I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Zero-Shot Image Classification

Muhammad Ferjad Naeem[*1], Muhammad Gul Zain Ali Khan[*2,3],   Yongqin Xian[5],   Muhammad Zeshan Afzal[2,3],
Didier Stricker[2,3],   Luc Van Gool[1],   Federico Tombari[4,5]
[1]ETH Zürich,   [2]TUKL   [3]DFKI   [4]TUM   [5]Google

Abstract

Recent works have shown that unstructured text (documents) from online sources can serve as useful auxiliary information for zero-shot image classification. However, these methods require access to a high-quality source like Wikipedia and are limited to a single source of information. Large Language Models (LLM) trained on web-scale text show impressive abilities to repurpose their learned knowledge for a multitude of tasks. In this work, we provide a novel perspective on using an LLM to provide text supervision for a zero-shot image classification model. The LLM is provided with a few text descriptions from different annotators as examples. The LLM is conditioned on these examples to generate multiple text descriptions for each class (referred to as views). Our proposed model, I2MVFormer, learns multi-view semantic embeddings for zero-shot image classification with these class views. We show that each text view of a class provides complementary information allowing a model to learn a highly discriminative class embedding. Moreover, we show that I2MVFormer is better at consuming the multi-view text supervision from LLM compared to baseline models. I2MVFormer establishes a new state-of-the-art on three public benchmark datasets for zero-shot image classification with unsupervised semantic embeddings. Code available at https://github.com/ferjad/I2DFormer

1. Introduction

In Zero-Shot Learning (ZSL), we task an image classification model trained on a set of seen classes to generalize to a disjoint set of unseen classes using shared auxiliary information. While there has been great progress made in the field, most works treat the auxiliary information to be fixed to a set of human-labeled attributes [16, 37, 48, 53]. While powerful, these attributes are hard to annotate and expensive to scale [46,58]. Unsupervised alternatives to attributes rely on pretrained word embeddings which provide limited
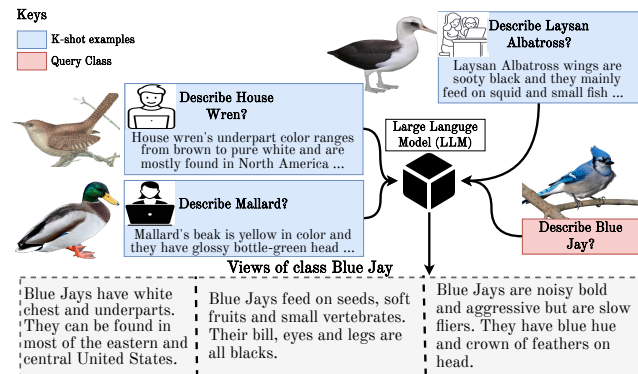

Figure 1. Different annotators focus on different attributes when describing a class. Large Language Models prompted with each of these annotations as k-shot examples can reveal complementary information about a class for zero-shot image classification. We refer to multiple LLM-generated descriptions as *views* of a class.

information about a class. Recent works [7,20,35,39] show that text documents from internet sources like Wikipedia can provide great auxiliary information for ZSL. Since these web documents describe a queried class in detail, they provide more information for the ZSL model compared to word embeddings. However, these methods only rely on a single source of text documents like Wikipedia, which might not sufficiently represent all classes a model is faced with. Multiple sources of text documents of a class can provide complementary information for the ZSL model. For example, in the case of birds, one source might focus more on the patterns of the feather, while another source might better describe the belly and the face of the bird. However, finding multiple good sources of text documents for each class requires additional annotation effort.

Large Language Models (LLM) [5, 12, 63] trained on web-scale text have shown impressive abilities of using their learned information to solve a multitude of tasks. These models can be conditioned with a k-shot prompt to generalize to a wide set of applications [5, 31, 60] using knowledge from multiple sources they were trained on. In

---

[1]First and second author contributed equally.

this work, we aim to generate multiple text descriptions of a class, that we recall as "views" hereinafter, with an LLM using a k-shot prompting strategy. We show that the LLM can act as a mixture of annotators conditioned on different annotation styles to generate complementary information about a class. Moreover, we propose a novel model, I2MVFormer, which utilizes our memory-efficient summary modules to extract discriminative information from each view of a class with the aim of learning a multi-view class embedding.

Our contributions in this work are as follows. 1) We provide the first study into using an LLM to generate auxiliary information for zero-shot image classification. Moreover, we propose a prompting strategy to extract multiple descriptions from an LLM that reveal complementary information about a class. 2) We propose I2MVFormer, a novel transformer-based model for zero-shot image classification which exploits multiple complementary sources of text supervision to learn a class embedding. I2MVFormer utilizes our Single-View Summary (SVSummary) module to extract rich discriminative information from each class view. This information is utilized by our Multi-View Summary (MVSummary) module to represent a class-level set of tokens from multiple views. The multi-view tokens are aligned with the image to maximize global and local compatibility between the images and the multiple views. 3) Our I2MVFormer achieves significant performance gains to establish a new state-of-the-art (SOTA) in unsupervised class embeddings in ZSL on three public benchmarks AWA2 [22], CUB [48] and FLO [36].

