2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00365

Open-Vocabulary Multi-label Image Classification with Pretrained Vision-Language Model

Son D.Dao
Faculty of Information and Technology
Monash University
Melbourne, Australia
duy.dao@monash.edu

Dat Huynh
Northeastern University
United States
huynh.dat@husky.neu.edu

He Zhao
CSIRO's Data61
Australia
he.zhao@ieee.org

Dinh Phung
Faculty of Information and Technology
Monash University
Melbourne, Australia
dinh.phung@monash.edu

Jianfei Cai
Faculty of Information and Technology
Monash University
Melbourne, Australia
jianfei.cai@monash.edu

***Abstract*—We design an open-vocabulary multi-label image classification model to predict multiple novel concepts in an image based on a powerful language-image pretrained model i.e. CLIP. While CLIP achieves a remarkable performance on single-label zero-shot image classification, it only utilizes global image feature which is less applicable for predicting multiple labels. To address the problem, we propose a novel method that contains an Image-Text attention module to extract multiple class-specific image features from CLIP. In addition, we introduce a new training method with contrastive loss to help the attention module find diverse attention masks for all classes. During testing, the class-specific features are interpolated with CLIP features to boost the performance. Extensive experiments show that our proposed method achieves state-of-the-art performance on zero-shot learning tasks for multi-label image classifications on two benchmark datasets.**

***Index Terms*—Zero-Shot Learning, Open-Vocabulary Multi-Label Classification.**

I. Introduction

Recognition of multiple labels in an image, referred to as multi-label image classification, is a fundamental problem in computer vision with many applications. Although current multi-label classification methods show effectiveness in dealing with a large number of annotated (seen) labels provided with the training dataset, most of them cannot predict new (unseen) labels in the inference phase, which is a practical and important problem in real-world scenarios. To tackle this issue, many approaches for multi-label zero-shot learning (MLZSL) [1]–[4] have been recently proposed to recognize multiple new (unseen) categories in images at test time, without having seen the corresponding visual examples during training (referred as ZSL setting). While in the generalized ZSL (GZSL) setting, test images can simultaneously contain multiple seen and unseen classes. In addition, there is a relaxed version of MLZSL, called open-vocabulary setting, when a model has access to a source of low-cost supervision data during the training step. That data source may have an implicit intersection but have a different form with the unseen label set, for example, image-caption pairs. Although ZSL/GZSL/open-vocabulary problems have been extensively studied for single-label classifications, they are new and emerging areas for multi-label image classifications. In those emerging areas, one important idea is to project images and labels into a joint visual-label space that captures the relationships between images and labels and the relationships between seen and unseen labels.

On the other hand, CLIP [5] is a recently vision-language pretrained model that learns the alignment between images and texts in the embedding spaces. Trained on huge datasets, CLIP's rich feature space shared by both image and text data enables zero-shot transfer to a range of down-stream tasks including image classification [5], object detection [6] and semantic segmentation [7]. Despite the impressive performance of CLIP on other problems, to our knowledge, how to leverage its power to benefit the multi-label case has not been carefully studied. This is a nontrivial task as CLIP uses only one global image feature to calculate the similarity score with text labels for ZSL predictions. This is suitable for the single-label problem but not for the multi-label problem. The global feature can capture the overall context of the image that is usually dominated by big or common objects/concepts, but small objects/concepts from diverse locations can be ignored.

In this paper, we introduce a simple yet effective method to leverage information from CLIP model to benefit multi-label ZSL. Specifically, with CLIP as our backbone model, we introduce a novel module that extracts multiple class-specific features from the original CLIP feature, which can better generalize for MLZSL. The proposed module relies on the self-attention layer, with new modifications that capture the relationship between text embeddings and image features, outputting multiple class-specific features. Moreover, when training the module, we employ the contrastive loss [8] to help the module exploit diverse regions from the image features.

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00365

During testing, class-specific features are combined with CLIP feature, which already has rich image information, to boost the model performance. Extensive experimental results on two well-known datasets show that our proposed method outperforms previous works by a large margin in MLZSL tasks and achieves state-of-the-art performances. Our contributions are summarized as follows: (1) We introduce a novel image-text attention module built upon CLIP backbone and a new training procedure with contrastive loss to obtain diverse image locations for class-specific features for MLZSL problems. (2) We comprehensively show that CLIP's generalization power can be a significant benefit for MLZSL tasks and demonstrate that our proposed approach outperforms other methods significantly on two popular datasets NUS-WIDE and Open-Image.

