding space that captures images and reasonably reflects similarities using a defined distance metric. In this regard, Siamese neural networks (SNNs) [9,10] emerge as a powerful approach to similarity learning. These networks have parallel neural networks with identical weights and parameter settings. Each network has a different input (image), and their outputs are combined to obtain a prediction. Therefore, SNNs learn embedding vectors used to compare sub-network inputs. The cost function determines the distance between the embedding vectors obtained by passing images through the system. To make these systems more robust, several variants of these architectures have been proposed, e.g., Facebook's DeepFace [11], Google's Facenet [1], and ArcFace [12], among others [13,14].

Although these computer vision models are quite powerful, they are considered black-box models, i.e., their internal reasoning and decision-making processes are not easily interpretable by humans. To overcome this issue, several explanation methods have been proposed [15–17]. The rationale of these methods consists of explaining why neural systems and other black-box AI models make their decisions. These methods include model-agnostic approaches [18–20], i.e., approaches that do not depend on the neural structure to be explained, only on its input and output. There are also example-based methods, such as

* Corresponding author.
E-mail address: mbgarcia@ugr.es (M. Bello).

https://doi.org/10.1016/j.inffus.2024.102407
Received 1 December 2023; Received in revised form 2 April 2024; Accepted 3 April 2024
Available online 5 April 2024

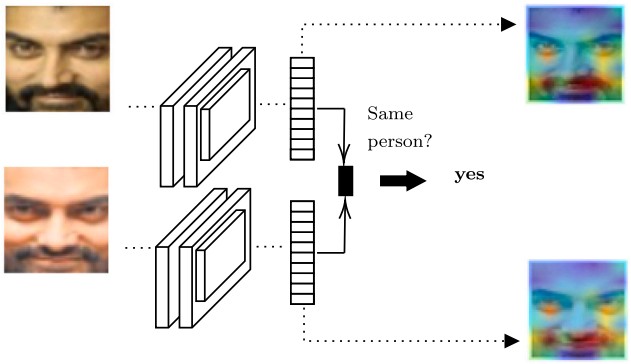


Fig. 1. Face recognition: the method explains why a neural model infers the similarity between pairs of images.

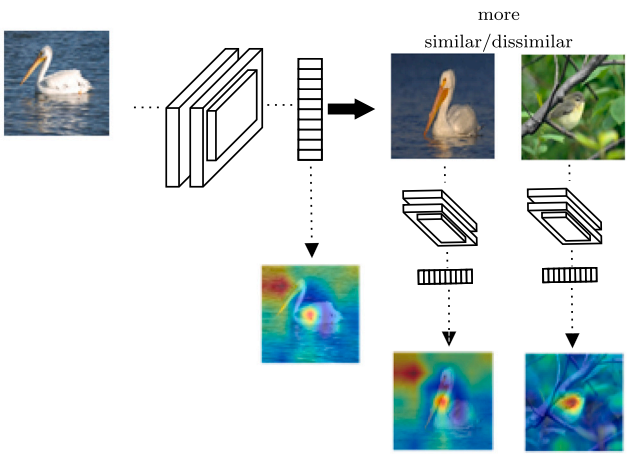


Fig. 2. Image retrieval: the method determines the features needed to retrieve the most similar and dissimilar image for a query image.

those based on data prototypes [21], counterfactual explanations [22], and adversarial examples [23]. Among other proposals, we found those in [24,25] that focus on visualizing the contribution of each pixel of an input image in the prediction using attention maps. These approaches make it possible to explain the predictions of classification networks. However, few works [26–31] have focused on explaining the inference process of similarity networks. These networks learn an embedding that maps similar images into nearby vectors of the abstract feature space and dissimilar images into distant vectors. Consequently, there is no inverse mapping between the embedding vectors and the corresponding input images, i.e., we do not know the subsets of features in the input space that correspond to some features of the embedding vector.

This paper introduces a new method, named ExplainLFS (*Explaining Neural Models from Latent Feature Space*), to explain the inference process of neural architectures for similarity learning between images. To do that, we rely on the embedding vectors obtained by the neural architecture for a pair of images. Each latent feature in the embedding vector is given a weight representing the relevance of that feature in the output predicted by the neural model. If the images are similar, the most similar features of the embedding vectors will have a higher relevance. If the images differ, the most dissimilar features will have a higher relevance. Subsequently, a process of perturbation and decoding is performed to establish a relationship between the latent features of the embedding vectors and the input images. In other words, the method can explain the features of the input images that most influence the final decision from the latent feature space of a pair of images.

The computer vision tasks targeted in our paper concern face recognition [1] and image retrieval [5] tasks. Figs. 1 and 2 illustrate the visual explanations generated by ExplainLFS for these computer vision tasks. Therefore, the main contribution of our paper is that, to the best of our knowledge, the most successful explanation methods cannot directly be applied to neural architectures used for similarity learning between images. The reason is that the input space for these computer vision tasks is composed of two images instead of a single one, as happens in image classification problems.

To measure the quality of generated explanations, we use the criteria presented in [29,32–35]. The authors assess quality based on two properties: selectivity, quantified by measuring how fast the model output changes when perturbing the features with the highest relevance scores, and continuity, quantified by measuring the variation of the explanations in an input domain. Furthermore, the explanations are contrasted with the explanations generated by three approaches adopted from the literature [26,28,31]. The former is a neural network that directly visualizes the similarities between a pair of images. It should be noted that this network does not infer whether one image is similar to another, nor does it infer the similarity score between two images. However, it produces similarity maps between similar images that can be used for comparison purposes. The latter are approaches explicitly designed to explain SNNs.

The structure of this paper is as follows. Section 2 reviews the state-of-the-art related to the explainability of neural models for similarity learning. Section 3 presents the proposed explanation method. Section 4 analyzes the quality of the explanations proposed by our method for facial recognition and image retrieval tasks. Section 5 concludes the paper.

2. Related work

Most of the existing visual explanation techniques in the literature focus on explaining the inference process of neural models intended for image classification tasks [18,19,24,25]. These techniques explain why an object is detected in an image. This is typically done by mapping the regions of interest in an image leading to a given class. However, when dealing with neural models for similarity learning tasks, it is necessary to explain what makes image *A* similar to image *B* but dissimilar to image *C*. Although the literature to explain this specific type of neural learning is relatively sparse, recent techniques to explain neural models for similarity learning between images have been proposed [3,26–28,30,31,36–38]. Other approaches have also been proposed to explain similarity learning networks, but aimed at other types of data, such as audio [29] or graphs [39].

The approaches in [3,26,27] are self-explaining neural networks able to learn visual explanations from their inference process. While the proposals in [3,26] apply only in pairwise similarity learning, the approach in [27] also generates attention visualizations by similarity in triplet and quadruplet neural models through gradient-based attention maps. Stylianou et al. [26] proposed a method to highlight the regions of the images that contribute most to pairwise similarity. In addition, Tummala et al. [31] proposed a Siamese twin network with a novel visual saliency mapping scheme to interpret their decisions.

