Expert Systems With Applications 211 (2023) 118635

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Zero-shot face recognition: Improving the discriminability of visual face features using a Semantic-Guided Attention Model

Cristiano Patrício *, João C. Neves

University of Beira Interior, Portugal
NOVA LINCS, Portugal

ARTICLE INFO

Keywords:
Face recognition
Zero-shot learning
Attention-based models

ABSTRACT

Zero-shot learning enables the recognition of classes not seen during training through the use of semantic information comprising a visual description of the class either in textual or attribute form. Despite the advances in the performance of zero-shot learning methods, most of the works do not explicitly exploit the correlation between the visual attributes of the image and their corresponding semantic attributes for learning discriminative visual features. In this paper, we introduce an attention-based strategy for deriving features from the image regions regarding the most prominent attributes of the image class. In particular, we train a Convolutional Neural Network (CNN) for image attribute prediction and use a gradient-weighted method for deriving the attention activation maps of the most salient image attributes. These maps are then incorporated into the feature extraction process of Zero-Shot Learning (ZSL) approaches for improving the discriminability of the features produced through the implicit inclusion of semantic information. For experimental validation, the performance of state-of-the-art ZSL methods was determined using features with and without the proposed attention model. Surprisingly, we discover that the proposed strategy degrades the performance of ZSL methods in classical ZSL datasets (AWA2), but it can significantly improve performance when using face datasets. Our experiments show that these results are a consequence of the interpretability of the dataset attributes, suggesting that existing ZSL datasets attributes are, in most cases, difficult to be identifiable in the image. Source code is available at https://github.com/CristianoPatricio/SGAM.

1. Introduction

In the supervised learning paradigm, classifiers are trained with a large number of images containing all the classes intended to be recognized during the test phase. However, for some problems, collecting a wide variety of samples for all the classes is unpractical (e.g., the problem of bird species recognition). Also, it may even be unfeasible to obtain samples for all target classes, being impossible for standard classifiers to identify samples not represented in the training set correctly. To overcome these problems, ZSL has been introduced as an alternative learning paradigm for recognizing unseen objects using solely the semantic description of the target classes.

Since the seminal work of Larochelle et al. (2008), where the initial concept of identifying classes not represented in the training set was introduced, several approaches have been introduced to improve the performance of ZSL methods. However, these approaches focused primarily on improving the process of learning a mapping between visual features and class attributes. The question of whether the extracted features possess discriminative information regarding semantic information of the image class has hardly ever been considered. Even though some approaches attempted to improve the discriminability of image features through the use of attention models, these approaches focused primarily on fine-grained datasets, whereas general purpose ZSL datasets have been disregarded in the evaluation of these methods.

Accordingly, in this paper, we build on the work of Liu et al. (2021) to introduce a semantic guided attention model for increasing the discriminability of visual features regarding the semantic information in the ZSL problem. In particular, we train a CNN for estimating the semantic information of an image and contrary to the work of Liu et al. (2021), we use a gradient-weighted method for deriving the attention activation maps for each attribute of the test image. These maps are then combined with the activation maps of the feature extraction network, aiming to increase the discriminability of the features provided to ZSL methods. Our work differs from the Liu et al. (2021) in two points: first, we propose a semantic-guided attention model to improve the discriminability of features extracted from images; second, these semantic-guided features can be used in arbitrary ZSL methods in

* Corresponding author at: University of Beira Interior, Portugal.
E-mail addresses: cristiano.ppatricio@gmail.com, cristiano.patricio@ubi.pt (C. Patrício), jcneves@di.ubi.pt (J.C. Neves).

https://doi.org/10.1016/j.eswa.2022.118635
Received 10 February 2022; Received in revised form 16 July 2022; Accepted 17 August 2022
Available online 24 August 2022

order to improve their recognition performance, while Liu et al. (2021) focused only on a particular ZSL method.

We show that the proposed method can increase the performance of zero-shot recognition in face imagery, but the same does hold in classical ZSL datasets. Based on this observation, we argue that the class attributes and visual attributes are weakly correlated in general-purpose ZSL datasets, meaning that these attributes are hardly interpretable from visual inspection of the image. To validate this hypothesis, we carry out an extensive visual analysis of the attribute attention maps obtained both in face datasets and in standard ZSL datasets. It is possible to observe that the attention maps obtained using face images highlight relevant image regions corresponding to the class attributes. In contrast, in the case of ZSL datasets, attention maps tend to focus on arbitrary regions.

Accordingly, the major contributions are the following: (1) An attention-based strategy for improving the performance of zero-shot face recognition methods; (2) The proposed approach can be used for assessing the interpretability of the semantic information used in ZSL datasets. (3) Our experiments show that some ZSL datasets have been annotated with hard to interpret attributes, whereas the CelebA (Liu et al., 2015) and LFWA (Liu et al., 2015) face datasets comprises attributes that can be easily correlated with specific image regions. This is a significant conclusion to guide further developments in zero-shot recognition.

The remainder of the paper is organized as follows: Section 2 summarizes the related work in the scope of our work. Section 3 introduces the proposed Semantic-Guided Attention Model. Section 4 is devoted to analyze and discuss the experimental results. Finally, conclusions and future work are drawn in Section 5.

2. Related work

2.1. ZSL problem

Zero-shot learning aims to recognize object classes not seen during the training phase, using a shared semantic space to transfer the knowledge from seen classes to unseen classes. Formally, there are $Y = Y^s \cup Y^u$ classes, where Y^s corresponds to the **s**een classes and Y^u the set of **u**nseen classes, with $Y^s \cap Y^u = \emptyset$. Similarly, $S = S^s \cup S^u$ denotes the corresponding seen and unseen class semantic representations (e.g. attribute vector), used as ancillary information to map from visual to semantic space. Given a set of images $X = X^s \cup X^u$, the goal of ZSL is to learn a classifier $f : X^u \rightarrow Y^u$ to predict the label of the image sampled from unseen classes, without having access to Y^u during training.

Early works of ZSL focus on estimating a set of attributes regarding the input image and then inferring the class label by measuring the probability that the predicted attributes match the semantic attributes of unseen classes. Lampert et al. (2009) introduced one of the first ZSL approaches where a classifier is trained for predicting a specific image attribute, such that in the test phase, class prediction is performed by matching the estimated attributes with the semantic information of the target classes.

Later, the notion of projection-based methods was introduced where the goal was to learn a projection function to map visual features directly to semantic space or vice-versa. Palatucci et al. (2009) proposed a Semantic Output Codes (SOC) classifier for the neural decoding task by learning a regression function to map from the visual feature space to the semantic space. The classification of the projected point is carried out in the semantic space using the nearest neighbour classifier. The learning of compatibility functions between visual space and semantic space was explored in Akata et al. (2013), Frome et al. (2013). The adoption of the encoder–decoder paradigm to map between visual feature space and semantic space was introduced by Kodirov et al. (2017). Since both the encoder and decoder are symmetric and linear models, they enable an extremely efficient learning algorithm with a low computational cost. Zhang et al. (2017) proposed the first end-to-end ZSL model by training a CNN for producing visual descriptors related to its correspondent class semantic representation vector. In order to improve the separability of seen and unseen samples in the latent space, Ding et al. (2021) proposed a semantic encoding out-of-distribution classifier which learns a bounded manifold for each seen class to repel the unseen samples from the seen classes in a joint embedding space.

