FAM: Visual Explanations for the Feature Representations from Deep Convolutional Networks

Yuxi Wu, Changhuai Chen, Jun Che, Shiliang Pu*
Hikvision Research Institute, China
{wuyuxi, chenchanghuai, chejun, pushiliang.hri}@hikvision.com

Abstract

In recent years, increasing attention has been drawn to the internal mechanisms of representation models. Traditional methods are inapplicable to fully explain the feature representations, especially if the images do not fit into any category. In this case, employing an existing class or the similarity with other image is unable to provide a complete and reliable visual explanation. To handle this task, we propose a novel visual explanation paradigm called Feature Activation Mapping (FAM) in this paper. Under this paradigm, Grad-FAM and Score-FAM are designed for visualizing feature representations. Unlike the previous approaches, FAM locates the regions of images that contribute most to the feature vector itself. Extensive experiments and evaluations, both subjective and objective, showed that Score-FAM provided most promising interpretable visual explanations for feature representations in Person Re-Identification. Furthermore, FAM also can be employed to analyze other vision tasks, such as self-supervised representation learning.

1. Introduction

Over the last few years, the model explanation increasingly draws attention due to the wide applications of **C**onvolutional **N**eural **N**etworks (CNNs). To this end, various visual explanation methods have been proposed.

For interpreting classification problem, **C**lass **A**ctivation **M**ap (CAM) [1, 33] is proposed to locate which regions of input image were looked at by the model for assigning the label. Recent CAM-based works could be divided into two branches, one is gradient-based CAMs [14] which use gradient of class confidence to incorporate the importance of inputs, the other is gradient-free CAMs [12, 22] which capture the importance by the change of class confidence. Meanwhile, a number of methods, e.g., DeepLift [16] and integrated gradients [20], approximate the contribution of inputs based on the element-wise product with gradients.

It is noteworthy that a target neuron or score must be specified to evaluate gradient for the above approaches. In classification, class confidence usually is the target. However, in some tasks, the class of a test sample might not exist in train set. As a zero-shot learning problem, the identities of test set in Person Re-Identification (Re-ID) are totally different from train set. As showed in Figure 1 (b), Score-CAM merely concerns on the similar part to the given train identity, such as umbrella, pants and shoes, rather than the person body in the test image. In this case, employing an existing class cannot provide a reliable explanation.

Much less works have focused on understanding the feature representation [34]. RAM [23] and CG-RAM [15] are proposed to reveal the associated visual cues between a pair of images, and other gradient-based methods also can be applied to the gradient of similarity value. However, the similarity-based methods still are unable to explain the feature representation completely. As showed in Figure 1 (c), the visualization results of RAM based on different images were quite different, and mostly depend on the selection of another image. The shoes got high activations in the result based on the rank-2 gallery image, but got low activations in others. Based on these conflicting results, it is hard to interpret the contribution degree of shoes on feature representation in this image. Other similarity-based approaches are also unable to explain this question. Besides that, a suitable image for comparison sometimes might be unavailable in practical work. These issues have restricted interpreting Re-ID model to assure the reliability.

To ameliorate the aforementioned flaws, we propose a novel visual explanation paradigm called **F**eature **A**ctivation **M**apping (**FAM**). FAM highlights the regions of images that contribute most to the global feature representation. Specifically, we first proposed Gradient-weighted FAM (**Grad-FAM**), which eliminates the dependency of another image in RAM. Inspired by Score-CAM, a different gradient-free FAM method Score-weighted FAM (**Score-FAM**) is then proposed. Objective evaluation of Score-FAM outperformed the other methods by large scale on public datasets. The experimental results demonstrate that