Alternatively, in [36], an explanation method is proposed based on textual information of the most salient attribute in image matching. In contrast to previous approaches, Plummer et al. presented an explanation method for image matching independent of the neural model learning process. Their explanations are based on the feature that best explains the match between attention maps generated from the embedding vectors of the pair images. The authors in [30,37] adapted the Grad-CAM method [25] for embedding networks. Their approach does not try to explain the similarities in a pair of images, only to visualize their latent features. In fact, the approach proposed by [30] is only useful when explaining the embeddings separately,

without considering the shared features used in the inference process. However, their visualization can help domain experts to deduce the reason for their similarities.

Utkin et al. [28] proposed an interesting explanation method explicitly tailored for SNNs. While effective in elucidating why an image belongs to a specific category, the main drawback lies in its inability to explain the distinctions between two pairs of images. The weakness of constructing a prototype based on existing image categories becomes apparent, as it solely addresses the individual explanation of an image's category, neglecting consideration of paired features between the embedding vectors of image pairs. Also, this explanation method requires a set of images belonging to the same category (e.g., of the same person) or at least one image in the dataset belonging to the category of the image pair. This limitation restricts the method's utility in scenarios like face recognition [8], where systems typically rely on a single reference image for comparison, such as a user ID. A similar proposal to Utkin et al. is presented by Ye et al. in [38], but suffers from the same limitation in face recognition scenarios. Both, are useful for datasets such as MNIST [40]. In addition, it has a high computational cost.

Many of these approaches derive their explanations from the category to which the images belong rather than identifying latent features in the embedding space that contribute to their similarity or dissimilarity. Some approaches may be more practical, as they necessitate additional information from the input space. However, the majority of these techniques do not qualify as *post-hoc* agnostic explanation methods, which are independent of the neural model's internal structure. Our proposal aims to address these gaps in the existing literature.

3. Explanation of neural models for similarity learning

The agnostic post-hoc method proposed in this section is devoted to generating explanations for similarity learning networks. Therefore, given two images to be explained, it outputs a heatmap visualizing the most important features in the latent feature space by which an image is similar to another. The method comprises three main steps: (1) the construction of a decoder from the latent feature space of each image, (2) the relevance assignment to each latent feature according to its importance in distinguishing between similarity and dissimilarity, and (3) the mapping process between the most important latent features and the input features of each image. These steps are detailed in the following sub-sections.

3.1. Decoding process

The decoding associates the latent features (i.e., embedding vectors obtained by the neural model in the feature extraction process) with the original input image. The intuition behind this step is to learn a decoding network that can reconstruct an image from its corresponding embedding vector. This decoding process acts as an inverse mapping that converts the embedding vectors into a pixel representation of the original image. This structure has the following distinctive characteristics:

- The input layer has as many input neurons as latent features in the embedding vector.
- The output layer has as many output neurons as input features (pixels) in the original image.
- The loss function measures the differences between the original image and the reconstruction obtained. Eq. (1) defines this function,

$$\mathcal{L} = \sum_{i=1}^{M} \|x_i - \tilde{x}_i\|_2^2 \tag{1}$$

where M is the number of features of the original image, and x_i and $\tilde{x}_i$ are the ith pixels that compose the original and reconstructed images, respectively.

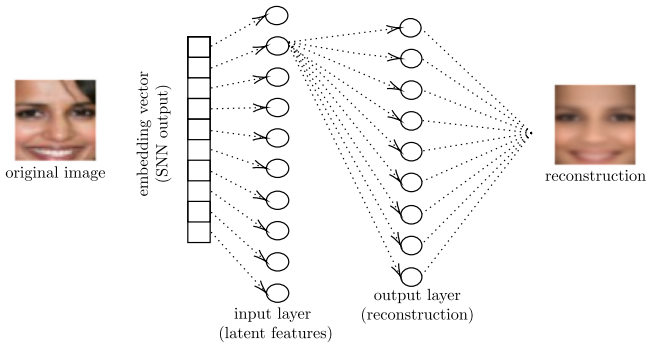


Fig. 3. Example of a decoding neural network without hidden layers. The input layer represents the embedding vector. The output layer represents the reconstruction neurons of the input image. It should be highlighted that this representation only uses 10 latent features in the embedding vector and 9 units in the output layer for visualization purposes.

Fig. 3 illustrates the proposed decoding process to reconstruct an image from its embedding vector when the neural model is an SNN such as FaceNet [1].

3.2. Relevance assignment process

This step identifies the latent features of a pair of images by which a neural model discriminates between similar or dissimilar images. By doing so, we calculate the similarity between the latent feature that composes the embedding vectors of two images and thus deduce their influence on the neural model decision. That is, the most similar features of the embedding vectors of two images should also be those that contribute the most weight to infer the similarity between the images. Contrariwise, the most different features should be the most relevant to conclude that the images are dissimilar.

Eq. (2) shows how to measure the relevance of the jth latent feature of an embedding vector,

$$\mathcal{R}_j = \begin{cases} 1 - \bar{d}_j, & \text{if images are similar} \\ \bar{d}_j, & \text{if images are dissimilar} \end{cases} \tag{2}$$

such that, $\bar{d}_j$ (defined in Eq. (3)) represents the normalized distance value between the jth feature of the embedding vectors h^1 and h^2 of a pair of images, respectively,

$$\bar{d}_j = \frac{|h_j^1 - h_j^2| - D^-}{D^+ - D^-} \tag{3}$$

where h_j^1 and h_j^2 are the jth values of the h^1 and h^2 embedding vectors, and D^+ and D^- are the maximum and minimum distance values between h^1 and h^2, respectively.

Next, we will develop a toy example to illustrate the inner working of our method and the accompanying equations:

- Suppose we have the following embeddings obtained for two similar images, $h^1 = [1.2, 1.9, 2.3, 3.1, 2.0, 2.1, 2.2]$ and $h^2 = [1.1, 3.0, 2.5, 4.0, 1.3, 1.4, 1.9]$, whose associated distance vector is $d = [0.1, 1.1, 0.2, 0.9, 0.7, 0.7, 0.3]$. Then, according to Eq. (3), the normalized distance vector is $\bar{d} = [0, 1, 0.1, 0.8, 0.6, 0.6, 0.2]$, and consequently, the relevance vector computed according to Eq. (2) is $\mathcal{R} = [1, 0, 0.9, 0.2, 0.4, 0.4, 0.8]$.
- Suppose we have the following embeddings obtained for two dissimilar images, $h^1 = [1.2, 1.9, 2.3, 3.1, 2.0, 2.1, 2.2]$ and $h^2 = [2.0, 1.5, 2.7, 0.3, 2.5, 1.6, 2.5]$, whose associated distance vector is $d = [0.8, 0.4, 0.4, 2.8, 0.5, 0.5, 0.3]$. Then, according to Eq. (3), the normalized distance vector is $\bar{d} = [0.2, 0.04, 0.04, 1, 0.08, 0.08, 0]$, and consequently, the relevance vector computed according to Eq. (2) is $\mathcal{R} = [0.2, 0.04, 0.04, 1, 0.08, 0.08, 0]$.