Despite the success of the projection-based methods in addressing the zero-shot learning problem, most of them fail to achieve reasonable results for the generalized zero-shot learning, where both seen and unseen classes are available in the test phase. Moreover, since the projection function is learned using data solely from seen classes during training, the learned projection function is usually biased towards these classes, hampering the generalization capability of the classifier. To mitigate these problems, recent ZSL methods rely on generative approaches to synthesize image features for unseen classes conditioned on their semantic representations. This is achieved by using a Conditional Generative Adversarial Network (CGAN), a Conditional Variational Autoencoder (CVAE) or a fusion of both (VAE-GAN).

2.2. Generative models

Generative approaches have proven to be effective in addressing the zero-shot learning problem. Xian, Lorenz, et al. (2018) introduced the idea of casting the ZSL problem as a traditional supervised classification problem by training a CGAN for synthesizing visual features for the unseen classes conditioned on semantic information of these classes. At the test phase, a softmax classifier can be used to perform the class predictions. Following the same idea, but using a different generative model, Mishra et al. (2018) proposed to learn a CVAE to generate visual features for the unseen classes based on the given attributes. Xian et al. (2019) developed a conditional generative model that combines the strength of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) by bringing them together to generate visual features from semantic information. In the same fashion, Ma and Hu (2020) proposed a modified VAE composed of two decoders to achieve a cross-modal alignment between the class embeddings and image features, which proved to be effective in learning a more discriminative latent space. Recently, following these ideas, several approaches have been proposed in order to improve the consistency of generated features by using feedback loops (Narayan et al., 2020) or contrastive embedding (Han et al., 2021).

2.3. Semantic-guided attention localization

The learning of discriminative visual features, especially in fine-grained scenarios, has been shown to significantly improve the zero-shot classification (Zhu et al., 2019). The work of Zhu et al. (2019) introduces a new approach for learning discriminative visual features guided by the semantic attributes annotated per image. They propose a multi-attention localization model for producing attention maps that capture the discriminative parts of the image regarding the semantic attributes. Then, these crucial regions are cropped and fed into a CNN-based network to extract the visual feature vector for each discriminative part. However, cropping such parts requires an additional network for learning the approximate region to perform the cropping operation. Contrary to this approach, we suggest the use of Grad-CAM (Selvaraju et al., 2017) for holistically producing class-discriminative localization maps. Therefore, we can obtain an Attribute Attention Map (AAM) through the maximum operation over each produced Grad-CAM map. Then, the new semantic-guided feature map is obtained with both the global category-level and the local attribute-level discrimination feature map.

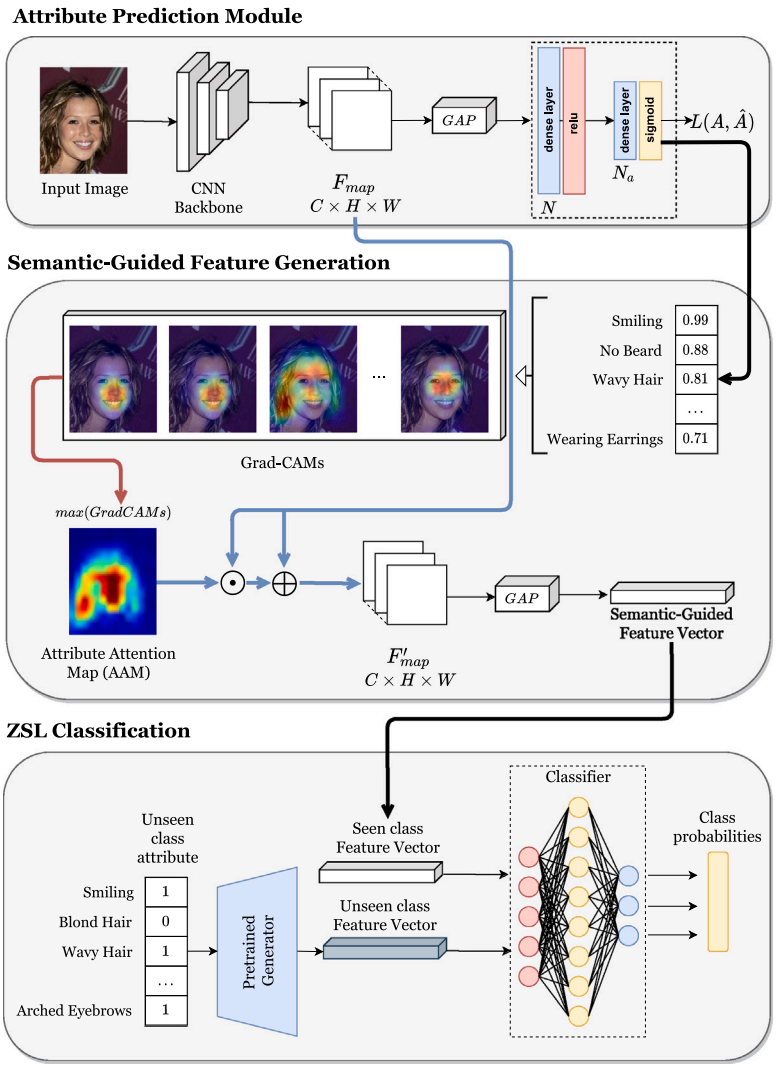

Fig. 1. Proposed Semantic-Guided Attention Model. The input image is fed into a CNN for estimating its semantic information, represented through N attributes. The Grad-CAM maps are generated regarding the k highest confidence attributes, and the Attribute Attention Map (AAM) is obtained as the maximum operation over the k Grad-CAM maps. The AAM is incorporated into the network by weighting the original feature maps over the produced AAM. The final semantic-guided feature vector is the result of the average pooling operation over the new feature maps. A generative ZSL method is trained using the semantic-guided feature vectors, comprising a generator to synthesize the class-discriminative samples for the unseen classes and a softmax classifier for final class prediction.

3. Semantic-guided attention model

The proposed Semantic-Guided Attention model is depicted in Fig. 1, and it consists essentially of three main phases: (1) Attribute prediction, (2) Attribute Attention Map (AAM) generation, and (3) Training of a generative ZSL approach using semantic-guided features. In the following sections, we will discuss the details behind each phase.

3.1. Attribute prediction module

The first stage of our proposed method consists of learning an attribute prediction model. The rationale behind this stage is twofold: (1) obtain the most prominent attributes in the image; (2) indirectly infer from the network the image regions that contribute the most to the attribute score prediction.

For this, we adopt the VGGFace model as the backbone, while the final layers are modified to match the number of attributes. The first layer has the size of the channels of the last convolutional layer, followed by a sigmoid activation layer with N neurons, where N corresponds to the number of the attributes to be predicted.

As depicted in Fig. 1, the input image is fed into the network, yielding a $C \times H \times W$ feature map. The feature map is down-sampled to $C \times 1 \times 1$ by average-pooling operation and reshaped into $C \times 1$ to match the dimension of the first layer of the attribute prediction model. The output vector is then normalized by a sigmoid activation layer, producing the probability confidence score of each attribute.

The model was optimized by the Adadelta algorithm with the default hyperparameters. Furthermore, we performed data augmentation on both the training and validation set. The chosen loss function was the binary cross-entropy loss, which is defined as:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (1)$$

where N is the output size, $\hat{y}_i$ is the predicted value, and y_i is the corresponding target value.