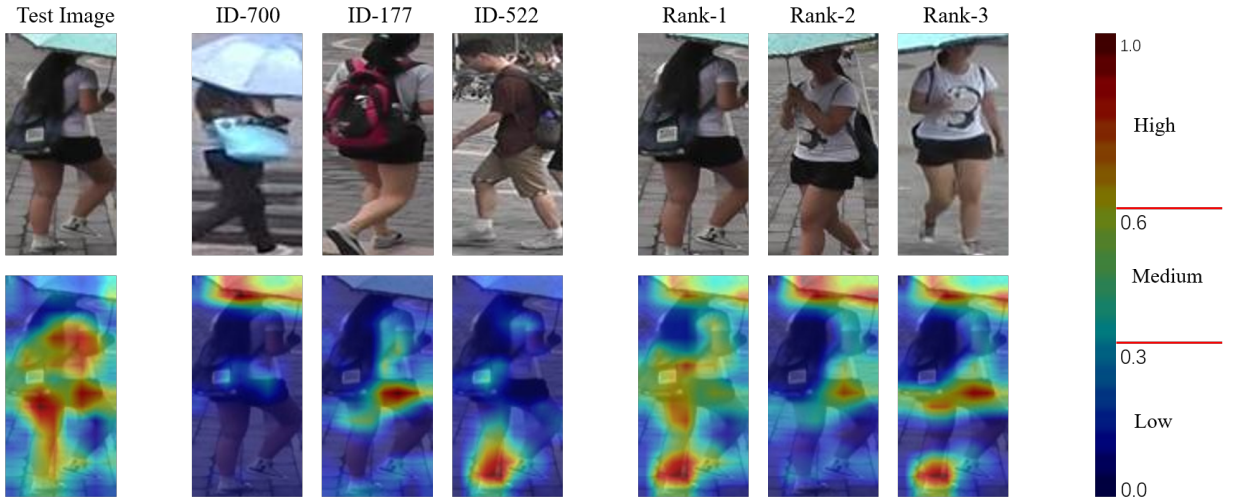


Figure 1. (a) A test image in Market1501 and the corresponding result of Score-FAM; (b) Score-CAM results w.r.t. the top three similar identities in train set according to the class confidences; (c) RAM results w.r.t. the top three similar gallery images according to ranking results. (d) The different colors in the visualizaition results represent corresponding activation values, which are shown in the color bar. Score-CAM merely concerned on the similar part to the given train identity. RAM got quite different results for different pairs. Score-FAM provided a complete visual explanation for the feature vector.

Score-FAM provides a faithful visual explanation of the embedding process on Person Re-ID.

Furthermore, our proposed FAM can be applied to more vision tasks including representation learning [2] and open-set recognition [5]. For a model that outputs feature vectors, the traditional visual explanation approaches require an additional classifier. Instead, FAM-based methods can interpret the feature representations directly.

Our key contributions are summarized as follows:

- We propose a novel localization technique FAM, including Grad-FAM and Score-FAM, to generate complete visual explanations for the feature vectors. Unlike the previous approaches, FAM is applicable to vision tasks that output feature vectors, such as Re-ID and representation learning.
- We propose new metrics in this work to objectively evaluate the faithfulness of visual explanations on Person Re-ID, i.e., whether the visualization result directly correlates with feature representation. Our results with these metrics show superior performance of Score-FAM over other approaches on dataset Market1501 [30] and CUHK03 [7].
- FAM can be employed to visualize self-supervised representation learning models without separately training linear classifiers.

The rest of this paper is organized as follows. Section 2 refers to recent works of Visualizing CNNs and Person Re-ID. In Section 3, we propose Grad-FAM and Score-FAM respectively. Experimental results in Section 4 demonstrate the effectiveness and outperformance of our proposed methods. We draw conclusions in Section 5.

2. Related Work

2.1. Visualizing CNNs

As one of the first efforts to interpret CNNs, Zeiler & Fergus [27] used deconvolution approach to get the regions of input image responsible for one neuron activation. Simonyan et al. [17] produced class-specific saliency maps by the partial derivatives of predicted class confidences w.r.t. inputs. Further, Guided Backpropagation [18] modified the backpropagation gradients to improve the quality of saliency maps. Yosinski et al. [26] visualized the functionality of a specific unit in networks, by synthesizing input image that cause the unit to have high activation. Sundararajan et al. [20] employed integrated gradients to attribute the prediction to inputs.

For a CNN with Global Average Pooling (GAP) layer, Zhou et al. [33] demonstrated that the class confidence Y^c for the class of interest c could be written as a linear combination of its global average pooled last feature maps A^k,

$$Y^c = \sum_k w_c^k \cdot \left(\frac{1}{Z} \sum_i \sum_j A_{ij}^k \right), \tag{1}$$