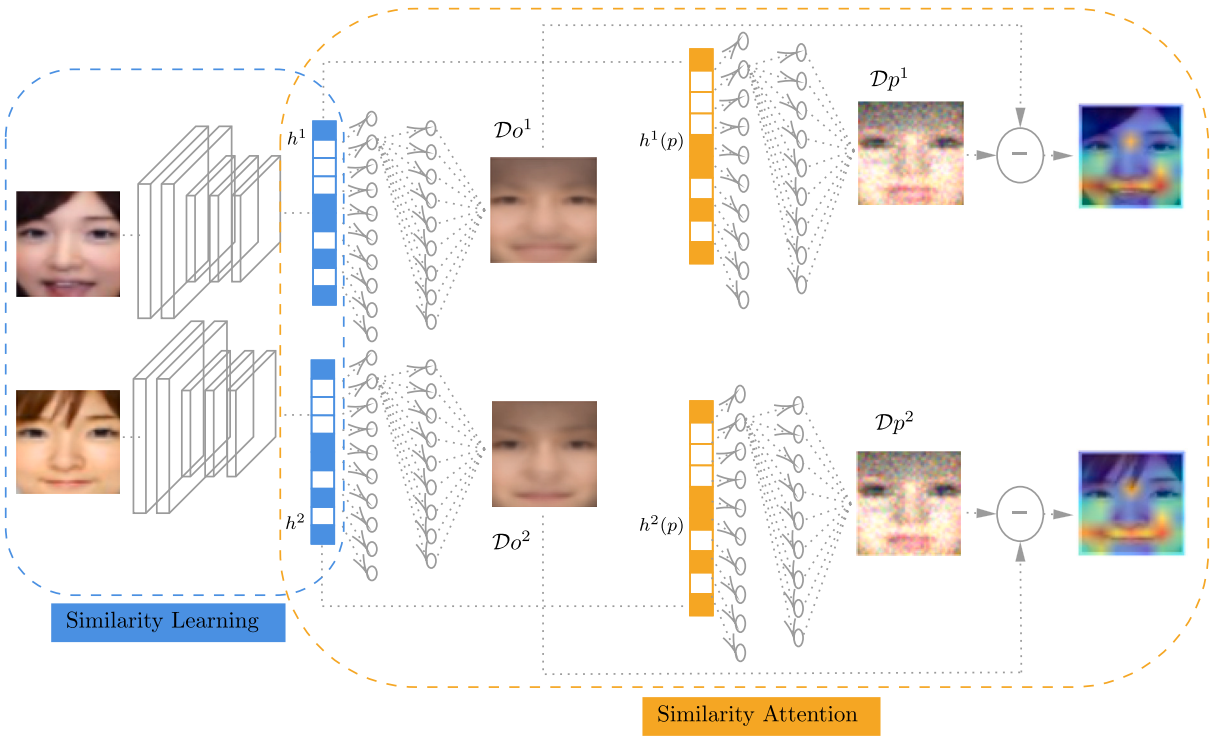


Fig. 4. The basic scheme of the ExplainLFS method for an SNN architecture that compares two images. The blue cells in the embedding vectors represent the most similar features between the embedding vectors, and the orange cells represent the most perturbed features according to the assigned relevance. In red (or orange color) are the pixels that, according to our approach, most influence the decision that these two images are "similar". These pixels result from the main differences between the reconstructions of the SNN output embeddings (Do) and the reconstructions of the perturbed embeddings (Dp).

The explanation underlying Eq. (2) lies in the assumption that, in the case of dissimilar images, the most important features are those that are more distant from each other. In the case of similar images, the most relevant features are those that are most similar to each other. Fig. 9 in the Experiments section demonstrates this assumption.

3.3. Mapping process

This process aims to infer which original image features are encoded in the most important latent features of the embedding vectors. For this purpose, we perform a perturbation process of the feature values of the embedding vector according to their relevance. The pixel areas of the reconstructed image that undergo the most significant change have the closest correspondence to the most perturbed features, i.e., more relevant latent features, since the higher the relevance of the feature, the greater its perturbation. The perturbation process is based on the relevance assigned to each latent feature. Thus, when the neural model infers that the images are similar, the more similar features are perturbed to a greater extent. Otherwise, the more dissimilar features undergo a more significant perturbation. Eq. (4) defines how to perturb the jth feature of an embedding vector from its relevance value $\mathcal{R}_j$,

$$h_j(p) = h_j + \mathcal{R}_j * p,\ j = 1, 2, \ldots, N \tag{4}$$

where N is the length of the embedding vector. The p parameter takes values in the $[0, 1]$ interval and is the same for perturbing all h_j values of the latent features that compose the embedding vectors. Its value cannot be too small (close to zero) because it vanishes the influence of the relevance ($\mathcal{R}_j$). Also, it cannot be too high (close to one) because it can cause the values $h_j(p)$ to surpass the upper boundary of the $[0, 1]$ interval. However, the variability of the p value will not affect the order of importance of the h_j value in the embedding vector and, consequently, will not affect the attention maps generated from them.

As a subsequent step in our method, the mapping process between the perturbed latent features and the original features of the image is performed as follows:

(a) A reconstruction of the perturbed embedding vector (Dp) and the original image embedding vector (Do) is obtained from the decoding neural network.
(b) The difference between the pixel values of Do and Dp is calculated. Thus, the pixels in which the most significant differences are evident (i.e., pixels where the difference exceeds a threshold ξ_2) correspond to the most perturbed features and, consequently, to the most relevant features. The threshold ξ_2 is conceived as a user-adjustable parameter. Although any value in the interval $[0, 1]$ is possible, it is recommended that this threshold have a value higher than 0.7. In this way, only pixels with a high relevance value in the inference process are displayed.
(c) An attention map (also called in the literature *saliency* map [41, 42]) that highlights the most relevant pixels obtained in Step 2 is built. Saliency maps provide an intuitive visual interpretation by highlighting the regions of an image that most influence the decision of a neural model.

Fig. 4 illustrates an example of the proposed explanation process for an SNN that compares two images as "similar". The results are based on the output of the FaceNet architecture [1] for a pair of images whose embedding vectors have a similarity of 0.78. In addition, Algorithm 1 presents a pseudocode summarizing the technical details of our method.