3.2. Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) is a technique that allows generating a class-discriminative localization map using any CNN-based network. By using the gradient information of any target class, Grad-CAM produces a map with the importance of each region in the final prediction, which can be

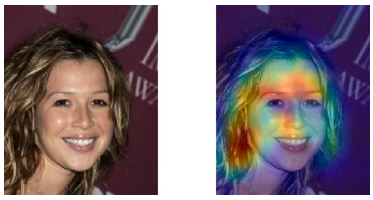

Fig. 2. Grad-CAM Original image (at left) representing a woman's face. Modified image (at right) overlapped with the generated Grad-CAM highlighting the face attribute "Wavy Hair". The highlighted regions on the heat-map evidence that the network "focused" on the "hair" region to predict the "Wavy Hair" class.

interpreted as a visual explanation of the decision process used by the network for predicting a specific target class. This technique is a generalization of the Class Activation Mapping (CAM) (Zhou et al., 2016) approach. The main difference between Grad-CAM and CAM is that the CAM requires a particular kind of CNN architecture. Specifically, an architecture that performs global average pooling (GAP) over convolutional maps, followed by the prediction layer (i.e., conv feature maps → GAP → softmax layer) (Selvaraju et al., 2017). In contrast, Grad-CAM can be applied to a wide range of CNN networks without requiring modifications in the network architecture.

The pipeline for producing a class-discriminative localization map Grad-CAM for any image embraces three main steps, which are summarized below.

1. **Calculate the gradient**. The gradient is calculated for the score y^c of the class c with respect to feature maps A^k of the last convolutional layer. The gradient matrix G for the class c is defined as follows:

$$G_k = \frac{\partial y^c}{\partial A^k} \tag{2}$$

with $G_k \in \mathbb{R}^{u \times v}$, $A^k \in \mathbb{R}^{u \times v}$, and $k \in \mathbb{N}$.

2. **Calculate the neuron importance weights** α_k^c. The gradients are then average-pooled in order to obtain the neuron importance weights α_k^c that captures the "importance" of the feature map k regarding the class c.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k} \tag{3}$$

and $\alpha_k^c \in \mathbb{R}^{1 \times 1}$.

3. **Calculate the final Grad-CAM map**. The computed weights α_k^c are then weighted with each of the feature maps, followed by a (Rectified Linear Unit) ReLU operation to obtain the final Grad-CAM map:

$$GradCAM_c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

and $GradCAM_c \in \mathbb{R}^{u \times v}$.

ReLU is applied in Eq. (4) because the final Grad-CAM map only captures the features that "have a positive influence on the class of interest" (Selvaraju et al., 2017).

Moreover, the final Grad-CAM map can be viewed as a heat-map that highlights the regions where the target class is present. Fig. 2 shows the generated Grad-CAM map for an image taken from CelebA (Liu et al., 2015) dataset.

However, the process of generating Grad-CAMs for each of the attributes can be computationally expensive. A suitable way to obtain good results is to select only the top-k most influential attributes (Liu et al., 2021). Fig. 3 depicts the generated Grad-CAMs for the top-10 most influential attributes regarding the input image. It is interesting to observe that the final generated AAM is an almost perfect outline of the face image due to the maximum operation performed along with the produced Grad-CAM maps.

3.3. Discriminative feature calculation

The Attribute Attention Map (AAM) is used jointly with the feature maps of the last convolutional layer to obtain a new feature map to obtain the final semantic-guided feature vector.

Actually, the new feature maps $F'_{map} = (F_{map} \odot AAM) + F_{map}$ are an improvement of the original feature maps F_{map}, since discriminative information is added by the AAM. Thus, the semantic-guided feature vector f' is calculated by the global average pooling operation to F'_{map} in order to match the dimension of the input layer of the generative method ($C \times 1$). More formally:

$$f' = GAP\left(\left(F_{map} \odot AAM\right) + F_{map}\right) \tag{5}$$

where GAP stands for Global Average Pooling operation.

3.4. Generative method

After computing the discriminative visual features, these can be used as input for a ZSL model. As discussed before, generative methods are currently state-of-the-art in ZSL due to their effectiveness, particularly in the generalized scenario. For this reason, our model is flexible to incorporate any generative method. However, a non-generative method can also be used. Despite the original purpose of the GAN networks being related to the process of generating images, in ZSL problems, the generative approaches aim to synthesize visual feature vectors conditioned on the given attributes. Furthermore, Xian, Lorenz, et al. (2018) proved that generating CNN features instead of images leads to a significant increase in performance of both ZSL and GZSL settings since visual features retain more discriminative information than a synthesized image produced by a GAN.

Finally, the synthesized feature vectors are then assigned to each unseen class to train a softmax classifier jointly with the samples of seen classes to infer the final predictions.

4. Experimental methodology

4.1. Datasets

The proposed model was tested on four datasets: *CelebFaces Attributes* (CelebA) (Liu et al., 2015), LFWA (Liu et al., 2015), *Animals with Attributes 2* (AWA2) (Xian, Lampert, et al., 2018), *CUB Birds 200-2011* (CUB) (Wah et al., 2011), and *Large-scale Attribute Dataset* (LAD) (Zhao et al., 2019). CelebA is a large-scale face attributes dataset with 202,599 face images divided by 10,177 identities. Each image is annotated with 40 binary attributes. Furthermore, we decided to split the dataset into two different splits, each one containing a different number of classes. LFWA has 13,233 images of 5,749 identities, each annotated with 40 binary attributes. AWA2 consists of 37,322 images of 50 classes of animals, each one annotated with 85 attributes. CUB is a fine-grained dataset comprising 200 classes of bird species, at a total of 11,788 images annotated with 312 attributes. LAD has 78,017 images of 5 super-classes, totalling 230 classes. Each instance is annotated with 359-dimensional binary attributes of visual, semantic and subjective properties. Table 1 provides the detailed statistics for the CelebA dataset.

Moreover, we adopt the Proposed Split (PS) (Xian, Lampert, et al., 2018) to ensure that none of the test classes appear in the dataset used to pre-train the CNN models. For semantic embeddings, we adopt the class-level attributes provided by Xian, Lampert, et al. (2018) for AWA2 (85-dim), CUB (312-dim). In the case of the CelebA and LAD datasets, we use the 40-dimensional and 359-dimensional binary attribute annotations provided by the authors, respectively. We conduct all experiments under the inductive setting, in which only labelled instances of seen classes are considered at the training phase (Xian, Lampert, et al., 2018).

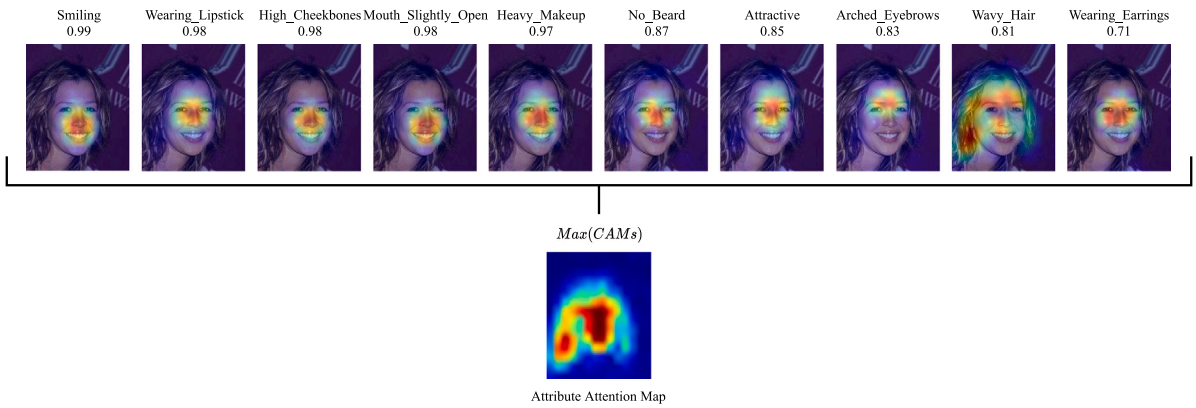

Fig. 3. Pipeline for generating the Attribute Attention Map (AAM) for an example face image of the CelebA (Liu et al., 2015) dataset. Each of the generated Grad-CAM focuses on the face region that is visually closest to the attribute prediction. The final AAM captures all the Grad-CAM highlighted regions, resulting in an almost perfect face outline.