where w_c^k is the weight for k-th neuron after GAP layer, A_{ij}^k

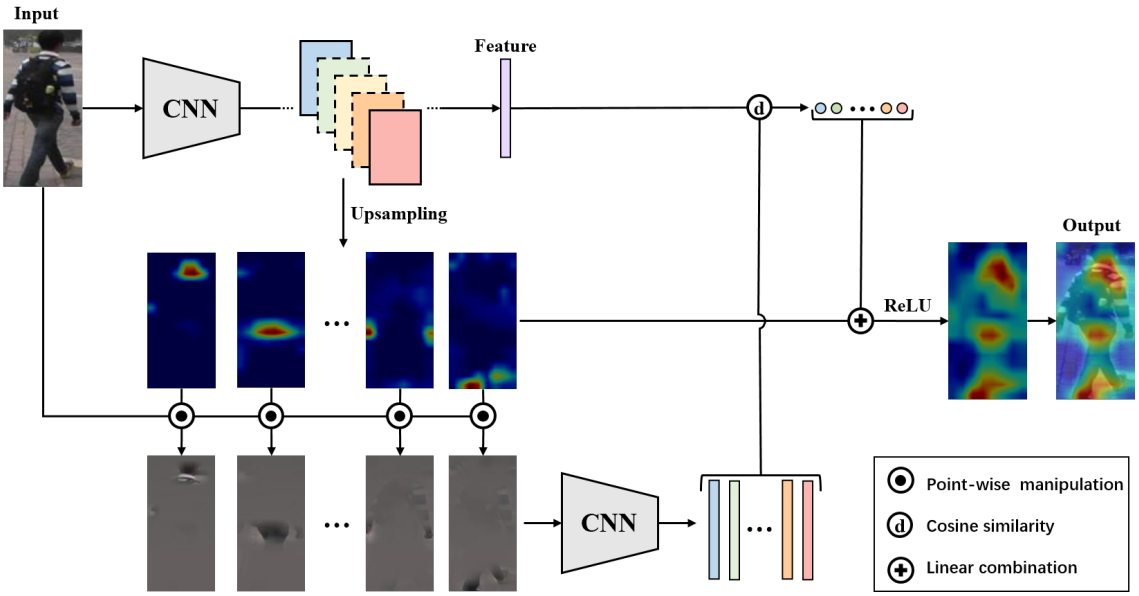


Figure 2. Pipeline of our proposed Score-FAM. Firstly, the given model extracts feature vector from the input image. In forward propagation, feature maps in the specified layer are taken out as masks on the input image after upsampling and normalization. Then, the generated images are input into the same model to obtain feature vectors, and corresponding cosine similarities to the original feature vector. Finally, the visualization result would be generated by linear combination of feature maps and similarity-based weights.

is the value of pixel (i, j) in the k-th channel of feature map, and Z is the number of pixels. Then the class activation map (CAM) for class c could be defined as:

$$L^c_{\text{CAM}} = \sum_k w^k_c \cdot A^k. \tag{2}$$

Similar methods were proposed for other pooling layers, such as global max pooling [9] and logsum-exp pooling [11]. But these approaches were restricted to CNNs with special architecture where the penultimate layer is a specified pooling.

As a generalization of CAM, Grad-CAM [14] extends the definition of w^k_c as the gradient of class confidence Y^c w.r.t. A^k to remove this restriction of architecture.

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial Y^c}{\partial A^k_{ij}}\right) \cdot A^k\right). \tag{3}$$

ReLU was applied to focus on the regions that had positive influence on the class of interest. Due to its applicability and well-understood visualization results, Grad-CAM has been widely applied in many tasks including classification, image captioning and visual question answering.

Considering the risk of gradient saturation problem, which would cause the gradients to diminish, Du et al. [22] proposed a gradient-free method Score-CAM. Score-CAM first upsampled feature map to original input shape, and then perturbed the input image with it. After the forward propagation with the masked input, the importance of that feature map is obtained by class confidence on the target category. Further, RISE [10] and mask [4] interpreted black-box models based on randomized and meaningful perturbations on inputs, respectively. From a different perspective, Desai & Ramaswamy [12] proposed another gradient-free approach Ablation-CAM based on Ablation studies.

However, the above approaches are all built on class confidence. Which means the visualization result is biased towards the selected class, and meaningless for samples not in train classes or models without classifiers. To handle these tasks, Yang et al. [23] proposed a method to visualize the similarity between a pair of images for models with GAP layer. For an image q, the similarity with another image g is proportional to $\frac{\langle v_g, v_q\rangle}{|v_g|}$, which could be formulated as

$$\frac{1}{|v_g|}\sum_k v^k_g \cdot \frac{1}{Z}\sum_i\sum_j A^k_{ij} = \frac{1}{Z}\sum_i\sum_j\sum_k \frac{1}{|v_g|} v^k_g \cdot A^k_{ij}, \tag{4}$$

where v_q, v_g are the feature representations of q and g respectively, v^k_g is the k-th element, $\langle , \rangle$ is the inner product of two vectors, and $|\cdot|$ is magnitude of the vector. Then Ranking Activation Map (RAM) is defined as

$$L^g_{\text{RAM}} = \sum_k \frac{1}{|v_g|} v^k_g \cdot A^k, \tag{5}$$

to reveal the associated visual cues between the query and gallery images in Re-ID task.

To remove the restriction of model structure, Confidence Gradient-weighted RAM (CG-RAM) [15] is proposed based on the gradient of cosine similarity. Nevertheless, the visualization results of these methods are biased towards the other chosen image, and still cannot provide a complete visual explanation for the input image.

2.2. Person Re-identification

Re-ID has been widely studied as a specific retrieval problem across non-overlapping cameras [3]. The key point is obtaining a feature representation with robust identity-discriminative information. With the superiority of deep learning, CNNs have been generally recognized as the most efficient method in Re-ID tasks. The widely-used ID-discriminative Embedding (IDE) model [31] constructed the training process as a classification problem by treating each training identity as a distinct class. In forward propagation with test example, the feature vector fed into the classification layer is regarded as feature representation of the input image. Given a query of interest, images that might belong to the same identity could be retrieved by the distances between corresponding feature vectors. It is now widely used in Re-ID community [24, 25].

