International Journal of Computer Vision (2024) 132:3806–3822
https://doi.org/10.1007/s11263-024-02053-3

I2DFormer+: Learning Image to Document Summary Attention for Zero-Shot Image Classification

Muhammad Ferjad Naeem[1] · Yongqin Xian[2] · Luc Van Gool[1] · Federico Tombari[2]

Received: 14 March 2023 / Accepted: 8 March 2024 / Published online: 24 April 2024

Abstract

Despite the tremendous progress in zero-shot learning (ZSL), the majority of existing methods still rely on human-annotated attributes, which are difficult to annotate and scale. An unsupervised alternative is to represent each class using the word embedding associated with its semantic class name. However, word embeddings extracted from pre-trained language models do not necessarily capture visual similarities, resulting in poor zero-shot performance. In this work, we argue that online textual documents, e.g., Wikipedia, contain rich visual descriptions about object classes, therefore can be used as powerful unsupervised side information for ZSL. To this end, we propose I2DFormer+, a novel transformer-based ZSL framework that jointly learn to encode images and documents by aligning both modalities in a shared embedding space. I2DFormer+ utilizes our novel Document Summary Transformer (DSTransformer), a text transformer, that learns to encode a sequence of text into a fixed set of summary tokens. These summary tokens are utilized by a cross-model attention module that learns finegrained interactions between image patches and the summary of the document. Consequently, our I2DFormer+ not only learns highly discriminative document embeddings that capture visual similarities but also gains the ability to explain what regions of the image are important for the decision. Quantitatively, we demonstrate that I2DFormer+ significantly outperforms previous unsupervised semantic embeddings under both zero-shot and generalized zero-shot learning settings on three public datasets. Qualitatively, we show that our methods lead to highly interpretable results. Furthermore, we scale our model to the large scale zero-shot learning setting and show state-of-the-art performance on two challenging ImageNet benchmarks.

Keywords Zero-shot Learning · Multimodal Learning · Transformers

1 Introduction

“What does a tiger look like? It is a fierce animal that looks like a scary, big cat with stripes." Tigers are not native to Japan, yet when the travelers coming from China described them in relation to native animals, it inspired a range of historic paintings depicting tigers in Japan. Humans possess an impressive ability to imagine and identify unseen objects from pure language descriptions. In computer vision, the ability to predict unseen classes is called zero-shot learning, which can be achieved by transferring knowledge from seen classes using auxiliary side information (or semantic embeddings) e.g., attributes (Vyas et al., 2020), word embeddings (Frome et al., 2013), etc. Although remarkable progress has been made, most of prior works (Vyas et al., 2020; Akata et al., 2015; Xian et al., 2018; Zhu et al., 2019; Narayan et al., 2020; Chen et al., 2021) rely on human annotated attributes as the side information. While attributes are appealing, they are often costly to annotate (Song et al., 2018; Yu et al., 2013; Wah et al., 2011) and scale to large datasets. Towards unsupervised semantic embeddings (Socher et al., 2013; Frome et al., 2013; Akata et al., 2015), word embeddings can be easily obtained from pre-trained language models (Pennington et al., 2014). Yet, they often do not reflect fine-grained

Communicated by Vittorio Murino.

✉ Yongqin Xian
yxian@google.com

Muhammad Ferjad Naeem
ferjad.naeem@vision.ee.ethz.ch

Luc Van Gool
vangool@vision.ee.ethz.ch

Federico Tombari
tombari@google.com

[1] Computer Vision Lab, ETH Zürich, Zurich, Switzerland

[2] Google, Zurich, Switzerland

visual similarities, thus limiting the performance (Akata et al., 2015).

The goal of this work is to learn visually aligned unsupervised semantic embeddings from online textual documents for zero-shot image classification. With the advent of the Internet, the collective knowledge of humans about the world has been distilled into online encyclopedias like Wikipedia. These encyclopedias present a rich source of fine-grained auxiliary information for a model. While the entries (referred to as documents) may describe an object class with rich visual details, they tend to contain a lot of noise. For example, an entry for 'horse' can define its appearance as well as interesting historic events it participated in. While the former is helpful for a visual model, the latter might introduce noise making it challenging to fully exploit this knowledge.

This work is an extension of our previous NeurIPS22 work I2DFormer (Naeem et al., 2022). In this work, we propose an extension of I2DFormer called Image to Document Transformer+ (I2DFormer+) that learns to align image and document pairs with their global representations as well as with token-wise representations of its summary features, i.e., image patches and summary tokens. As a result, without any image-level language supervision, our model is able to develop an understanding of different parts of an animal, its habitat, etc, leading to a more discriminative semantic embedding. We summarise our contributions as:

1. We propose a novel text transformer Document Summary Transformer (DSTransformer). DSTransformer takes as input a text sequence and a set of learnable tokens. DSTransformer learns to summarize the text into a fixed set of output tokens.
2. With DSTransformer, we improve upon our novel Image to Document Attention (I2D Attention) module (Naeem et al., 2022) that learns to identify visually discriminative properties in a document leading to a more discriminative semantic embedding. With the introduction of DSTransformer, the memory footprint of attention remains constant allowing for scalability to large datasets like ImageNet.
3. Our model I2DFormer+ consistently improves the SOTA in unsupervised semantic embeddings on four challenging datasets, i.e., AWA2, CUB, FLO and ImageNet. Moreover, we qualitatively demonstrate that our model learns highly interpretable results.
4. We show that the learned document embedding can be used with any existing ZSL model to significantly improve its performance. To the best of our knowledge, I2DFormer Naeem et al. (2022) and I2DFormer+ are the first methods to learn an attention-based embedding from noisy documents for ZSL without relying on any pre-trained part localization model or attribute vocabulary.
5. We adapt our model on the ImageNet scale. I2DFormer+ sets a new SOTA on ImageNet scale zero-shot learning on two challenging dataset splits.

2 Related Works

Zero-shot Learning aims to generalize a model trained on seen classes onto a disjoint set of unseen classes using shared auxiliary information available for both sets (Vyas et al., 2020). Several methods in this direction learn a compatibility function between the image and the class embedding space (Romera-Paredes and Torr, 2015; Naeem et al., 2021; Changpinyo et al., 2016; Mancini et al., 2022; Akata et al., 2015; Zhang et al., 2017; Xian et al., 2016; Mancini et al., 2021). Another competing line of work uses generative models like GANs to learn the feature space of seen and unseen classes (Xian et al., 2018, 2019; Zhu et al., 2019, 2018; Verma et al., 2018; Schonfeld et al., 2019). A complementary line of work focuses on learning improved visual-semantic embeddings (Liu et al., 2018; Zhang et al., 2017; Jiang et al., 2016; Cacheux et al., 2019) and training better image encoders (Ji et al., 2018; Zhu et al., 2019; Xu et al., 2020). Semantic embeddings are a crucial building block for all of these methods. However, despite its importance, it is a less studied topic. Human labeled attributes (Xian et al., 2018; Patterson et al., 2014; Wah et al., 2011; Farhadi et al., 2009; Naeem et al., 2022) have become the de-facto semantic embedding for most methods. However, they are hard and expensive to scale as they require human experts (Song et al., 2018; Yu et al., 2013; Wah et al., 2011).

Zero-shot dataset transfer Zero-shot dataset transfer has emerged as a popular topic since the success of CLIP (Radford et al., 2021). CLIP trains a two tower transformer model for vision and language. These two transformers interact at the output layer with a dot product. Being trained on web-scale dataset of 400 million image and captions, CLIP shows great dataset transfer properties. Several works have followed up on CLIP with better training strategies (Cui et al., 2022), incorporating patch to word attention (Yao et al., 2022) and incorporating unsuperivsed training (Li et al., 2021). Several works have built upon the generalization abilities of CLIP by extending it to segmentation (Ghiasi et al., 2022; Lüddecke and Ecker, 2022) and detection (Gu et al., 2021). These works train a student network to learn the feature distribution of the visual encoder of CLIP and use the text encoder of CLIP to generate the classifiers resulting in inheriting some of open set abilities of the CLIP model. While zero-shot dataset transfer is a very promising topic, it differs from our task in a major way. The CLIP model is trained on a web-scale dataset and hence observes almost all visual concepts while training. In zero-shot image classification, we have a strong constraint that the model should not have observed any instance of the zero-shot classes.

Learning semantic embeddings with minimal supervision aims to use cheap to obtain side information to learn a semantic embedding with minimal label information. Several works have explored using text corpora as an alternative source of semantic embeddings. Some approaches include using word embeddings from pretrained language models (Yamada et al., 2020; Pennington et al., 2014; Mikolov et al., 2013) and knowledge graphs (Wang et al., 2018; Kampffmeyer et al., 2019; Bucher et al., 2017; Naeem et al., 2021; Mancini et al., 2022) to encode semantic similarities. Another line of work aims to directly learn semantic embeddings from documents containing information about classes. Earlier works in this direction used TF-IDF (Salton and Buckley, 1988) to directly embed the document in a joint image space (Elhoseiny et al., 2013). Successive works have focused on reducing the noise in the document by using predefined attribute vocabulary (Al-Halah and Stiefelhagen, 2017), learning better weights for TF-IDF embeddings (Qiao et al., 2016) or complementing these embeddings with a part detection network (Elhoseiny et al., 2017; Zhu et al., 2018). Recent works have incorporated Transformer based language models to directly embed a document to a semantic embedding (Kil and Chao, 2021; Bujwid and Sullivan, 2021). However, all these works either learn the semantic embedding against the global image representation or use a pretrained part detector for the human-labeled attributes to filter the relevant details. VGSE (Xu et al., 2022) instead proposes to directly learn semantic embeddings from images of seen classes and extrapolate them to the unseen classes by measuring their class name similarities. Our model, I2DFormer instead uses both the knowledge in text documents and the images of seen classes to learn a semantic embedding and ZSL model.