II. Related Works

A. Multi-label Zero-shot Learning

In multi-label zero-shot learning, The goal is to learn from the seen label set and generalize to the unseen label set. The ZSL problem for single-label classifications has been widely researched in [9], [10], while ZSL for multi-label image classification is a relatively new and emerging direction [1], [2], [11]. The common idea behind the existing works is that an external source of semantic information, such as attribute vectors or text embeddings, is paired with image features for measuring the compatible scores of certain labels. These scores are then ranked to determine the existence of the labels in the testing image.

For instance, previous works [11]–[13] try to optimize a joint embedding space for image features and label features using a modified zero-shot ranking loss. In a different direction, [14] suggested incorporating a structured knowledge graph to describe the relationships between multiple labels and the interdependence between seen and unseen labels. An information propagation mechanism in the knowledge graph is proposed for solving multi-label classification and ML-ZSL tasks. However, these works abandon the discriminative power from image regions by using only a global image feature for classification. As aforementioned, the global feature is useful to describe the overall image content, which is beneficial for single-label classification, but it is suboptimal to use for multi-label classification as small objects or different instances in different locations can be ignored. Some previous studies [15], [16] address the problem by utilizing the region proposal method to generate multiple features for each label. However, the proposal requires bounding box information and is inapplicable to abstract concepts.

Recently, [1] introduced shared attention-based MLZSL approaches, where the shared attention features are class-agnostic and then combined with specific text embeddings of evaluation classes to produce compatible scores for ranking. Since the attentions are shared among all classes, they propose 3 regularization terms to help the attention module locate diverse image regions corresponding to each class. The losses encourage different attention modules to focus on diverse regions of an image, while enforce the seen labels to effectively use all attention modules and to find relevant regions that lead to high compatible scores. On the other hand, [2] proposed a region-based classification framework that enhances the region-based features through region and scene information. Then, the enriched feature is multiplied with the text embeddings of each seen class to obtain the attention mask for each seen class. Afterward, top-k activation is chosen to produce classification scores. The top-k selection allows the model to find diverse image regions for different classes. While we also extract class-specific features using attention mechanism, we leverage a powerful multi-modal pretrained model for generalizing to unseen labels. We further utilize a contrastive learning loss function to help the attention module find diverse and meaningful image locations for different classes.

B. Learning from pretrained Vision-Language.

While it is hard to obtain exact labels for each image, it is easier to have image-captioning pairs from the Internet. [5] trained a multi-modal model which aligns image representations and text representations by a contrastive framework. The pretrained model (CLIP) is transferable to many downstream tasks such as image classification, object detection, semantic segmentation or image generation [6], [7], [17]. Our framework leverages the power of CLIP to the MLZSL problem, which has not been carefully studied. By using the availability of label information, we extract multiple class-specific features from CLIP using its text embeddings and its image features.

III. Proposed Method

In this section, we first introduce the problem formulation of MLZSL and its tasks as well as the definition of open-vocabulary multi-label classification. Next, we present a naive way of using CLIP to solve the tasks. After that, we describe our method including the proposed Image-Text attention module, contrastive loss during training, and feature interpolation during testing.

A. Problem Formulation.

Multi-label Zero-shot Learning (MLZSL). Let $\boldsymbol{X}_i \in \mathbf{X}$ be the feature of the input image i and $\boldsymbol{y} \in \{0,1\}^{\mathbb{S}}$ be the corresponding multi-hot ground truth label among $\mathbb{S}$ seen labels of the training label set $\boldsymbol{Y}^{\mathbb{S}}$. The goal of MLZSL is to learn a mapping $f(\boldsymbol{X}) : \boldsymbol{X} \rightarrow \{0,1\}^{\mathbb{S}}$ which can be generalized to $\mathbb{U}$ unseen labels from the testing label set $\boldsymbol{Y}^{\mathbb{U}}$. There are two strategies for testing. One is Zero-shot Learning (ZSL) predicting unseen label i.e., $f(\boldsymbol{X}) : \boldsymbol{X} \rightarrow \{0,1\}^{\mathbb{U}}$, and the other is Generalized Zero-Shot Learning (GZSL) predicting both seen and unseen label i.e. $f(\boldsymbol{X}) : \boldsymbol{X} \rightarrow \{0,1\}^{\mathbb{Y}}$, where $|\mathbb{Y}| = |\mathbb{S}| + |\mathbb{U}|$.

