Implicit and Explicit Attention for Zero-Shot Learning

Faisal Alamri[0000−0001−6695−2504] and Anjan Dutta[0000−0002−1667−2245]

University of Exeter, Streatham Campus, Exeter, EX4 4RN, United Kingdom
{F.Alamri2,A.Dutta}@exeter.ac.uk

Abstract. Most of the existing Zero-Shot Learning (ZSL) methods focus on learning a compatibility function between the image representation and class attributes. Few others concentrate on learning image representation combining local and global features. However, the existing approaches still fail to address the bias issue towards the seen classes. In this paper, we propose implicit and explicit attention mechanisms to address the existing bias problem in ZSL models. We formulate the implicit attention mechanism with a self-supervised image angle rotation task, which focuses on specific image features aiding to solve the task. The explicit attention mechanism is composed with the consideration of a multi-headed self-attention mechanism via Vision Transformer model, which learns to map image features to semantic space during the training stage. We conduct comprehensive experiments on three popular benchmarks: AWA2, CUB and SUN. The performance of our proposed attention mechanisms has proved its effectiveness, and has achieved the state-of-the-art harmonic mean on all the three datasets.

Keywords: Zero-shot Learning · Attention Mechanism · Self-Supervised Learning · Vision Transformer.

1 Introduction

Most of the existing Zero-Shot Learning (ZSL) methods [44,38] depend on pretrained visual features and necessarily focus on learning a compatibility function between the visual features and semantic attributes. Recently, attention-based approaches have got a lot of popularity, as they allow to obtain an image representation by directly recognising object parts in an image that correspond to a given set of attributes [50,53]. Therefore, models capturing global and local visual information have been quite successful [50,51]. Although visual attention models quite accurately focus on object parts, it has been observed that often recognised parts in image and attributes are biased towards training (or *seen*) classes due to the learned correlations [51]. This is mainly because the model fails to decorrelate the visual attributes in images.

Therefore, to alleviate these difficulties, in this paper, we consider two alternative attention mechanisms for reducing the effect of bias towards training classes in ZSL models. The first mechanism is via the self-supervised pretext task, which implicitly attends to specific parts of an image to solve the pretext task, such as recognition of image rotation angle [27]. For solving the pretext task, the model essentially focuses on learning image features that lead to solving the pretext task. Specifically, in this work, we consider rotating the input image concurrently by four different angles

$(0°, 90°, 180°, 270°)$ and then predicting the rotation class. Since pretext tasks do not involve attributes or class-specific information, the model does not learn the correlation between visual features and attributes. Our second mechanism employs the Vision Transformer (ViT) [13] for mapping the visual features to semantic space. ViT having a rich multi-headed self-attention mechanism explicitly attends to those image parts related to class attributes. In a different setting, we combine the implicit with the explicit attention mechanism to learn and attend to the necessary object parts in a decorrelated or independent way. We attest that incorporating the rotation angle recognition in a self-supervised approach with the use of ViT does not only improve the ZSL performance significantly, but also and more importantly, contributes to reducing the bias towards seen classes, which is still an open challenge in the Generalised Zero-Shot Learning (GZSL) task [43]. Explicit use of attention mechanism is also examined, where the model is shown to enhance the visual feature localisation and attends to both global and discriminative local features guided by the semantic information given during training. As illustrated in Fig. 1, images fed into the model are taken from two different sources: 1) labelled images, which are the training images taken from the *seen* classes, shown in green colour, and 2) other images, which could be taken from any source, shown in blue. The model is donated as ($\mathcal{F}(.)$), in this paper, we implement $\mathcal{F}(.)$ either by ViT or by ResNet-101 [22] backbones. The first set of images is used to train the model to predict class attributes leading to the class labels via nearest search. However, the second set of images is used for rotation angle recognition, guiding the model to learn visual representations via implicit attention mechanism.

To summarise, in this paper, we make the following contributions: (1) We propose the utilisation of alternative attention mechanisms for reducing the bias towards the seen classes in zero-shot learning. By involving self-supervised pretext task, our model implicitly attends decorrelated image parts aiding to solve the pretext task, which learns image features independent of the training classes. (2) We perform extensive experiments on three challenging benchmark datasets, i.e. AWA2, CUB and SUN, in the generalised zero-shot learning setting and demonstrate the effectiveness of our proposed alternative attention mechanisms. We also achieve consistent improvement over the state-of-the-art methods. (3) The proposed method is evaluated with two backbone models: ResNet-101 and ViT, and shows significant improvement in the model performances, and reduces the issue of bias towards seen classes. We also show the effectiveness of our model qualitatively by plotting the attention maps.

2 Related Work

In this section we briefly review the related arts on zero-shot learning, Vision Transformer and self-supervised learning.