3. Approach

In this section, we introduce the mechanism of proposed Grad-FAM and Score-FAM respectively for interpreting feature representation.

3.1. Grad-FAM

The visual result of RAM locates the similar regions between two images for the model. When the two images are almost same, such as the Rank 1 in Figure 1 (c), the entire regions could be regarded as associated visual cues. In this case, the activation degree of different sub-regions would be closely relevant to the importance of feature representation. Furthermore, employing exactly the same image might make the relevance closer. Therefore, we assume that RAM for an image with itself could point out the sub-regions that are critical to feature representation.

Meanwhile, we remove the architecture restriction of RAM in a similar manner as Grad-CAM. Grad-RAM is defined as:

$$L^{g}_{\text{Grad-RAM}} = \text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial \frac{\langle f(X_g), f(X)\rangle}{|f(X_g)|}}{\partial A^k_{ij}}\right)\cdot A^k\right), \tag{6}$$

where X, X_g is the input and the other image, $f(X)$ is the feature representation of X. The ReLU layer ensures that only the regions with positive influence on the similarity would be retained. For CNNs with GAP layer, Grad-RAM is a strict generalization of RAM.

It is noteworthy that Grad-RAM is different with CG-RAM, which is based on the gradient of cosine similarity:

$$d_{\cos}(f(X), v_g) = \frac{\langle f(X), f(X_g)\rangle}{|f(X)|\cdot|f(X_g)|}. \tag{7}$$

The cosine similarity would be close to 1 for any model, when the two images are very similar. In this case, the gradient of similarity would become vanishing, and the visual result is unreliable.

According to the previous assumption, we replace the other image X_g in Equation (6) with X to get rid of extra input. Then we could obtain the equation:

$$\frac{\langle f(X), f(X)\rangle}{|f(X)|} = |f(X)|, \tag{8}$$

and thus define **Grad-FAM** as:

$$L_{\text{Grad-FAM}} = \text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial |f(X)|}{\partial A^k_{ij}}\right)\cdot A^k\right). \tag{9}$$

The proposed framework could be achieved by slight modification to the derivative part of Grad-CAM.

Obviously, Grad-FAM is only related to the given image, and provide a complete visual explanation for the feature representation. Furthermore, other gradient-based methods can be also applied in this paradigm.

3.2. Score-FAM

In Re-ID tasks, the retrieval results are based on the ranking of similarities among gallery images, which are usually measured by cosine similarity between corresponding feature vectors.

Inspired by CIC [22], we generate **C**hannel-wise **I**ncrease of **S**imilarity (CIS) based on cosine similarity, in order to measure the importance of each feature map to feature representation.

Channel-wise Increase of Similarity: *Given a CNN model $v = f(X)$ that takes an input image X and outputs a vector v. The k-th channel of a feature map A_l for an internal convolutional layer l in f is denoted as feature map A^k_l. For a known baseline input B, the contribution of A^k_l towards v is defined as*

$$S(A^k_l) = d_{\cos}\left(f(X\circ M^k_l), f(X)\right) - d_{\cos}\left(f(B), f(X)\right), \tag{10}$$

$$M^k_l = \text{Norm}\left(\text{Up}(A^k_l)\right),$$

where $d_{cos}(\cdot,\cdot)$ is the cosine similarity between two vectors, $\text{Up}(\cdot)$ *denotes the operation that upsamples A^k_l into input size, and* $\text{Norm}(\cdot)$ *is a normalization function that maps elements within $[0, 1]$ range,*

$$\text{Norm}(M) = \frac{M - \min M}{\max M - \min M}. \tag{11}$$

Algorithm 1 Score-FAM algorithm

Require: Image X, Baseline Image B, Model $f(\cdot)$, layer l
Ensure: Saliency Map $L_{\text{Score-FAM}}$
initialization;
// get activation of layer l and feature vector;
$A_l, v_0 \leftarrow f(X), v_b \leftarrow f(B)$
$M \leftarrow []$, $C \leftarrow$ the number of channels in A_l
for $k \in [0, C-1]$ **do**
  // get the mask from feature map;
  $M_l^k \leftarrow \text{Norm}\left(\text{Upsample}(A_l^k)\right)$
  // Hadamard product;
  M.append$(X \circ M_l^k)$
end for
$M \leftarrow \text{Batchify}(M)$
// extract feature vectors from generated images;
$v_l \leftarrow f(M)$
// compute the similarity-based weights;
for $k \in [0, C-1]$ **do**
  $S_l^k \leftarrow d_{\cos}(v_l^k, v_0) - d_{\cos}(v_b, v_0)$
end for
$a_k \leftarrow \frac{\exp(S_l^k)}{\sum_h \exp(S_l^h)}$
$L_{\text{Score-FAM}} \leftarrow \text{ReLU}\left(\sum_k a_k \cdot A_l^k\right)$

The upsampled feature map works as a mask to perturb the input image with only the regions of interest retained. The normalized function is employed to make the mask smoother. To avoid sharp boundaries between masked and salient regions, we employ the blurred image of input to replace the masked regions. Further, the blurred image is also employed as the baseline image B in this work.