2. Related Work

Zero-shot learning aims to learn a model that can generalize beyond the seen classes it was trained on. This is accomplished by using side information that is shared with a set of disjoint unseen classes. Towards this, several methods learn a compatibility function between the image feature and a class embedding representing the auxiliary information [9, 19, 28, 41, 51, 62]. These methods often suffer from bias against unseen classes. To address this, another family of methods additionally learns the distribution of the features of images using a generative model [4, 21, 43, 53, 67, 68]. Approaches in this category focus on learning a class conditional generator that can generate features of unseen classes [4, 21] or exploit semantic information about a class to generate features of unseen classes directly [44, 67]. Once learned on seen classes, these models generate features of unseen classes using its auxiliary information to tackle the bias issue. Other works focus on learning improved visual-semantic embeddings [8, 17, 26, 62] and training better image feature extractors [54, 59, 69]. However, all these methods assume that the set of auxiliary information is fixed to human-labeled attributes [16, 33, 37, 48, 53]. However, labeling attributes is expensive and hard to scale on large datasets as it requires expert annotators [46, 48, 58].

Unsupervised semantic embedding aims to learn the semantic embedding of seen and unseen classes using side information that does not require human intervention. The most influential works in this direction use word embeddings from a pretrained model to encode semantic similarities [30, 38, 45, 56] and refine them using knowledge graphs [6, 18, 29, 34, 49]. VGSE [55] learns a class embedding by using image patches and class embedding vectors. Several works have explored leveraging text documents from sources like Wikipedia to learn class embeddings since they contain rich information about a class. The literature in this direction exploits pretrained language models [1, 7, 14, 15, 20, 39, 67] in addition to noise minimization through a predefined vocabulary [39] or part detection network [15, 67]. However, these works treat the embedding of the document with a pretrained model as fixed. Recently I2DFormer [35] propose a transformer-based model that learns a class embedding from raw text. Unlike zero-shot transfer models like CLIP [40], which only maximizes the global compatibility between an image and text embedding, I2DFormer maximizes both the global and the local compatibility of the text features against the image. However, I2DFormer relies on expensive local attention between each image patch and document token, which does not scale well to large text sources.

Large Language Models (LLM) like GPT-3 [5], OPT [63] and PaLM [12] are trained on very large web-scale datasets. Once trained, these models have impressive abilities towards zero-shot and few-shot inference on a multitude of tasks such as Open Question and Answering [66], generating code [11], text summarization [32], etc. These models rely on a k-shot prompt defining the problem they have to solve along with 0 or k examples. Once prompted for the target class, they use their stored knowledge from web-scale training to generate text for the target class. Recently some works have tried to pair these models with vision models using generated text [60] or adding vision as a modality [2]. Other works explore prompting vision-language models [25, 31, 50] for continual learning, image caption generation, and action understanding. However, no work has yet explored leveraging an LLM to generate auxiliary information for zero-shot image classification.

3. Generating multiple text views with an LLM

We define a text description of a class as its "view". This is analogous to a document used in previous work [35]. A single view from a source like Wikipedia as used in previous works [1, 7, 35, 39], while powerful, can present knowledge gaps for less frequent classes. These classes might be described better in other, more specialized databases on
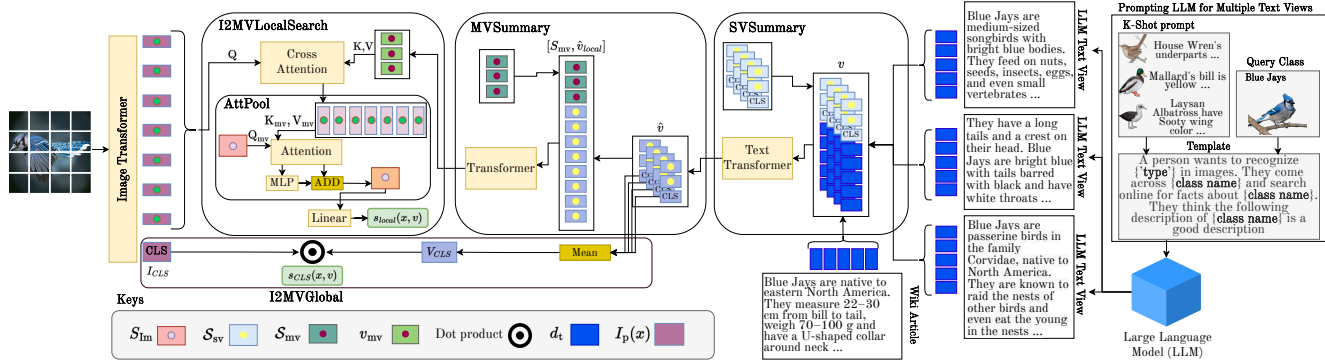
Figure 2. **I2MVFormer**, our novel Transformer based model uses LLM-generated document supervision conditioned on multiple annotators and noisy Wiki documents as multiple views of a class to learn a zero-shot model. Our SVSummary module generates a single view level `CLS` representation and local summary. The CLS summaries over multiple views are used to align the global features against the image. Our MVSummary module uses the summary tokens of each view to generate a multi-view summary of the class. This multi-view summary is aligned with the per-patch feature of the image in our I2MVLocalSearch module. Together the two modules learn a highly discriminative multi-view class embedding.

the internet. We propose to use LLMs to generate multiple views of a class, exploiting the property of these models of storing the knowledge from multiple internet sources. In this section, we describe our novel prompting method to automatically generate multi-view text descriptions of object classes using a pretrained LLM.