Zero-Shot Learning (ZSL): Zero-Shot Learning (ZSL) uses semantic side information such as attributes and word embeddings [47,32,36,14,16,4] to predict classes that have never been presented during training. Early ZSL models train different attribute classifiers assuming independence of attributes and then estimate the posterior of the test classes by combining attribute prediction probabilities [28]. Others do not follow the independence assumption and learn a linear [17,3,2] or non-linear [45] compatibility

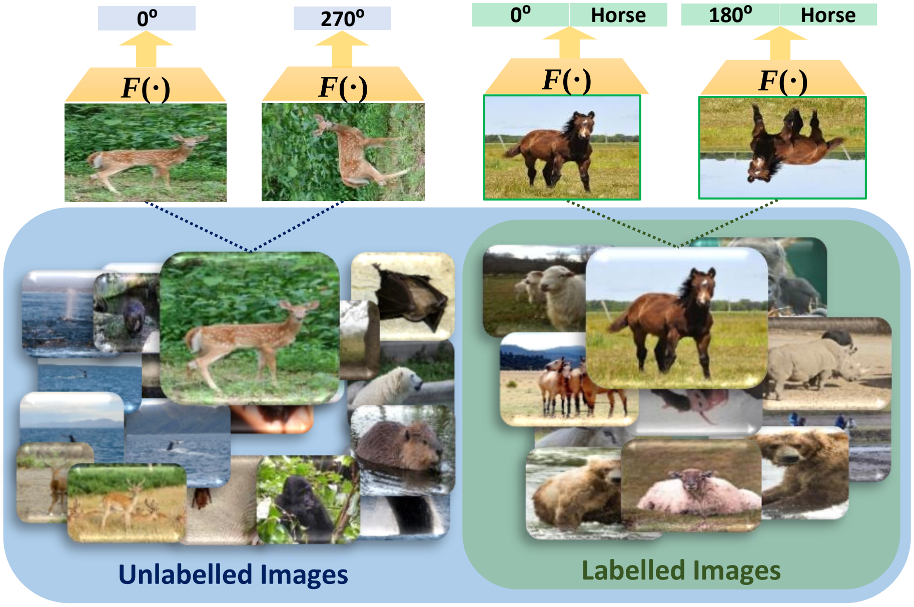

Fig. 1. Our method maps the visual features to the semantic space provided with two different input images (unlabelled and labelled data). Green represents the labelled images provided to train the model to capture visual features and predict object classes. Blue represents the unlabelled images that are rotated and attached to the former set of images to recognise rotated image angles in a self-supervised task. The model learns the visual representations of the rotated images implicitly via the use of attention. (Best viewed in colour)

function from visual features to semantic space. There are some other works that learn an inverse mapping from semantic to visual feature space [39,55]. Learning a joint mapping function for each space into a common space (i.e. a shared latent embedding) is also investigated in [45,23,20]. Different from the above approaches, generative models synthesise samples of unseen classes based on information learned from seen classes and their semantic information, to tackle the issue of bias towards the seen classes [44,58,38]. Unlike other models, which focus on the global visual features, attention-based methods aim to learn discriminative local visual features and then combine with the global information [53,59]. Examples include S^{2}GA [53] and AREN [50] that apply an attention-based network to incorporate discriminative regions to provide rich visual expression automatically. In addition, GEN [49] proposes a graph reasoning method to learn relationships among multiple image regions. Others focus on improving localisation by adapting the human gaze behaviour [30], exploiting a global average pooling scheme as an aggregation mechanism [52] or by jointly learning both global and local features [59]. Inspired by the success of the recent attention-based ZSL models, in this paper, we propose two alternative attention mechanisms to capture robust image features suitable to ZSL task. Our first attention mechanism is implicit and is based on self-supervised pretext task [27], whereas the second attention mechanism is explicit and is based on ViT [13]. To the best of our knowledge, both of these attention models are still unexplored in the context of ZSL. Here we also point out that the inferential comprehension of visual representations upon the use of SSL and ViT is a future direction to consider for ZSL task.

Vision Transformer (ViT): The Transformer [41] adopts the self-attention mechanism to weigh the relevance of each element in the input data. Inspired by its success, it has been implemented to solve many computer vision tasks [5,13,25] and many enhancements and modifications of Vision Transformer (ViT) have been introduced. For example, CaiT [40] introduces deeper transformer networks, Swin Transformer [31] proposes a hierarchical Transformer capturing visual representation by computing self-attention via shifted windows, and TNT [21] applies the Transformer to compute the visual representations using both patch-level and pixel-level information. In addition, CrossViT [9] proposes a dual-branch Transformer with different sized image patches. Recently, TransGAN [24] proposes a completely free of convolutions generative adversarial network solely based on pure transformer-based architectures. Readers are referred to [25], for further reading about ViT based approaches. The applicability of ViT-based models is growing, but it has remained relatively unexplored to the zero-shot image recognition tasks where attention based models have already attracted a lot of attention. Therefore employing robust attention based models, such as ViT is absolutely timely and justified for improving the ZSL performance.

Self-Supervised Learning (SSL): Self-Supervised Learning (SSL) is widely used for unsupervised representation learning to obtain robust representations of samples from raw data without expensive labels or annotations. Although the recent SSL methods use contrastive objectives [10,19], early works used to focus on defining pretext tasks, which typically involves defining a surrogate task on a domain with ample weak supervision labels, such as predicting the rotation of images [27], relative positions of patches in an image [11,33], image colours [29,56] etc. Encoders trained to solve such pretext tasks are expected to learn general features that might be useful for other downstream tasks requiring expensive annotations (e.g. image classification). Furthermore, SSL has been widely used in various applications, such as few-shot learning [18], domain generalisation [7] etc. In contrast, in this paper, we utilise the self-supervised pretext task of image rotation prediction for obtaining implicit image attention to solve ZSL.