Table 1
Statistics for CelebA (Liu et al., 2015) dataset in terms of number of attributes (Att), number of classes in Y_{tr} and Y_{ts}, and number of images at training and test time.

Dataset	Att	Y	Number of images			At training time	At test time	
			Y_{tr}	Y_{ts}	Total	Y_{tr}	Y_{tr}	Y_{ts}
CelebA(All)	40	10177	8142 + 814	1221	202599	143597	34912	24090
CelebA(500)	40	500	300 + 100	100	15038	11232	800	3006

4.2. Data preprocessing

Preprocessing phase involved normalizing each input image to a pixel value range of $[0, 1]$ and performing data augmentation, including random transformations, such as rotation, horizontal flip and zoom. Regarding the CelebA dataset, each image was cropped using the bounding boxes coordinates provided by Liu et al. (2015). For the LFWA dataset, deep funneled images (Huang et al., 2012) were used over the original images since deep funneled images are aligned versions of original images that proved to improve the performance on recognition tasks (Huang et al., 2012).

4.3. Methods

To evaluate the proposed approach, four ZSL methods were selected, including ESZSL (Romera-Paredes & Torr, 2015), SAE (Kodirov et al., 2017), DEM (Zhang et al., 2017), and the state-of-the-art TF-VAEGAN (Narayan et al., 2020) generative method.

ESZSL (Romera-Paredes & Torr, 2015) is a representative approach of the use of inner product similarity for determining the similarity between the image features and image classes. Similarity scores for test classes are inferred through a set of linear transformations encoded by matrix multiplication operations. The reduced complexity of these transformations, and the possibility of parallelizing matrix multiplication, ensures a short inference time.

SAE (Kodirov et al., 2017) treats ZSL as an encoding–decoding problem by using an auto-encoder (AE) to transform the visual features to the semantic space and subsequently recover the same features from the class semantic representation. This strategy significantly outperformed state-of-the-art and allowed to classify an image either by inferring its semantic representation from the visual features or transforming the semantic representation of target classes to the visual feature space and then using k-NN in the feature space. The use of a linear AE allows a very fast attribute estimation. However, the overall inference time depends mainly on the size of unseen classes due to the use of k-NN.

DEM (Zhang et al., 2017) was one of the first end-to-end ZSL models where the CNN and the feature projection function were jointly optimized. This strategy significantly increases training time, but the classification inference phase only depends on the size and number of fully connected layers.

TF-VAEGAN (Narayan et al., 2020) constitutes an improvement over traditional generative approaches since it increases the semantic consistency of generated features by enforcing the decoding of generated features to be similar to original semantic embeddings. This semantic embedding decoder is also used during the inference stage for providing the feature classifier with both extracted features and their decoded semantic embeddings, which slightly increases the inference time over f-CLSWGAN.

4.4. Evaluation protocols

In order to measure the performance of ZSL methods, we adopt the top-1 accuracy to measure the average per-class top-1 accuracy under the ZSL setting, defined as:

$$acc_y = \frac{1}{|Y|} \sum_{c=1}^{|Y|} \frac{\#correct\ predictions\ in\ c}{\#samples\ in\ c} \tag{6}$$

where Y is the total number of classes.

5. Results and discussion

This section provides the results obtained in the four datasets considered in our experiments: CelebA, AWA2, LAD and CUB. For each dataset, we provide a quantitative evaluation of the proposed strategy when coupled with different ZSL methods. Additionally, for some datasets, a qualitative evaluation of the proposed strategy is also provided.

5.1. CelebA: Quantitative evaluation

Based on the results presented in Table 2, we can infer that the use of the proposed Semantic Guided Attention model (SGAM) increases the classification accuracy by 13.7%, on average, if we consider the most populated 500 classes. On the other hand, the classification accuracy is boosted by 6.45%, on average, if the total of classes is considered. Due to the richness and visual interpretability of attributes, the attention model is very suitable for this dataset.

Table 2
Comparative analysis of the performance of ZSL methods when using the proposed strategy on the CelebA dataset.

Methods	CelebA(500)		CelebA(All)	
	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM
SAE (Kodirov et al., 2017)	7.2	**19.1**	1.3	**3.8**
ESZSL (Romera-Paredes & Torr, 2015)	15.7	**32.5**	4.1	**10.2**
DEM (Zhang et al., 2017)	13.1	**25.6**	2.0	**8.5**
TF-VAEGAN (Narayan et al., 2020)	7.9	**21.5**	6.8	**17.5**

The original performance of each method, i.e., without the proposed SGAM, is reported in the column marked with "w/o. SGAM", whereas the results using the SGAM are provided in the column marked with "w. SGAM". SGAM refers to Semantic Guided Attention Model. Also, the results are provided for the complete version of the dataset "CelebA(All)" and for a version comprising only the most populated 500 classes denoted as "500". The results report top-1 accuracy in %.

Table 3
Comparative analysis of the performance of ZSL methods when using the proposed strategy on the LFWA dataset.

Method	w/o. SGAM	w. SGAM
SAE (Kodirov et al., 2017)	1.9	**2.9**
ESZSL (Romera-Paredes & Torr, 2015)	2.9	**3.5**
DEM (Zhang et al., 2017)	2.7	**4.7**
TF-VAEGAN (Narayan et al., 2020)	4.7	**5.1**

The original performance of each method, i.e., without the proposed SGAM, is reported in the column marked with "w/o. SGAM", whereas the results using the SGAM are provided in the column marked with "w. SGAM". SGAM refers to Semantic Guided Attention Model. The results report top-1 accuracy in %.

Table 4
Zero-shot learning results on AWA2 using the proposed method.

Method	w/o. SGAM	w. SGAM
SAE (Kodirov et al., 2017)	52.89	**52.93**
ESZSL (Romera-Paredes & Torr, 2015)	55.89	51.94
DEM (Zhang et al., 2017)	59.68	48.07
TF-VAEGAN (Narayan et al., 2020)	66.32	**66.56**

The results obtained without the proposed SGAM are reported in the column marked with "w/o. SGAM", whereas the results using the SGAM are along the column marked with "w. SGAM". SGAM refers to Semantic Guided Attention Model. The results report top-1 accuracy in %.

5.2. CelebA: Qualitative evaluation

Fig. 3 provides a visual explanation of the performance improvement observed when using the proposed approach. Considering that the Grad-CAMs generated to highlight the most relevant regions for predicting the attributes of the image, it is only natural that the AAM improves the discriminability of the features produced by the network and consequently the ZSL accuracy, as evidenced by the results present in Table 2.

5.3. LFWA: Quantitative evaluation

As evidenced by the results in Table 3, the proposed SGAM increase the performance in all evaluated ZSL methods. It is noted that only the most 50 populated classes were considered for the unseen classes since more than 70% of all classes contain only one sample. This class imbalance issue justifies the low accuracy in recognizing unseen classes and the low margin of the performance improvement between the results obtained with the SGAM and without the SGAM.

5.4. LFWA: Qualitative evaluation

The visual inspection of Fig. 8 confirms the usefulness of the generated AAM in highlighting relevant facial regions regarding the predicted attributes compared to the resulting heatmap from the top-1 predicting using the VGGFace feature extractor. In a similar fashion to CelebA, this enables the increase of the recognition performance since the final feature vector includes more discriminative information than its original version using only the top-1 prediction of the VGGFace.