Learning cross-modal attention between image and text to ground text in images without region level supervision has been a long-studied problem in visual question answering, image captioning, etc. Das et al. (2017); De Vries et al. (2017); Rohrbach et al. (2016, 2017). Methods in this line of work learn a mapping between the region level features from an image and its caption. More recently, Transformers (Vaswani et al., 2017) have made a breakthrough in this field with models like ViLBERT (Lu et al., 2019) and FILIP (Rohrbach et al., 2017) that learn a cross-modal attention to learn cross modal embeddings. They show that the grounding of text in the image naturally emerges as a by-product (Xu et al., 2022). However, these works rely on having access to image-level text which is expensive to obtain. Our model instead addresses the much more challenging problem of learning a cross-modal embedding and attention from images and their class-level text document.

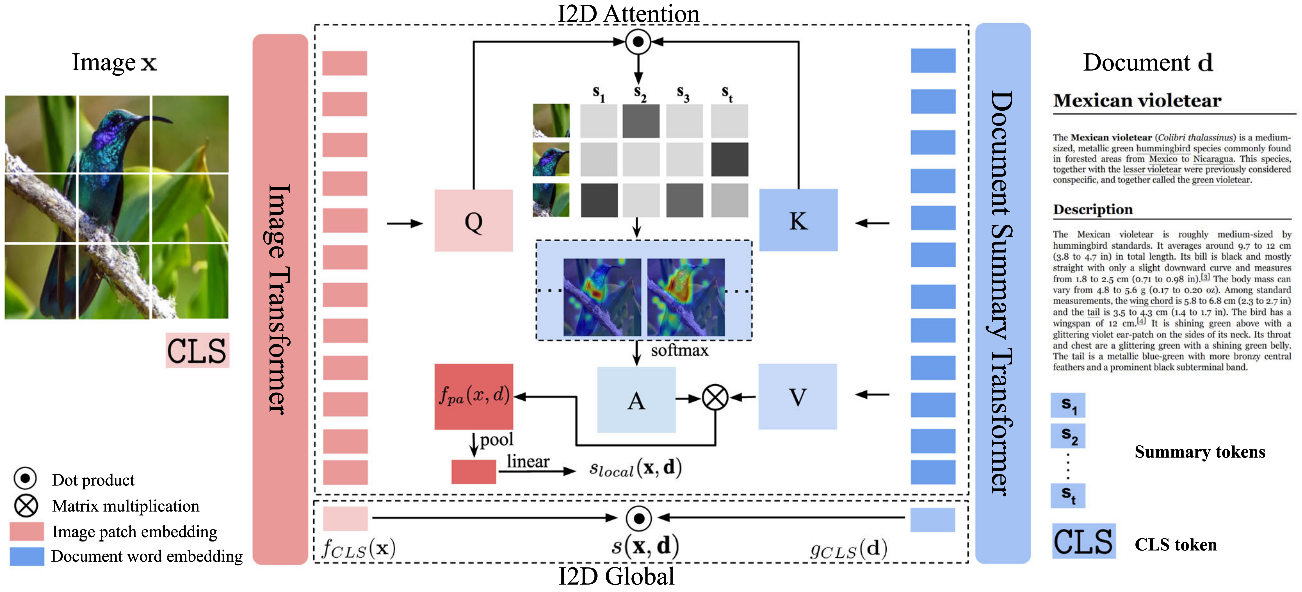
3 Image to Document Transformer (I2DFormer)

In this section, we re-introduce our previous work I2DFormer Naeem et al. (2022) for clarity since it is the base for I2DFormer+. The vast majority of existing ZSL works utilize either human-annotated attributes or word embeddings as auxiliary information. We instead utilize the textual collection of encyclopedia (wiki) entries of classes as side information given the wealth of free document collections describing object classes available on the internet. I2DFormer is a pure-transformer based ZSL framework that learns to align image and document pairs with their global representations and with token-wise representations i.e., image patches and document words. In the following section we introduce I2DFormer+, that improves the scalability of I2DFormer by learning summary tokens to encode documents. We show an overview of our method in Fig. 1.

Notations We define the classes that are included in the training set as seen classes $\mathcal{Y}^s$, and the classes that are excluded from training as unseen classes $\mathcal{Y}^u$. Let $\mathcal{T} = \{(\mathbf{x}, \mathbf{y}, \mathbf{d}) \| \mathbf{x} \in \mathcal{X}^s, \mathbf{y} \in \mathcal{Y}^s, \mathbf{d} \in \mathcal{D}^s\}$ be our training set where $\mathbf{x}$ denotes an RGB image from the training images $\mathcal{X}^s$, $\mathbf{y}$ is its label belonging to the seen classes $\mathcal{Y}^s$, $\mathbf{d}$ is a document e.g., Wikipedia article, containing textual descriptions of the object class $\mathbf{y}$, and $\mathcal{D}^s$ is a collection of documents describing seen classes. At test time, another collection of documents $\mathcal{D}^u$ describing the unseen classes $\mathcal{Y}^u$ will be made available to the model. This simulates an internet query to fetch extra information about an unseen class. Those documents will be used as the side information to connect seen and unseen classes. The task of ZSL is to make a prediction among only unseen classes, while GZSL needs to predict both seen and unseen classes.

3.1 I2D Global: Learning Joint Image-Document Embeddings with Transformer

Our model is a dual-stream transformer architecture. The model learns an embedding function $\mathcal{F}$, an image transformer (Dosovitskiy et al., 2021), for images, and $\mathcal{G}$, a document transformer (Vaswani et al., 2017), for text documents. The first part of our model learns a global compatibility between the Image and the Document by our Image to Document(I2D) Global module. On the image side, given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we reshape it into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \times C)}$, where (H, W) is the size of an input image with C as the RGB channels, (P, P) is the size of each image patch, and $N = HW/P^2$ is the resultant number of patches. Moreover, we append a `CLS` token to $\mathbf{x}_p$ as the input to the image transformer to learn a global image representation. Inspired by LiT (Zhai et al., 2022), we use a pretrained frozen image transformer (Dosovitskiy et al.,



Mexican violetear

The **Mexican violetear** (*Colibri thalassinus*) is a medium-sized, metallic green hummingbird species commonly found in forested areas from Mexico to Nicaragua. This species, together with the lesser violetear were previously considered conspecific, and together called the green violetear.

Description

The Mexican violetear is roughly medium-sized by hummingbird standards. It averages around 9.7 to 12 cm (3.8 to 4.7 in) in total length. Its bill is black and mostly straight with only a slight downward curve and measures from 1.8 to 2.5 cm (0.71 to 0.98 in).[3] The body mass can vary from 4.8 to 5.6 g (0.17 to 0.20 oz). Among standard measurements, the wing chord is 5.8 to 6.8 cm (2.3 to 2.7 in) and the tail is 3.5 to 4.3 cm (1.4 to 1.7 in). The bird has a wingspan of 12 cm.[4] It is shining green above with a glittering violet ear-patch on the sides of its neck. Its throat and chest are a glittering green with a shining green belly. The tail is a metallic blue-green with more bronzy central feathers and a prominent black subterminal band.

Fig. 1 I2DFormer+, our novel Transformer based model, uses noisy documents as auxiliary information to learn a zero-shot model. Our Document Summary Transformer takes a document and a set of learnable summary and CLS tokens to extract the local summary and global class information contained in the document. The first part of the model, I2D Global, learns to encode images and noisy documents to a shared embedding space using the output embedding corresponding to the CLS token. In order to distill discriminative local information from the document, our I2D Attention module(Naeem et al., 2022) learns fine-grained interactions between image patches and document words. I2DFormer uses the output of the tokenized input document for I2DAttention while I2DFormer+ uses the output corresponding to the learnable summary tokens. Together, the two modules learn a highly discriminative document semantic embedding `I2DEmb`

2021). This is followed by a learnable feature projection layer that maps the image embeddings to a joint image-document embedding space with dimensionality r. The image encoder $\mathcal{F}$ outputs $f_{CLS}(\mathbf{x}) \in \mathbb{R}^r$ as the global image feature and $f_p(\mathbf{x}) \in \mathbb{R}^{N \times r}$ as the patch-wise image embedding for the input image where r is the feature dimension.

On the document side, given a document $\mathbf{d}$ consisting of M words, we get its token-wise input feature representation with a pretrained word embedding model. Note that we use words and tokens interchangeably as we use GloVe word features as tokens (Pennington et al., 2014). Since each document consists of a typically long sequence of words, we further pass this feature representation through a learnable MLP as a token projection layer to reduce the feature dimension and the memory footprint, yielding $\mathbf{d}_t \in \mathbb{R}^{M \times r}$, where r is the feature dimension as the output of the token projection layer. Our learnable document transformer consists of transformer encoder blocks with multi-head attention. We append a `CLS` $\in \mathbb{R}^r$ token to this sequence and pass it through the document transformer to get $g_{CLS}(\mathbf{d}) \in \mathbb{R}^r$ as the global document embedding and $g_t(\mathbf{d}) \in \mathbb{R}^{M \times r}$ as the word-wise text embedding for the input document. We later refer to the learned $g_{CLS}(\mathbf{d})$ as a document embedding (semantic embedding) `I2DEmb` that can be used by any ZSL method.