While LLMs have impressive zero-shot abilities, practically, they require significant prompt engineering to get good zero-shot output, as noted in several works [23–25, 27, 50, 64, 65]. Since LLMs are few-shot learners [5], given k-shot examples, they can generate output for any class they are prompted for. We exploit this few-shot ability of LLMs to generate multiple text descriptions of each class representing multiple views. Given a set of annotated examples of how to describe a few classes of a dataset in natural language, the LLM can be prompted to mimic the labeling style to generate text supervision for all classes in a dataset. We require $f + 1$ annotated examples to generate f views from multiple inferences of the LLM. The extra example is reserved for replacement whenever the query class is present in the k-shot examples. We observed that without this, the LLM is susceptible to producing the exact class description as the one in the examples.

To curate $f + 1$ examples without requiring significant annotation cost, we rely on the Wiki articles released by [35]. These articles are filtered only to contain sections of Wikipedia labeled to contain visual information. This is done as sources like Wikipedia contain a lot of noise in the form of non-visual information. We notice that different wiki articles denote different annotation styles as the content is sourced from a pool of volunteers. In our work, we use $f = 3$ as the number of views generated by LLM for each class. We randomly select four classes from each dataset and curate their wiki articles from [35] to only focus

on the visually relevant details. This allows us to obtain a relatively rich source of auxiliary information representing multiple sources without drastically increasing annotation effort. We now want the LLM to use these as examples of class descriptions to generate similar examples over all classes in the dataset. We append each labeled example with the following prompt.

*"A person wants to recognize {**'type'**} in images. They come across {**class name**} and search online for facts about {**class name**}. They think the following description of {**class name**} is a good description."*

In the given template, we use type as "animals", "birds" and "flowers" for AWA2 [52], CUB [48] and FLO [36], respectively, moreover {**class name**} defines the name of the labelled class. The target class is then entered into the template and fed as the input to the LLM with the k-shot examples. The LLM generates a description for this target class conditioned on the labeling style of our k-shot examples. We use $k = 2$ for most of our experiments, i.e., each view is generated conditioned on 2 labeled examples. These 2 shots are a combination of the $3 + 1$ labeled examples for the 3 views of each class. We later show in supplementary that the LLM is fairly robust to the choice of prompt given the same k-shot example.

4. I2MVFormer

Most methods in zero-shot literature either rely on human-labeled attributes [16, 33, 37, 48, 53] or pretrained word embeddings [6, 18, 29, 34, 45, 49, 56] as auxiliary information. Recently I2DFormer [35] has shown that text can be a powerful substitute without requiring significant labeling effort. Multiple descriptions, referred to as views in this work, of a class can provide complementary infor-

mation of each class to the model and potentially lead to better performance. Our proposed model **Image to Multi-View Transformer (I2MVFormer)** is designed to exploit the knowledge in multiple views of each class to learn a zero-shot model. I2MVFormer consists of two streams of transformers to process the visual and the textual data as shown in Figure 2. I2MVFormer aligns the global information available in each view with the image along with the fine-grained information. Moreover, I2MVFormer reduces the memory footprint of I2DFormer while significantly outperforming it on three public benchmarks.

Notations. We represent the classes in the training set as $\mathcal{Y}^s$ and the unseen classes only available at test time as $\mathcal{Y}^u$. Let $\mathcal{T} = \{(\mathbf{x}, \mathbf{y}, \mathbf{v}) | \mathbf{x} \in \mathcal{X}^s, \mathbf{y} \in \mathcal{Y}^s, \mathbf{v} \in \mathcal{V}^s\}$ 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 $\mathcal{Y}^s$, $\mathbf{v} = \{d_1, d_2....d_q\}$ is the set of documents representing q views of the class and $\mathcal{V}_s$ the set of views of the seen classes. At test time an additional set $\mathcal{V}_u$ is made available as the views of novel classes where $\mathcal{V} = \mathcal{V}_s + \mathcal{V}_u$ and $\mathcal{V}_s \cap \mathcal{V}_u = \emptyset$. The ZSL task requires the model to predict a class from the set of unseen classes $\mathcal{Y}_u$, and the Generalized ZSL (GZSL) requires the model to predict a class over both seen and unseen classes $\mathcal{Y} = \mathcal{Y}^s + \mathcal{Y}^u$.

4.1. Image Transformer

On the image side, our model learns $\mathcal{F}$, an image transformer as an embedding function. Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we reshape it into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 C)}$, 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 recent advances in multimodal learning [61], we use a pretrained frozen image transformer [13] followed by a learnable MLP layer which maps the features to a joint image-text embedding space with dimentionality r. $\mathcal{F}$ outputs $I_{CLS}(\mathbf{x}) \in \mathbb{R}^r$ as the global image feature and $I_p(\mathbf{x}) \in \mathbb{R}^{N \times r}$ as the patch-wise image embedding for the input image.

4.2. SVSummary: Extracting class level summary from each view.

In our multi-view setup, each training class $\mathbf{y}$ is associated with $\mathbf{v}$, the set of documents representing multiple views of the class. The model is now faced with an increased amount of text compared to existing works that rely on a single view. This makes existing solutions [35] that rely on exhaustive attention on all document tokens against the image computationally very expensive. Our Single-ViewSummary (SVSummary) module is a text transformer that aims to represent the most discriminative information

available in each text description (view) of a class into a fixed set of Summary tokens. It is important to note that we define a summary as the most important features described in the text rather than its classical definition of a human consumable shortened description of the text.