3 Implicit and Explicit Attention for Zero-Shot Learning

In this work, we propose an Implicit and Explicit Attention mechanism-based Model for solving image recognition in Zero-Shot Learning (IEAM-ZSL). We utilise self-supervised pretext tasks, such as image rotation angle recognition, for obtaining image attention in an implicit way. Here the main rational is for predicting the correct image rotation angle, the model needs to focus on image features with discriminative textures, colours etc., which implicitly attend to specific regions in an image. For having explicit image attention, we utilise the multi-headed self-attention mechanism involved in Vision Transform model.

From ZSL perspective, we follow the inductive approach for training our model, i.e. during training, the model only has access to the training set (*seen* classes), consisting of

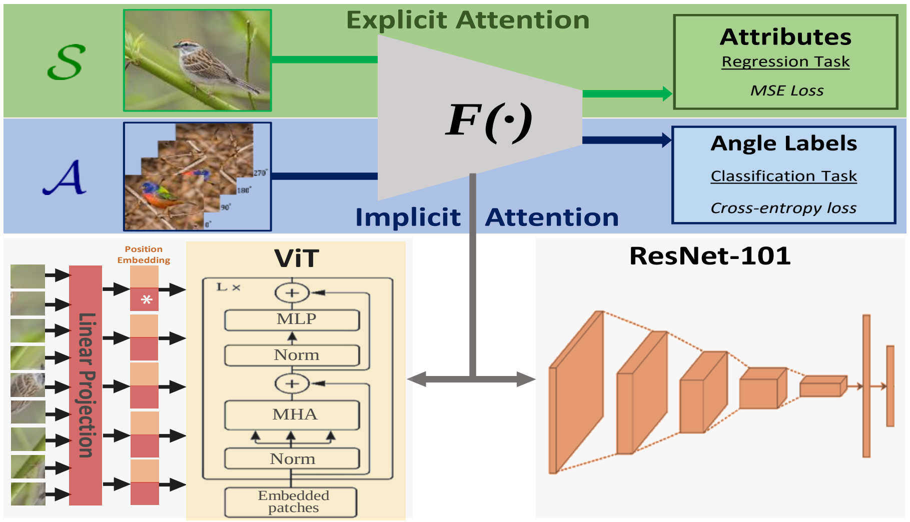

Fig. 2. IEAM-ZSL Architecture. IEAM-ZSL consists of two pipelines represented in Green and Blue colours, respectively. The former takes images from the ZSL datasets with their class-level information input to the Transformer Encoder for attributes predictions. Outputs are compared with semantic information of the corresponding images using MSE loss as a regression task. The latter, shown in Blue colour, is fed with images after generating four rotations for each (i.e. $0°$, $90°$, $180°$, and $270°$), to predict the rotation angle. At inference, solely the ZSL test datasets, with no data augmentation, are inputted to the model to predict the class-level attributes. A search for the nearest class label is then conducted.

only the labelled images and continuous attributes of the seen classes ($\mathcal{S} = \{\mathbf{x}, \mathbf{y} | \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}^s\}$). An RGB image in image space $\mathcal{X}$ is denoted as $\mathbf{x}$, where $\mathbf{y} \in \mathcal{Y}$ is the class-level semantic vector annotated with M different attributes. As depicted in Fig. 2, a 224×224 image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ with resolution $H \times W$ and C channels is fed into the model. Addition to $\mathcal{S}$, we also use an auxiliary set of unlabelled images $\mathcal{A} = \{\mathbf{x} \in \mathcal{X}\}$ for predicting the image rotation angle to obtain implicit attention. Note, here the images from $\mathcal{S}$ and $\mathcal{A}$ may or may not overlap, however, the method does not utilise the categorical or semantic label information of the images from the set $\mathcal{A}$.

3.1 Implicit Attention

Self-supervised pretext tasks provide a surrogate supervision signal for feature learning without any manual annotations [27,12,1] and it is well known that this type of supervision focuses on image features that help to solve the considered pretext task. It has also been shown that these pretext tasks focus on meaningful image features and effectively avoid learning correlation between visual features [27]. As self-supervised learning avoids considering semantic class labels, spurious correlation among visual features are not learnt. Therefore, motivated by the above facts, we employ an image rotation angle prediction task to obtain implicitly attended image features. For that, we rotate an image by $0°$, $90°$, $180°$ and $270°$, and train the model to correctly classify the rotated images. Let $g(\cdot|a)$ be an operator that rotates an image $\mathbf{x}$ by an angle $90° \times a$, where $a \in \{0, 1, 2, 3\}$. Now let $\hat{\mathbf{y}}_a$ be the predicted probability for the rotated image $\mathbf{x}_a$ with label a, then the loss for training the underlying model is computed as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{a=1}^{4} \log(\hat{\mathbf{y}}_a) \tag{1}$$

In our case, the task of predicting image rotation angle trains the model to focus on specific image regions having rich visual features (for example, textures or colours). This procedure implicitly learns to attend image features.

