Improving Multi-label Recognition using Class Co-Occurrence Probabilities

Samyak Rawlekar[1] ⋆, Shubhang Bhatnagar[1] ⋆, Vishnuvardhan Pogunulu Srinivasulu[2], and Narendra Ahuja[1]

[1] University of Illinois Urbana-Champaign, IL, USA
{samyakr2,sb56,n-ahuja}@illinois.edu
[2] Vizzhy, USA vishnu@vizzhy.com

Abstract. Multi-label Recognition (MLR) involves the identification of multiple objects within an image. To address the additional complexity of this problem, recent works have leveraged information from vision-language models (VLMs) trained on large text-images datasets for the task. These methods learn an independent classifier for each object (class), overlooking correlations in their occurrences. Such co-occurrences can be captured from the training data as conditional probabilities between a pair of classes. We propose a framework to extend the independent classifiers by incorporating the co-occurrence information for object pairs to improve the performance of independent classifiers. We use a Graph Convolutional Network (GCN) to enforce the conditional probabilities between classes, by refining the initial estimates derived from image and text sources obtained using VLMs. We validate our method on four MLR datasets, where our approach outperforms all state-of-the-art methods.

Keywords: Multi-label Recognition · Graph Convolution Networks · Vision-Language Models ·

1 Introduction

Multi-label recognition (MLR) involves identifying each of the multiple classes from which objects are present in an image. It has many applications such as identifying all different: diseases evident in a chest x-ray [22], products in a query image for e-commerce [5], and food items in a plate for diet monitoring systems [36,2]. MLR is more challenging than classifying images having a single object [18,15] because an image may contain combinatorially large mix of classes, for which learning would require exponentially larger number of images (O(2^N) images for N classes) than for Single label recognition (SLR). The objects may also occur in different layouts so recognition either requires object segmentation and recognition of each segmented object independently, or recognizing the object mix from the features measured over the entire image.

⋆ Equal contribution.

Several approaches [41,9] have taken the latter approach, namely recognition from image level features. Further, these approaches also reduce complexity by recognizing the presence of each object in an image independent of the others that may also be present. Training here amounts to learning independent classifiers for each object, which are then used to detect the corresponding objects, both using image level features. Apart from not segmenting the image and features into those for different objects, these methods also neglect the evidence for the presence of an object provided by the context of other objects. In practice, many objects are in sets, making their occurrences interdependent. Using independent classifiers neglects the mutual information present, which, if used, could enhance the performance of individual classifiers. This is particularly important in view of the already larger amount of data needed for MLR, and the relatively small sizes (vs SLR) of available annotated MLR datasets (because multiple classes present in an image require more annotation effort than required for SL images).

To mitigate the paucity of labeled data, recent MLR approaches [41,17] have focused on adapting large Vision Language Models (VLMs) e.g. CLIP [39], ALIGN [24], OpenCLIP [23] for the task. Instead of finetuning the VLM on small MLR datasets, these approaches [41,17] rely on learning a pair of positive/negative text prompts associated with each class; the prompts associated with a class are learned by maximizing/ minimizing the similarities of their embeddings with those of the embeddings extracted from images containing objects of the class. The additional information captured by the prompts constrains possible matches between the data and the desired labels, in turn limiting the number of parameters needed to be learned. This helps mitigate overfitting on small datasets. At test time, whether a class is present in a given image is ascertained by measuring the similarity of the image embeddings with embeddings extracted from previously learned positive/negative prompts for the class. However, the learning of prompts here is done independently for each class, again neglecting the class co-occurrences.

We propose a two-stage framework that leverages the knowledge of VLMs but injects the co-occurrence information into the models. We obtain an initial estimate of the evidence logits for each class in a subimage in terms of the match between the embeddings of the subimage and those of the positive and negative text prompts associated with the class. Results from the subimages are aggregated to obtain logits for all classes across the image. These independently extracted subimage logits are refined to enforce prior knowledge of the joint probabilities of pairs of classes present in different subimages. Given that a class is present in a training image, the conditional probabilities of other classes being also present are measured from the frequencies of observing those classes in the image. Since the prior probability of a class is known from the VLM output, we can combine it with the conditional probabilities of other classes to obtain joint probabilities of class pairs. Enhancement of the outputs of the independent classifiers by using conditional probabilities is carried out through a graph convolutional network (GCN). GCN thus utilizes the conditional probabilities to

improve upon the outputs of the independently obtained class estimates using the vision-language model. As commonly done, we compensate for the differences in the frequencies of different classes occurring in the training images by reweighting the estimated probabilities of different classes to remove the class bias in the training data.