Given $\mathbf{v} = \{d_1, d_2, ..., d_n\}$, the set of documents representing views of class $\mathbf{y}$, we pre-tokenize each document and represent it by a pretrained word embedding model similar to [35]. We learn a shallow MLP on top to improve upon the pretrained word embedding model. The output of this yields $d_t \in \mathbb{R}^{M \times r}$ for each document d_t in the view set where M is the length of the document and r is the feature dimension. We introduce Single-View Summary tokens $\mathcal{S}_{sv} \in \mathbb{R}^{T \times r}$ as a set of T learnable tokens. These tokens are introduced to specialize them for the downstream image compatibility task instead of using all M tokens of a view. Given $T < M$, this results in a reduced constant memory cost of the later cross-modal alignment independent of the length of each view. $\mathcal{S}_{sv}$ is appended to each view and passed through a learnable text transformer with several Transformer encoding blocks. On the output side, we take the token representations corresponding to $\mathcal{S}_{sv}$ for every single view d_t to get its summary $S_{sv}(d_t)$. For $\mathbf{v}$, the set of views of a class, this step yields $\hat{\mathbf{v}} = \{\mathcal{S}_{sv}(d_1), \mathcal{S}_{sv}(d_2), ..., \mathcal{S}_{sv}(d_q)\} \in \mathbb{R}^{q \times T \times r}$ as the learnable summary for each view of a class. The first summary token of each view is specialized as the CLS token, which contains the global feature of this view for global alignment in the next step.

4.3. I2MVGlobal: Learning global compatibility between image and multiple views

I2MVGlobal aims to align the global feature of an image $\mathbf{x}$ with the ground truth class $\mathbf{y}$ using the view set $\mathbf{v}$. To accomplish this, we use the CLS token of each view in Figure 2 and take a mean over the view dimension q to get a global feature across all views defined as $V_{CLS}(\mathbf{v}) \in \mathbb{R}^r$. We align this feature with the global image-level feature I_{CLS} from the image encoding function $\mathcal{F}$. To accomplish this, we define a scoring function s_{CLS} that measures the compatibility between I_{CLS} and V_{CLS} as a dot product:

$$s_{CLS}(\mathbf{x}, \mathbf{v}) = I_{CLS}(\mathbf{x}) \cdot V_{CLS}(\mathbf{v}). \tag{1}$$

The learning objective aims to assign high values to correct pairs and low scores otherwise. For a particular training instance $(\mathbf{x}, \mathbf{y}, \mathbf{v})$, we minimize the following cross-entropy loss over $\mathcal{V}_s$ the set of views for the seen classes.

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

4.4. MVSummary: Generating multi-view summary from all views

Previous work in ZSL [35] has shown that only aligning the global feature of an image to a text description can lead to over-fitting to seen classes. The existing solution to mitigate this relies on expensive cross-attention between all image and document tokens which becomes increasingly expensive as increased text is available for each class. Our SVSummary module allows this at reduced memory cost as it generates a fixed number of summary tokens irrespective of the length of a view.

We concatenate the summary tokens of each view (except the CLS token, which is used in I2MVGlobal) in the view dimension q to define $\hat{\mathbf{v}}_{local} \in \mathbb{R}^{q(T-1)\times r}$, the summary of all views. This set will grow linearly, given increased views of each class, and can become expensive to align against the local features of an image. We mitigate this through our Multi-View Summary Transformer (MVSummary) module, which introduces a set of multi-view summary tokens $\mathcal{S}_{mv} \in \mathbb{R}^{T\times r}$ with T learnable tokens. These tokens are aimed at summarizing the local summaries of all views into one class-level multi-view summary. We concatenate $\mathcal{S}_{mv}$ with $\hat{\mathbf{v}}_{local}$ and use it as input to a learnable Transformer encoder block. The output feature representation corresponding to tokens in $\mathcal{S}_{mv}$ represent the multi-view local summary $\mathbf{v}_{mv} \in \mathbb{R}^{T\times r}$ for a given class.

4.5. I2MVLocalSearch: Fine-grained alignment between image and summary

Our I2MVLocalSearch aims to align the patch-wise image features $I_p(\mathbf{x}) \in \mathbb{R}^{N\times r}$ with the multi-view summary tokens of each class $\mathbf{v}_{mv}$. The core idea is that each summary token specializes in an aspect of a class defined by the multiple views of the class. Aligning an image to these encourage the model to focus on local features that are important for fine-grained classification. We define this as a query search problem where $Q = I_p(\mathbf{x})W_q$ defines the visual query set, and $K = \mathbf{v}_{mv}(\mathbf{v})W_k$, $V = \mathbf{v}_{mv}(\mathbf{v})W_v$ define the keys to compare with and values to combine with. W_q, W_k and W_v are learnable linear mappings $\in \mathbb{R}^{r\times r}$. The first step involves computing an attention matrix between Q and K and using it to calculate a new multi-view representation of image patches $I_{mvpatch} = \text{softmax}(\frac{QK^T}{\sqrt{r}})V \in \mathbb{R}^{N\times r}$.