Open-vocabulary multi-label classification. The setting is similar to ZSL and GZSL, except during training, the model has access to a large source of low-cost supervision data that may have an implicit intersection with the unseen label set

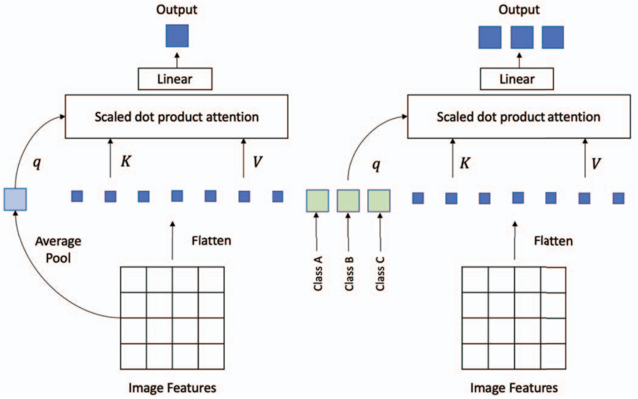

Fig. 1: Illustrations of **left:** CLIP's attention pooling layer, which uses global average-pooling feature as the query, and **right:** Our proposed Image-Text attention module, which uses text embedding as query. The proposed module can extract class-specific features from diverse spatial locations.

i.e image-caption pair. More specifically, the model is trained using an image-caption dataset that covers a diverse number of words, and a much smaller dataset with seen multi-label annotation. Note that in this process, unseen labels might have overlapped with the entire language vocabulary, but they remain unknown during training. In our opinion, previous works on MLZSL [1], [2] that use ImageNet pretrained model can also be categorized as open-vocabulary multi-label classification because ImageNet labels may have intersection with other datasets such as NUS-WIDE or Open-Image.

B. A Naive Adaptation of CLIP for Open-Vocabulary multi-label classification

As CLIP can learn aligned embeddings of texts and images, it has achieved promising results for single-label zero-shot classification [5]. Its inference mechanism naturally fits MLZSL. Specifically, by using the CLIP text encoder, we can obtain the L1-normalized text embeddings of all class labels in the dataset (including both seen and unseen classes), denoted as $\boldsymbol{T} \in \mathbb{R}^{N_C \times d}$, where N_C is the number of classes and d is the feature's dimension. Given the well trained CLIP image encoder $E_{img}(\cdot)$, we extract the image features $\boldsymbol{X}$ for the image set $\mathbf{I}$ as follows: $\boldsymbol{X} = E_{img}(\mathbf{I}); \boldsymbol{X} \in \mathbb{R}^{N \times d}$, where N is the batch size.

The task of Open-vocabulary prediction fits naturally to CLIP pretrained model as it is trained to match between image features and text embeddings. During inference, we can compare the similarity between image features and text embeddings extracted from the class names. For instance, we can compute the similarity score between an image feature $\boldsymbol{x} \in \boldsymbol{X}, \boldsymbol{x} \in \mathbb{R}^d$ and the j^{th} class text embeddings $\boldsymbol{t}_j \in \boldsymbol{T}; \boldsymbol{t}_j \in \mathbb{R}^d$ by this function:

$$p_j = < \boldsymbol{x}, \boldsymbol{t}_j >; \boldsymbol{p} = \{p_j | j \in \{1, \cdots, N_C\}\}, \qquad (1)$$

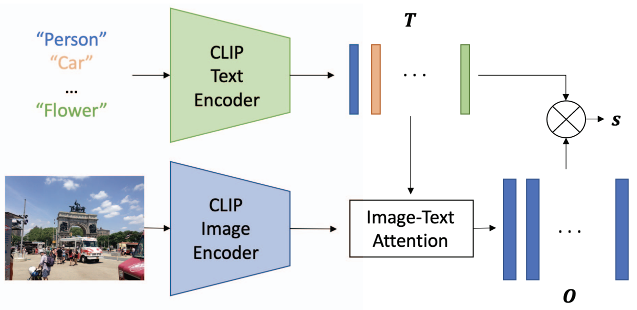

Fig. 2: The overall structure of our proposed method, utilizing CLIP pretrained text and image encoders. The proposed Image-Text attention module is the only part that needs to be trained. The set of class names is encoded through the text encoder to get the text embeddings $\boldsymbol{T}$. The Image-Text attention module takes $\boldsymbol{T}$ and image features from the image encoder as input and returns class-specific features $\boldsymbol{O}$. We then compute the row-wise dot production $\boldsymbol{s} = < \boldsymbol{O}, \boldsymbol{T} >$ as class scores.

where $< \cdot >$ is the dot production, and $\boldsymbol{p}$ is the total score for all classes. Then we can pick the top-k highest score from $\boldsymbol{p}$. We name this naive adaption of CLIP as CLIP-ML.