3.2 Explicit Attention

For obtaining explicit attention, we employ Vision Transformer model [13], where each image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ with resolution $H \times W$ and C channels is fed into the model after resizing it to 224×224. Afterwards, the image is split into a sequence of N patches denoted as $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 . C)}$, where $N = \frac{H.W}{P^2}$. Patch embeddings (small red boxes in Fig. 2) are encoded by applying a trainable 2D convolution layer with kernel size=(16, 16) and stride=(16, 16)). An extra learnable classification token ($\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$) is appended at the beginning of the sequence to encode the global image representation, which is donated as $(*)$. Position embeddings (orange boxes) are then attached to the patch embeddings to obtain the relative positional information. Patch embeddings ($\mathbf{z}$) are then projected through a linear projection $\mathbf{E}$ to D dimension (i.e. $D = 1024$) as in Eq. 2. Embeddings are then passed to the Transformer encoder, which consists of Multi-Head Attention (MHA) (Eq. 3) and MLP blocks (Eq. 4). A layer normalisation (Norm) is applied before every block, and residual connections after every block. The image representation ($\hat{\mathbf{y}}$) is then produced as in Eq. 5.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad \mathbf{E} \in \mathbb{R}^{(P^2.C) \times D}, \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{2}$$

$$\mathbf{z}'_\ell = \text{MHA}(\text{Norm}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L \; (L = 24) \tag{3}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{Norm}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1 \ldots L \tag{4}$$

$$\hat{\mathbf{y}} = \text{Norm}(\mathbf{z}_L^0) \tag{5}$$

Below we provide details of our multi-head attention mechanism within the ViT model.

Multi-Head Attention (MHA): Patch embeddings are fed into the transformer encoder, where the multi-head attention takes place. Self-attention is performed for every patch in the sequence of the patch embeddings independently; thus, attention works simultaneously for all the patches, leading to multi-headed self-attention. This is computed by creating three vectors, namely Query (Q), Key (K) and Value (V). They are created by multiplying the patch embeddings by three trainable weight matrices (i.e. W^Q, W^K and W^V) applied to compute the self-attention. A dot-product operation is performed on the Q and K vectors, calculating a scoring matrix that measures how much a patch embedding has to attend to every other patch in the input sequence.

The score matrix is then scaled down and converted into probabilities using a softmax. Probabilities are then multiplied by the V vectors, as in Eq. 6, where d_k is the dimension of the vector K. Multi-headed self-attention mechanism produces a number of self-attention matrices which are concatenated and fed into a linear layer and passed sequentially to 1) regression head and 2) classification head.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

The multi-headed self-attention mechanism involved in the Vision Transformer guides our model to learn both the global and local visual features. It is worth noting that the standard ViT has only one classification head implemented by an MLP, which is changed in our model to two heads to meet the two different underlying objectives. The first head is a regression head applied to predict M different class attributes, whereas the second head is added for rotation angle classification. For the former task, the objective function employed is the Mean Squared Error (MSE) loss as in Eq. 7, where $\mathbf{y}_i$ is the target attributes, and $\hat{\mathbf{y}}_i$ is the predicted ones. For the latter task, cross-entropy (Eq. 1) objective is applied.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \tag{7}$$

The total loss used for training our model is defined in Eq. 8, where $\lambda_1 = 1$ and $\lambda_2 = 1$.

$$\mathcal{L}_{\text{TOT}} = \lambda_1\mathcal{L}_{\text{CE}} + \lambda_2\mathcal{L}_{\text{MSE}} \tag{8}$$

During the inference phase, original test images from the seen and unseen classes are inputted. Class labels are then determined using the cosine similarity between the predicted attributes and every target class embeddings predicted by our model.

4 Experiments

Datasets: We have conducted our experiments on three popular ZSL datasets: AWA2, CUB, and SUN, whose details are presented in Table 1. The main aim of this experimentation is to validate our proposed method IEAM-ZSL, demonstrating its effectiveness and comparing it with the existing state-of-the-art methods. Among these datasets, AWA2 [47] consists of $37,322$ images of 50 categories (40 seen + 10 unseen). Each category contains 85 binary as well as continuous class attributes. CUB [42] contains $11,788$ images forming 200 different types of birds, among them 150 classes are considered as seen, and the other 50 as unseen, which is split by [2]. Together with images CUB dataset also contains 312 attributes describing birds. Finally, SUN [35] has the largest number of classes among others. It consists of 717 types of scene images, which divided into 645 seen and 72 unseen classes. The SUN dataset contains $14,340$ images with 102 annotated attributes.

Implementation Details: In our experiment, we have used two different backbones: (1) ResNet-101 and (2) Vision Transformer (ViT), both of which are pretrained on ImageNet and then finetuned for the ZSL tasks on the datasets mentioned above. We resize

Table 1. Dataset statistics: The number of classes (seen + unseen classes shown within parenthesis), the number of attributes and the number of images per dataset.

Datasets	AWA2 [47]	CUB [42]	SUN [35]
Number of Classes	50	200	717
(Seen + Unseen)	$(40 + 10)$	$(150 + 50)$	$(645 + 72)$
Number of Attributes	85	312	102
Number of Images	$37,322$	$11,788$	$14,340$

the image to 224×224 before inputting it into the model. For ViT, the primary baseline model employed uses an input patch size 16×16, with 1024 hidden dimension, and having 24 layers and 16 heads on each layer, and 24 series encoder. We use the Adam optimiser for training our model with a fixed learning rate of 0.0001 and a batch size of 64. In the setting where we use self-supervised pretext task, we construct the batch with 32 *seen* training images from set $\mathcal{S}$ and 32 rotated images (i.e. eight images, where each image is rotated to $0°$, $90°$, $180°$ and $270°$) from set $\mathcal{A}$. We have implemented our model with PyTorch[1] deep learning framework and trained the model on a GeForce RTX 3090 GPU on a workstation with Xeon processor and 32GB of memory.