We want to align this representation with the ground truth class $\mathbf{y}$. We define a learnable attention pooling to get an image-level feature from these patch representations. We use a learnable Image summary token $\mathcal{S}_{Im} \in \mathbb{R}^{r}$ as the query Q_{mv}. The keys K_{mv} and values V_{mv} are defined by passing $I_{mvpatch}$ through a linear layer. We compute attention between the query and key and use these values to compute intermediate representation $I'_{mv} \in \mathbb{R}^{r}$. This is passed through a learnable two-layer MLP and a skip connection

to get the multi-view image feature $I_{mv} \in \mathbb{R}^{r}$ as:

$$I'\text{mv}(\mathbf{x}, \mathbf{v}) = \text{softmax}(\frac{Q_{mv}K_{mv}^T}{\sqrt{r}})V_{mv}$$
$$I_{mv}(\mathbf{x}, \mathbf{v}) = I'_{mv}(\mathbf{x}, \mathbf{v}) + \text{MLP}(I'_{mv}(\mathbf{x}, \mathbf{v})) \quad (3)$$

The multi-view image feature is used to predict a local alignment score s_{local} by a learnable linear layer $J \in \mathbb{R}^{1\times r}$. This score is optimized with a cross-entropy loss L_{local},

$$s_{local}(\mathbf{x}, \mathbf{v}) = J(I_{mv})$$
$$L_{Local} = -\log(\frac{\exp s_{local}(\mathbf{x}, \mathbf{v})}{\sum_{\mathbf{v}'\in\mathcal{V}^s} \exp s_{local}(\mathbf{x}, \mathbf{v}')}) \quad (4)$$

To summarize, SVSummary allows each document view a chance to describe the class. The MVSummary uses these to generate a multi-view class summary which is subsequently used by our I2MVLocalSearch. This multi-view summary is aligned with the patch-level image features to promote fine-grained feature learning.

4.6. Inference.

Given an input image $\mathbf{x}$, a prediction $\hat{\mathbf{y}}$ corresponds to the view set that yields the highest compatibility score among unseen classes for ZSL and among both seen and unseen classes for GZSL:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{v}'\in\mathcal{V}} s_{CLS}(\mathbf{x}, \mathbf{v}'). \quad (5)$$

5. Experiments

We conduct extensive experimentation on three popular ZSL datasets Animal with Attributes 2 (AWA2) [22], Caltech-UCSD Birds (CUB) [48] and Oxford Flowers (FLO) [36] using the evaluation protocol and data splits proposed by Xian et al. [53]. We do not use any human-labeled attributes similar to other works in unsupervised class embeddings. In the following, we discuss implementation details, detailed experiments and their conclusions.

Implementation Details. We use PaLM540B [12] as the Large Language Model (LLM) for our main experiments prompted with two shots per view at a temperature value of 0.9. We use 3 LLM generated views in addition to the wiki articles released by [35] for our main experiments. The SV-Summary is implemented as a two-block deep text transformer. The MVSummary transformer is implemented with a similar configuration to SVSummary. T the number of summary tokens is set as 64 for CUB and 128 for AWA and FLO. We use GloVe [38] as the initial token representation similar to [35]. We use the VIT/B16 checkpoint trained on ImageNet1K as the pretrained Image Transformer to be consistent with previous work. The patch projection and

Model	Auxiliary Information	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 [38]	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 [38]	Wiki	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 [3]	Wiki	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 [47]	Wiki	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 [42]	Wiki	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 [55]	IMG + CLSN	69.6	37.1	-	56.9	82.8	67.4	27.6	70.6	39.7	-	-	-
I2DFormer [35]	Wiki	76.4	45.4	40.0	66.8	76.8	71.5	35.3	57.6	43.8	35.8	91.9	51.5
	3-LLM (ours)	69.7	46.0	41.9	65.2	80.4	72.0	36.6	59.5	45.3	37.4	94.2	53.5
	3-LLM + Wiki (ours)	<u>77.3</u>	<u>47.0</u>	<u>43.0</u>	<u>68.6</u>	77.4	<u>72.7</u>	<u>38.5</u>	59.3	<u>46.7</u>	<u>40.4</u>	80.1	<u>53.8</u>
I2MVFormer (ours)	Wiki	73.6	42.1	41.3	66.6	<u>82.9</u>	73.8	32.4	<u>63.1</u>	42.8	34.9	<u>96.1</u>	51.2
	3-LLM (ours)	76.4	47.8	44.4	72.7	81.3	76.8	40.1	58.0	47.4	41.1	91.1	56.6
	3-LLM + Wiki (ours)	**<u>79.6</u>**	**<u>51.1</u>**	**<u>46.2</u>**	**<u>75.7</u>**	79.6	**<u>77.6</u>**	**<u>42.5</u>**	59.9	**<u>49.7</u>**	**<u>41.6</u>**	91.0	**<u>57.1</u>**

Table 1. **Comparing our I2MVFormer with baseline.** Our I2MVFormer significantly improves on the baselines to set a new SOTA for unsupervised class embeddings. We report top-1 accuracy (**T1**) on unseen classes for ZSL, and seen/unseen (**s/u**) classes and their harmonic mean (**H**) for GZSL. We see that the 3-LLM generated views provide complementary information to the wiki articles and significantly improve the performance. Moreover, we see that I2MVFormer is better at consuming multi-view knowledge compared I2DFormer. Finally, we see that I2MVFormer with LLM supervision alone can outperform I2DFormer with Wiki article indicating that the LLM alone can generate rich class descriptions. Best results within a method are <u>underlined</u>. Best results overall are in **bold**.