4. Experiments

This section evaluates the proposed explanation method in two similarity learning tasks: face recognition [1] and image retrieval [5]. The road map of our experimental methodology is as follows. In the first experiment, we show the effectiveness of the ExplainLFS method for both computer vision tasks using representative examples. Secondly, we assess the quality of the explanations generated by our method w.r.t. two desired explanation properties: *selectivity* and *continuity*. Finally, we compare the quality of explanations produced by our method against

Algorithm 1 ExplainLFS

Require: Trained Decoding Neural Network (TDNN), Trained SNN model (TSNN), Image pair (I^1, I^2), and Hyperparameters $\{\xi_1, \xi_2, p\}$.
Ensure: Similarities Attention Maps.

STEP 1: {Generate Embeddings and Measure Similarity.}
$similarity,\ h^1,\ h^2 \leftarrow \text{TSNN}(I^1,\ I^2)$

STEP 2: {Determine Relevance of Latent Features.}
for j from 1 to N **do**
  $\bar{d}_j \leftarrow normalized\_distance(h_j^1, h_j^2)$
  if $similarity > \xi_1$ **then**
    $\mathcal{R}_j \leftarrow 1 - \bar{d}_j$
  else
    $\mathcal{R}_j \leftarrow \bar{d}_j$
  end if
end for

STEP 3: {Mapping Latent Features to Input Space.}
for j from 1 to N **do**
  $h_j^1(p) \leftarrow h_j^1 + \mathcal{R}_j * p$
  $h_j^2(p) \leftarrow h_j^2 + \mathcal{R}_j * p$
end for
$\mathcal{D}o^1,\ \mathcal{D}o^2 \leftarrow \text{TDNN}\ (h^1,\ h^2)$
$\mathcal{D}p^1,\ \mathcal{D}p^2 \leftarrow \text{TDNN}\ (h^1(p),\ h^2(p))$
$\mathcal{H}^1 \leftarrow difference(\mathcal{D}o^1, \mathcal{D}p^1) > \xi_2$
$\mathcal{H}^2 \leftarrow difference(\mathcal{D}o^2, \mathcal{D}p^2) > \xi_2$
return $\mathcal{H}^1,\ \mathcal{H}^2$

other state-of-the-art approaches and empirically demonstrate the superiority of our proposal. To do that, we will resort to the "pixel-flipping experiment", which is perhaps the most widely used for assessing the quality of explanation methods that generate feature importance scores [29,32–35,43].

4.1. Face recognition

To illustrate the effectiveness of ExplainLFS for this task, we relied on FaceNet [1], one of the most widely used SNN architectures for face recognition, with an accuracy of 99% on the VGGFace2 dataset [44]. During our experimentation, we used the FaceNet neural model (see https://github.com/davidsandberg/facenet for training details), which is pre-trained to generate embedding vectors of length equal to 512 over three million images of 9,131 people at different angles, different times (i.e., in terms of age), various accessories, hairstyles, and background variability.

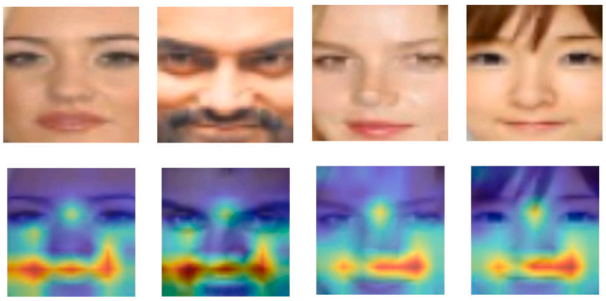
It should be noted that training the decoder is entirely independent of the SNN training. We take 24,000 images from the CelebFaces Attributes dataset [45] to train the decoder. The image size in this dataset is $40 \times 40 \times 3$ (size of the decoder output layer). The FaceNet architecture requires a re-scaling process of the image size (i.e., $160 \times 160 \times 3$) to generate the embedding vectors (i.e., the input layer of the decoder). The decoding error measured according to Eq. (1) is 0.02.

In the experiments, we arbitrarily used $p = 0.3$ and $\xi_2 = 0.7$. However, our proposal is not sensitive to these hyperparameters, due to their inherent nature. The first is a value applied equally to all features of an embedding vector. Therefore, any variation of this parameter will result in embeddings that have a linear relationship between them. The second is a threshold for displaying purposes, conceived as a user-adjustable parameter during the explainability phase. In addition, FaceNet infers as "similar" images those image pairs whose similarity value is greater than $\xi_1 = 0.3$, and as "dissimilar" images, the opposite.

Figs. 5 and 6 depict the outcomes of ExplainLFS applied to the inferences generated by FaceNet for four pairs of similar and dissimilar images. In these figures, the first row corresponds to the pair of images subjected to the inference process. The second row displays the corresponding heatmaps, revealing pixels of higher relevance in red or orange and those of lower relevance in blue. The similarity of the images of the four pairs is 0.84, 0.45, −0.13, and −0.19, respectively.

(a) first example (b) second example

Fig. 5. Examples of similar images with their corresponding visual explanations highlighting what makes these images similar.



(a) first example (b) second example

Fig. 6. Examples of dissimilar images with their corresponding visual explanations highlighting what makes these images dissimilar.

These illustrative examples show that the explanations for both similar and dissimilar images indicate that the FaceNet neural model prioritizes certain facial features (such as the contour of the mouth, eyes, and nose in that sequence) when distinguishing between different categories. Additionally, as the similarity value diminishes for similar images, the SNN model necessitates considering a larger number of pixels to distinguish between categories effectively.

4.2. Image retrieval

To assess the effectiveness of the ExplainLFS method for this task, we use the graphical image retrieval interface in https://github.com/Chien-Hung/ImageRetrievalGUI, which follows the image retrieval protocols in [5,46]. The experiments are performed using the models in https://github.com/Confusezius/Deep-Metric-Learning-Baselines, which report a Recall@K of 86% and 92.1% [47] on the CUB200 and Cars-196 datasets. The former includes 11,788 images and 200 bird classes [48], while the latter is comprised of 16,185 images and 196 car classes [49].

When it comes to training the decoder, we used 5,924 and 8,131 images from the CUB200 and Cars-196 datasets, respectively, and embedding vectors of length equal to 512 to train the decoder. The decoding error after the training phase is done is 0.03 and 0.05, respectively. It should be noted that, although the lowest possible decoding error is desired, the performance of the proposed method should not be affected by this error. In other words, ExplainLFS estimates the differences between reconstructions produced by the same decoder, i.e., between the reconstruction of the original image embedding and the reconstruction of the perturbed embedding.