When the mask captures more characteristics that the feature representation focuses on, the similarity to input image would be higher. Therefore, the CIS score of feature map indicates the importance to feature representation.

Finally, our proposed visual explanation method **Score-FAM** is described as:

$$L_{\text{Score-FAM}} = \text{ReLU}\left(\sum_k a_k \cdot A_l^k\right), \tag{12}$$

$$a_k = \frac{\exp\left(S(A_l^k)\right)}{\sum_h \exp\left(S(A_l^h)\right)}. \tag{13}$$

Similar to [14,22], a ReLU is also applied to the linear combination of feature maps. Because only the regions that have a positive influence on feature representation are interested in. Meanwhile, the CIS scores of different feature maps have different amplitude. Thus it is reasonable to represent the weights of Score-FAM as post-softmax value.

The pipeline of the proposed framework is illustrated in Figure 2, and complete details of the implementation are described in Algorithm 1.

methods	time (seconds per image)
Grad-FAM	0.05
Score-FAM	9.69

Table 1. The computational time of Grad-FAM and Score-FAM.

As showed in Table 1, Score-FAM requires much more computational cost than Grad-FAM, but the gap is acceptable for few samples.

The last convolution layer is usually the preferable choice because it is end point of feature extraction [14]. However, all other convolutional layers also could be employed in both Grad-FAM and Score-FAM.

4. Experiment

In this section, we conduct experiments to evaluate the effectiveness of the proposed FAM methods. First, Section 4.2 introduce the experimental setup in this work. Second, we objectively compare the performance against existing state-of-the-art methods on Person Re-ID in Section 4.2. Then the sanity check of Score-FAM is followed in Section 4.3. Finally, we employ Grad-FAM to analyze the self-supervised representation learning in Section 4.4.

4.1. Experimental Setup

All experiments are implemented with the Pytorch 1.6 framework on a NVIDIA Tesla P40 GPU.

In the following experiments, we employed ResNet-50 as a base model on two public Re-ID benchmark datasets.

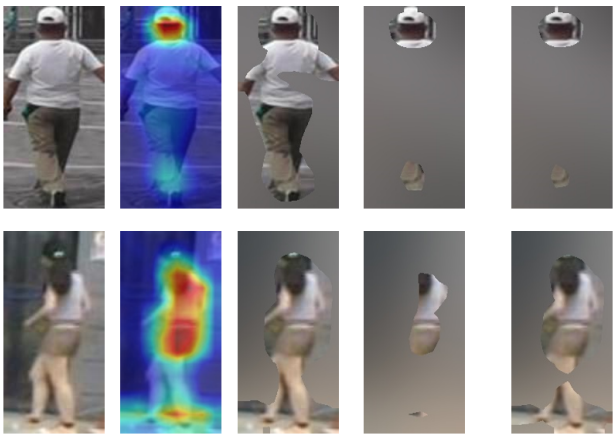
Market1501 [30] dataset contains 32,668 person images of 1,501 identities captured by six cameras. Train set is composed of 12,936 images of 751 identities, and test data is composed of images of other 750 identities.

CUHK03 [7] dataset contains 14,096 images of 1,467 different identities. Each person is captured from two cameras in the CUHK campus, and the protocol proposed in [32] providing fixed train/test splits with 767 and 700 disjoint identities, respectively. The dataset provides both manually annotated and DPM-detected bounding boxes.

The Re-ID models were trained accordingly for each dataset with the bag of tricks [8], which is a strong baseline with ImageNet [13] pre-training for Person Re-ID.

The input images of Re-ID models are resized to [256, 128]. We blur the original input with Gaussian Blur to generate the baseline image for Score-CAM and Score-FAM. Following [28], the parameters of Gaussian Blur, radius and sigma are set as 51 and 50 respectively.

Our experiments involve 4 state-of-the-art visual explanation methods, including Grad-CAM, Score-CAM, RAM and CG-RAM. Grad-CAM and Score-CAM require selecting a class to generate saliency map, which would be the



(a) Image (b) Result (c) Top 50% (d) Top 10% (e) Our Proposed

Figure 3. (a) 2 test images in Market1501; (b) the visualization results; (c) explanation maps by retaining pixels with top 50% values; (c) explanation maps by retaining pixels with top 10% values; (e) our proposed explanation maps. The traditional explanation maps based on a fixed ratio cannot match the visualization results for all examples. Meanwhile, the explanation maps generated by our proposed method are closely related to the saliency maps in visual sense.

predicted train class for test images. RAM and CG-RAM are based on a pair of images, in which we choose the Rank-1 gallery image for comparison.