MLP in SVSummary are two layers with ReLU and LayerNorm. For GZSL, we apply calibrated stacking [10] to calibrate the activations of unseen classes on a held-out set. We use the Adam optimizer with a learning rate of $1e^{-3}$ and the model converges in $\approx$ 24 hours. L_{CLS} and L_{Local} are combined with weights ablated on the validation set. Detailed training details and examples of LLM-generated views are available in the supplementary. Our experimentation framework is implemented in PyTorch and the model can be trained on a single A100 40GB GPU. For the previous SOTA, I2DFormer, we concatenate the text in different views as it is designed for a single view (document) of a class. Performance of pretrained/ classical embedding baselines like GloVe [38], TF-IDF [42], VGSE [55] etc. are taken from [35] with their training setup. We report the top-1 per-class mean accuracy in ZSL. In GZSL, we report the top-1 per-class mean accuracy on seen (s) and unseen (u) classes separately along with their harmonic mean (H).

5.1. Comparing with State-of-the-Art.

We compare our results with state-of-the-art in unsupervised class embeddings in Table 1 and show that LLM-generated multiple views can significantly improve the performance in ZSL. Moreover, we show that I2MVFormer significantly outperforms all the previous methods to set a new state-of-the-art in unsupervised class embeddings on all three datasets. Our detailed observations are as follows.

LLM documents vs Wiki documents. We observe that Wiki articles from [35] and the 3-LLM generated views provide complementary information to consistently improve the ZSL and GZSL performance on all datasets across all metrics. This validates our hypothesis that ZSL models can benefit from multiple perspectives of a class and the LLM is able to generate them without significant annotation effort. Compared to previous SOTA results of I2DFormer with Wiki, I2MVFormer achieves an absolute improvement of 3.2% on AWA, 5.7% on CUB and 6.2% on FLO in ZSL. Similar improvements are seen in the GZSL setting, where we see consistent improvements. We observe that I2MVFormer is better at consuming multi-view knowledge compared to the previous SOTA I2DFormer validating our hypothesis that per-view processing of text allows for extracting richer information from each view. Since our model is specifically developed for multi-view documents, we see that it is on par with I2DFormer across a single view of the Wiki article but achieves significant improvement once multiple views are introduced. Finally, we see that I2MVFormer with only LLM generated views surpasses I2DFormer with Wiki documents on the three datasets indicating that LLM alone can generate highly discriminative class descriptions for zero-shot image classification. This confirms our hypothesis that LLM with targeted prompting can provide multiple highly discriminative views of a class.

Learning per-view summary vs text concatenation. I2MVFormer processes each document in the multiple views independently before generating class level V_{CLS}

| | Components | | | | AWA | CUB | FLO |
	L_{CLS}	L_{Local}	SVS	MVS	T1	T1	T1
a)	✓				73.6	45.6	38.9
b)	✓		✓		74.1	48.5	39.1
c)		✓	✓	✓	57.7	32.5	24.2
d)	✓	✓	✓		78.4	49.0	43.2
e)	✓	✓	✓	✓	**79.6**	**51.1**	**46.2**

Table 2. **Ablating over I2MVFormer**, we confirm the importance of each component of our model. We observe that L_{CLS} and L_{Local} are complementary to each other. Moreover, SVSummary (SVS) and MVSummary(MVS) reduce the complexity of cross-modal attention while improving performance.

and local multi-view summary tokens $\mathbf{v}_{mv}$. This is in contrast to I2DFormer, which concatenates all views into a single text sequence for global and local alignment. We see that the I2MVFormer strategy remains superior as the model can first extract highly discriminative facts from each view and later combine them together in the learned CLS and local summary tokens. When all views are concatenated, the large text sequence can contain repeated information. Moreover, learning local alignment on such a large sequence works less optimally than allowing the model to first extract a set of highly discriminative local summary tokens. We see that I2MVFormer consistently outperforms I2DFormer at LLM across 3 and LLM+Wiki supervision across 4 views. These improvements are in addition to the reduced memory cost as I2MVFormer requires half the GPU memory compared to I2DFormer across 4 views.

5.2. Ablation over I2MVFormer.

We ablate over the various components of our model in Table 2 using LLM+Wiki views. Rows a) and b) only optimize for the global feature between the image and text using L_{CLS}. Row a) optimizes for the V_{CLS} generated by the concatenation of all views, while Row b) introduces our SVSummary module. We see that learning a per-view summary while reducing the cost of attention in the text, also offers a performance improvement. Row c) only optimizes the fine-grained alignment between image patches and multi-view summary tokens using L_{Local}. We see that this alone performs worse than the global head as fine-grained alignment is a hard problem to optimize as also noted in previous works [35, 57]. Row d) and e) optimize for both the L_{CLS} and L_{Local}. We observe that the two losses are complementary and result in a significant improvement in performance as the model aligns the image and text modality with global as well as local features. Row d) uses a concatenation of per-view summary tokens while Row e) uses our MVSummary module to first learn a set of tokens representing a multi-view summary of the class for learning local alignment. We confirm that MVSummary improves the perfor-

| | | Zero-Shot Learning | | Generalized Zero-Shot Learning | | | | | | |
| | Shots | AWA2 | FLO | AWA2 | | | FLO | | | |
		T1	T1	u	s	H	u	s	H
a)	0 shot	73.0	40.7	66.6	79.1	72.3	38.0	85.7	52.7
b)	1 shot unique	74.2	42.1	68.8	**82.8**	75.1	39.8	89.9	55.2
c)	2 shots repeated	73.1	43.1	67.8	79.9	73.4	39.7	90.1	55.1
d)	2 shots unique	**76.4**	**44.4**	**72.7**	81.3	**76.8**	**41.1**	**91.1**	**56.6**

Table 3. **Ablating over different prompting strategies**, we observe that k-shot prompting works better than 0 shot prompting resulting in richer class descriptions. Moreover, unique k-shot examples serve better at generating multiple views than repeated k-shot examples for each view.

mance of the model while reducing its memory complexity in cross-modal attention.