We test our approach on four benchmarks: MS-COCO-small (5% of the training set), PASCAL VOC, FoodSeg103, and UNIMIB-2016. The first two are commonly used for MLR, whereas we have two additional datasets that have been used in other contexts [44,10]. The number of images in each of these datasets is small compared to even many SLR datasets, although MLR calls for larger datasets. This makes MLR here even harder. Our experiments show that the use of inter-class influence in the second stage of our framework significantly improves performance over the state-of-the-art methods, which detect each class independently. As expected, the advantage is higher for classes for which VLM yields low accuracy but which frequently co-occur with other classes for which VLM accuracy is higher. We also show that our loss re-weighing greatly improves the performance on datasets where there is a significant class bias.

Our contributions:

- We propose a two-stage framework to adapt VLMs for MLR with limited annotated data, by enhancing the VLM based independent class estimates obtained in the first stage, with conditional probability priors extracted from the dataset, using a GCN.
- We validate our algorithm quantitatively using mean average precision (mAP) on four MLR datasets. Our method surpasses the previous SOTA MLR approaches by more than 2% (COCO-14-small), 0.4% (VOC-2007), 3.9% (FoodSeg103) and 11% (UNIMIB2016).

2 Related Works

2.1 Multi-Label Recognition

Multi-label recognition is an important, well-studied problem in computer vision, with a wide range of approaches being proposed to tackle it [32]. An important line of work has focused on learning disjoint binary classifiers for identifying each class of objects in an image [31,37,11,9]. These approaches require large labeled datasets for training. They make no use of information that can be derived from modeling the co-occurrence of different classes, which is especially important when only a small amount of annotated data is available for training.

Other works have proposed to use recurrent neural networks (RNNs) to model label dependencies in an image [42,30,48]. Specifically, they cast MLR into a sequence prediction problem, using beam search to find a sequence of objects having the highest likelihood of being present in the image. Like our method, these methods also model label dependencies, but do so implicitly via the hidden state of RNNs. This requires vast amounts of labeled data to learn them. Recent approaches for MLR have proposed the use of VLMs to improve MLR performance when only a limited amount of labeled data is available.

2.2 Vision-Language Models for MLR

VLMs learn representations that are transferable to a wide range of downstream tasks such as classification [21,43,47,51], retrieval [3,26], and segmentation[45,49,50] by aligning hundreds of millions of image-text(prompts) pairs. Such approaches commonly focus on learning prompts suitable for these downstream tasks [53]. [41] adapts VLMs for MLR, proposing to learn a pair of prompts associated with the presence/ absence of each class. Text embeddings extracted from the learned prompts are used to gather local evidence from the image features extracted by the VLM, which is then aggregated and combined. On the other hand, SCPNet [17] is another VLM based approach for MLR which uses the similarity between the class prompt embedding from VLMs to train a classifier for each class. In these methods, classifiers for each class are independent of those for other classes, not making use of any contextual information during inference on an image.

2.3 Long tailed Learning

The distribution of frequency with which objects belong to a class in real-world images often follows a long-tailed distribution[25,35,34,13]. Networks trained for multi-class classification on such data tend to perform poorly on the tail classes which have less data available. Several approaches have been proposed to mitigate this issue including data augmentation (augment the tail classes) [46], data re-sampling (sample images to obtain balanced distribution) [6,19,33,52], adjusting classifier margins (classification thresholds vary for every class)[27,4] and loss re-weighing [38,14] being popular. We use loss re-weighing to mitigate the effect of label imbalance on our method when it is trained to model label conditional probabilities.

3 Method

Suppose in a given set of images, $\mathcal{D} = \{\mathbf{x}_i\}, i \in \{1 \dots |\mathcal{D}|\}$, every image $\mathbf{x_i}$ may contain objects from up to N classes. The image is thus associated with N labels $\mathbf{y_i} \in \{0,1\}^N$ where y_i^j denotes the presence or absence (1 or 0) of the j-th class in the image. Then the MLR problem requires identification of all labels associated with any input image.