We define a scoring function $s : \mathcal{X} \times \mathcal{D} \rightarrow \mathbb{R}$ that measures the similarity of any image $\mathbf{x}$ and document $\mathbf{d}$ pair. The scoring function computes the dot product between global image embedding $f_{CLS}(\mathbf{x})$ and document embedding $g_{CLS}(\mathbf{d})$, formulated as

$$s(\mathbf{x}, \mathbf{d}) = f_{CLS}(\mathbf{x}) \cdot g_{CLS}(\mathbf{d}). \tag{1}$$

The learning objective is to make the scoring function assign high scores to correct image and document pairs and low scores to incorrect ones. Therefore, for a particular training instance $(\mathbf{x}, \mathbf{y}, \mathbf{d})$, and $\mathcal{D}^s$ the collection of documents belonging to seen classes, we minimize the following cross-entropy loss,

$$L_{CLS} = -\log\left(\frac{\exp s(\mathbf{x}, \mathbf{d})}{\sum_{\mathbf{d}' \in \mathcal{D}^s} \exp s(\mathbf{x}, \mathbf{d}')}\right) \tag{2}$$

3.2 I2D Attention: Learning Image Patch to Document Word Attention

Our I2D Global module essentially aligns image-document pairs using their global representations. While this paradigm has been popularized by influential works like CLIP (Radford et al., 2021), it relies on a large amount of image-text pairs to learn all discriminative local features and represent them in the output of `CLS` token. However, we are dealing with a more challenging problem where the number of training images is small (a few thousand) and there is only one document associated with each class. Aligning two modalities at a global level will be prone to overfitting and hard to generalize to unseen classes at test time. Moreover, our documents are directly collected from the Internet and therefore are noisy e.g., a large portion of the words are irrelevant to visual appearance. To address these challenges, we proposed I2D Attention (Naeem et al., 2022), a novel cross-modality attention module, to learn fine-grained interaction between image patches and document words, capturing local features defined in the document such as body parts of an animal, their habitat in the form of image background, etc. We argue that learning these local mappings allows a model to generalize beyond the seen classes.

Our I2D Attention module takes as inputs the patch-wise embeddings $f_p(\mathbf{x}) \in \mathbb{R}^{N \times r}$ of the image and the token-wise embeddings $g_t(\mathbf{d}) \in \mathbb{R}^{M \times r}$ of the document. We task the model with searching for the visually-relevant words in the documents using image patches as the queries. More specifically, we define $Q = f_p(\mathbf{x})W_q$ as the image queries, $K = g_t(\mathbf{d})W_k$ as the text keys to compare with, and $V = g_t(\mathbf{d})W_v$ as the text values to mix with after the search, where W_q, W_k and W_v are learnable linear transformations, all in size $r \times r$. The I2D Attention module estimates the cross-modal attention $A(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{N \times M}$ by computing a dot product between every image patch and word pair followed by a softmax,

$$A(\mathbf{x}, \mathbf{d}) = softmax\left(\frac{QK^T}{\sqrt{r}}\right) \tag{3}$$

This attention matrix is used to compute new feature representations $f_{pa}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{N \times r}$ for all image patches as linear combinations of rows of the value matrix V i.e., $f_{pa}(\mathbf{x}, \mathbf{d}) = A(\mathbf{x}, \mathbf{d})V$. Intuitively, this operation recomputes the image patch embeddings using the token-wise embeddings of relevant words in a document. To obtain the image-level embedding, we apply global pooling on the patch dimension N of the new patch embeddings $f_{pa}(\mathbf{x}, \mathbf{d})$, yielding $\hat{f}_{pa}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{1 \times r}$. Afterwards, we compute the local alignment score between an image-document pair by applying a simple linear layer,

$$s_{local}(\mathbf{x}, \mathbf{d}) = H(\hat{f}_{pa}) \tag{4}$$

where $H \in \mathbb{R}^{r \times 1}$ is a learnable linear layer. Given a particular training example $(\mathbf{x}, \mathbf{y}, \mathbf{d})$, we optimize the following cross-entropy loss,

$$L_{local} = -log\left(\frac{\exp s_{local}(\mathbf{x}, \mathbf{d})}{\sum_{\mathbf{d}' \in \mathcal{D}^s} \exp s_{local}(\mathbf{x}, \mathbf{d}')}\right) \tag{5}$$

We do not use any skip connection similar to previous cross-modal attention blocks like ViLBERT (Lu et al., 2019) as we want the attention weighted embedding to directly give us a linearly separable representation. Our I2D Attention Module searches for relevant patch features in the document for each training class $y \in \mathcal{Y}_s$ and learns to associate visual concepts with the noisy text in the document. The classification loss L_{local} maximizes the contribution of discriminative words in the document and minimizes the contribution of irrelevant details about a class. Furthermore, calculating cross-entropy over the full seen set ensures that the model is aware of similar attributes between fine-grained classes and can pick additional cues to separate such classes. I2D Attention introduces minimal learnable parameters with only 4 additional linear layers and rather forces $\mathcal{F}$ and $\mathcal{G}$ to minimize irrelevant details in the tokenwise $f_p(\mathbf{x})$ and $g_t(\mathbf{d})$ as well as the global $f_{CLS}(\mathbf{x})$ and $g_{CLS}(\mathbf{d})$ embeddings. This is in contrast to architectures like ViLBERT (Lu et al., 2019) where several self-attention layers are stacked on top of the cross-modal module to further learn the output embedding with self-attention. We later show in our experiments that this can hurt the performance in our data constrained zero-shot learning setup. Although documents have been explored before in ZSL, prior work uses fixed document embeddings that are encoded with TF-IDF (Elhoseiny et al., 2013; Zhu et al., 2018; Elhoseiny et al., 2017) or extracted from a pre-trained language model (Bujwid and Sullivan, 2021; Kil and Chao, 2021). In contrast, our document embeddings are aided by the attention module to identify important details and thus are also assisted by this additional visual information.

3.3 I2DFormer+: Improving Noise Robustness and Compute Efficiency of Local Attention

I2DFormer (Naeem et al., 2022) allows for directly learning the class embeddings from documents by optimising for a global and a local alignment. While the global alignment is only dependent on the dot product with the CLS token, the local alignment relies on an expensive cross modality attention. The computational complexity of this grows with both the number of classes and the length of the class documents. Since Transformer memory requirement grows quadratically with the input sequence, this can become prohibitively expensive on classes with long documents or datasets with a large number of classes such as ImageNet. Moreover, it tries to align the full text with an image which contains noise. We address these limitations by proposing I2DFormer+ which extracts the local information available in a class document into a fixed set of summary tokens. These learnable summary tokens allow for a fixed computation cost of cross modality alignment independent of the length of a document.

Generating Summary features from Document Each class $\mathbf{y}$ is associated with a document $\mathbf{d}$ describing the class.

An exhaustive search over all words of documents of each class in I2DAttention becomes increasingly expensive as the length of documents increase. We propose Document Summary Transformer (DSTransformer), a text transformer that aims to learn to summarize the highly discriminative features available in a document of a class into a fixed set of tokens (Jaegle et al., 2021). It is important to note that we define summary as a set of features corresponding to learnable tokens rather than a human interpretable summary. DSTransformer replaces the Document Transformer used in I2DFormer Naeem et al. (2022).

Given a text document $\mathbf{d}$ consisting of M words, we get its token-wise input representation with a pretrained word embedding model. Similar to I2DFormer, we pass this through a learnable MLP as a token projection layer to reduce the feature dimension and memory footprint to yield $\mathbf{d}_t \in \mathbb{R}^{M \times r}$. We introduce $\mathcal{S} = \{s_1, s_2, ...s_T\} \in \mathbb{R}^{T \times r}$ as a set of T learnable tokens that are appended to the earlier tokenized representation of the document. These tokens are introduced to summarize the discriminative information available in each document. Given $T < M$, this results in significantly reduced constant memory requirement of the later I2D Attention independent from the length of the input document. Moreover, we introduce a CLS $\in \mathbb{R}^r$ token that is tasked with summarizing the global information available in a document. $\mathcal{S}$ and CLS are appended to the document of each class and passed through DSTransformer consisting of several Transformer encoder blocks. We take the output representation corresponding to tokens in $\mathcal{S}$ to get $g_S(d)$, the local summary of the document and $g_{CLS}(d)$ as the global feature of the document. The $g_{CLS}(d)$ is used in the I2D Global module and $g_S(d)$ is used in the I2D Attention module to learn global and local alignment between the image and the document.

3.4 Inference

Given an input image $\mathbf{x}$, we search for the document $\hat{\mathbf{d}}$ that yields the highest compatibility score,

$$\hat{\mathbf{d}} = \arg\max_{\mathbf{d}' \in \mathcal{D}} s(\mathbf{x}, \mathbf{d}'). \tag{6}$$

The search space includes only documents of unseen classes in zero-shot learning i.e., $\mathcal{D} = \mathcal{D}^u$, and all classes in generalized zero-shot learning (GZSL) i.e., $\mathcal{D} = \mathcal{D}^s \cup \mathcal{D}^u$. The final prediction is simply the class label associated with the document $\hat{\mathbf{d}}$. For GZSL, we apply calibrated stacking (Chao et al., 2016) to calibrate the activations of unseen classes on a held-out set to reduce the bias towards seen classes. We only use the output of the global prediction as it is computationally cheaper and has distilled the knowledge of patch-to-token interactions while training. The attention between image patch and summary tokens is computed as the explainability of the model's decision when required.