5.3. Ablation on generating views from LLM.

In this section, we study how to generate good text supervision from an LLM for zero-shot image classification. Unless mentioned, we do not use the Wiki article as a view to only study the impact of the LLM.

Influence of LLM k-shot prompting on performance. We study the impact of k-shot prompting in Table 3 on generating 3 views per class. We observe that even in zero-shot prompting in row a), the views generated by the LLM allow for a very competitive model further validating that LLM can serve as a search engine for generating class supervision in ZSL. We observe that with 1-shot prompting in row b), we see an improvement over zero-shot as the LLM is now aware of what sort of information we require per class. For our 2-shot prompting, we test repeatedly querying the model three times with the same 2-shot example in row c) vs providing a unique combination of the 4 examples in row d). We observe that while repeating the same 2-shot example generates competitive views, these provide limited additional information. We see the best performance in row d) where we prompt the model with unique 2-shot examples. This allows the model to combine two labeling styles to generate a combined perspective that contains more information as evidenced by the improved accuracy numbers. We expect a further increase in performance if the LLM has access to more unique k-shot examples. However, this would require labeling 10 k-shot examples which are already 20% of the classes for AWA. Since we are interested in learning semantic embeddings with minimal supervision, we leave studying this for future works.

Impact of Multiple views on performance. We study the impact of introducing multiple views as supervision for I2MVFormer in Table 4. We observe that increased views generated with each k-shot example consistently improve the performance. The best performance is achieved by the three views and LLM for a total of 4 perspectives about each class available to the model. This further validates our hy-

Views from LLM	Zero-Shot Learning		Generalized Zero-Shot Learning					
	AWA2	FLO	AWA2			FLO		
	T1	T1	u	s	H	u	s	H
1	71.6	39.0	67.5	75.2	71.2	34.6	88.0	49.6
2	74.8	43.6	70.5	80.2	75.0	37.7	91.0	53.3
3	76.4	44.4	72.7	81.3	76.8	41.1	**91.1**	56.6
3 + Wiki	**79.6**	**46.2**	**75.7**	79.6	**77.6**	**41.6**	91.0	**57.1**
4	76.6	44.5	72.9	81.2	76.8	40.5	89.6	55.8

Table 4. **Ablating over number of views**, we observe that each view provides a useful source of information for I2MVFormer and improves the model performance.

LLM	Zero-Shot Learning		Generalized Zero-Shot Learning					
	AWA2	FLO	AWA2			FLO		
	T1	T1	u	s	H	u	s	H
PaLM 62B	74.0	38.6	66.1	**82.3**	73.3	37.1	70.3	48.6
GPT3 175B	74.2	44.2	68.8	81.0	74.2	40.4	83.4	54.5
PaLM 540B	**76.4**	**44.4**	**72.7**	81.3	**76.8**	**41.1**	**91.1**	**56.6**

Table 5. **Ablating over different LLM**, we observe that our prompting strategy can be used with different LLMs for generating powerful supervision for zero-shot image classification.

pothesis that increased views representing different annotator biases can improve the zero-shot performance. The introduction of LLM as an annotator enables it without requiring actual human annotators as the LLM has stored the knowledge available online and can use it to mimic the annotators available in the k-shot example. Comparing the performance of 2 views in Table 4 with using 3 views across the repeated 2-shot example in row c) of Table 3, we observe that the model generally benefits more from increased perspectives used in the k-shot example of each view than increased views without new perspective in the k-shot example. Moreover, we observe that the fourth view from LLM which repeats the k-shot example performs worse than using the Wiki article as the fourth view. This hints that the knowledge retrieval ability of an LLM is impacted by the information available in its k-shot examples. We expect there to be a potential further increase in performance if the LLM has access to more unique k-shot examples but this again comes at increased labeling cost.
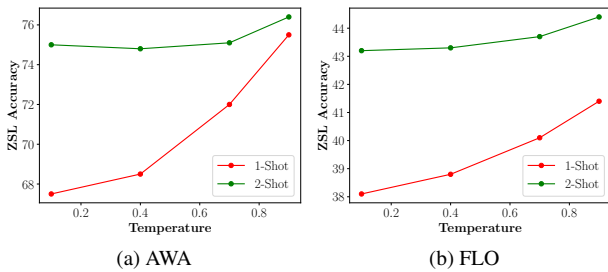

(a) AWA

(b) FLO

Figure 3. **Ablation over Temperature values.** We observe a performance increase with the temperature value of the LLM. However, at 2-shots, the model is fairly robust to changing temperature and achieves impressive performance at all temperature values.

Impact of temperature value of LLM on performance. We study the impact of the temperature value of the LLM in Figure 3. The temperature value of the LLM controls the distribution it explores. A high value means the language model will sample from further away from the mean and introduce more stochasticity. We observe from Figure 3 that the performance of I2MVFormer improves with the temperature value as the LLM can represent more diverse text views of the class with its effect most profound in

one-shot prompting. However, this temperature value can end up being yet another hyperparameter to ablate over. We notice that in two-shot prompting, this temperature value has a smaller impact on the performance of I2MVFormer as the 2 examples provide better conditioning for each view to constrict the model in what kind of information we require for each class. We see a smaller change in performance across different temperature values with the best performance achieved at a temperature of 0.9 on both datasets.