4.2. Objective Evaluation for Person Re-ID

In this section, we objectively evaluate the faithfulness of visual explanations for Re-ID tasks.

The reliability of visualization results for classification is judged by the explanation map, which usually retains a fixed ratio (like 50%) of pixels with top activation values for the input. However, the explanation maps generated in this way cannot match the visualization results for all examples, as showed in Figure 3. In this work, we propose a more adaptive way to generate explanation maps.

From the view of visual perception, the color bar in Figure 1 (d) can be divided into three parts as red, green and blue. For an observer, the red sub-regions seem to be significant, and the blue parts show little influence.

Based on the division of saliency map, we create a mask M for every image.

$$M_{ij} = \begin{cases} 1, & L_{ij} \geq B_{\text{High}} \\ \delta, & B_{\text{Low}} \leq L_{ij} < B_{\text{High}} \\ 0, & L_{ij} < B_{\text{Low}} \end{cases} \tag{14}$$

where L is the saliency map, B_{High}, B_{Low} are the boundaries for high activation (red) and low activation (blue) respectively, and δ is the retained proportion of inputs in the medium activation region (green). Then we define explanation map E by masking M on the input.

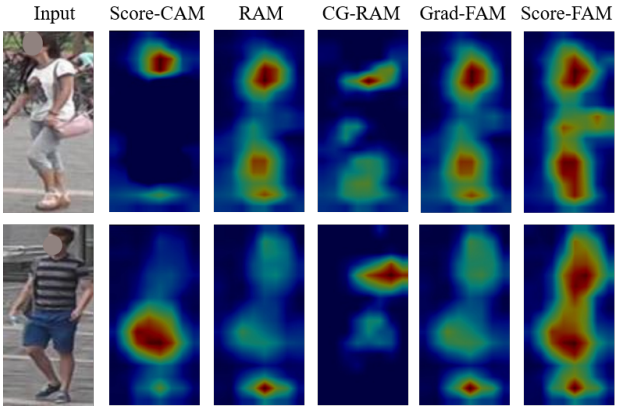


Figure 4. Visualization results of two samples in Market1501 by Score-CAM, RAM, CG-RAM and our proposed Grad-FAM, Score-FAM.

The basic idea behind explanation map is preserving the emphasized sub-regions, and partly retaining the parts with medium activation. As showed in the Figure 1 (d), we set 0.3 and 0.6 as the boundaries respectively. δ is set as 0.9 in this work, because the results of Person Re-ID are very sensitive to disturbance. In our experiments, masking inputs with 10% would have real impact on the retrieval results. The masked sub-regions would be replaced by the blurred image of original input. Sample qualitative result is shown in Figure 3.

The feature representation of explanation map would be similar to original input, if the sub-regions with high activation really contribute most in the embedding process. Although the selection of hyper-parameters is a little subjective, we believe it is fair to compare different approaches with the same evaluation method.

By observing the similarities and retrieval results of explanation map among gallery images, we could measure whether the saliency map captures the sub-regions that are critical to feature representation. Inspired by the metrics used in [1], we studied the performance objectively with three new metrics: (i) AS Drop; (ii) AP Drop; and (iii) AP Increase, which would be described below.

(i) AS Drop: Average Similarity Drop.

A good explanation map should contain the regions that are most relevant to feature representation, and it is expected to be close to not only the original input but also other images in the same identity. We employ the change of **A**verage **S**imilarity (**AS**) to all images in the same identity, as compared to the original input, to measure the performance of explanation map. The saliency map would be more relevant to the corresponding feature representation, if AS drops less.

methods	Market1501 [30]			CUHK03 [7]		
	AS Drop (%)	AP Drop (%)	AP Increase (%)	AS Drop (%)	AP Drop (%)	AP Increase (%)
Grad-CAM	52.08	77.18	2.49	56.97	74.53	8.57
Score-CAM	54.07	67.41	7.66	51.27	62.62	18.07
RAM	29.65	38.45	14.61	35.88	51.32	22.00
CG-RAM	57.95	79.05	2.70	59.63	77.73	7.64
ours/Grad-FAM	28.28	35.14	15.97	33.11	47.55	24.86
ours/Score-FAM	**16.57**	**17.01**	**30.26**	**17.49**	**28.68**	**38.86**

Table 2. Evaluation results on Person Re-ID (lower is better in AS Drop and AP Drop, higher is better in AP Increase). The best records are marked in **bold**.