4 Experiments on Small Scale Datasets

We conduct extensive experiments on Animals with Attributes2 (AWA2) (Xian et al., 2018), Caltech-UCSD Birds (CUB) (Wah et al., 2011) and Oxford Flowers (FLO) (Nilsback and Zisserman, 2008), which are widely used datasets in ZSL. We follow the evaluation protocol and data splits proposed by Xian et al. (2018). Since the main focus of this work is to learn unsupervised semantic embeddings, we do not use any human-annotated attributes. In the following, we first describe how documents are collected and implementation details. Then, we quantitatively compare against SOTA unsupervised semantic embeddings methods and ZSL methods. Finally, we show quantitative results to demonstrate the interpretability of our method.

Collecting documents We use online sources for documents that can be queried with minimal human supervision. These sources contain useful knowledge about each class but might have a lot of noise as irrelevant textual details. For AWA2, we use A-Z Animals [69], an animal encyclopedia. For CUB, we use Wikipedia [70]. For FLO, we use a collection of gardening blogs and Wikipedia [70] to collect documents for these classes. However, we found documents for flowers to be less focused on the patterns of petals and pistils and rather more focused on the general description of the plant and its taxonomic biological classification. FLO is therefore a challenging dataset to generalize from document-based embeddings. We adopt a simple filtering step on these collected articles similar to Kil and Chao (2021). We look at the documents for 10% of classes of each dataset and identify sections that contain relevant information about the class. The rest of the documents are filtered to only contain these sections. The average size of a document is ≈400 words. To put this into perspective, models like CLIP (Radford et al., 2021; Yao et al., 2022) use image captions of at max 77 tokens (Radford et al., 2021; Pham et al., 2023). The long length of the documents presents an additional challenge. The collected documents are attached with this submission as supplementary material.

Training Details We implement our model in PyTorch and train on an Nvidia A100 GPU. We use the VIT/B16 checkpoint trained on ImageNet 1k by Dosovitskiy et al. (2021) as the pretrained Image Transformer. The image patch projection and token projection layers are implemented as a shallow MLP. Maxpool or Meanpool are chosen as global pooling by ablation. The model is trained with Adam optimizer with a learning rate of $1e^{-3}$ and takes ≈24 h to converge. L_{CLS} and L_{local} relative weights are chosen by ablation. More details are available in the supplementary. For baseline methods, we use the CLS features from the same VIT/B16 checkpoint with author's implementations. We ablate these methods over multiple hyperparameters to report the best run. For VGSE, we use the semantic embeddings released by the original authors (not available for FLO).

4.1 Comparison with SOTA Unsupervised Semantic Embeddings

In this section, we compare with existing unsupervised semantic embeddings where they are obtained without using human supervision using the same ZSL method (our I2D global module).

Compared semantic embeddings For GloVe (classname) (Pennington et al., 2014), we simply extract GloVe vectors of class names. This method has been adopted by many prior ZSL methods (Norouzi et al., 2014; Frome et al., 2013; Akata et al., 2015; Naeem et al., 2021) due to its simplicity. For GloVe (Document) (Pennington et al., 2014), we average over the feature vectors of each word in the document. LongFormer (Beltagy et al., 2020) is a text transformer model trained for documents and outputs a CLS embedding given a document. MPNet(Song et al., 2020) is the current SOTA Sentence Transformer model(Reimers and Gurevych, 2019) trained to optimize embeddings for natural language classification tasks. Since the original model is trained for short sequences, we average over the individual sentence embeddings similar to Kil and Chao (2021); Bujwid and Sullivan (2021). TF-IDF (Salton and Buckley, 1988) stands for Term Frequency-Inverse Document Frequency, which has been used by some prior ZSL methods (Elhoseiny et al., 2013; Lei Ba et al., 2015). VGSE (Xu et al., 2022) learns the semantic embeddings from image patches and word embeddings of class names. Since these embedding models generate one embedding for the whole document, we replace the Document Transformer with an equally deep MLP.

Results From Table 1, we observe that our method I2DFormer+ consistently outperforms all semantic embedding methods in both ZSL and GZSL. Compared to GloVe (Document) (Pennington et al., 2014), which also serves as an input to our method (without the average over words), the learned embedding of our model achieves an impressive 77.3% accuracy vs 61.6% on AWA2, 45.9 % vs 29.0 % on CUB and 41.3% vs 25.8% on FLO with a relative 1.25×, 1.5× and 1.6× improvement each. This shows that our learned document embedding assisted by our I2D attention module significantly improves the zero-shot performance. We see that these improvements are also consistent in GZSL where we see a significant improvement in the HM. Similar results are observed for other pretrained language semantic embeddings Longformer, MPNet and TF-IDF (Beltagy et al., 2020; Song et al., 2020; Salton and Buckley, 1988). Since the original embedding models for these baselines were only trained on language data, the generated semantic embedding is unlikely to capture the most visually discriminative features described in the document. Our model however is able to learn a more informed semantic embedding thanks to supervision from the images of the seen classes. Comparing rows 1 and 2, we see that the use of documents over classnames leads to a major improvement as documents capture better class similarities.

Compared to VGSE (Xu et al., 2022), a strong unsupervised semantic embedding baseline, we observe that our model again substantially outperforms it. While both VGSE and our model exploit patch-wise similarities in images of different classes to learn a class embedding, our model is additionally able to complement this embedding with localized information available from the documents thanks to our I2D Attention. Finally, I2DFormer+ achieves better performance than I2DFormer (Naeem et al., 2022) thanks to our DSTransformer which reduces the noise in the summary tokens and allows for easier visual alignment while reducing computational cost.

4.2 Comparing with SOTA ZSL Methods

In this section, we compare our full model I2DFormer+ with existing SOTA zero-shot models across baseline embeddings and our learned document embedding. For a fair comparison, we evaluate those methods with the same VIT/B16 image features. The GloVe baseline refers to encoding the document with the average over the per token GloVe embeddings. We show in Table 2 that our new method I2DFormer+ and I2DFormer (Naeem et al., 2022), and their learned document embeddings I2DEmb+ and I2DEmb (Naeem et al., 2022) achieve SOTA performance.

Compared to baselines, our model I2DFormer+ or our learned embedding consistently outperform all baseline ZSL methods and embeddings to establish a new SOTA. I2DFormer+ achieves SOTA ZSL performance on CUB and FLO, the fine-grained datasets. On CUB, I2DFormer+ achieve an impressive 45.9% compared to the closest 43.7% of APN that also uses our I2DEmb+. On FLO, I2DFormer+ achieves 41.3% compared to the closest 40.1% of f-VAEGAN-D2 that again uses our I2DEmb+. In GZSL, on CUB, I2DFormer+ achieves 45.3% HM compared to the closest 42.2% of f-VAEGAN-D2 (I2DEmb). On FLO, I2DFormer+ achieves an impressive 51.8% HM compared to the closest 50.5% of APN (I2DEmb+). We would like to emphasize that our model is outperforming both the generative baselines in GZSL on these two datasets. Generative models have previously been shown to be the most competitive baselines in these datasets. However, since I2DFormer+ learn a fine-grained attention between the image patches and the words in the article, it is able to outperform these baselines with this extra knowledge without requiring feature

Table 1 Comparing our I2DFormer+ with unsupervised semantic embedding methods using the same image feature and method (our I2D Global module)

Semantic Embedding	Source	Zero-Shot Learning			Generalized Zero-Shot Learning								
		AWA2	CUB	FLO	AWA2			CUB			FLO		
		T1	T1	T1	u	s	H	u	s	H	u	s	H
GloVe(Pennington et al., 2014)	CLSN	52.1	20.4	21.6	42.1	75.3	54.0	16.2	43.6	23.6	14.4	88.3	24.8
GloVe(Pennington et al., 2014)	DOC	61.6	29.0	25.8	49.5	78.1	60.6	23.8	62.6	34.5	14.7	91.0	25.3
LongFormer (Beltagy et al., 2020)	DOC	44.2	22.6	8.8	41.6	81.8	55.2	19.9	41.0	26.8	8.8	89.8	16.0
MPNet(Song et al., 2020)	DOC	61.8	25.8	26.3	58.0	76.4	66.0	20.6	44.3	28.2	22.2	**96.7**	36.1
TF-IDF(Salton and Buckley, 1988)	DOC	46.4	39.9	34.0	29.6	**87.6**	44.2	29.0	52.1	37.3	28.9	94.8	44.3
VGSE(Xu et al., 2022)	IMG + CLSN	69.6	37.1	-	56.9	82.8	67.4	27.6	**70.6**	39.7	-	-	-
I2DFormer(Ours) (Naeem et al., 2022)	IMG + DOC	76.4	45.4	40.0	66.8	76.8	71.5	35.3	57.6	43.8	35.8	91.9	51.5
I2DFormer+ (Ours)	IMG + DOC	**77.3**	**45.9**	**41.3**	**69.8**	83.2	**75.9**	**38.3**	55.2	**45.3**	**36.9**	86.9	**51.8**

In ZSL, we report top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**). We consider semantic embeddings that are either directly extracted (with a pretrained language model) or learned from different sources including classnames (CLSN), document (DOC), a combination of image and classnames (IMG+CLSN), and a combination of image and document (IMG+DOC). Our new model I2DFormer+ significantly improves on the baselines to set a new SOTA for unsupervised class embeddings.
Bold refers to the best performance result as normal in Computer Vision community

generation. On AWA2, a coarse classification dataset, we see that I2DFormer+ achieves SOTA performance among the Discriminative baselines. However, the best performance is achieved by the Generative baseline f-VAEGAN-D2 using I2DEmb on this dataset. f-VAEGAN-D2 with I2DEmb achieves the best ZSL accuracy of a remarkable 85.1% vs. the closest 84.0% achieved by the same method with I2DEmb+. In GZSL, f-VAEGAN-D2 with I2DEmb achieves SOTA with an impressive HM of 77.2% followed by 75.9% of the same method with I2DFormer+ embeddings. These baselines are only able to outperform I2DFormer with our learned I2DEmb+ and I2DEmb.