Evaluation: The proposed model is evaluated on the three above mentioned datasets. We have followed the inductive approach for training our model, i.e. our model has no access to neither visual nor side-information of unseen classes during training. During the evaluation, we have followed the GZSL protocol. Following [46], we compute the top-1 accuracy for both seen and unseen classes. In addition, the harmonic mean of the top-1 accuracies on the seen and unseen classes is used as the main evaluation criterion. Inspired by the recent works [52,50,8], we have used the Calibrated Stacking [8] for evaluating our model under GZSL setting. The calibration factor γ is dataset-dependent and decided based on a validation set. For AWA2 and CUB, the calibration factor γ is set to 0.9 and for SUN, it is set to 0.4.

4.1 Quantitative Results

Table 2 illustrates a quantitative comparison between the state-of-the-art methods and the proposed method using two different backbones: (1) ResNet-101 [22] and (2) ViT [13]. The baseline models performance without the employment of the SSL approach is also reported. The performance of each model is shown in % in terms of Seen (S) and Unseen (U) classes and their harmonic mean (H). As reported, the classical ZSL models [28,17,34,34,45,2] show good performance in terms of seen classes. However, they perform poorly on unseen classes and encounter the bias issue, resulting in a very low harmonic mean. Among the classical approaches, [2] performs the best on all the three datasets, as it overcomes the shortcomings of the previous models and considers the dependency between attributes. Among generative approaches, Xian2019FVAEGAND2AF [48] performs the best. Although f-CLSWGAN [44] achieves the highest score on AWA2 unseen classes, it shows lower harmonic means on all the

[1] Our code is available at: https://github.com/FaisalAlamri0/IEAM-ZSL

Table 2. Generalised zero-shot classification performance on AWA2, CUB and SUN. Reported models are ordered in terms of their publishing dates. Results are reported in %.

Models	AWA2			CUB			SUN		
	S	U	H	S	U	H	S	U	H
DAP [28]	84.7	0.0	0.0	67.9	1.7	3.3	25.1	4.2	7.2
IAP [28]	87.6	0.9	1.8	72.8	0.2	0.4	37.8	1.0	1.8
DeViSE [17]	74.7	17.1	27.8	53.0	23.8	32.8	30.5	14.7	19.8
ConSE [34]	90.6	0.5	1.0	72.2	1.6	3.1	39.9	6.8	11.6
ESZSL [37]	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8
SJE [3]	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
SSE [57]	82.5	8.1	14.8	46.9	8.5	14.4	36.4	2.1	4.0
LATEM [45]	77.3	11.5	20.0	57.3	15.2	24.0	28.8	14.7	19.5
ALE [2]	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3
*GAZSL [58]	86.5	19.2	31.4	60.6	23.9	34.3	34.5	21.7	26.7
SAE [26]	82.2	1.1	2.2	54.0	7.8	13.6	18.0	8.8	11.8
*f-CLSWGAN [44]	64.4	57.9	59.6	57.7	43.7	49.7	36.6	42.6	39.4
AREN [50]	79.1	54.7	64.7	63.2	69.0	66.0	40.3	32.3	35.9
*Xian2019FVAEGAND2AF [48]	76.1	57.1	65.2	75.6	63.2	68.9	50.1	37.8	43.1
SGMA [59]	87.1	37.6	52.5	71.3	36.7	48.5	-	-	-
IIR [6]	83.2	48.5	61.3	52.3	55.8	53.0	30.4	47.9	36.8
*E-PGN [54]	83.5	52.6	64.6	61.1	52.0	56.2	-	-	-
SELAR [52]	78.7	32.9	46.4	76.3	43.0	55.0	37.2	23.8	29.0
ResNet-101 [22]	66.7	40.1	50.1	59.5	52.3	55.7	35.5	28.8	31.8
ResNet-101 with Implicit Attention	**74.1**	**45.9**	**56.8**	**62.7**	**54.5**	**58.3**	**36.3**	**31.9**	**33.9**
Our model (ViT)	**90.0**	**51.9**	**65.8**	**75.2**	**67.3**	**71.0**	**55.3**	**44.5**	**49.3**
Our model (ViT) with Implicit Attention	**89.9**	**53.7**	**67.2**	**73.8**	**68.6**	**71.1**	**54.7**	**48.2**	**51.3**

S, U, H denote Seen classes ($\mathcal{Y}^s$), Unseen classes ($\mathcal{Y}^u$), and the Harmonic mean, respectively. For each scenario, the best is in red and the second-best is in blue. * indicates generative representation learning methods.

datasets than [48]. As noticed, the first top scores for the AWA2 unseen classes accuracy are obtained by generative models [44,48], which we assume is because they include both seen and synthesised unseen features during the training phase. Moreover, attention-based models, such as [59,50] are the closest to our proposed model, perform better than the other models due to the inclusion of global and local representations. [50] outperforms all reported models on the unseen classes of the CUB dataset, but still has low harmonic means on all the datasets. SGMA [59] performs poorly on both AWA2 and CUB, and it clearly suffers from the bias issue, where its performance on unseen classes is considered deficient compared to other models. Recent models such as SELAR [52] uses global maximum pooling as an aggregation method and achieves the best scores on CUB seen classes, but achieves low harmonic means. In addition, its performance is seen to be considerably impacted by the bias issue.