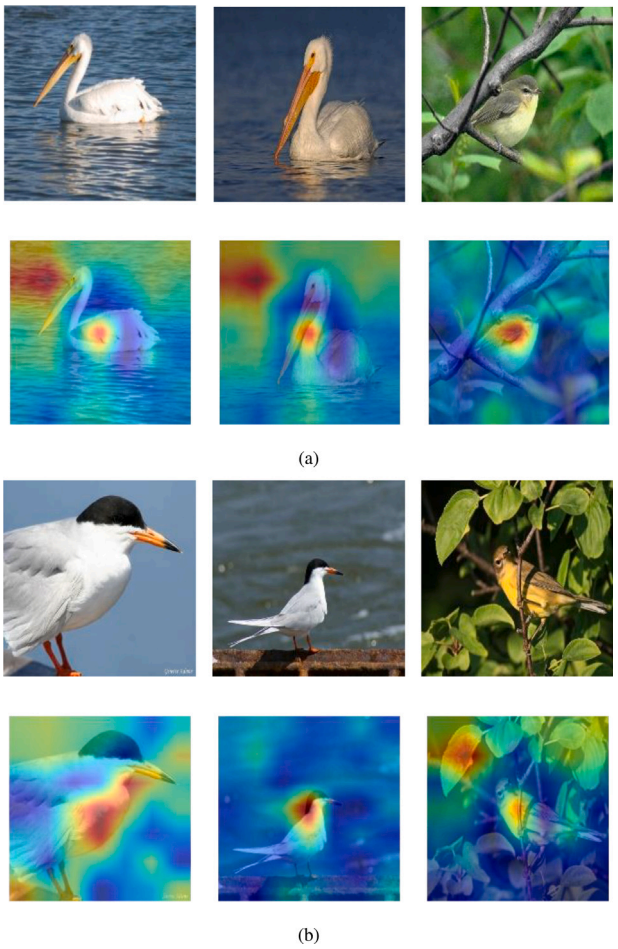


Fig. 7. Explanations provided by ExplainLFS for the results produced by ImageRetrievalGUI on the CUB200 dataset. On the one hand, the similarity of the first and second query images with their corresponding most similar image is 0.94 and 0.96, respectively. On the other hand, the similarity with its retrieved most dissimilar image is 0.28 and 0.38, respectively.

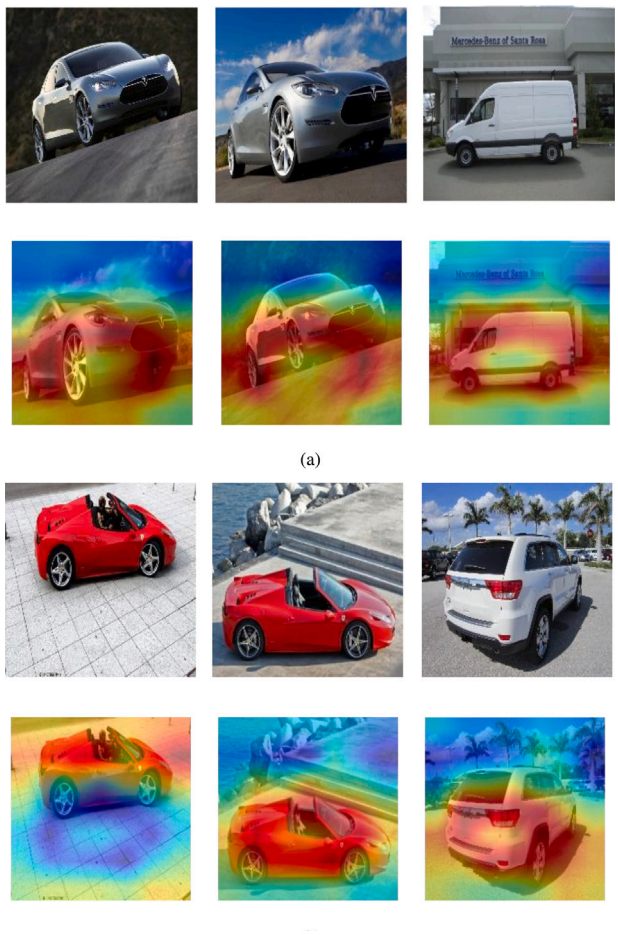


Fig. 8. Explanations provided by ExplainLFS for the results produced by ImageRetrievalGUI on the Cars-196 dataset. On the one hand, the similarity of the first and second query images with their corresponding most similar image is 0.96 and 0.98, respectively. On the other hand, the similarity with its retrieved most dissimilar image is 0.28 and 0.31, respectively.

Figs. 7 and 8 showcase the outcomes of ExplainLFS in elucidating the output produced by ImageRetrievalGUI for four images sourced from the CUB200 and Cars-196 datasets. In these figures, the initial column illustrates the query image, while the second and third columns display the most similar and dissimilar images retrieved in each scenario.

A close inspection of the visual explanations generated by our method suggests that the neural model relies on distinctive features such as the beak and plumage, contingent on the type of bird, to discern between various bird categories. It is also noticeable that environmental cues occasionally contribute to the network predictions. When it comes to distinguishing between different car classes, the pivotal areas of interest involve the shapes of the wheels and bumpers.

4.3. Evaluating the quality of explanations

The explanations of the proposed method are evaluated according to two properties [33]: selectivity ("pixel-flipping" experiment [32] or area under the cumulative feature importance curve [29,35]) and continuity ("robustness" or local *Lipschitz* estimation for other authors as [35]). In addition, as a third evaluation criterion, the neural system proposed in [26] is used as a substitute for an expert's judgment, another of the criteria most commonly used in the literature to measure the quality of the explanations [50,51]. Note that, the decision not to conduct a direct user study is based on the fact that the experimentation performed is not directly oriented to a real application case in which we can interact with domain experts.

4.3.1. Explanation selectivity

"A perturbation in a relevant input variable will cause a steeper decrease in the prediction score than that of a less relevant one " [29,33,35,43]. The hypothesis to be tested is that a perturbation (or noise injection by insertion/deletion) on the most relevant pixels in the output of the explanation method can cause a considerable change in the inference of the neural model. To make this process more rational, we associate areas of pixels with superpixels (i.e., of size 5×5). The relevance of a superpixel is calculated based on the relevance of the pixels containing it. Thus, perturbations in the superpixels composed of the most relevant pixels are expected to cause significant fluctuations in the similarity values inferred by the neural model for a pair of images.