5.5. AWA2: Quantitative evaluation

The proposed strategy was also applied to AWA2, a traditional ZSL dataset. In our experiments, we considered the MobileNetV2 architecture for being the CNN network with an acceptable accuracy/speed trade-off (Patrício & Neves, 2021). Surprisingly, the results reported in Table 4 evidence that there is no significant improvement over the standard features, whereas, in some methods, the proposed strategy degrades the method accuracy. We argue that these results are mainly due to the properties of the semantic information provided in the dataset. The proposed method is effective when the annotated attributes are visually interpretable, i.e., possess some correlation with specific image regions. For example, AWA2 comprises attributes such as "smart", "solitary" and "domestic", which represent high-level information about the animal in the image, and are unrelated to any specific image region. This lack of interpretability in some of the attributes of AWA2 leads to the generation of sparse attention maps, which consequently degrade the discriminability of the features generated by the proposed approach, justifying thus the performance degradation in this dataset. Also, these results suggest that this classical benchmark of ZSL methods may be a source of bias in the evaluation of ZSL approaches since there is no clear relation between visual attributes and the semantic embeddings provided with the data.

5.6. AWA2: Qualitative evaluation

To justify the results obtained on the AWA2 dataset, Fig. 4 depicts the generated Grad-CAM maps for a horse image, whose visual inspection shows that the attribute predictions are quite random and meaningless regarding the horse class. As a result, the obtained AAM has a lot of noise information instead of having discriminative information of the regions related to the attributes. As a consequence, the features discriminability decreases, compromising the classifier accuracy.

5.7. LAD: Quantitative evaluation

The evaluation protocol of LAD diverges from the remaining datasets since the performance of a method corresponds to the average accuracy on all super-classes (Animals, Fruits, Vehicles, Electronics, and Hairstyles) on the given five splits. The results obtained for the LAD dataset are provided in Table 5. Comparing the original performance of the ZSL methods with the performance obtained when using SGAM shows that the proposed strategy could not improve the original results. We argue that these results can be partially explained by the interpretability of the semantic information provided in the dataset and simultaneously due to the use of an over-complete set of attributes, i.e., the attributes of all super-classes are used globally for all images independently of their super-class. This use of a highly diverse set of attributes increases the probability of incorrect predictions, as can be observed by Fig. 9, where the Vehicles attributes have overpowered the Fruits attributes. This creates for some classes an incorrect AAM, explaining why the performance decreases for most ZSL methods.

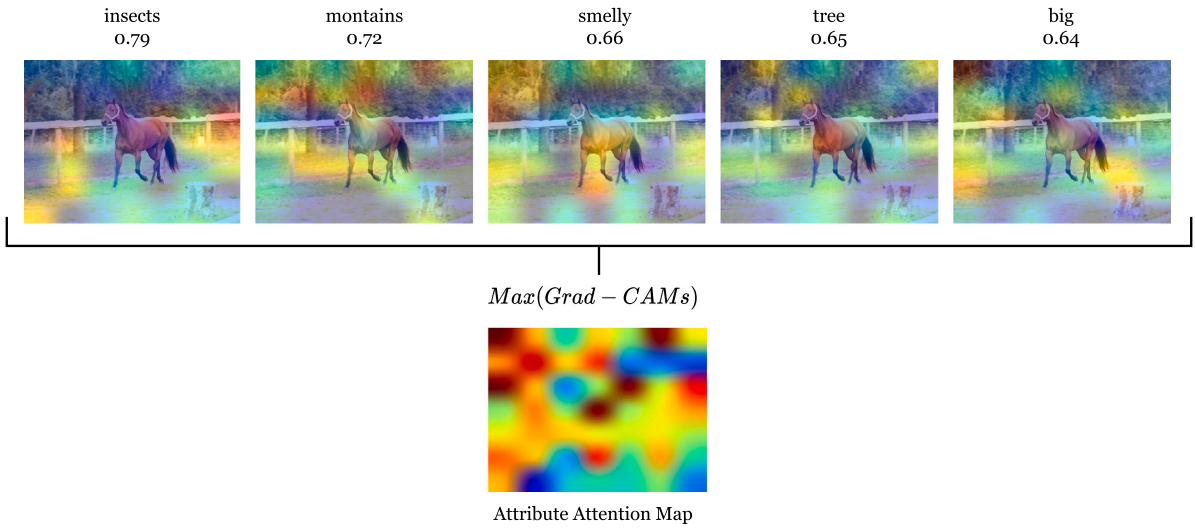

Fig. 4. Attribute Attention Map for a horse image. The predicted attributes are not visually interpretable enough for generating a class-discriminative Grad-CAM. Instead of the generated AAM focusing on the horse class important regions, it highlights other useless parts.

Table 5
Zero-shot learning results on LAD using the proposed method.

Methods	Animals		Fruits		Vehicles		Electronics		Hairstyles		Average	
	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM
SAE (Kodirov et al., 2017)	45.09	50.47	27.67	29.06	55.73	58.37	34.76	34.86	38.04	37.79	40.26	**42.11**
ESZSL (Romera-Paredes & Torr, 2015)	**65.72**	64.84	40.03	38.43	65.90	63.74	37.80	38.15	**41.19**	41.07	50.13	49.25
DEM (Zhang et al., 2017)	62.10	49.80	42.12	29.95	65.50	51.55	40.53	32.94	41.71	35.12	50.39	39.87
TF-VAEGAN (Narayan et al., 2020)	64.53	63.99	**48.07**	42.57	**67.06**	62.81	**40.62**	36.45	40.93	36.80	52.24	48.52

The results obtained without the proposed SGAM are reported in the column marked with "w/o. SGAM", whereas the results using the SGAM are along the column marked with "w. SGAM". SGAM refers to Semantic Guided Attention Model. The results report top-1 accuracy in %.

5.8. CUB: Quantitative evaluation

CUB Birds is a fine-grained dataset comprising 200 bird species. Each image is annotated with 312 attributes. However, we can divide these 312 attributes into 27 super-group of attributes. For example, the attributes "*has_head_pattern::eyebrow*", "*has_head_pattern::eyering*", "*has_head_pattern::plain*", "*has_head_pattern::eyeline*", "*has_head_pattern::striped*", and "*has_head_pattern::capped*" are all concern to the super attribute "*has_head_pattern*", the only difference between them is the type of head pattern; thereby, we can consider these six attributes as a super-attribute "*has_head_pattern*". Since variability is reduced in the dataset, the Semantic-Guided Attention Model is expected to work similarly to the CelebA in this dataset. Consequently, we have applied the SGAM strategy to the CUB Birds dataset.

The CNN architecture used to extract features of the CelebA dataset was the VGG-Face, which is trained on face images. To mimic this behaviour, we fine-tuned the top layers of the DenseNet (Huang et al., 2017) architecture to recognize the bird species of the NABirds (Van Horn et al., 2015) dataset, which is a large-scale dataset with 48,000 annotated images of the 400 species of birds that are commonly observed in North America.

In our experiments, we performed central cropping on the images so that the height and width were equal, and additional importance was given to the central part of the attention maps by applying a Gaussian mask to the AAM. This allowed smoothing of the areas that were not important for the prediction. Finally, the attribute prediction network was trained using a custom loss. Each super attribute shares the same weight in the loss value, rather than weighing all attributes similarly. The results obtained are reported in Table 6.

The attribute classifier achieved a top-1 accuracy of 91%. Fig. 5 shows the predicted attributes for a given bird image and the corresponding generated Grad-CAM. Through visual inspection, it is possible to observe that the predicted attributes are highly correlated with the highlighted regions on the generated Grad-CAM.

Table 6
Zero-shot learning results on CUB using the proposed method.