C. Proposed Method

While the naive adaptation is a reasonable model, we can still improve it for MLZSL. Specifically, as shown in the left sub-figure of Figure 1, the image encoder of CLIP adopts a Transformer-style multi-head attention layer where the globally average-pooled feature works as the query and the feature at each spatial location generates a key-value pair. In this way, the global features are accumulated with other spatial features for the final output feature, which may be suitable for capturing big objects and common objects in the image, but the fine-grained information from diverse locations can be ignored.

In this paper, we propose a new method with CLIP as the backbone, whose overall architecture is described in Figure 2. In general, we observe that in CLIP, the image feature computed at each spatial location already captures a rich response of local semantics that correspond well with tokens in the text embeddings. The use of the average pooling in CLIP destroys such local semantics. Taking that into account, we propose an attention module to extract class-level features from CLIP. Our modification is simple and takes into account diverse locations from the image features. Specifically, given an image with a set of multiple labels, by viewing a label as a word or phrase, we can get the image feature and the label embeddings from CLIP, which are denoted as $\boldsymbol{X}' \in \mathbb{R}^{HW \times d}$ and $\boldsymbol{t}_j \in \mathbb{R}^d$. Note that $\boldsymbol{X}'$ is the feature before the attention pool layer from CLIP. Here we would like to learn the class-specific image feature $\boldsymbol{o}_j$ with an attention module with $\boldsymbol{X}'$

and $\boldsymbol{t}_j$, by using the following attention mechanism:

$$\boldsymbol{o}_j = \mathcal{F}_o(\sum_i softmax(\frac{\boldsymbol{q}_j \boldsymbol{k}_i}{C})\boldsymbol{v}_i), \tag{2}$$

$$\boldsymbol{q}_j = \mathcal{F}_q(\boldsymbol{t}_j); \boldsymbol{K} = \mathcal{F}_k(\boldsymbol{X}'); \boldsymbol{V} = \mathcal{F}_v(\boldsymbol{X}'), \tag{3}$$

where $\mathcal{F}_*$ are the linear layers, $\boldsymbol{q}_j \in \mathbb{R}^d$, $\boldsymbol{k}_i, \boldsymbol{v}_i$ are the feature at i^{th} location (i^{th} row) of $\boldsymbol{K} \in \mathbb{R}^{HW \times d}$ and $\boldsymbol{V} \in \mathbb{R}^{HW \times d}$ respectively, and C is a constant scaling factor. Here we have simplified the formula by ignoring the channel-wise splitting and concatenation.

We denote the class-specific features for all the classes as $\boldsymbol{O} \in \mathbb{R}^{N_C \times d}$, where N_C is the number of classes, and each row $\boldsymbol{o}_j (\forall j \in \{1, \cdots, N_C\})$ is the image feature for class j. Here, $N_C = |\mathbb{S}|$ during training, and $N_C = |\mathbb{S}| + |\mathbb{U}|$ during testing. Given the output feature $\boldsymbol{O}$, we can calculate its similarity score for each class j as follows:

$$s_j =< \boldsymbol{o}_j, \boldsymbol{t}_j >; \boldsymbol{o}_i \in \boldsymbol{O}; \boldsymbol{t}_j \in \boldsymbol{T}, \tag{4}$$

Note that both $\boldsymbol{o}_j$ and $\boldsymbol{t}_j$ are L1-normalized and the scores for all classes is denoted as $\boldsymbol{s} = \{s_j | j \in \{1, \cdots, N_C\}\}$.

Although attention mechanisms have been used in previous works [1], [2] to generalize to unseen classes, our method is quite different from theirs. Specifically, our attention module is specially designed on top of CLIP to extract the class-specific features for images. Since CLIP is trained on a huge amount of image-captioning pairs, image features are well aligned with text embeddings, which are leveraged in our method to unseen classes.