Fig. 9 shows the average similarity values given by FaceNet after performing multiple perturbations (exactly 20, as suggested in [24]) on those most relevant superpixels according to the proposed method. For pairs of similar images (a), it is necessary to perform the perturbations on one of the images of the pair. In contrast, for dissimilar pairs (b), the perturbations must be performed on both images due to their dissimilar nature. In both cases, we observe how the similarity value between a pair of images is affected by perturbing the most relevant superpixels, i.e., it decreases between pairs of similar images and increases between pairs of dissimilar images. Both behaviors are

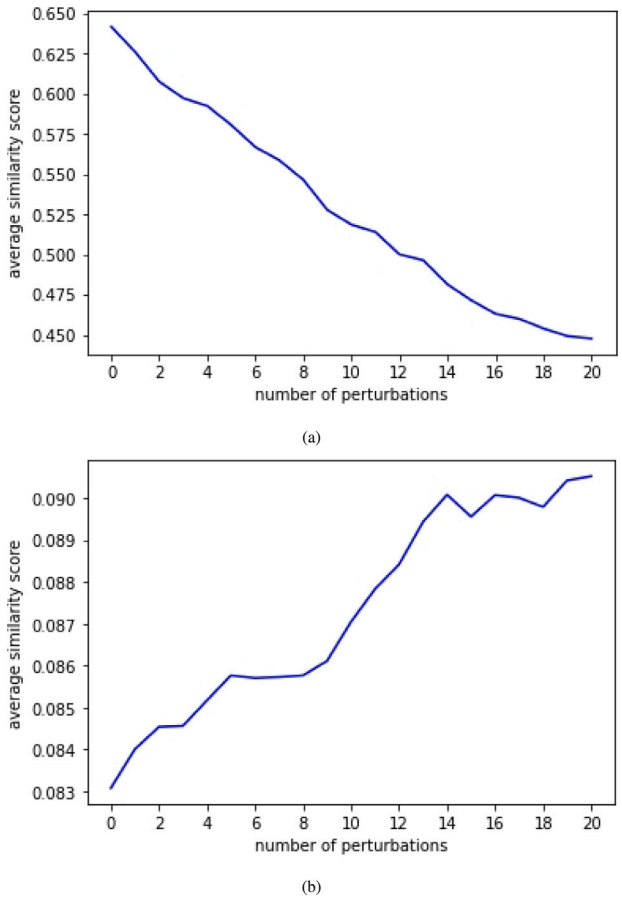


Fig. 9. Average similarity values of FaceNet on (a) a set of similar image pairs, and (b) a set of dissimilar image pairs. The x-axis represents the number of perturbations performed, taking into account the order of relevance of each superpixel with respect to the obtained inference. The y-axis represents the average similarity value obtained after completing x perturbations.

as expected. First, perturbing the features of one image that make it identical to another causes its similarity to the other to decrease. Second, introducing equivalent transformations on the features of two images that determine their dissimilarity causes it to decrease and, consequently, their similarity to each other to increase slightly. This effect is accumulative as a function of the number of perturbations.

4.3.2. Explanation continuity

"If two data points are nearly equivalent, then the explanations of their predictions should also be nearly equivalent" [33]. The hypothesis to be tested is that if two images x and x' are similar, the explanations given by the method for their predictions should also be similar. To carry out this experimentation, an image x is displaced from left to right and right to left in its input space (see Fig. 11).

From the generated image pairs, we obtain the relevance vectors $\mathcal{H}(x)$ and $\mathcal{H}(x')$, i.e., the output of applying the proposed method to the FaceNet inference on a pair of images composed of the original image x and one of its displacements x', respectively. Then, the difference between $\mathcal{H}(x)$ and $\mathcal{H}(x')$ is calculated as the L2 norm among them.

Fig. 10 shows the continuity curve associated with the proposed ExplainLFS explanation method. Observe that the difference between the relevance vectors increases until it reaches a maximum difference (where the image disappears entirely) as the image is displaced from left to right (1–10). In addition, the difference between the relevance vectors (11–20) gradually decreases until reaching a displacement where the original image is visualized almost entirely. This behavior evidences the continuity of the generated explanations, i.e., the more similar the images of the pair are, the closer their relevance vectors are. Conversely, the distance between their relevance vectors increases as their dissimilarity increases.

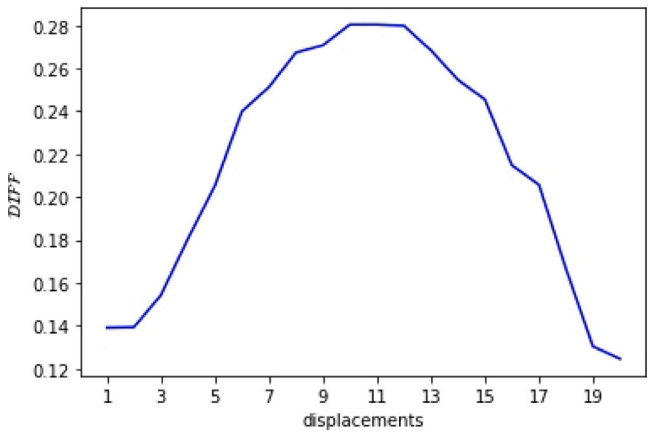


Fig. 10. Continuity curve of the proposed method. The x-axis represents the x' displacements (in Fig. 11), where the original image x is the coordinate origin. The y-axis represents the value of $DIFF = \|\mathcal{H}(x) - \mathcal{H}(x')\|_2$.

4.3.3. Relevance maps vs. similarity maps

Next, we compare the explanations resulting from the proposed method with a network that visualizes the similarities between a pair of images [26]. In this experiment, we use this network as a substitute for an expert's judgment, which is another criterion used in the literature to evaluate the quality of an explanation [50,51]. This network (from now on referred to as *SIMExpert*) takes a pair of images and generates similarity maps where the image regions responsible for the pairwise similarity are visualized. In this sense, the similarity maps provided by SIMExpert are compared with the relevance maps produced by ExplainLFS, the method of Utkin et al. [28], and the method of Tummala et al. [31].

Figs. 12 and 13 show examples of these maps for six image pairs. The similarity and relevance maps are presented as a heatmap, where pixels with higher relevance (or similarity) are shown in red (or orange color), and those with lower relevance are visualized in blue. Visually, our approach agrees with SIMExpert that the most distinctive regions when identifying facial similarities are the eyes and the mouth-nose contour (i.e., the mustache area). For Tummala et al.'s method, this area is also the most important, although, its attention map is broader. However, in some cases, Utkin et al.'s method focuses on the edges of the image, which is one of its weaknesses.

4.4. Comparison with the state-of-the-art

Samek et al. [32] suggested evaluating explanation method quality through the "pixel-flipping" experiment. This experiment gauges the speed at which the prediction score decreases as the features with the highest relevance for prediction undergo perturbations. Consequently, a rapid drop in the function value, reflected in a low area under the curve (AUC), signals the accurate identification of relevant features [29,34,35].

Fig. 14 illustrates performance curves for the SIMExpert, Utkin, Tummala, and ExplainLFS methods associated with the "pixel-flipping" experiment. In short, these curves depict the performance of the FaceNet architecture as an image pair is progressively perturbed (by random insertion/deletion in the most influential pixels) according to the explanation given by each XAI method. The quicker the classifier performance declines following input perturbation guided by the relevance analysis, the more effective the XAI method is at pinpointing the input features responsible for the neural model's output.