4.3 Ablation Study

What kind of Patch to Word Attention is required in ZSL? We study the importance of learning patch to word attention for Document based embeddings in Table 3. We see that while only training I2DGlobal can learn a competitive ZSL model, it significantly improves and achieves SOTA performance with the introduction of our I2D Attention module (Naeem et al., 2022) in I2DFormer+. We see a relative 14%, 16%, and 8% improvement over I2DGlobal. This validates our hypothesis that the patch to word attention distills its knowledge to the global I2DEmb+, improving its performance. In the same table, we also ablate over 2 competing cross-modal attention modules. FILIP (Yao et al., 2022) is a recent method that proposes to associate each image patch to its most attended word. We see that this hurts the performance when using noisy Documents. ViLBERT (Lu et al., 2019) proposes a cross-modal attention module which is paired with a self-attention block (Vaswani et al., 2017) to learn an image embedding. We see that while this improves the performance over I2DGlobal on AWA2, it leads to worse performance on our fine-grained datasets CUB and FLO potentially due to the bigger model requiring more training data. Our I2D Attention outperforms both these baselines and achieves SOTA performance.

What kind of input text representation works best for I2DFormer+? We ablate over several pretrained word/ token representations to be used as an input to our Document Summary Transformer (DSTransformer) in Table 4 and note that GloVe (Pennington et al., 2014) achieves the best result. We observe that the Transformer based language models LongFormer (Beltagy et al., 2020) perform much worse than older baselines Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). We conjecture that this is due to the DSTransformer having limited text data for the seen classes while training. Transformer-based models generate different

Table 2 Comparing I2DFormer+ with baseline ZSL methods, under various unsupervised semantic embeddings we see that our model and embeddings (I2DEmb+ and I2DEmb) set a new SOTA

Type	ZSL Model	Semantic Embeddings	Zero-Shot Learning			Generalized Zero-Shot Learning								
			AWA2	CUB	FLO	AWA2			CUB			FLO		
			T1	T1	T1	u	s	H	u	s	H	u	s	H
Generative	GAZSL(Zhu et al., 2018)	GloVe	63.7	37.5	20.9	22.2	90.8	35.6	5.93	36.2	10.2	8.38	97.3	15.4
		VGSE	74.7	35.7	-	29.5	93.8	44.9	10.5	51.8	10.5	-	-	-
		I2DEmb(Ours)	83.1	42.9	34.2	56.8	**94.7**	71.0	15.9	50.4	24.1	28.8	90.1	43.7
		I2DEmb+ (Ours)	77.0	43.0	36.2	45.9	86.7	60.0	16.8	59.0	26.2	29.1	98.3	44.9
	f-VAEGAN-D2 (Xian et al., 2019)	GloVe	70.7	31.8	32.1	65.7	69.5	67.6	23.9	55.7	33.5	25.0	**99.0**	39.9
		VGSE	75.0	40.7	-	70.8	79.0	74.7	32.7	57.5	41.7	-	-	-
		I2DEmb(Ours)	**85.1**	41.9	36.9	73.2	81.7	77.2	33.4	57.3	42.2	30.0	97.3	45.8
		I2DEmb+ (Ours)	84.0	35.5	40.1	67.8	86.2	75.9	33.2	57.4	42.1	32.4	98.7	48.8
Discriminative	SJE(Akata et al., 2015)	GloVe	56.6	27.1	13.1	41.3	83.4	55.3	14.4	51.6	22.5	4.6	93.2	8.7
		VGSE	70.1	31.6	-	49.9	84.8	62.8	23.1	57.5	33.0	-	-	-
		I2DEmb(Ours)	72.6	38.2	33.4	55.8	82.6	66.6	25.0	56.2	34.6	18.5	87.1	30.5
		I2DEmb+ (Ours)	72.8	40.2	39.8	66.8	80.2	72.9	30.3	52.1	38.3	34.6	92.7	50.4
	APN(Xu et al., 2020)	GloVe	73.8	20.7	15.2	57.6	84.6	68.5	19.6	32.6	24.5	12.8	39.4	19.3
		VGSE	74.0	34.3	-	65.0	72.4	68.5	23.2	52.9	32.1	-	-	-
		I2DEmb(Ours)	74.5	40.6	35.4	65.5	76.9	70.7	30.0	49.9	37.5	32.0	85.3	46.5
		I2DEmb+ (Ours)	74.8	43.7	39.7	70.9	71.2	71.1	37.2	45.2	40.8	36.3	83.0	50.5
	I2DFormer(Ours) (Naeem et al., 2022)	I2DEmb (Ours)	76.4	45.4	40.0	66.8	76.8	71.5	35.3	57.6	43.8	35.8	91.9	51.5
	I2DFormer+ (Ours)	I2DEmb+ (Ours)	77.3	**45.9**	**41.3**	69.8	83.2	75.9	**38.3**	55.2	**45.3**	**36.9**	86.9	**51.8**

In ZSL, we report top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**). Best embedding results within a method are underlined
Best results overall are **bolded**

Table 3 Ablation over I2DFormer+

Model	AWA2	CUB	FLO
I2DGlobal	69.4	37.2	37.2
I2DGlobal + FILIP(Yao et al., 2022)	67.3	35.7	38.3
ViLBERT(Lu et al., 2019)	75.0	29.9	21.3
I2DFormer+	**77.3**	**45.9**	**41.3**

The proposed I2DGlobal module greatly benefits from the addition of I2D Attention to achieve SOTA performance. Comparing against FILIP and VilBERT cross-modal attention, we see that I2D Attention achieves SOTA
Bold refers to the best performance result as normal in Computer Vision community

Table 4 Ablating over input embeddings for our Document Summary Transformer we see that older models like Word2Vec and GloVe serve as better input representation than modern Transformer-based language models

Input Embedding	AWA2	CUB	FLO
LongFormer(Beltagy et al., 2020)	53.9	39.8	26.1
Word2Vec(Mikolov et al., 2013)	75.0	44.6	39.0
GloVe(Pennington et al., 2014)	**77.3**	**45.9**	**41.3**

Bold refers to the best performance result as normal in Computer Vision community

word features for the same word with self-attention (Vaswani et al., 2017). Documents of unseen classes use the same and additional vocabulary in new sentences causing a distribution shift in their input representation.

Ablation between Global and Local Scores I2DFormer+ learns a global score s in the I2DGlobal module and a local score s_{local} in the I2DAttention module. We additionally report the performance and ablation with s_{local} in Table 5. Comparing row 1 and row 2, we see that only training the individual block already results in a competitive model. However, I2D Global achieves better performance as learning

Table 5 Ablation on the scoring heads of I2DFormer+ on the I2DGlobal and I2DAttention modules

Model	Scoring	Zero-Shot Learning			Generalized Zero-Shot Learning								
		AWA2	CUB	FLO	AWA2			CUB			FLO		
		T1	T1	T1	u	s	H	u	s	H	u	s	H
I2D Global	s	69.4	37.2	37.2	59.1	79.7	67.8	28.5	**59.1**	38.4	28.4	88.2	43.0
I2D Attention	s_{local}	65.7	37.1	25.8	61.1	73.1	66.6	27.5	50.2	35.5	24.1	93.7	38.3
I2DFormer+	s_{local}	75.3	45.0	39.2	65.1	81.9	72.5	35.2	51.9	41.7	33.3	95.8	49.5
I2DFormer+	s	**77.3**	**45.9**	**41.3**	**69.8**	83.2	**75.9**	**38.3**	55.2	**45.3**	**36.9**	86.9	**51.8**

I2DGlobal computes the global score s and I2DAttention computes the local score s_{local}. We observe that jointly training both leads to the best performance for both s_{local} and s setting a SOTA. In ZSL, we report top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**)

Bold refers to the best performance result as normal in Computer Vision community

Table 6 Ablating over number of summary tokens for our Document Summary Transformer we see that our AWA achieves the best results at 128 while CUB and FLO achieve the best results at 256 tokens

Summary Tokens	AWA2	CUB	FLO
64	75.2	45.8	39.5
128	**77.3**	45.3	39.6
256	75.6	**45.9**	**40.9**

Bold refers to the best performance result as normal in Computer Vision community

cross-modal attention is a harder task than matching global embeddings. Comparing row 2 and 3, we see that combining both modules lead to a major improvement in s_{local} as it distills the knowledge of the global embedding. We see that the two modules of I2DFormer+ have a symbiotic relationship where both greatly benefit from joint training and achieve a boost in performance. Comparing rows 3 and 4, we see that the global score s achieves better performance as it additionally uses global information of the image and the document and sets the state-of-the-art.

Ablation over Summary Tokens We ablate over the number of summary tokens in Document Summary Transformer in Table 6. These tokens summarise the local information available in the document into a fixed set of tokens used for cross modal alignment in our I2DAttention module. We notice that for AWA, we get the best performance at 128 tokens. For the two finegrained datasets CUB and FLO, the best performance requires more tokens and is achieved at 256 tokens.