Method	w/o. SGAM	w. SGAM
SAE (Kodirov et al., 2017)	48.29	**48.75**
ESZSL (Romera-Paredes & Torr, 2015)	62.47	**62.56**
DEM (Zhang et al., 2017)	66.94	59.05
TF-VAEGAN (Narayan et al., 2020)	74.42	73.73

The results obtained without the proposed SGAM are reported in the column marked with "w/o. SGAM", whereas the results using the SGAM are along the column marked with "w. SGAM". SGAM refers to Semantic Guided Attention Model. The results report top-1 accuracy in %.

6. Ablation study

In this section, a more detailed visual examination will be conducted to understand the effectiveness of the Attribute Attention Maps (AAM), obtained through the Semantic-Guided Attention Model, in producing discriminative visual features.

6.1. Effect of the Attribute Attention Map

AWA2. As previously discussed in Section 5.5, in the case of the AWA2 dataset, the obtained AAM focuses on randomly distributed parts on the image which are outside of the relevant region regarding the image class. This way, the produced visual feature (semantic-guided feature vector) contains useless information for image classification. Fig. 6 shows the AAM generated by the proposed approach and the activation map obtained for the top-1 ImageNet class prediction of the pre-trained MobileNetV2. The comparison between the two heatmaps shows that the activation map obtained for the off-the-shelf CNN concentrates on regions related to the image class, whereas the generated AAM disregards discriminative regions of the image while simultaneously

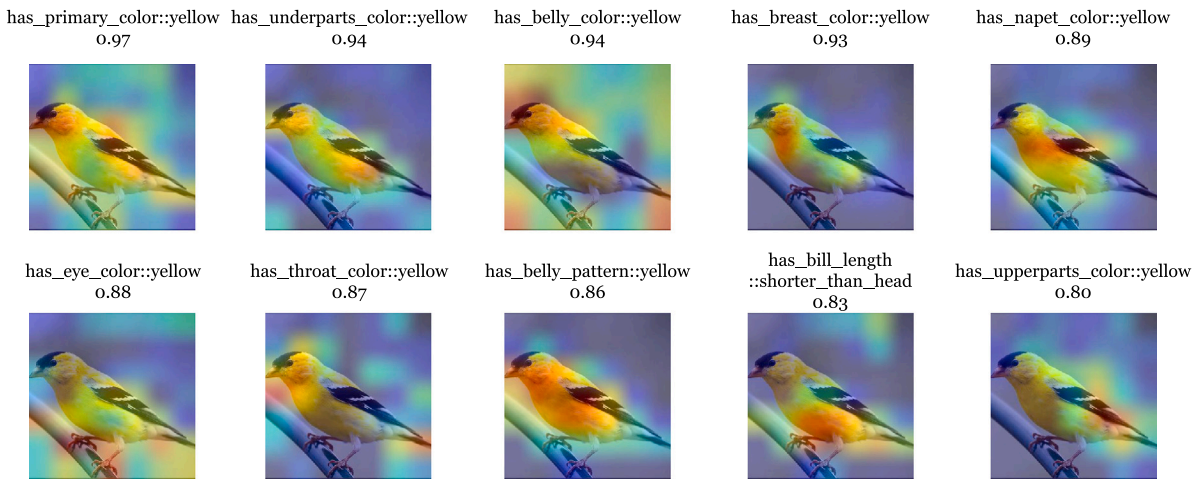

Fig. 5. Attribute visualization maps for a bird image. For predicted attributes for an image of a bird along with the generated Grad-CAM maps.

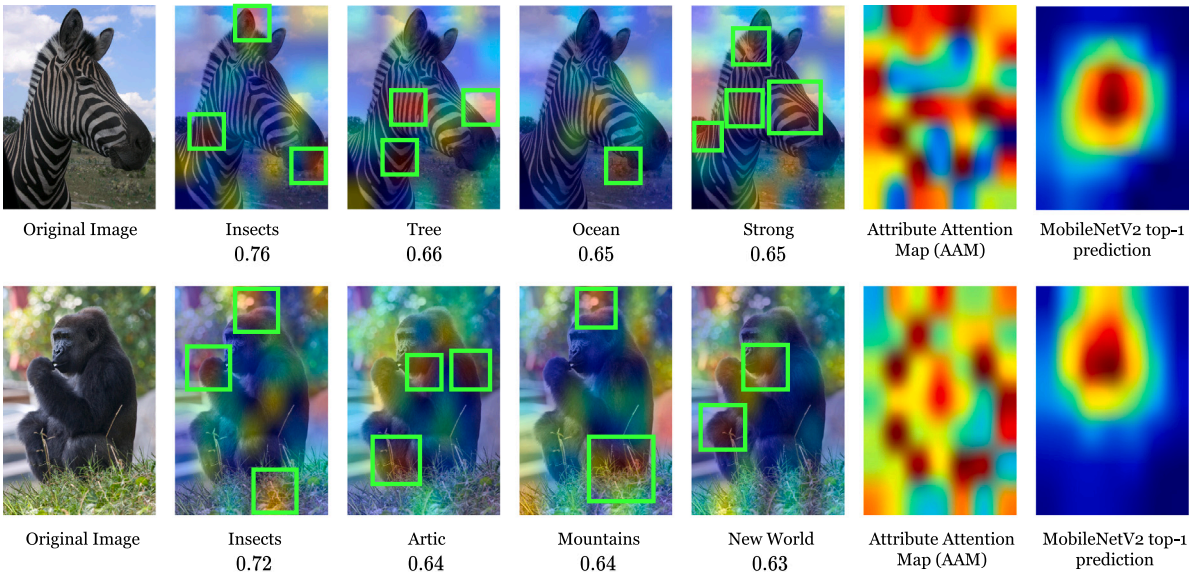

Fig. 6. Visualization of the generated Grad-CAMs for the top-4 predicted attributes of two classes of the AWA2 dataset along with the final AAM and the heat map regarding the top-1 ImageNet class prediction of the MobileNetV2. It is evidenced that the heat map from the top-1 prediction of the MobileNetV2 concentrates on more relevant parts regarding the image class in comparison to the generated AAM from the top-10 attributes presented in the image.

increasing the importance of background/noisy regions. We argue that this is a consequence of the subjectivity of the dataset semantic information, such as "Insects", "New World", "Strong". Considering this, the visual explanations in Fig. 6 corroborate the results presented in Table 4.

CelebA. In the case of the CelebA dataset, the use of the annotated attributes to improve the quality of the visual features through the AAM is explained by the visualizations depicted in Fig. 7. In contrast to the AWA2 dataset, the generated AAM for CelebA justifies the relevant improvement (10.0%, on average, considering the total of classes) over the baseline value (3.55%, on average) regarding the ZSL classification, as evidenced by the results in Table 2. Fig. 7 compares the original activation map of a pre-trained network with the generated AAM, where it is possible to observe that the proposed approach highlights a significantly higher region of the human face, which inherently improves the quality of the produced semantic-guided visual feature since it contains important information for the class prediction. On the other hand, the evidenced region in the "heated" (red) area of the activation map generated by the top-1 prediction of the VGGFace architecture lacks discriminative regions crucial to make a more accurate prediction.

LFWA. The qualitative results obtained with the LFWA dataset corroborate the conclusions drawn for the case of CelebA dataset. Similar to CelebA, LFWA dataset has 40 binary annotated attributes, which are visually depicted in the examples of Fig. 8. It is perceived that the generated AAM embraces a wider facial area than the heatmap obtained from the top-1 prediction of the VGGFace architecture, resulting in superior performance on recognition unseen classes due to the more informative extracted features, as evidenced by the results in Table 3. Compared to CelebA quantitative results, the lack of a considerable amount of examples in the case of the LFWA dataset (13233 vs. 202599 in CelebA) and the class imbalacement significantly affects the recognition performance in all evaluated ZSL methods.