The AS Drop is expressed as

$$\frac{1}{N_{\mathbb{Q}}}\sum_{q\in\mathbb{Q}}\frac{\max\left(0,\ D_q(X_q,E_q)\right)}{D_q(X_q,B_q)}, \tag{15}$$

$$D_q(I,J)=\sum_{g\in\mathbb{G}_q}\left[d_{\cos}\left(f(I),f(X_g)\right)-d_{\cos}\left(f(J),f(X_g)\right)\right],$$

where $\mathbb{Q}$ is the query set, $N_{\mathbb{Q}}$ is the size of $\mathbb{Q}$, $\mathbb{G}_q$ is composed of gallery samples from the same identity of the query q, X_q, E_q, B_q are the original image, corresponding explanation map and baseline image of q respectively. We employ $\max$ to handle the case where explanation map gets closer to other images than the original input.

(ii) AP Drop: Average Precision Drop.

AS Drop measures the absolute distance between the explanation map and corresponding identity in embedding space of the model. However, Re-ID tasks are more concerned with the relative distances among all samples for retrieval. Thus the explanation map is expected to be closer to the samples in the same identity than other identities. In Re-ID tasks, the retrieval result is usually evaluated by **A**verage **P**recision (**AP**) [30], which is the area under the Precision-Recall curve.

The definition of AP Drop is similar to AS-Drop:

$$\frac{1}{N_{\mathbb{Q}}}\sum_{q\in\mathbb{Q}}\frac{\max\left(0,\ \mathrm{AP}(X_q)-\mathrm{AP}(E_q)\right)}{\mathrm{AP}(X_q)-\mathrm{AP}(B_q)}, \tag{16}$$

where $\mathrm{AP}(I)$ is the AP of image I for the gallery set.

(iii)AP Increase: Increase in Average Precision.

Complementary to AP Drop, there might be scenarios where the explanation map get a better retrieval result than original input (especially when the query image gets serious background interference). In this metric, we measure the number of query samples, whose explanation map gets a higher AP than original image. Formally, the Increase in Average Precision (denoted as AP Increase) is defined as

$$\frac{1}{N_{\mathbb{Q}}}\sum_{q\in\mathbb{Q}}\mathrm{Sign}\left(\mathrm{AP}(X_q)<\mathrm{AP}(E_q)\right), \tag{17}$$

where $\mathrm{Sign}(\cdot)$ is indicator function, which returns 1 if the input is True.

The three metrics are calculated per query image and averaged over the entire query set. Results on Market1501 and CUHK03 are reported in Table 2.

As shown in Table 2, the experimental results of Grad-CAM and Score-CAM were much worse than RAM on all metrics. These results prove that the visual explanations based on train classes are meaningless for test images as expected. CG-RAM got the worst performance in this test, which means the gradient of similarity is not suitable for visual explanation as mentioned in Section 3.1.

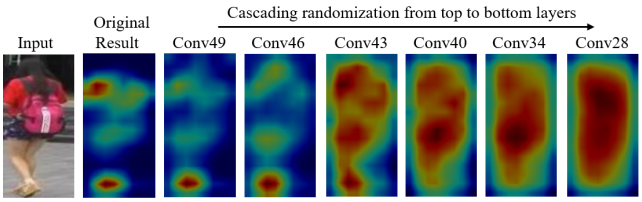
Meanwhile, Grad-FAM got a better performance than RAM, which verified the effectiveness of our method. The visualization result based on the original input itself is more reliable than the similarity with other image. Score-FAM outperformed the other methods on three metrics by large scale in both datasets. Especially in Market1501, the AP Drop and AP increase are 17.01% and 30.26% respectively. The masked explanation maps got a similar retrieval results to the original inputs, which means the feature vectors depend almost entirely on the sub-regions that are emphasized by Score-FAM.

Meanwhile, the visualization results shown in Figure 4 also support the superior performance of Score-FAM, which emphasizes the sub-regions that other methods easily overlook.

The good performance on Person Re-ID demonstrates that Score-FAM successfully points out the most distinguishable part of person body for feature representation, and reveal the embedding process of Re-ID model more faithfully than previous approaches.

4.3. Sanity Check

In this section, we employ sanity check [21] to check whether Grad-FAM and Score-FAM provide reliable explanation for model's behavior. In the cascading randomization, the weights of model are randomized from the top to bottom layers. An explanation method would fail the sanity check, if the outputs remain similar for networks with



(a) Grad-FAM Results for different model weights.

(b) Score-FAM Results for different model weights.

Figure 5. Sanity check results of Grad-FAM and Score-FAM by cascading randomization. The weights of ResNet50 trained on Market1501 are reinitialized randomly from the top layers to bottom. In this procedure, visualization results show sensitivity to model parameters randomization.

widely differing parameters.

As shown in Figure 5, the results of Grad-FAM and Score-FAM show sensitivity to model parameters randomization, and the quality of saliency maps could reflect the quality of model. Therefore, two types of FAM both pass the sanity check.

4.4. Application on Self-Supervised Learning

Besides Person Re-ID, the proposed FAM methods also can be applied to analyze the feature representations in other vision tasks.