D. Training the Proposed Method

Multi-label Ranking Loss: Our proposed method can be trained by the commonly used ranking loss function in MLZSL [3], [11], where we want the scores for the positive labels to be higher than the scores for the negative labels. In other words, for one image i, we want to minimize the cost c^i_{jk} defined as:

$$c^i_{jk} = s^-_k - s^+_j; k \in \bar{\mathbb{P}}^i; j \in \mathbb{P}^i, \tag{5}$$

where $\mathbb{P}^i$ is the set of positive labels and $\bar{\mathbb{P}}^i$ is the set of negative labels of image i, respectively. We have $|\mathbb{P}^i| + |\bar{\mathbb{P}}^i| = N_C$, which is the total number of classes. s^-_k and s^+_j are the scores for the negative label k and the positive label j, respectively. Following [11], the ranking loss function for one image is:

$$\mathcal{L}^i_{rank} = \frac{1}{|\mathbb{P}^i|} \sum_{j \in \mathbb{P}^i} \sum_{k \in \bar{\mathbb{P}}^i} \log(1 + \exp{(c^i_{jk})}) \tag{6}$$

The total ranking loss is the mean over all images within the mini-batch:

$$\mathcal{L}_{rank} = \frac{1}{N} \sum_{i \in N} \mathcal{L}^i_{rank} \tag{7}$$

Contrastive loss for diverse attention masks: In our study, we find that using ranking loss leads to reasonable results. However, the attention masks of the classes are less diverse and usually focus on the same locations of an image, as shown in the bottom row of Figure 3. The reason is that the rank loss tries to differentiate the positive and the negative labels, but ignores the difference within the positive labels.

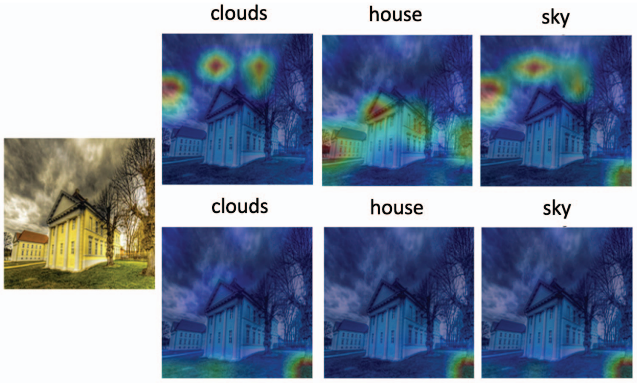

Fig. 3: The contrastive loss helps improve the quality of attention masks. Left most is the input image. Top row shows the attention masks extracted from the model with $\mathcal{L}_{Con}$. Bottom row shows the masks extracted from the model without $\mathcal{L}_{Con}$.

To help the model produce more meaningful and diverse attention masks, we introduce the supervised contrastive loss for positive class-specific features. Motivated by [8], the loss enforces features of the same class to be similar while be different from other classes. Given a mini-batch of N images, we first project all the class-specific image features $\boldsymbol{O}$ into a lower dimension embedding set $\mathbb{Z} = \{\boldsymbol{z}_{ij} \in \mathbb{R}^{d_z} | i \in \{1, \cdots, N\}; j \in \{1, \cdots, N_C\}\}$. Similarly, we define the set of the ground-truth labels of the class-specific features as: $\mathbb{Y}' = \{y'_{ij} \in \{0, 1\} | i \in \{1, \cdots, N\}; j \in \{1, \cdots, N_C\}\}$. If we view an image's class-specific embedding $\boldsymbol{z}_{ij}$ as an instance instead of the image itself, $\boldsymbol{z}_{ij}$ is associated with a single ground-truth label y_{ij}. We further define $\mathbb{I} = \{\boldsymbol{z}_{ij} \in \mathbb{Z} | y'_{ij} = 1\}$ as the set that contains the embeddings of all positive labels in the mini-batch, and $\mathbb{A}(i, j) = \mathbb{I} \setminus \boldsymbol{z}_{ij}$ as the set contains the embeddings in $\mathbb{I}$ with $\boldsymbol{z}_{ij}$ excluded.