LAD. As discussed in the previous section, there is no significant improvement over the baseline results. In order to perceive how the SGAM impacts the final classification accuracy, we selected two examples from the LAD dataset. We generated the Grad-CAMs for the top-4 predicted attributes and the corresponding AAM. The visual inspection of the heatmaps depicted in Fig. 9 allows us to infer that the generated AAM gives importance to non-informative regions, such as the corners of the image. On the other hand, the Grad-CAM produced by the top-1

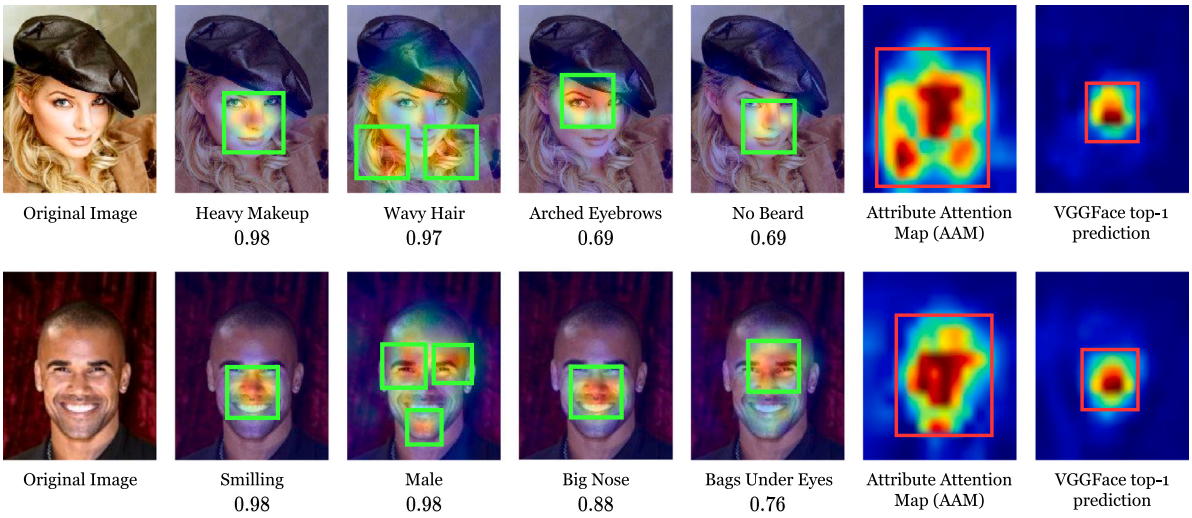

Fig. 7. Visualization of the generated Grad-CAMs for the top-4 predicted attributes of two celebrities of the CelebA dataset along with the final AAM and the heat map regarding the top-1 prediction of the VGGFace. The generated AAM contains more useful parts (wider area) concerning the face region than the reduced region illustrated in the heat map of the VGGFace top-1 prediction.

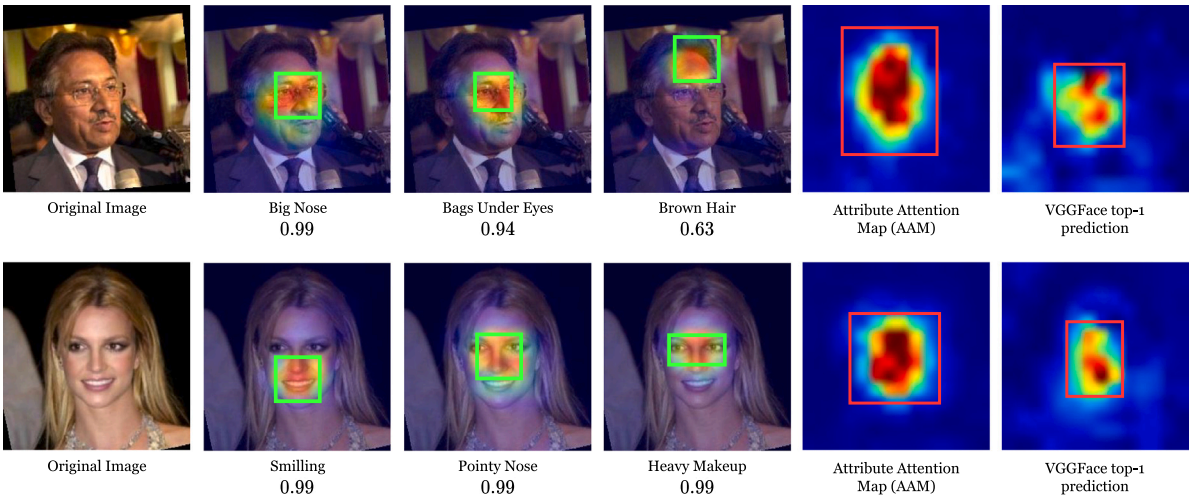

Fig. 8. Visualization of the generated Grad-CAMs for the top-3 predicted attributes of two celebrities of the LFWA dataset along with the final AAM and the heat map regarding the top-1 prediction of the VGGFace. The generated AAM contains more useful parts (wider area) concerning the face region than the reduced region illustrated in the heat map of the VGGFace top-1 prediction.

Table 7
Performance of different ZSL methods on CUB with and without the use of the proposed learning strategy.

Method	1 attribute		20 attributes	
	w/o. SGAM	w. SGAM	w/o. SGAM	w. SGAM
SAE (Kodirov et al., 2017)	48.29	46.81	48.29	48.75
ESZSL (Romera-Paredes & Torr, 2015)	62.47	58.84	62.47	62.56
DEM (Zhang et al., 2017)	67.41	29.50	67.41	59.05
TF-VAEGAN (Narayan et al., 2020)	**74.42**	70.38	**74.42**	73.73

The results obtained without the proposed SGAM are reported in the column marked with "w/o. SGAM", whereas the results using the SGAM are along the column marked with "w. SGAM". SGAM refers to Semantic Guided Attention Model. The results report top-1 accuracy in %.

prediction of the MobileNetV2 is more refined in identifying the informative regions that contribute to the prediction.

CUB. To measure the quantitative impact of using a single AAM rather than the average of the top-K, we compare the performance obtained on CUB using both approaches. The results are provided in Table 7, confirming the qualitative results observed in Figs. 6, 7 and 9.

To perceive the effect of the Gaussian mask in the overall AAM, Fig. 10 depicts the AAM without the application of the Gaussian mask, the AAM after applying the Gaussian mask, and the AAM obtained from the original DenseNet architecture, without any modification in its internal model.