ResNet-101: For a fair evaluation of the robustness and effectiveness of our proposed alternative attention-based approach, we consider the ResNet-101 [22] as one of our backbones, which is also used in prior related arts [17,2,26,54,52]. We have used the ResNet-101 backbone as a baseline model, where we only consider the global representation. Moreover, we also use this backbone with implicit attention, i.e. during training, we simultaneously impose a self-supervised image rotation angle prediction task for training the model. Note, for producing the results in Table 2, we only use the images from the seen classes as set $\mathcal{A}$, which is used for rotation angle prediction task. As presented in Table 2, our model with ResNet-101 backbone has performed inferiorly compared to our implicit and explicit variant, which will be discussed in the next paragraph. However, even with the ResNet-101 backbone, the contribution of our implicit attention mechanism should be noted, which provides a substantial boost to the model performance. For example, on AWA2, a considerable increment is observed on both seen and unseen classes, leading to a significant increase in the harmonic mean (i.e. 50.1% to 56.8%). The performance of the majority of the related arts seems to suffer from bias towards the seen classes. We argue that our method tends to mitigate this issue on all the three datasets. Our method enables the model to learn the visual representations of unseen classes implicitly; hence, the performance is increased, and the bias issue is alleviated. Similarly, on the SUN dataset, although this dataset consists of 717 classes, the proposed implicit attention mechanism illustrates the capability of providing ResNet-101 with an increase in the accuracy in terms of both seen and unseen classes, leading to an increase of ≈ 2 points in the harmonic mean, i.e. from 31.8% to 33.9%.

Vision Transformer (ViT): We have used Vision Transformer (ViT) as another backbone to enable explicit attention in our model. Similar to the ResNet-101 backbone, we use the implicit attention mechanism with ViT backbone as well. During training, we simultaneously impose self-supervised image rotation angle prediction task for training the model. Here also we only use the images from the seen classes for image rotation angle task. As shown in Table 2, consideration of explicit attention performs very well on all the three datasets and it outperforms all the previously reported results with a significant margin. Such results are expected due to the involvement of self-attention employed in ViT. It captures both the global and local features explicitly guided by the class attributes given during training. Furthermore, attention focuses to each element of the input patch embeddings after the image is split, which effectively weigh the relevance of different patches, resulting in more compact representations. Although explicit attention mechanism is seen to provide better visual understanding, the effectiveness of the implicit attention process in terms of recognising the image rotation angle is also quite important. It does not only improve the performance further but also reduces the bias issue considerably, which can be seen in the performance of the unseen classes. In addition, it allows the model via an implicit use of self-attention to encapsulate the visual features and regions that are semantically relevant to the class attributes. Our model achieves the highest harmonic mean among all the reported models on all the three datasets. In terms of AWA2, our approach scores the third highest accuracy on both seen and unseen classes, but the highest harmonic mean. Note that on AWA2 dataset, our model still suffers from bias towards seen classes. We speculate that is due to the lack of the co-occurrence of some vital and identical attributes between seen and unseen classes. For example, attributes *nocturnal* in bat, *longneck* in giraffe or *flippers* in seal score the highest attributes in the class-attribute vectors, but rarely appear among other classes. However, on CUB dataset, this issue seems to be mitigated, as our model scores the highest harmonic mean (i.e. $H = 71.1\%$), where the performance on unseen classes is increased compared to our model with explicit attention. Finally, our model with implicit and explicit attention achieves the highest scores on classes on the SUN dataset, resulting in the best achieved harmonic mean. In summary, our proposed implicit and explicit attention mechanism proves to be very effective across all the three considered datasets. Explicit attention using the ViT backbone with multi-head self-attention is quite important for the good performance of the ZSL model. Implicit attention in terms of self-supervised pretext task is another important mechanism to look at, as it boosts the performance on the unseen classes and provides better generalisation.