For instance, self-supervised representation learning can learn a good representation without human annotations. Since the self-supervised models do not have classifiers, the class-based visual explanation approaches cannot help explain the feature representations. To handle this task, Zhou et al. [29] separately trained a linear classifier on the feature vectors for class-specific gradients [17]. However, this method is limited to the existing classes, and it is inconvenient to train additional classifiers for every model. Instead, FAM can easily solve the above issues.

As an example, a HRNet-W30 [19] model is trained over 200 epochs on the ImageNet dataset [13] by BYOL [6], which is a state-of-the-art algorithm for self-supervised representation learning. As showed in Figure 6, we visualize the self-supervised models without additional classifiers at different training epochs by Grad-FAM. For the model at 80 epochs, the foreground objects have got high attention, but the salient regions are spread across the background at this point. At 140 epochs, the attention on the background get lower. Finally, the salient regions are localized to a small region at 200 epochs. See Appendix for more examples.

Analyzing the change of Grad-FAM results in different training stage can help researchers get a better understanding of self-supervised representation learning. Furthermore, the FAM methods can be employed for analyzing the difference among various pretext tasks and different model architectures in self-supervised representation learning, which will be explored in future work.

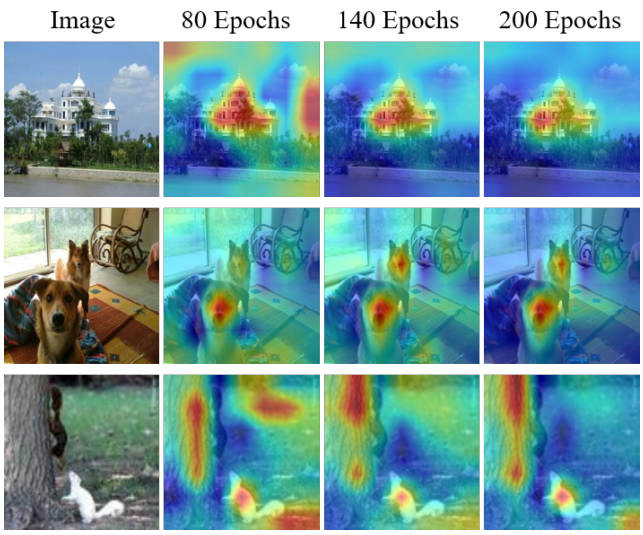


Figure 6. Visualizing the training process of the HRNet-W30 model trained by BYOL. Given images for the models at 80, 140, 200 epochs in self-supervised training, we visualize the change of salient regions generated by Grad-FAM. The model at 80 epochs has noticed the foreground objects, but is prone to distraction by backgrounds at this point. In the subsequent training process, the salient regions are gradually localized to a small region.

5. Conclusion

In this work, we proposed a novel visual explanation paradigm FAM, including Grad-FAM and Score-FAM, to explain the embedding process of CNN-based models. Our methods address a principal shortcoming of previous approaches, which are based on a target neuron or similarity with other image, and fail to interpret the feature vector completely. Contrarily, FAM methods locate the regions of inputs that contribute most to the global feature vectors. We validate the effectiveness of our methods both objectively and subjectively on Person Re-ID. Experimental results on Market1501 and CUHK03 demonstrate that Score-FAM achieves a much better performance than the current state-of-the-art explanation approaches. Furthermore, Score-FAM can be employed to analyze other vision tasks without classifiers, such as self-supervised representation learning.

Future work will be to explore deeper connection between the feature vectors and inputs, and combine FAM with other traditional approaches and more vision tasks.

References

[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 1, 6

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[3] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):392–408, 2017. 4

[4] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 3

[5] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 2

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 8

[7] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 2, 5, 7

[8] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5

[9] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015. 3

[10] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 3

[11] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015. 3

[12] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020. 1, 3

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5, 8

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 3, 5

[15] Dong Shen, Shuai Zhao, Jinming Hu, Hao Feng, Deng Cai, and Xiaofei He. Es-net: Erasing salient parts to learn more in re-identification. *IEEE Transactions on Image Processing*, 30:1676–1686, 2020. 1, 4

[16] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017. 1

[17] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014. 2, 8

[18] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2

[19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 8

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 1, 2

[21] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6021–6029, 2020. 7

[22] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 1, 3, 4, 5

[23] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019. 1, 3

[24] Mang Ye, Xiangyuan Lan, and Pong C Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–186, 2018. 4

[25] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-

identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[26] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015. 2

[27] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[28] Qinglong Zhang, Lu Rao, and Yubin Yang. A novel visual interpretability for deep neural networks by optimizing activation maps with perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3377–3384, 2021. 5

[29] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10990–10998, 2021. 8

[30] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2, 5, 7

[31] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 4

[32] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. 2017. 5

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2

[34] Sijie Zhu, Taojiannan Yang, and Chen Chen. Visual explanation for deep metric learning. *IEEE Transactions on Image Processing*, 30:7593–7607, 2021. 1