7. Conclusions and future work

In this paper, we proposed an attention-based strategy (SGAM) for improving the discriminability of visual features used in ZSL methods. The semantic information is exploited to infer the image regions

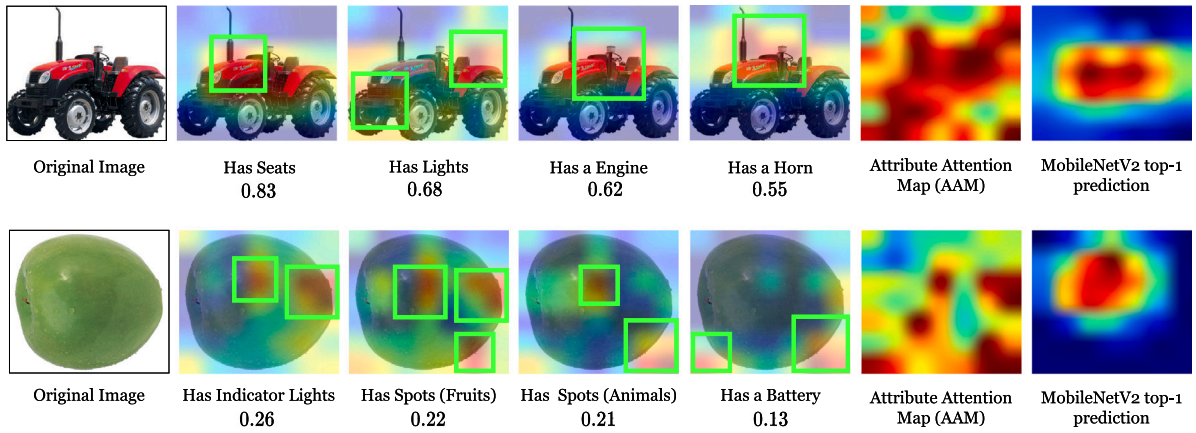

Fig. 9. Visualization of the generated Grad-CAMs for the top-4 predicted attributes of two images of the LAD dataset along with the final AAM and the heat map regarding the top-1 prediction of the MobileNetV2. The generated AAM contains non-useful regions regarding the class label, which impacts the final classification accuracy. On the other hand, the heat map of the MobileNetV2 top-1 prediction concentrates on more relevant regions that improve the classification accuracy.

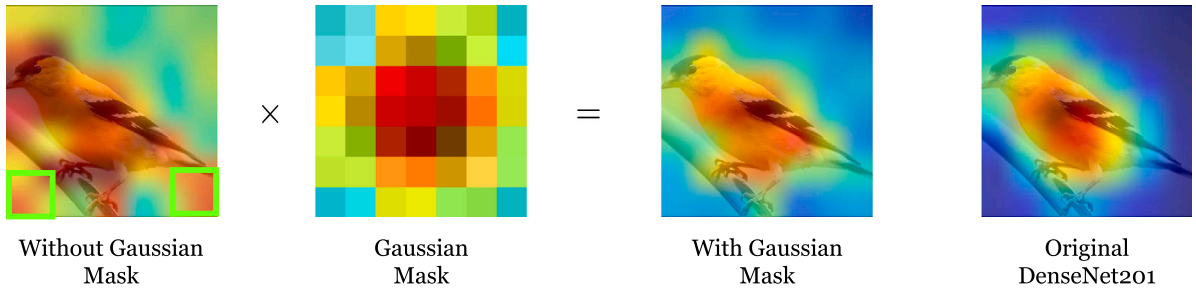

Fig. 10. Comparison of the generated AAMs. The produced AAM without applying the Gaussian mask contains non-informative regions, as highlighted by the green squares in the left image. On the other hand, after applying the Gaussian mask, the AAM focuses on more discriminative regions, as expected. However, without any architecture modification, the AAM generated by the original DenseNet contains more refined localization regions than the AAM obtained by the trained DenseNet. This fact justifies the slight improvement over the original results, in which the SGAM is not applied.

that correlate the most with semantic attributes. The experimental results reveal that the proposed method can improve ZSL methods performance when applied to datasets comprising visual interpretable attribute annotations, increasing the classification accuracy by a maximum of 13.7%, on average. However, the addition of the SGAM does not improve the classification accuracy when applied to the traditional ZSL datasets, namely AWA2. Furthermore, the qualitative assessment confirms that the predicted attributes are not visually interpretable enough to produce relevant Grad-CAM maps, causing the generated Attribute Attention Map to concentrate less on important regions, thus decreasing the discriminability of the extracted features. Although the generated AAMs based on the attributes of the CUB dataset manage to focus the meaningful regions quite well, the sparsity around the localized region impacts the final discriminative feature. By comparing the generated AAM with the Grad-CAM obtained through the original DenseNet network, it is possible to observe that the latter is much more refined and more concentrated in the region of interest. Nevertheless, the concatenation of features extracted from the original DenseNet with the produced feature vector weighted by the AAM can achieve superior performance when compared to the value without using the SGAM strategy.

As future work, we believe that a possible solution to improve the performance in classical ZSL datasets is building on strategies also explored in the facial attribute prediction domain, such as semantic segmentation (Kalayeh et al., 2017) or graph learning (Chen et al., 2021), in order to achieve an improvement in the attributes prediction task and consequently generating more refined localization maps relating to the predicted attribute.

CRediT authorship contribution statement

Cristiano Patrício: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft. **João C. Neves:** Conceptualization, Methodology, Supervision, Writing – review & editing.

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.

Acknowledgements

The authors would like to thank the support provided by FCT strategic project NOVA LINCS (UIDB/04516/2020).

References

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2013). Label-embedding for attribute-based classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 819–826).

Chen, Z., Gu, S., Zhu, F., Xu, J., & Zhao, R. (2021). Improving facial attribute recognition by group and graph learning. In *Proceedings of IEEE international conference on multimedia and expo*.

Ding, J., Hu, X., & Zhong, X. (2021). A semantic encoding out-of-distribution classifier for generalized zero-shot learning. *IEEE Signal Processing Letters*, *28*, 1395–1399.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Proceedings of the advances in neural information processing systems (NeurIPS)* (pp. 2121–2129).

Han, Z., Fu, Z., Chen, S., & Yang, J. (2021). Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2371–2381).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, G. B., Mattar, M., Lee, H., & Learned-Miller, E. (2012). Learning to align from scratch. In *Advances in neural information processing systems, vol. 25*.

Kalayeh, M. M., Gong, B., & Shah, M. (2017). Improving facial attribute prediction using semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6942–6950).

Kodirov, E., Xiang, T., & Gong, S. (2017). Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3174–3183).

Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 951–958).

Larochelle, H., Erhan, D., & Bengio, Y. (2008). Zero-data learning of new tasks. In *Proceedings of the international conference on artificial intelligence, vol. 1(2)* (p. 3).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of IEEE international conference on computer vision* (pp. 3730–3738).

Liu, Z., Zhang, X., Zhu, Z., Zheng, S., Zhao, Y., & Cheng, J. (2021). MFHI: Taking modality-free human identification as zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Ma, P., & Hu, X. (2020). A variational autoencoder with deep embedding model for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence, vol. 34(07)* (pp. 11733–11740).

Mishra, A., Krishna Reddy, S., Mittal, A., & Murthy, H. A. (2018). A generative model for zero-shot learning using conditional variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2188–2196).

Narayan, S., Gupta, A., Khan, F. S., Snoek, C. G., & Shao, L. (2020). Latent embedding feedback and discriminative features for zero-shot classification. In *Proceedings of the european conference on computer vision* (pp. 479–495).

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In *Proceedings of the advances in neural information processing systems (NeurIPS), vol. 22* (pp. 1410–1418).

Patrício, C., & Neves, J. (2021). Zspeedl - evaluating the performance of zero-shot learning methods using low-power devices. In *Proceedings of the IEEE conference on advanced video and signal based surveillance* (pp. 1–8).

Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *Proceedings of the international conference on machine learning* (pp. 2152–2161).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., & Belongie, S. (2015). Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 595–604).

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200–2011 Dataset*. California Institute of Technology.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *41*(9), 2251–2265.

Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018). Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5542–5551).

Xian, Y., Sharma, S., Schiele, B., & Akata, Z. (2019). F-VAEGAN-D2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10275–10284).

Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2021–2030).

Zhao, B., Fu, Y., Liang, R., Wu, J., Wang, Y., & Wang, Y. (2019). A large-scale attribute dataset for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).

Zhu, Y., Xie, J., Tang, Z., Peng, X., & Elgammal, A. (2019). Semantic-guided multi-attention localization for zero-shot learning. In *Proceedings of the advances in neural information processing systems (NeurIPS)* (pp. 14943–14953).