In the mini-batch, we now consider $\boldsymbol{z}_{ij} \in \mathbb{I}$ as the anchor, which is the feature of class j in image i. The set of the embeddings with the same label as the anchor is defined as $\mathbb{P}(i, j) = \{\boldsymbol{z}_{kj} \in \mathbb{A}(i, j) | y'_{kj} = y'_{ij} = 1\}$. We define the contrastive loss for the anchor $\boldsymbol{z}_{ij}$ as:

$$\mathcal{L}^{ij}_{Con} = \frac{-1}{|\mathbb{P}(i, j)|} \sum_{\boldsymbol{z}_p \in \mathbb{P}(i,j)} \log \frac{\exp(\boldsymbol{z}_{ij} \cdot \boldsymbol{z}_p / \tau)}{\sum_{\boldsymbol{z}_a \in \mathbb{A}(i,j)} \exp(\boldsymbol{z}_{ij} \cdot \boldsymbol{z}_a / \tau))}, \tag{8}$$

where τ is the scalar temperature. The contrastive loss for the whole mini-batch is simply the summation of all the anchors:

$$\mathcal{L}_{Con} = \sum_{\boldsymbol{z}_{ij} \in \mathbb{I}} \mathcal{L}^{ij}_{Con} \tag{9}$$

We train our model with the total loss as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rank} + \mathcal{L}_{Con} \tag{10}$$

Method	Task	NUS-WIDE (#seen/ #unseen = 925/81)							Open-Images (#seen/ #unseen = 7186/400)						
		K = 3			K = 5			mAP	K = 10			K = 20			mAP
		P	R	F1	P	R	F1		P	R	F1	P	R	F1	
CONSE [18]	ZSL	17.5	28.0	21.6	13.9	37.0	20.2	9.4	0.2	7.3	0.4	0.2	11.3	0.3	40.4
	GZSL	11.5	5.1	7.0	9.6	7.1	8.1	2.1	2.4	2.8	2.6	1.7	3.9	2.4	43.5
LabelEM [19]	ZSL	15.6	25.0	19.2	13.4	35.7	19.5	7.1	0.2	8.7	0.5	0.2	15.8	0.4	40.5
	GZSL	15.5	6.8	9.5	13.4	9.8	11.3	2.2	4.8	5.6	5.2	3.7	8.5	5.1	45.2
Fast0Tag [11]	ZSL	22.6	36.2	27.8	18.2	48.4	26.4	15.1	0.3	12.6	0.7	0.3	21.3	0.6	41.2
	GZSL	18.8	8.3	11.5	15.9	11.7	13.5	3.7	14.8	17.3	16.0	9.3	21.5	12.9	45.2
One Attention per Label [20]	ZSL	20.9	33.5	25.8	16.2	43.2	23.6	10.4	-	-	-	-	-	-	-
	GZSL	17.9	7.9	10.9	15.6	11.5	13.2	3.7	-	-	-	-	-	-	-
LESA [1]	ZSL	25.7	41.1	31.6	19.7	52.5	28.7	19.4	0.7	25.6	1.4	0.5	37.4	1.0	41.7
	GZSL	23.6	10.4	14.4	19.8	14.6	16.8	5.6	16.2	18.9	17.4	10.2	23.9	14.3	45.4
ZSSDL [3]	ZSL	24.2	41.3	30.5	18.8	53.4	27.8	25.9	6.1	47.0	10.7	4.4	68.1	8.3	62.9
	GZSL	27.7	**13.9**	**18.5**	23.0	**19.3**	21.0	12.1	**35.3**	**40.8**	**37.8**	**23.6**	**54.5**	**32.9**	75.3
BiAM [2]	ZSL	-	-	33.1	-	-	30.7	26.3	-	-	8.3	-	-	5.5	**73.6**
	GZSL	-	-	16.1	-	-	19.0	9.3	-	-	19.1	-	-	15.9	**84.5**
CLIP-ML (RN50) [5]	ZSL	32.4	39.8	35.7	25.4	52.1	34.1	36.4	7.0	54.5	12.4	4.9	75.6	9.1	69.1
	GZSL	27.2	10.5	15.1	22.3	14.3	17.4	12.9	15.2	17.5	16.3	11.2	25.8	15.6	75.3
OVML (Ours)	ZSL	**36.3**	**44.7**	**40.1**	**27.9**	**57.2**	**37.5**	**42.6**	**9.6**	**74.4**	**16.9**	**5.6**	**87.5**	**10.6**	68.4
	GZSL	**32.9**	12.6	18.3	**28.1**	18.0	**22.0**	**14.3**	33.9	39.2	36.4	23.2	53.5	32.3	77.8

TABLE I: MLZSL results on NUS-WIDE and Open-Image dataset. The best results are in boldfaced and the second best results are underlined.