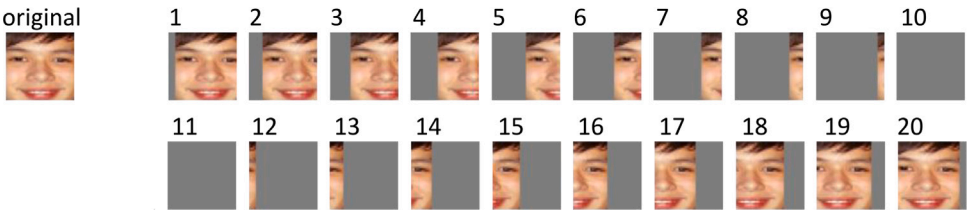


Fig. 11. Displacements of the original image along the x-axis, i.e., from left to right (i.e., 1–10) and from right to left (i.e., 20–11).

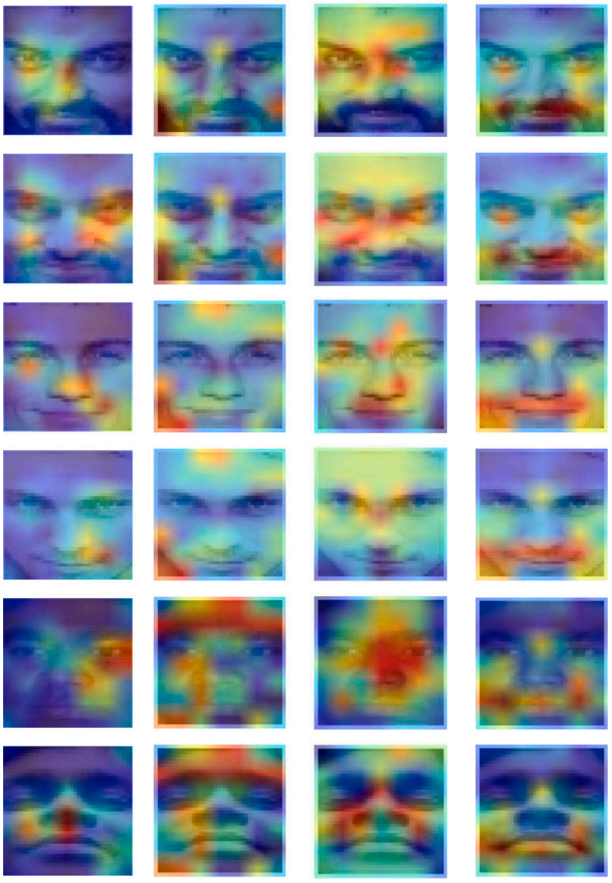


Fig. 12. Heatmaps (pairs 1–3) obtained with SIMExpert, Utkin, Tummala, and ExplainLFS methods. The first column represents the similarity maps given by SIMExpert, while the second, third, and fourth columns represent the relevance maps provided by Utkin, Tummala, and ExplainLFS, respectively.

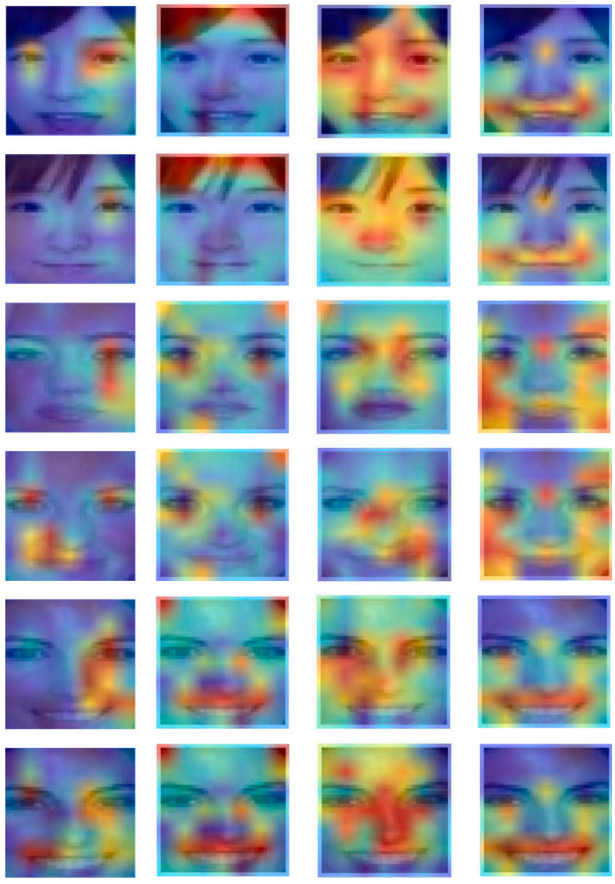


Fig. 13. Heatmaps (pairs 4–6) obtained with SIMExpert, Utkin, Tummala, and ExplainLFS methods. The first column represents the similarity maps given by SIMExpert, while the second, third, and fourth columns represent the relevance maps provided by Utkin, Tummala, and ExplainLFS, respectively.

The obtained AUC values clearly indicate our proposal's superiority over the Tummala, Utkin, and SIMExpert approaches (even taking into account that order in the evaluation ranking). Additionally, it is noteworthy that the similarity value experiences a significant drop when the six most significant superpixels, as per our approach, are perturbed.

5. Conclusions

The current research proposes an explainability method for similarity learning networks named ExplainLFS. The proposed method visualizes the most relevant pixels in an image pair on which these neural models rely to infer similarity or dissimilarity. We rely on latent features of the embedding space derived from the network learning process. Our proposal is a *post-hoc* agnostic technique and applies to any neural architecture for similarity learning on embeddings, e.g., face recognition and image retrieval tasks. The explanations obtained on the FaceNet architecture largely satisfy two very desirable properties:

- the "selectivity" property, i.e., the similarity degrees between a pair of images are also affected by perturbing the most relevant pixels in the inference of the neural model.
- the "continuity" property, i.e., similar explanations are obtained for predictions on similar images.

Furthermore, if we use SIMExpert [26] as a neural expert to compare the explanations given by ExplainLFS, Utkin [28], and Tummala [31], our method is closer to the neural expert's judgment than these methods. Likewise, a sharp drop in the value of the similarity function indicates the superiority of ExplainLFS over the Tummala, Utkin, and SIMExpert approaches in identifying relevant features when comparing a pair of images.

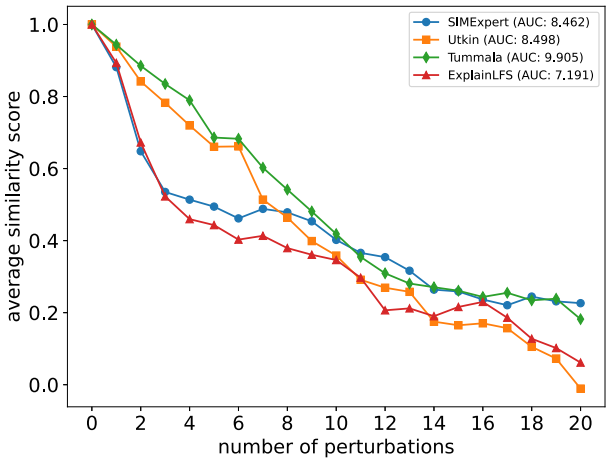


Fig. 14. ROC curves obtained by the SIMExpert, Utkin, Tummala, and ExplainLFS methods. The x-axis represents the number of perturbations performed, taking into account the order of relevance of each superpixel with respect to the obtained inference. The y-axis represents the average similarity value obtained after completing x perturbations.