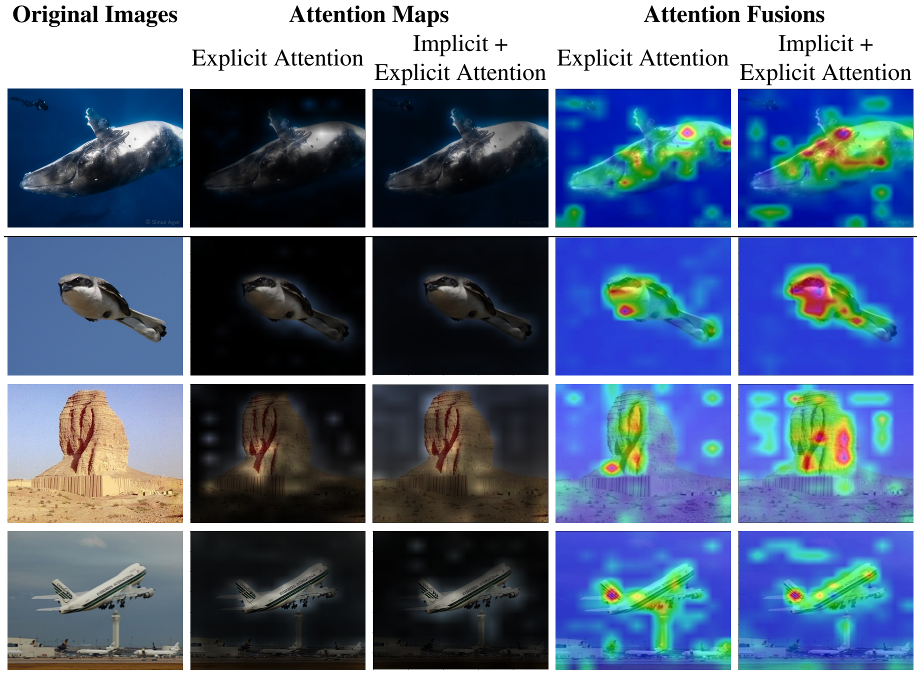

Fig. 3. Examples of implicit and explicit attention. First column: Original images, Second and Third: Attention maps without and with SSL, respectively, Four and Fifth: Attention Fusions without and with SSL, respectively. Our model benefits from using the attention mechanism and can implicitly learn object-level attributes and their discriminative features.

Attention Maps: Fig. 3 presents some qualitative results, i.e. attention maps and fusions obtained by our proposed implicit and explicit attention-based model. For generating these qualitative results, we have used our model with explicit attention mechanism, i.e. we have used the ViT backbone. Attention maps and fusions are presented on four randomly chosen images from the considered datasets. Explicit attention with ViT backbone seems to be quite important for the ZSL tasks as it can perfectly focus on the object appearing in the image, which justifies the better performance obtained by our model with ViT backbone. Inclusion of implicit attention mechanism in terms of self-supervised rotated image angle prediction further enhances the attention maps and particularly focuses on specific image parts important for that object class. For example, as shown in the first row of Fig. 3, our model with implicit and explicit attention mechanism focuses on both global and local features of the *Whale* (i.e. water, big, swims, hairless, bulbous, flippers, etc.). Similarly, on the CUB dataset, the model pays attention to objects' global features, and more importantly, the discriminative local features (i.e. *loggerhead shrike* has a white belly, breast and throat, and a black crown forehead and bill). For natural images taken from the SUN dataset, our model with implicit attention is seen to focus on the *ziggurat* paying more attention to its global features. Furthermore, as in the *airline* image illustrated in the last row, our model considers both global and discriminative features, leading to precise attention map that focuses accurately on the object.

Table 3. Ablation performance of our model with ResNet-101 and ViT backbone on AWA2, CUB and SUN datasets. Here we use the training images from the seen classes as $\mathcal{S}$ and varies $\mathcal{A}$ as noted in the first column of the following table. S, U and PASCAL respectively denote the training images from the seen classes, test images from the unseen classes, and PASCAL VOC2012 training set images.

Source of Rotated	Backbone	AWA2			CUB			SUN		
Images ($\mathcal{A}$)	(Implicit Attention)	S	U	H	S	U	H	S	U	H
S & U	ResNet-101	79.9	44.2	56.4	60.1	56.0	58.0	35.0	33.1	33.7
	ViT	87.3	56.8	68.8	74.2	68.9	71.1	54.7	50.0	52.2
PASCAL	ResNet-101	72.0	44.3	54.8	62.5	53.1	57.4	35.6	30.3	33.1
	ViT	88.1	51.8	65.2	73.4	68.0	70.6	55.2	46.3	50.6
PASCAL & U	ResNet-101	75.1	46.5	57.4	62.9	54.4	58.4	33.7	32.7	33.2
	ViT	89.8	53.2	66.8	73.02	69.7	71.3	53.9	51.0	52.4
PASCAL & S	ResNet-101	73.1	44.5	55.4	62.5	53.2	57.5	36.6	30.1	33.1
	ViT	91.2	51.6	65.9	73.7	68.8	71.1	54.19	46.9	50.9

4.2 Ablation Study

Our ablation study evaluates the effectiveness of our proposed implicit and explicit attention-based model for the ZSL tasks. Here we mainly analyse the outcome of our proposed approach if we change the set $\mathcal{A}$ which we use for sampling images for self-supervised image angle prediction task during training. In Section 4.1, we have only used the seen images for this purpose; however, we have also noted important observation if we change the set $\mathcal{A}$. Note, here we can use any collection of images as $\mathcal{A}$, since it does not need any annotation regarding its semantic class, because in this case, the only supervision used is the class corresponds to image angle rotation which can be generated online during training. In Table 3, we present results on all three considered datasets with the above mentioned evaluation metric, where we only vary $\mathcal{A}$ as noted in the first column of Table 3. Note, in all these settings $\mathcal{S}$ remains fixed, and it is set to the set of images from the seen classes. In all the settings, we observe that explicit attention in terms of ViT backbone performs significantly better than the classical CNN backbone, such as ResNet-101. We also observe that the inclusion of unlabelled images from unseen classes (can be considered as transductive ZSL [2]) significantly boosts the performance on all the datasets (see rows 1 and 3 in Table 3). Moreover, we also observe that including datasets that contain diverse images, such as PASCAL [15] improve the performance on unseen classes and increase generalisation.