4.4 Qualitative Results

Document Transformer attention for I2DEmb+ We look at the learned attention over documents of unseen classes in Table 7 and plot the top 8 most attended words across the Document Transformer attention heads for I2DEmb+. On AWA2, we see that class name is complemented with human-like labelled attributes for these classes such as the color of the animal, type of the feet, and habitat etc. For the fine-grained datasets, CUB and FLO, we see that for similar classes like the two warblers, the model learns similar attributes like "ruby-crowned" as well as discriminating "tiger stripes" vs "chestnut patterns". We confirm our hypothesis that a learned document embedding will focus on discriminating properties of the class from the noisy document.

Visualizing Document word to Image attention as the column of the attention matrix in Fig. 2a, we see the impressive localization ability of I2DFormer Naeem et al. (2022) for the top attended words in I2DEmb. We see that the model is able to localize the unseen classes horse and giraffe in the image despite never observing them while training. The discriminating properties like the hoofed legs are also localized in the image. For CUB, we see that between the two very similar images of two unseen classes, the model identifies the yellow bottom as an important property from the two different documents of the ground truth class. However, the model is further able to identify the discriminative tiger stripes of the Cape May Warbler to differentiate it from the Tropical Kingbird which has gray-green feathers leading to correct classification. Finally, on FLO, the localization ability of I2DFormer remains consistent where the Peruvian lily is identified by localizing it as a Lily and identifying its stripped and curved petals. Similarly for Globe Thistle, the model is able to differentiate the sharp teeth, soft and wrinkled parts of the flower. The prevalence of these words as top attended words in the document transformer and their impressive localization verifies our hypothesis that the attention module distills its knowledge to the CLS head. A model that does not learn patch to word attention can miss these properties if they are not deemed important among the seen classes.

Visualizing Summary Token to Image attention in I2DFormer+ reveals similar interpretability to I2DFormer Naeem et al. (2022). Since I2DFormer+ abstracts the documents information into a fixed set of tokens, this attention is computed against each token as shown in Fig. 2b. We visu-

Table 7 Top attended words for I2DEmb+ for unseen classes in the Document Transformer consist of discriminative properties available in the document

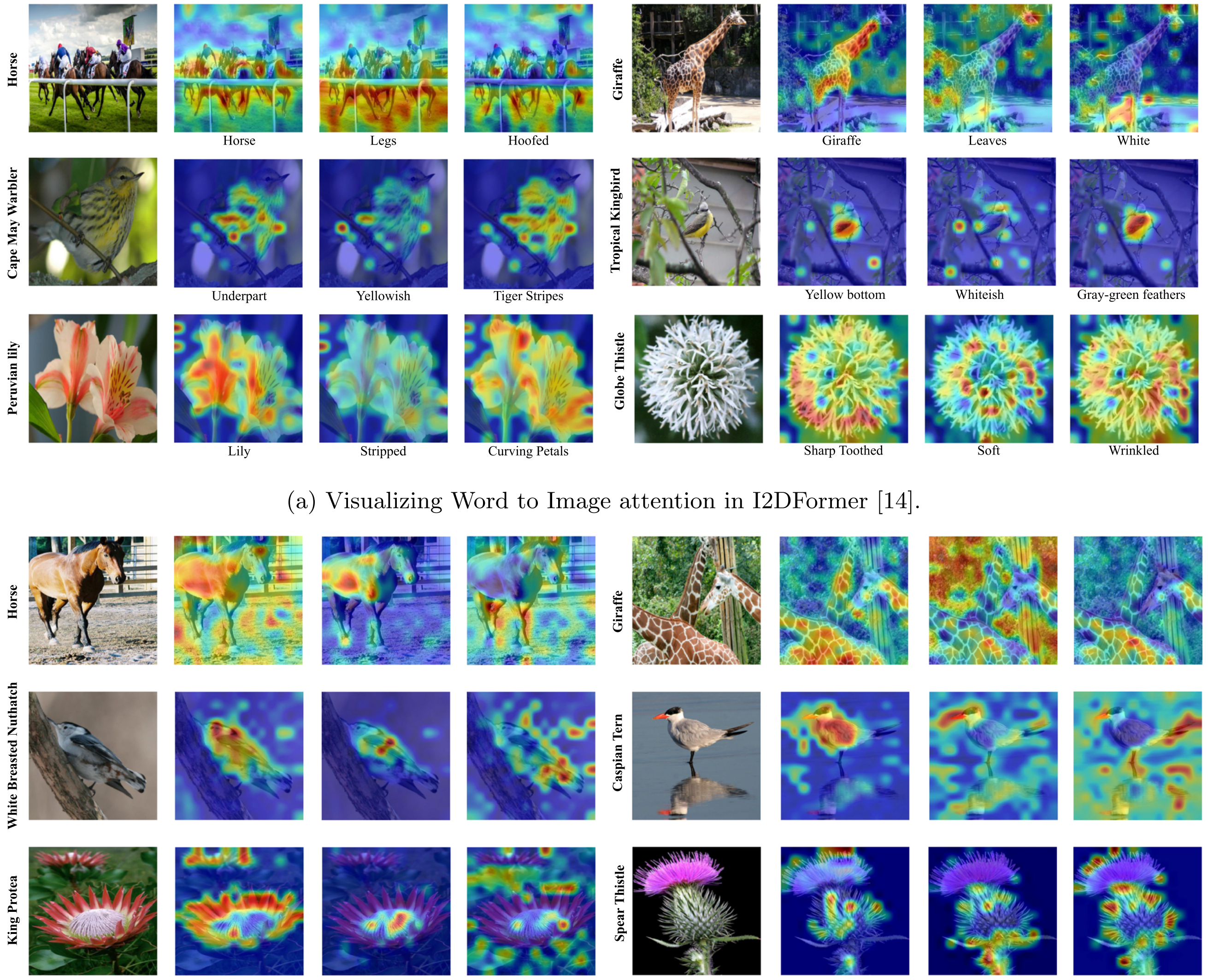
	Classname	Top attended words for I2DEmb
AWA2	Blue Whale	enourmous, polar, greyish, massive, blue, smooth, water, whale
	Sheep	fluffy, grass, wooly, horns, rams, stocky, hooves, sleeker
	Seal	blubber, saltwater, fur, seal, ocean, frigid, elongated, pinnipeds
	Giraffe	hoofed, mammals, woodlands, leaves, markings, reddish, pigmented, mane
CUB	Green Violetear	black, shining, bronzy, green chest, colibrim, canopy, glittering
	Tropical Kingbird	dark color, gray, rural, venezuela, gray-green feathers, flycatcher, gray-headed, kingbird
	Cape May Warbler	tiger stripes, short-tailed, decurved, ruby-crowned, white, olive, cape, yellowish
	Chestnutsided Warbler	crown, wingbars, markings, plumage, brushy, oak, warbler, green
FLO	Pink Primrose	wildflower, primrose, four-petaled, glabrous, bloom, buttercups, ranunculus, amapola
	Globe Thistle	sticky, weed, daisy, wooly, thistles, sharp toothed, wrinkled, florets
	Peruvian Lily	lobes, tuber, stripped, flecked, purple, streaked, curving petals, resupinate
	Tiger lily	ornamental, tiger, capsules, bulblets, lilium, tigrinum, pollinated, lily

alize three tokens per example and see that the model has abstracted various concepts in a summary token. For AWA, we see that the model focus on the horse, its background and identifying features like hoofed legs similar to I2DFormer. Similarly for Giraffe, the model looks at the giraffe in the image, the leaves in the background and the discriminative patterns on the fur. For CUB, the model focus on the bird and the patterns on the feather and the tail, and the environment the bird is found in. Finally, this interpretability remains consistent on FLO where the model focuses on the petals of King Protea, its large central part and the surrounding leaves. Similarly for Spear Thistile, the model focuses on the top flower and the various spiky structures on the body. It is important to note that this interpretability emerges without any paired image level supervision. Moreover, we see that the summary token encode the local information available in the document and provide similar cues to I2DFormer Naeem et al. (2022) at significantly reduced computational cost for attention.

5 Experiments on Large Scale Dataset

ImageNet is a challenging benchmark for zero-shot image classification methods as it tasks the model to train on 1000 classes and generalize to 500 additional classes. Due to the difficulty of the task, most works exclude this dataset and as a result the performance has saturated in recent years. We conduct large scale experiments on two different splits of ImageNet dataset. Both the splits consist of the same training classes from ImageNet1k. The first split called MP500 (Xian et al., 2018) consists of unseen classes from the most populated classes in ImageNet21k which are not part of ImageNet1k. The second split called Generic Object ZSL Dataset (GOZ) (Hascoet et al., 2019) takes a more structured approach to construct the evaluation set. Namely, they only contain test classes that are not direct neighbor of training classes in WordNet, have high quality word embeddings and have sufficient number of images. The GOZ split also contains a total of 500 unseen classes.

Collecting documents We use Wikipedia as the source of documents for ImageNet similar to baseline works Bujwid and Sullivan (2021); Kil and Chao (2021). However, since ImageNet consists of a large number of classes, some of which are very similar, we can not directly rely on the results of Wikipedia's python API as it can map multiple classes to the same document as also reported by Bujwid and Sullivan (2021). Moreover, some classes in ImageNet have vague names. For example, the class consisting of OLED Monitors with synset id “n03854506" has the english name OLED in WordNet; however, the wikipedia article for OLED describes the technical details of OLED technology and not the displays. Matching articles manually for a large number of



(a) Visualizing Word to Image attention in I2DFormer [14].

(b) Visualizing Summary token to Image attention in I2DFormer+.

Fig. 2 Visualizing Image Attention we see that our model I2DFormer has learned to localize words in the image without any paired patch-word supervision. This learned attention differentiates the two similar birds in the second row by identifying and localizing tiger stripes and gray-green as discriminative properties. Similar abilities emerge in I2DFormer+ where the model learns this interpretability against the summary tokens. These summary tokens have encoded the local information available in the document into a fixed set of tokens

classes is not scalable. We construct an automated pipeline to match a class in ImageNet with the most suitable document on Wikipedia.