Qualitatively, the top row of Figure 3 shows that with the proposed supervised contrastive loss, the attention masks are much more diverse and accurate than using only the rank loss.

IV. Experiments

Datasets. To demonstrate the effectiveness of the proposed method, we evaluate our method on two well-known datasets NUS-WIDE [21] and Open-Image [22]. The NUS-WIDE dataset contains 81 carefully annotated labels from human annotators and 925 additional labels extracted from Flicker user tags. Similar to other works [1]–[3], we use the 925 labels as seen and the other 81 labels as unseen. We also test our method on Open-Images (v4) dataset, which consists of 9 million training images, 41,260 validation images, and 125,436 testing images. The seen label set has 7,186 labels from the training set, where each label has at least 100 training samples. Following other works, we select 400 most frequent labels from the test set as the unseen labels during testing.

Evaluation Metrics. Similar to other work on MLZSL [1]–[3], for evaluation, we use the mean Average Precision (mAP) and Precision (P), Recall(R), F1-score (F1) at top K predictions in each image.

Implementation Details. We use the pretrained Resnet-50 (RN-50) of CLIP as our main backbone for comparison. The backbone is fixed during training. To train the attention module, we use Adam and 1-cycle learning rate schedule with the maximum learning rate set to 2e-4. The batch size is 64. The input image resolution is set to 224×224.

Main Results We compare our method with state-of-the-art methods. Table I shows the results. On the NUS-WIDE dataset, it can be seen that our approach achieves state-of-the-art on the ZSL task. Ours outperforms all the previous works on all the metrics by a large margin. Specifically, our method surpasses the second-best method (BiAM) by an absolute gain of 7.0%/ 6.8%/ 16.3% on F1@3, F1@5, and mAP respectively. Interestingly, the naive adaptation of CLIP for multi-label classification, CLIP-ML achieves remarkable performance on both ZSL and GZSL settings, which is even better than several advanced methods. Our proposed method based on CLIP improves on CLIP-ML with a significant margin on all the metrics. For the GZSL task, our method still leads on the majority of the metrics. On the other metrics, ours is the second best, whose performance is marginally below that of ZSSDL. Since the backbone of ours is fixed during training, it might be less effective to predict seen classes when compare to ZSSDL which has a whole network trained on the seen classes.

The results on Open-Image dataset show the same patterns as on NUS-WIDE in Table I. Concretely, OVML achieves the highest F1@3 and F1@5 scores on ZSL task. Among the existing methods, ZSSDL achieves the second-best F1@3 and F1@5 scores on ZSL task. Our method outperforms ZSSDL by increasing 6.2% and 2.3% on F1@3 and F1@5 respectively. When comparing OVML and CLIP-ML, our method has significant improvement in most of the metrics. The naive CLIP-ML also outperforms other existing methods on ZSL task on all metrics, except mAP. In this dataset, ZSSDL still has the highest performance in GZSL task, while BiAM has the best mAP on both tasks.

We believe that the use of CLIP and our specially designed architecture and training algorithm that fit CLIP are the sources of the performance gain. To examine whether using CLIP as the backbone improves over other methods, we replace the backbones of ZSSDL and BiAM with the pretrained RN-50 of CLIP, which is the same as ours. The results on NUS-WIDE are shown in Table II. It can be seen that both ZSSDL and BiAM with CLIP improve over their counterparts without CLIP, while ours outperforms their CLIP versions in general. This shows that our specially designed architecture and training algorithms can make the best use of CLIP's power.

Ablation Study. We provide the ablation study on each component of the proposed method in Table III. The results

Method	Task	F1@3	F1@5	mAP
ZSSDL	ZSL	30.5	27.8	25.9
	GZSL	18.5	21.0	12.1
CLIP-ZSSDL	ZSL	36.8	35.5	34.6
	GZSL	**19.2**	**23.2**	11.2
BiAM	ZSL	33.1	30.7	26.3
	GZSL	16.1	19.0	9.3
CLIP-BiAM	ZSL	34.4	32.8	29.5
	GZSL	15.5	19.0	10.5
OVML (Ours)	ZSL	**40.1**	**37.5**	**42.6**
	GZSL	18.3	22.0	**14.3**

TABLE II: Methods with pretrained CLIP as the backbones.