Impact of size and family of LLM on performance. We study the impact of different LLM in Table 5 and observe that the largest model PaLM540B provides the best auxiliary information leading to the best performance. Moreover, we observe that the smaller 60B version of PaLM achieves very promising performance indicating that while increased parameters do bring more performance, the "smaller" LLM can still achieve impressive results. As we compare the results of PaLM with GPT3, we observe that the larger PaLM540B model outperforms GPT3 for our 2-shot setup as also noted in the original PaLM manuscript [12].

6. Conclusion

We propose a novel perspective of using a Large Language Model as an oracle to reveal multiple views (text descriptions) of a class. Since an LLM is trained on web-scale data it only requires a few k-shot examples to generate multiple high-quality text descriptions. We show that these LLM-generated views provide complementary information to Wiki documents for learning a zero-shot image classification model. We propose I2MVFormer, a novel transformer-based model, that incorporates our SVSummary module to learn a per-view summary representing discriminative information about a class available in each view. These summaries are used by our MVSummary module to learn class-level multi-view summaries. The multi-view summaries are aligned with the global and local image information to learn a highly discriminative zero-shot image classification model. Our summary modules allow a reduction in the memory requirement of utilizing text in zero-shot image classification models. Moreover, I2MVFormer brings significant performance improvements to set a new state of the art in unsupervised semantic embeddings.

References

[1] Ziad Al-Halah and Rainer Stiefelhagen. Automatic discovery, association estimation and learning of semantic attributes for a thousand categories. In *CVPR*, 2017. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. In *arXiv:2004.05150*, 2020. 6

[4] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *ICLR*, 2019. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 1, 2, 3

[6] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *ICCV Workshops*, 2017. 2, 3

[7] Sebastian Bujwid and Josephine Sullivan. Large-scale zero-shot image classification from rich and diverse textual descriptions. In *LANTERN*, 2021. 1, 2

[8] Yannick Le Cacheux, Herve Le Borgne, and Michel Crucianu. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *ICCV*, 2019. 2

[9] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 2

[10] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. 6

[11] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 2

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 2, 5, 8

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[14] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *CVPR*, 2013.

[15] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. Link the head to the" beak": Zero shot learning from noisy text description at part precision. In *CVPR*, 2017. 2

[16] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 3

[17] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *ICCV*, 2019. 2

[18] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, 2019. 2, 3

[19] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Learning attention propagation for compositional zero-shot learning. *WACV*, 2023. 2

[20] Jihyung Kil and Wei-Lun Chao. Revisiting document representations for large-scale zero-shot learning. In *NAACL*, 2021. 1, 2

[21] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018. 2

[22] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2, 5

[23] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. In *EMNLP*, 2021. 3

[24] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. In *EMNLP*, 2021. 3

[25] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *CVPR*, 2022. 2, 3

[26] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *NeurIPS*, 2018. 2

[27] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022. 3

[28] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 2

[29] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. In *T-PAMI*. IEEE, 2022. 2, 3

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 2

[31] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning. *ArXiv*, 2021. 1, 2

[32] Yassine M'rabet and Dina Demner-Fushman. Holms: Alternative summary evaluation with large language models. In *COLING*, 2020. 2

[33] Muhammad Ferjad Naeem, Evin Pınar Örnek, Yongqin Xian, Luc Van Gool, and Federico Tombari. 3D Compositional Zero-shot Learning with DeCompositional Consensus. In *ECCV*, 2022. 2, 3

[34] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021. 2, 3

[35] Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. I2dformer: Learning image to document attention for zero-shot image classification. *arXiv preprint arXiv:2209.10304*, 2022. 1, 2, 3, 4, 5, 6, 7

[36] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGI*, 2008. 2, 3, 5

[37] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. In *IJCV*. Springer, 2014. 1, 2, 3

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 2, 5, 6

[39] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*, 2016. 1, 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021. 2

[41] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 2

[42] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information processing & management*, 1988. 6

[43] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, 2019. 2

[44] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, 2019. 2

[45] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 2, 3

[46] Jie Song, Chengchao Shen, Jie Lei, An-Xiang Zeng, Kairi Ou, Dacheng Tao, and Mingli Song. Selective zero-shot classification with augmented attributes. In *ECCV*, 2018. 1, 2

[47] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*, 2020. 6

[48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 2, 3, 5

[49] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018. 2, 3

[50] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022. 2, 3

[51] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 2

[52] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 3

[53] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. In *T-PAMI*, 2019. 1, 2, 3, 5

[54] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 2

[55] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *CVPR*, 2022. 2, 6

[56] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In *ACL*, 2020. 2, 3

[57] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 7

[58] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013. 1, 2

[59] Yunlong Yu, Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *NeurIPS*, 2018. 2

[60] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 1, 2

[61] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 4

[62] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 2

[63] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1, 2

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Zi-wei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 3

[66] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021. 2

[67] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 2

[68] Yizhe Zhu, Jianwen Xie, Bingchen Liu, and Ahmed Elgammal. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In *ICCV*, 2019. 2

[69] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*, 2019. 2