Given a class name in ImageNet, we query Wikipedia and store all documents corresponding to the search result after performing section filtering as proposed by Bujwid and Sullivan (2021). We utilize a pretrained CLIP model (Radford et al., 2021) to do unsupervised matching between a class and the document that best describes it. Since CLIP is trained for short text sequences, its text encoder only supports a maximum sentence length of 77 tokens. For a given document, we split it into its sentences. If a sentence is longer than 77 tokens, we further split it into chunks smaller than the maximum input tokens. We get the text embedding for each of these and mean over them to get the embedding of the document. We sample 100 images from the class and compute their visual embedding using the visual embedder of the pretrained CLIP model. We compute a dot product between each image and the document embeddings to measure their compatibility with the images of the class. We average over the number of images and match the class with the document that results in the highest compatibility across the 100 images. Moreover, we also ensure that if a document is matched with a class, it is not used for matching with a subsequent class.

Table 8 Comparing our I2DFormer+ with unsupervised semantic embedding methods using the same image feature and method (our I2D Global module)

Semantic Embedding	Source	Zero-Shot Learning		Generalized Zero-Shot Learning					
		MP500	GOZ	MP500			GOZ		
		T1	T1	u	s	H	u	s	H
`GloVe` (Pennington et al., 2014)	`CLSN`	16.9	9.1	15.3	40.7	22.2	8.7	37.4	14.1
`GloVe` (Pennington et al., 2014)	`DOC`	19.4	11.6	17.3	45.1	25.0	10.6	42.4	17.0
`LongFormer` (Beltagy et al., 2020)	`DOC`	18.8	11.4	16.2	43.5	23.6	10.5	38.5	16.5
`MPNet` (Song et al., 2020)	`DOC`	19.7	10.3	17.1	40.2	24.0	9.5	37.9	15.2
`TF-IDF` (Salton and Buckley, 1988)	`DOC`	19.0	10.3	16.3	37.3	22.6	9.3	36.0	14.8
I2DFormer(Ours) (Naeem et al., 2022)	`IMG + DOC`	23.2	15.5	19.5	43.7	26.7	13.4	41.5	20.3
I2DFormer+ (Ours)	`IMG + DOC`	**24.5**	**17.6**	**20.9**	**46.0**	**28.7**	**15.5**	**45.2**	**23.1**

In ZSL, we report top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**). We consider semantic embeddings that are either directly extracted (with a pretrained language model) or learned from different sources including classnames (CLSN), document (DOC), and a combination of image and document (IMG+DOC). Our models I2DFormer+ and I2DFormer significantly improves on the baselines to set a new SOTA for unsupervised class embeddings

Bold refers to the best performance result as normal in Computer Vision community

We repeat this process for all classes in our ImageNet splits to collect their documents. We have attached the collected documents with the source files of the manuscript.

Computing the Loss over sampled negatives We compute L_{CLS}, the global alignment loss and L_{Local}, the local alignment loss over the set of seen classes $\mathcal{Y}^s$ available while training on small scale datasets. However, this becomes computationally expensive on Large Scale datasets as the number of classes increases. We address this by proposing a sampled negatives based loss rather than computing the loss over all the classes in $\mathcal{Y}^s$. Given a batch of randomly sampled training examples from label set $\mathcal{Y}^s$, we define $\mathcal{Y}^b$ as the set of labels that are represented in the batch. We additionally sample negatives not present in the batch to define $\mathcal{Y}^n$. For each training batch, these negatives are randomly sampled and the loss is computed against the label set $\mathcal{Y}^b \cup \mathcal{Y}^n$ for both L_{CLS} and L_{Local}.

Training Details We train I2DFormer+ and I2DFormer with a compute budget of a single A100 GPU similar to our small scale experiments. We use a batch size of 16 for training and sample 200 additional negatives for batch-wise loss computation. The number of summary tokens are fixed to 128. I2DFormer Naeem et al. (2022) is only able to be trained with 20 negatives due to needing costly local attention over the full document. The model is trained with Adam Optimizer with a learning rate of $1e^{-3}$ and takes 7 days to converge to the reported numbers. L_{CLS} and L_{Local} relative weights chosen by ablation.

5.1 Comparison with SOTA Unsupervised Semantic Embeddings on ImageNet

In this section, we compare with existing unsupervised semantic embeddings where they are obtained without using human supervision using the same ZSL method (our I2D global module). We report the results in Table 8.

Results From Table 8, we observe that our method I2DFormer+ consistently outperforms all semantic embedding methods in both ZSL and GZSL. We see similar conclusion to the small scale setting where replacing the `GloVe` (Classname) with `GloVe` (Document) results in a large boost in performance. This further validates our hypothesis that documents serve as better auxiliary information compared to class names. I2DFormer+ significantly outperforms `GloVe` (Document), the initialization embedding of I2DFormer+. I2DFormer+ achieves 24.5% ZSL accuracy on MP500 vs 19.4 of `GloVe` and 17.6% ZSL accuracy on GOZ compared to 11.6% of `GloVe`. We see similar consistent improvements in the GZSL setting where I2DFormer+ shows impressive gains in the HM. This further validates our hypothesis that a learnable document embedding will outperform a frozen embedding model like GloVe. We see similar results in other pretrained language semantic embeddings `Longformer`, `MPNet` and `TF-IDF` (Beltagy et al., 2020; Song et al., 2020; Salton and Buckley, 1988). These language only models are not trained on any visual data and hence they are less likely to capture the most visually relevant features described in text. Our model however benefits from both our learnable text transformer as well as our cross model attention block I2DAttention which learns to extract the most visually relevant information. Finally, comparing I2DFormer+ to I2DFormer Naeem et al. (2022), we see that the I2DFormer+

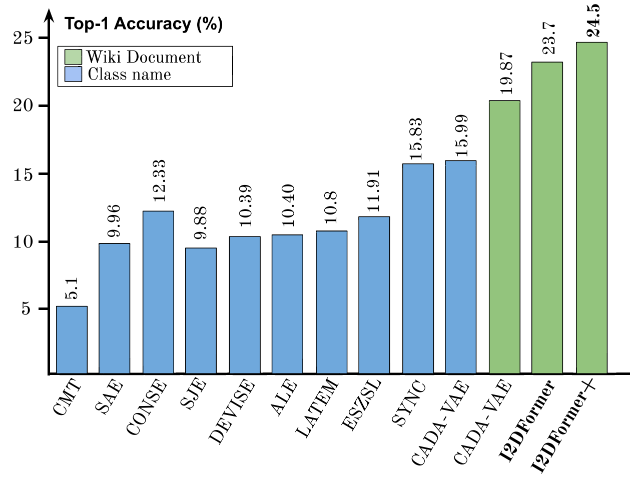


Fig. 3 Comparing I2DFormer with baseline ZSL methods on ImageNet Mp500 split, we observe that our model consistently outperforms all baselines to set a new state-of-the-art in the challenging large scale zero-shot learning setting

achieves better performance across both the splits of ImageNet. We attribute this improvement to DSTransformer's ability to reduce noise in the document into a fixed set of summary tokens and the computational cost improvements of I2DFormer+. This allows I2DFormer+ to process more negatives for each training sample compared to I2DFormer at the same computational budget.

5.2 Comparison with SOTA Models on MP500

We compare I2DFormer+ and I2DFormer Naeem et al. (2022) with ZSL results reported by Bujwid and Sullivan (2021) on the MP500 splits.

Results From Fig. 3, we observe that I2DFormer+ consistently outperforms all baseline methods to set a new state-of-the-art on ImageNet scale zero-shot image classification. `I2DEmb+` achieve an impressive zero-shot accuracy of 24.5 % compared to the previous best reported result of 19.87% of CADA-VAE with wikipedia article (Bujwid and Sullivan, 2021). We attribute this improvement to the ability of our model to directly learn a class embedding from the document text with global and fine-grained alignment with the two modules of our model. Baseline models are limited them to the information available in the pretrained embedding and can not further extract fine-grained knowledge. We make the same observation as in the small scale experiments where I2DFormer+ outperforms the generative baseline. Finally as we compare I2DFormer+ with I2DFormer Naeem et al. (2022), we observe that our architecture improvements translate to improvement in performance. I2DFormer+ achieves a ZSL accuracy of 24.5% compared to 23.7% of I2DFormer.

Table 9 Comparing I2DFormer+ with baseline ZSL models on challenging GOZ split of ImageNet, we observe a significant improvement in the GZSL setting

Model	Aux Info	ZSL (%)	GZSL (%)
CONSE (Norouzi et al., 2014)	Classname	10.65	0.12
DEVISE (Frome et al., 2013)	Classname	11.15	7.87
ESZSL (Romera-Paredes and Torr, 2015)	Classname	13.54	4.59
GCN-6 (Wang et al., 2018)	Classname	9.58	4.81
GCN-2 (Kampffmeyer et al., 2019)	Classname	14.09	4.96
ADGPM (Kampffmeyer et al., 2019)	Classname	14.10	4.9
I2DFormer+ (Ours)	Wiki	**17.6**	**23.1**

I2DFormer+ achieves a 3× improvement over the closest GCN based baseline ADGPM

Bold refers to the best performance result as normal in Computer Vision community

5.3 Comparison with SOTA Models on GOZ

Most works in Large Scale zero-shot learning have focused on the older splits of ImageNet proposed by Xian et al. (2018). While these splits have driven progress in the field, they come with some major limitations. Splits such as MP500 do not take parent and child relations in WordNet tree. As a result, several zero-shot classes are direct children or parent of training classes. In zero-shot prediction, these classes are picked as nearest neighbors to seen classes and gives a false sense of improvement. Moreover, ImageNet21k consists of several classes that have low quality images or are not well represented in Wikipedia which impacts the quality of their pretrained word embedding. As a result, these classes are bound to have low accuracy due to data or embedding quality. These issues are analysed by Hascoet et al. (2019) to propose a new zero-shot split of ImageNet that consists of unseen classes that are sufficiently different from training classes, have good quality of image data and have good quality word embeddings. We compare I2DFormer+ with the results reported by author in Hascoet et al. (2019) in Table 9. We report the zero-shot accuracy and the harmonic mean in generalized zero-shot setting.