Finetune	Attention	$\mathcal{L}_{Con}$	F1@3	F1@5	mAP
			35.7	34.1	36.4
✓			29.3	29.1	41.6
	✓		36.9	34.8	36.3
	✓	✓	**40.1**	**37.5**	**42.6**

TABLE III: Ablation study of the ZSL task on the NUS-WIDE dataset.

are from the ZSL task on the NUS-WIDE dataset. The first row is CLIP-ML which is the naive adaptation of CLIP while the last row corresponds to our OVML method. First, we look at whether finetuning the CLIP backbone on the training set helps. When finetuned, the results drop on F1@3 and F1@5 while increasing on mAP. This is as expected as in the ZSL setting, we focus on the generalization power of the model for the unseen labels. When finetuned on the training set, the model will focus more on the seen labels, which may hurt its performance on unseen labels. Next, we examine how the proposed attention module improves. We can see that training with the proposed image-text attention module leads to slightly better F1@3 and F1@5 scores while having a similar mAP score compare to the original CLIP-ML. However, as shown in Figure 3 and described in section III-C, the attention masks tend to focus on the same locations, which is not ideal to extract different class information. Finally, it can be observed that adding the contrastive loss improves all three metrics and achieves the best results.

V. Conclusion

In this paper, we have introduced OVML, an open-vocabulary multi-label image classification model. Specifically, the proposed method has an Image-Text attention module to exploit class-specific information from the rich semantic feature of CLIP, a powerful pre-trained vision-language model, which can be effectively used for predicting multiple classes. To properly train the model, we have adopted contrastive loss during the training process to help the attention module find informative and diverse spatial information. Experiments on two popular datasets show that the proposed approach achieves state-of-the-art results on ZSL. We believe that our paper will inspire more future works on leveraging the power of CLIP for the MLZSL and MLGZSL problems.

References

[1] D. Huynh and E. Elhamifar, “A shared multi-attention framework for multi-label zero-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8776–8786.

[2] S. Narayan, A. Gupta, S. Khan, F. S. Khan, L. Shao, and M. Shah, “Discriminative region-based multi-label zero-shot learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8731–8740.

[3] A. Ben-Cohen, N. Zamir, E. B. Baruch, I. Friedman, and L. Zelnik-Manor, “Semantic diversity learning for zero-shot multi-label classification,” *arXiv preprint arXiv:2105.05926*, 2021.

[4] N. Hayat, H. Lashen, and F. E. Shamout, “Multi-label generalized zero shot learning for the classiffcation of disease in chest radiographs,” in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 461–477.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.

[6] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Zero-shot detection via vision and language knowledge distillation,” *arXiv preprint arXiv:2104.13921*, 2021.

[7] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, “Denseclip: Language-guided dense prediction with context-aware prompting,” *arXiv preprint arXiv:2112.01518*, 2021.

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[9] H. Jiang, R. Wang, S. Shan, and X. Chen, “Transferable contrastive network for generalized zero-shot learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9765–9774.

[10] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, “Generalized zero-shot learning via aligned variational autoencoders,” *red*, vol. 2, p. D2, 2019.

[11] Y. Zhang, B. Gong, and M. Shah, “Fast zero-shot image tagging,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 5985–5994.

[12] M. Ye and Y. Guo, “Multi-label zero-shot learning with transfer-aware label embedding projection,” in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3671–3675.

[13] Z. Lu, J. Zeng, S. Shan, and X. Chen, “Zero-shot facial expression recognition with multi-label label propagation,” in *Asian Conference on Computer Vision*. Springer, 2018, pp. 19–34.

[14] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. F. Wang, “Multi-label zero-shot learning with structured knowledge graphs,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1576–1585.

[15] S. Rahman, S. Khan, and N. Barnes, “Deep0tag: Deep multiple instance learning for zero-shot image tagging,” *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 242–255, 2019.

[16] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. L. Yuille, “Multiple instance visual-semantic embedding.” in *BMVC*, 2017.

[17] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, “Styleclip: Text-driven manipulation of stylegan imagery,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094.

[18] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, “Zero-shot learning by convex combination of semantic embeddings,” in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[19] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for image classification,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.

[20] J.-H. Kim, J. Jun, and B.-T. Zhang, “Bilinear attention networks,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 1571–1581.

[21] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “Nus-wide: a real-world web image database from national university of singapore,” in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.

[22] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit *et al.*, “Openimages: A public dataset for large-scale multi-label and multi-class image classification,” *Dataset available from https://github. com/openimages*, vol. 2, no. 3, p. 18, 2017.