5 Conclusion

This paper has proposed implicit and explicit attention mechanisms for solving the zero-shot learning task. For implicit attention, our proposed model has imposed self-supervised rotated image angle prediction task, and for the purpose of explicit attention, the model employs the multi-head self-attention mechanism via the Vision Transformer model to map visual features to the semantic space. We have considered three publicly available datasets: AWA2, CUB and SUN, to show the effectiveness of our proposed model. Throughout our extensive experiments, explicit attention via the multi-head self-attention mechanism of ViT is revealed to be very important for the ZSL task. Additionally, the implicit attention mechanism is also proved to be effective for learning image representation for zero-shot image recognition, as it boosts the performance on unseen classes and provides better generalisation. Our proposed model based on implicit and explicit attention mechanism has provided very encouraging results for the ZSL task and particularly has achieved state-of-the-art performance in terms of harmonic mean on all the three considered benchmarks, which shows the importance of attention-based models for ZSL task.

Acknowledgement

This work was supported by the Defence Science and Technology Laboratory (Dstl) and the Alan Turing Institute (ATI). The TITAN Xp and TITAN V used for this research were donated by the NVIDIA Corporation.

References

1. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: ICCV (2015)
2. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE TPAMI (2016)
3. Akata, Z., Reed, S.E., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR (2015)
4. Alamri, F., Dutta, A.: Multi-Head Self-Attention via Vision Transformer for Zero-Shot Learning. In: IMVIP (2021)
5. Alamri, F., Kalkan, S., Pugeault, N.: Transformer-encoder detector module: Using context to improve robustness to adversarial attacks on object detection. In: ICPR (2021)

6. Cacheux, Y.L., Borgne, H.L., Crucianu, M.: Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In: ICCV (2019)
7. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain Generalization by Solving Jigsaw Puzzles. In: CVPR (2019)
8. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: ECCV (2016)
9. Chen, C., Fan, Q., Panda, R.: CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In: ICCV (2021)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: ICML (2020)
11. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
12. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M.A., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
14. Dutta, A., Akata, Z.: Semantically Tied Paired Cycle Consistency for Any-Shot Sketch-Based Image Retrieval. IJCV (2020)
15. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (2012)
16. Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z.: Learning Robust Representations via Multi-View Information Bottleneck. In: ICLR (2020)
17. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NIPS (2013)
18. Gidaris, S., Bursuc, A., Komodakis, N., Perez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: ICCV (2019)
19. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised Learning. In: NeurIPS (2020)
20. Gune, O., Banerjee, B., Chaudhuri, S.: Structure aligning discriminative latent embedding for zero-shot learning. In: BMVC (2018)
21. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. arXiv (2021)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
23. Jiang, H., Wang, R., Shan, S., Yang, Y., Chen, X.: Learning discriminative latent attributes for zero-shot classification. In: ICCV (2017)
24. Jiang, Y., Chang, S., Wang, Z.: TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. In: CVPR (2021)
25. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F., Shah, M.: Transformers in vision: A survey. arXiv (2021)
26. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR (2017)
27. Komodakis, N., Gidaris, S.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
28. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
29. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV (2016)

30. Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., Harada, T.: Goal-oriented gaze estimation for zero-shot learning. In: CVPR (2021)
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv (2021)
32. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
33. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
34. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. In: ICLR (2014)
35. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR (2012)
36. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
37. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: ICML (2015)
38. Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders. CVPR (2019)
39. Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y.: Ridge regression, hubness, and zero-shot learning. In: ECML/PKDD (2015)
40. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. arXiv (2021)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
42. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. rep., California Institute of Technology (2011)
43. Wang, W., Zheng, V., Yu, H., Miao, C.: A survey of zero-shot learning. ACM-TIST (2019)
44. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR (2018)
45. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: CVPR (2016)
46. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE TPAMI (2019)
47. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: CVPR (2017)
48. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning. In: CVPR (2019)
49. Xie, G.S., Liu, L., Zhu, F., Zhao, F., Zhang, Z., Yao, Y., Qin, J., Shao, L.: Region graph embedding network for zero-shot learning. In: ECCV (2020)
50. Xie, G.S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: Attentive region embedding network for zero-shot learning. In: CVPR (2019)
51. Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z.: Attribute prototype network for zero-shot learning. In: NIPS (2020)
52. Yang, S., Wang, K., Herranz, L., van de Weijer, J.: On implicit attribute localization for generalized zero-shot learning. IEEE SPL (2021)
53. Yu, Y., Ji, Z., Fu, Y., Guo, J., Pang, Y., Zhang, Z.M.: Stacked semantics-guided attention model for fine-grained zero-shot learning. In: NeurIPS (2018)
54. Yu, Y., Ji, Z., Han, J., Zhang, Z.: Episode-based prototype generating network for zero-shot learning. In: CVPR (2020)
55. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: CVPR (2017)

56. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
57. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. In: ICCV (2015)
58. Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A.: Imagine it for me: Generative adversarial approach for zero-shot learning from noisy texts. In: CVPR (2018)
59. Zhu, Y., Xie, J., Tang, Z., Peng, X., Elgammal, A.: Semantic-guided multi-attention localization for zero-shot learning. In: NeurIPS (2019)