Results We observe from Table 9 that I2DFormer+ outperforms all baselines to set a new state-of-the-art on this more challenging split too. I2DFormer+ significantly outperforms the previous DEVISE (Frome et al., 2013) with a 3× improvement to achieve a HM of 23.1% compared to 7.87%

of the baseline. Similarly on ZSL accuracy, I2DFormer+ achieves 17.6% compared to 14.10% of ADGPM.

6 Conclusion

We propose I2DFormer+, a fully Transformer based framework for learning semantic embeddings from noisy documents. Our I2D Global module learns a shared embedding space between an image and document embeddings. This is assisted by our I2D Attention module learns local features about the class defined in the document without any paired image-level captions. Our DSTransformer summarizes the most discriminative global and local information available in a document into a fixed set of learnable tokens. This leads to performance improvement while reducing the computational complexity of our local attention. As a result, our full model I2DFormer+ achieves SOTA performance on both ZSL and GZSL with respect to baseline semantic embedding baselines and zero-shot models. In addition, our model develops an impressive ability to identify and localize discriminative properties of a class in the image. Finally, we show that the learned embeddings from our model can further improve all zero-shot methods.

References

Akata, Z., Reed, S., Walter, D., Lee, H. & Schiele, B. (2015). Evaluation of output embeddings for fine-grained image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936.

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2015). Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence, 38*, 1425–1438.

Al-Halah, Z., & Stiefelhagen, R. (2017). Automatic discovery, association estimation and learning of semantic attributes for a thousand categories. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 614–623.

Beltagy, I., Peters, M.E., & Cohan, A. (2020). Longformer: The long-document transformer. In: arXiv:2004.05150

Bucher, M., Herbin, S., & Jurie, F. (2017). Generating visual representations for zero-shot classification. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2666–2673.

Bujwid, S., & Sullivan, J. (2021). Large-scale zero-shot image classification from rich and diverse textual descriptions. In: *LANTERN*.

Cacheux, Y.L., Borgne, H.L., & Crucianu, M. (2019). Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10333–10342.

Changpinyo, S., Chao, W.-L., Gong, B., & Sha, F. (2016). Synthesized classifiers for zero-shot learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336.

Chao, W.-L., Changpinyo, S., Gong, B., & Sha, F. (2016). An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pp. 52–68. Springer.

Chen, S., Wang, W., Xia, B., Peng, Q., You, X., Zheng, F., & Shao, L. (2021). Free: Feature refinement for generalized zero-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 122–131.

Cui, Y., Zhao, L., Liang, F., Li, Y., & Shao, J. (2022). Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. arXiv preprint arXiv:2203.05796

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., & Batra, D. (2017). Visual dialog. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335.

De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., & Courville, A. C. (2017). Modulating early visual processing by language. *Advances in Neural Information Processing Systems, 30*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR*.

Elhoseiny, M., Saleh, B., & Elgammal, A. (2013). Write a classifier: Zero-shot learning using purely textual descriptions. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2584–2591.

Elhoseiny, M., Zhu, Y., Zhang, H., & Elgammal, A. (2017). Link the head to the" beak": Zero shot learning from noisy text description at part precision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5640–5649.

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785. IEEE.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems, 26*.

Ghiasi, G., Gu, X., Cui, Y., & Lin, T.-Y. (2022). Scaling open-vocabulary image segmentation with image-level labels. In: *Computer Vision–ECCV 2022: 17th European Conference*, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, pp. 540–557. Springer.

Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921

Hascoet, T., Ariki, Y., & Takiguchi, T. (2019). On zero-shot recognition of generic objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9553–9561.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., & Shelhamer, E. (2021). Perceiver io: A general architecture for structured inputs & outputs. In: *ICLR*.

Ji, Z., Fu, Y., Guo, J., Pang, Y., & Zhang, Z. M. (2018). Stacked semantics-guided attention model for fine-grained zero-shot learning. *Advances in Neural Information Processing Systems, 31*.

Jiang, H., Wang, R., Shan, S., & Chen, X. (2019). Transferable contrastive network for generalized zero-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9765–9774.

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., & Xing, E.P. (2019). Rethinking knowledge graph propagation for zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11487–11496.

Kil, J., & Chao, W.-L. (2021). Revisiting document representations for large-scale zero-shot learning. In: *NAACL*.

Lei Ba, J., Swersky, K., & Fidler, S. (2015). Predicting deep zero-shot convolutional neural networks using textual descriptions. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4247–4255.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., & Yan, J. (2021). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208

Liu, S., Long, M., Wang, J., & Jordan, M. I. (2018). Generalized zero-shot learning with deep calibration network. *Advances in Neural Information Processing Systems, 31*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. 32

Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096.

Mancini, M., Naeem, M.F., Xian, Y., & Akata, Z. (2021). Open world compositional zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5222–5230.

Mancini, M., Naeem, M. F., Xian, Y., & Akata, Z. (2022). Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*.

Naeem, M.F., Örnek, E.P., Xian, Y., Van Gool, L., & Tombari, F. (2022). 3d compositional zero-shot learning with decompositional consensus. In: *European Conference on Computer Vision*, pp. 713–730. Springer.

Naeem, M.F., Xian, Y., Tombari, F., & Akata, Z. (2021). Learning graph embeddings for compositional zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 953–962.

Naeem, M. F., Xian, Y., Gool, L. V., & Tombari, F. (2022). I2dformer: Learning image to document attention for zero-shot image classification. *Advances in Neural Information Processing Systems, 35*, 12283–12294.

Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., & Shao, L. (2020). Latent embedding feedback and discriminative features for zero-shot classification. In: *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pp. 479–495. Springer.

Nilsback, M.-E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G., & Dean, J. (2014). Zero-shot learning by convex combination of semantic embeddings. In: *ICLR*.

Patterson, G., Xu, C., Su, H., & Hays, J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision, 108*, 59–81.

Pennington, J., Socher, R. & Manning, C.D. (2014). Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., & Wu, Y. (2023). Combined scaling for zero-shot transfer learning. *Neurocomputing, 555*, 126658.

Qiao, R., Liu, L., Shen, C., & Van Den Hengel, A. (2016). Less is more: zero-shot learning from online textual documents with noise suppression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2249–2257.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*, pp. 8748–8763. PMLR.

Reimers, N., Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In: *EMNLP*.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016). Grounding of textual phrases in images by reconstruction. In: *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 817–834. Springer.

Rohrbach, A., Rohrbach, M., Tang, S., Joon Oh, S., & Schiele, B. (2017). Generating descriptions with grounded and co-referenced people. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4979–4989.

Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In: *International Conference on Machine Learning*, pp. 2152–2161. PMLR.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In: *Information Processing & Management*.

Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., & Akata, Z. (2019). Generalized zero-and few-shot learning via aligned variational autoencoders. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8247–8255.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems, 26*.

Song, J., Shen, C., Lei, J., Zeng, A.-X., Ou, K., Tao, D., & Song, M. (2018). Selective zero-shot classification with augmented attributes. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 468–483.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems, 33*, 16857–16867.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł.,& Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Verma, V.K., Arora, G., Mishra, A., & Rai, P. (2018). Generalized zero-shot learning via synthesized examples. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4281–4289.

Vyas, M.R., Venkateswara, H., & Panchanathan, S. (2020). Leveraging seen and unseen semantic relationships for generative zero-shot learning. In: *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, pp. 70–86. Springer.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, X., Ye, Y., & Gupta, A. (2018). Zero-shot recognition via semantic embeddings and knowledge graphs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6857–6866.

Website: A-Z Animals. https://a-z-animals.com/

Website: Wikipedia. https://en.wikipedia.org/

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., & Schiele, B. (2016). Latent embeddings for zero-shot classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 69–77.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10275–10284.

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, 41*, 2251–2265.

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., & Wang, X. (2022). Groupvit: Semantic segmentation emerges from text supervision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144.

Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2022). Vgse: Visually-grounded semantic embeddings for zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9316–9325.

Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2020). Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems, 33*, 21969–21980.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In: *ACL*.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., & Xu, C. (2022). FILIP: Fine-grained interactive language-image pre-training. In: *ICLR*.

Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., & Chang, S.-F. (2013). Designing category-level attributes for discriminative visual recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 771–778.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., & Beyer, L. (2022). Lit: Zero-shot transfer with locked-image text tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133.

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.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., & Elgammal, A. (2018). A generative adversarial approach for zero-shot learning from noisy texts. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1004–1013.

Zhu, Y., Xie, J., Liu, B., & Elgammal, A. (2019). Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9844–9854.

Zhu, Y., Xie, J., Tang, Z., Peng, X., & Elgammal, A. (2019). Semanticguided multi-attention localization for zero-shot learning. *Advances in Neural Information Processing Systems, 32*.

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