CRediT authorship contribution statement

Marilyn Bello: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Pablo Costa:** Validation, Software, Investigation, Data curation. **Gonzalo Nápoles:** Writing – review & editing, Writing – original draft, Supervision. **Pablo Mesejo:** Supervision, Project administration. **Óscar Cordón:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research has been developed within the R&D project CONFIA (PID2021-122916NB-I00), funded by MICIU/AEI/10.13039/ 5011000 11033/ and FEDER, EU. Also, funding for open access charges is covered by Universidad de Granada / CBUA.

References

[1] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[2] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: Deep hypersphere embedding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[3] M. Zheng, S. Karanam, Z. Wu, R.J. Radke, Re-identification with consistent attentive siamese networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5735–5744.
[4] W. Wu, D. Tao, H. Li, Z. Yang, J. Cheng, Deep features for person re-identification on metric learning, Pattern Recognit. 110 (2021) 107424.
[5] C.-Y. Wu, R. Manmatha, A.J. Smola, P. Krahenbuhl, Sampling matters in deep embedding learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2840–2848.
[6] B. Chen, W. Deng, Hybrid-attention based decoupled metric learning for zero-shot image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2750–2759.
[7] A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, One-shot learning for semantic segmentation, 2017, arXiv preprint arXiv:1709.03410.
[8] J. Sen, B. Sarkar, M.A. Hena, M.H. Rahman, Face recognition using deep convolutional network and one-shot learning, Int. J. Comput. Sci. Eng. 7 (4) (2020) 23–29.
[9] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a " siamese" time delay neural network, Adv. Neural Inf. Process. Syst. 6 (1993) 737–744.
[10] D. Chicco, Siamese neural networks: An overview, Methods Mol. Biol. 2190 (2021) 73–94.
[11] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[12] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[13] Z. Bukovčiková, D. Sopiak, M. Oravec, J. Pavlovičová, Face verification using convolutional neural networks with siamese architecture, in: International Symposium on Electronics in Marine, IEEE, 2017, pp. 205–208.
[14] W. Hayale, P.S. Negi, M. Mahoor, Deep siamese neural networks for facial expression recognition in the wild, IEEE Trans. Affect. Comput. (2021).
[15] A. Barredo, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[16] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI methods-a brief overview, in: International Workshop on Extending Explainable AI beyond Deep Models and Classifiers, Springer, 2022, pp. 13–38.
[17] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Min. Knowl. Discov. (2023) 1–59.
[18] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[19] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[20] M. Bello, G. Nápoles, L. Concepción, R. Bello, P. Mesejo, Ó. Cordón, REPROT: Explaining the predictions of complex deep learning architectures for object detection through reducts of an image, Inform. Sci. 654 (2024) 119851.
[21] J. Bien, R. Tibshirani, Prototype selection for interpretable classification, Ann. Appl. Stat. 5 (4) (2011) 2403–2424.
[22] R.M. Byrne, Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, in: International Joint Conference on Artificial Intelligence, 2019, pp. 6276–6282.
[23] J. Su, D.V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Trans. Evol. Comput. 23 (5) (2019) 828–841.
[24] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.
[25] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 618–626.
[26] A. Stylianou, R. Souvenir, R. Pless, Visualizing deep similarity networks, in: IEEE Winter Conference on Applications of Computer Vision, IEEE, 2019, pp. 2029–2037.
[27] M. Zheng, S. Karanam, T. Chen, R.J. Radke, Z. Wu, Visual similarity attention, 2019, arXiv preprint arXiv:1911.07381.
[28] L. Utkin, M. Kovalev, E. Kasimov, An explanation method for siamese neural networks, in: International Scientific Conference on Telecommunications, Computing and Control, Springer, 2021, pp. 219–230.
[29] A. Fedele, R. Guidotti, D. Pedreschi, Explaining siamese networks in few-shot learning for audio data, in: International Conference on Discovery Science, Springer, 2022, pp. 509–524.
[30] I.E. Livieris, E. Pintelas, N. Kiriakidou, P. Pintelas, Explainable image similarity: Integrating siamese networks and grad-CAM, J. Imaging 9 (10) (2023) 224.
[31] S. Tummala, A.K. Suresh, Few-shot learning using explainable siamese twin network for the automated classification of blood cells, Med. Biol. Eng. Comput. (2023) 1–15.
[32] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2016) 2660–2673.
[33] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15.
[34] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, 2018, arXiv preprint arXiv:1806.07421.
[35] E. Doumard, J. Aligon, E. Escriva, J.-B. Excoffier, P. Monsarrat, C. Soulé-Dupuy, A quantitative approach for the comparison of additive local explanation methods, Inf. Syst. 114 (2023) 102162.
[36] B.A. Plummer, M.I. Vasileva, V. Petsiuk, K. Saenko, D. Forsyth, Why do these match? explaining the behavior of image similarity models, in: European Conference on Computer Vision, Springer, 2020, pp. 652–669.

[37] L. Chen, J. Chen, H. Hajimirsadeghi, G. Mori, Adapting grad-cam for embedding networks, in: IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2794–2803.
[38] X. Ye, D. Leake, W. Huibregtse, M. Dalkilic, Applying class-to-class siamese networks to explain classifications with supportive and contrastive cases, in: International Conference on Case-Based Reasoning, Springer, 2020, pp. 245–260.
[39] C. Chen, Y. Shen, G. Ma, X. Kong, S. Rangarajan, X. Zhang, S. Xie, Self-learn to explain siamese networks robustly, in: IEEE International Conference on Data Mining, IEEE, 2021, pp. 1018–1023.
[40] L. Deng, The mnist database of handwritten digit images for machine learning research [best of the web], IEEE Signal Process. Mag. 29 (6) (2012) 141–142.
[41] E. Mohamed, K. Sirlantzis, G. Howells, A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation, Displays 73 (2022) 102239.
[42] C. Molnar, Interpretable machine learning: a guide for making black box models explainable, 2019, https://christophm.github.io/interpretable-ml-book.
[43] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis, 2017, arXiv preprint arXiv:1706.07206.
[44] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2018, pp. 67–74.
[45] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3730–3738.
[46] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.
[47] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, J.P. Cohen, Revisiting training strategies and generalization performance in deep metric learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 8242–8252.
[48] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-Ucsd Birds-200–2011 Dataset, California Institute of Technology, 2011.
[49] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: IEEE Workshop on 3D Representation and Recognition, 2013, pp. 554–561.
[50] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint arXiv:1702.08608.
[51] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Comput. Surv. 55 (13s) (2023) 1–42.