Our approach in this paper uses a VLM g_ϕ, parameterized using weights ϕ (a pair of encoders $g_{\phi,img}$, $g_{\phi,text}$). VLMs are pretrained to align image and textual features over large datasets to learn features suited for various tasks/domains. As mentioned in Sec. 1, these models associate a pair of positive and negative text prompts $\{\mathbf{t_{j,+}}, \mathbf{t_{j,-}}\}$ with each class j (complete set denoted by ψ). A text encoder $g_{\phi,text}$ extracts text embeddings from each of these prompts, gives them to an image-text feature aggregation head p, which matches visual features extracted from different subimages with the text embeddings, and combines them to obtain an initial set of logits for each class in the image. We then use a GCN f_θ with weights θ to refine the logits output by p, by leveraging the statistical

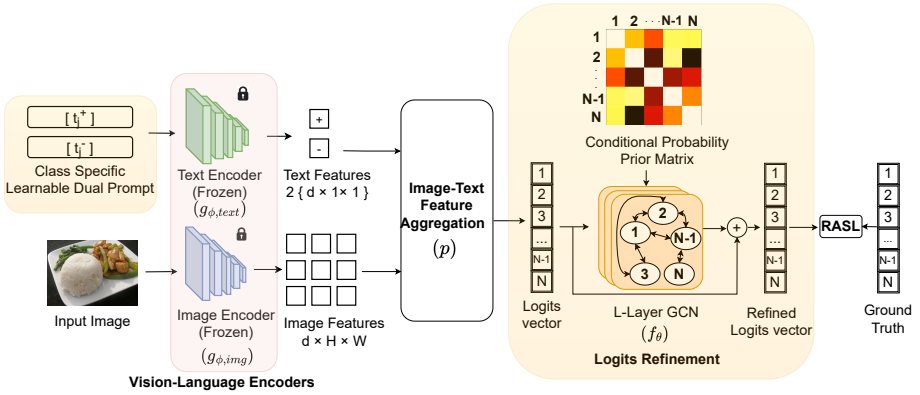

Fig. 1: Method Overview: Given an image with multiple objects, we extract image features and text features from the subimages using a vision-language model (CLIP). An image-text feature aggregation module (Sec. 3.1) combines these features to identify all classes present in the image as a union of the classes present in the subimages, giving an initial set of image level class logits. These logits are passed to a GCN, that uses conditional probabilities between classes to refine these initial predictions (Sec. 3.2). We train this framework while reweighting the loss generated by classes to address any class imbalance in the training data using a Reweighted Asymmetric Loss (RASL), a weighted version of the familiar ASL.

co-occurrence of classes observed in the training dataset. An overview of our proposed method is given in Figure 1.

The following subsections present the various parts of our method.

3.1 Initial Logits Estimation

We use a VLM as a feature extractor and initial classifier for our method. The VLM's image encoder does spatial pooling of windows in the final layer to obtain a single d dimensional feature vector for a single-label image x_i. This is not suitable for MLR case as spatial pooling operation combines the features of multiple objects in different regions of the image, with overall features being dominated by those extracted from a single object. We remove the pooling layer of the image encoder, using it to get features $g_{\phi,img}(\mathbf{x_i}) = \mathbf{z_i}$ (of shape $d \times H \times W$) for a given image x_i, hence preserving information from the individual windows.

Our image-text feature aggregation head p is similar to [41]. For each class j, it learns a pair of text prompts $\{\mathbf{t_{j,+}}, \mathbf{t_{j,-}}\}$, which are projected to d-dimensional embeddings $\mathbf{r_{j,+}}, \mathbf{r_{j,-}}$ using $g_{\phi,text}$. Cosine similarity of the d-dimensional image features at a particular point (h, w) with $\mathbf{r_{j,+}}$ indicates the presence of the class, while similarity with $\mathbf{r_{j,-}}$ indicates its absence. These similarities are aggregated and used by p to give logits $p(\mathbf{z_i})$ for the image.

3.2 Refining Logits using Conditional Prior

We refine the logits $p(\mathbf{z_i})$ using a GCN, which ensures conformity with the conditional probability priors extracted from the training dataset.

Estimating Label Conditional Probability Prior: To derive the label prior from the dataset, we first calculate the label co-occurrence matrix $C_{N\times N} = (c_{mn})$ over training dataset $\mathcal{D}$. Each entry c_{mn} is given by

$$c_{mn} = \sum_{i=0}^{|\mathcal{D}|} y_i^m \times y_i^n \tag{1}$$

The c_{mn} denotes the number of times that objects belonging to classes m and n occur together in an image. Using this, we calculate an estimate of conditional probability matrix $A_{N\times N} = (a_{mn})$, where each entry gives an estimate for the probability $P(y_i^n = 1|y_i^m = 1)$:

$$a_{mn} = \hat{P}(y_i^n = 1|y_i^m = 1) = \frac{c_{mn}}{c_{mm}} \tag{2}$$

GCN for Refinement: We refine the logits $p_\psi(\mathbf{z_i})$ using the GCN f_θ which uses the conditional probability matrix A to define the connection weights between its N nodes. Specifically, given that f_θ has L layers, each layer calculates:

$$H^l = \rho(AH^{l-1}W^l) \tag{3}$$

where H^{l-1} is the output vector of the previous layer and ρ is a non linearity (Leaky ReLU). H^0 is defined as the input logits $p_\psi(\mathbf{z_i})$. W^l are learnable weights for each layer. Using a GCN layer ensures that the logits are refined using information from only those nodes used for computing the logits, reducing the number of parameters learned while also taking advantage of the conditional probability estimates. After passing through multiple layers of the GCN, we get the updated predictions $f_\theta(p_\psi(\mathbf{z_i}))$. We add the initial logits to the updated logits to obtain our refined logits prediction $p_\psi(\mathbf{z_i}) + f_\theta(p_\psi(\mathbf{z_i}))$

3.3 Training

We train the image-text feature aggregation module p and the GCN f_θ, while freezing the VLM g_ϕ. We adopt the widely used Asymmetric Loss (ASL) [40], a modified version of the focal loss, to train our network for MLR.

ASL [40] addresses the inherent imbalance in MLR caused by the prevalence of negative examples compared to positive ones in training images. Similar to focal loss [28], ASL underweighs the loss term due to negative examples. However, it does so using two focusing parameters (γ_+ and γ_-) instead of one (γ) used by focal loss. However, ASL does not address the issue of sample imbalance, caused

by some classes having fewer examples in the dataset. Towards this, we add a loss re-weighting term (α) to ASL. Our re-weighed ASL (RASL) is defined as:

$$\mathcal{L}_{RASL}(\hat{y}_i^j) = \begin{cases} \alpha_j \left(1 - \hat{y}_i^j\right)^{\gamma_+} \log\left(\hat{y}_i^j\right) & \text{when } y_i^j = 1 \\ \alpha_j \left(\hat{y}_{i,\delta}^j\right)^{\gamma_-} \log\left(1 - \hat{y}_{i,\delta}^j\right) & \text{when } y_i^j = 0 \end{cases} \tag{4}$$

where $\hat{y}_i^j$ represents the corresponding prediction associated with label y_i^j; $\hat{y}_{i,\delta}^j = \max(\hat{y} - \delta, 0)$, with δ representing the shifting parameter defined in ASL; and α_j is the re-weighting parameter for class j, defined as:

$$\alpha = \left(\frac{a_{jj}}{\sum_{j=1}^{N} a_{jj}}\right)^{-1} \tag{5}$$

Then the total loss over the dataset $|\mathcal{D}|$ is given by:

$$\mathcal{L}_{total} = \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{N} \mathcal{L}_{RASL}(\hat{y}_i^j) \tag{6}$$

4 Experiments

In this section, we discuss the datasets used, implementation details of our approach, evaluation metrics and provide a thorough analysis of our approach.

4.1 Datasets

We evaluate our method on four different MLR benchmarks: MS-COCO 2014-small and PASCAL VOC 2007, which are widely used MLR benchmarks, as well as FoodSeg103 and UNIMIB 2016, which are smaller MLR datasets suitable for testing in the low data regime. Details of these datasets are given below:
MS-COCO 2014-small: MS-COCO [29] is another popular MLR dataset and consists of 82,081 training images and 40,504 validation images with objects belonging to 80 classes. To evaluate our methods performance in the low data regime, we use MS-COCO 2014-small, which is a small, randomly selected subset comprising 5% of MS-COCO 2014 which amounts to 4014 images. During testing, we use the complete validation set.
PASCAL VOC 2007: VOC [20] is a widely used outdoor scene MLR dataset consisting of 9,963 images from 20 classes. We follow the standard trainval set for training and use the test set for testing.
FoodSeg103: FoodSeg103 [44] serves as a benchmark dataset for food segmentation and multi-label food recognition. It consists of 4983 training images and 2135 test images, with a total of 32,097 food instances belonging to 103 different food classes. The number of images per class follows a long-tail distribution typical of real-world datasets. We use the standard train-test data split.

Dataset	Method	CP	CR	CF1	OP	OR	OF1	mAP
COCO-small [29]	DualCoOp[41]	53.3	73.5	59.8	47.1	79.5	59.2	70.2
	SCPNet[17]	51.9	70.3	59.7	47.2	78.9	59.1	69.3
	Ours	**54.1**	**74.3**	**62.6**	**47.7**	**82.6**	**60.5**	**72.6**
VOC [20]	SSGRL*[8]	-	-	-	-	-	-	93.4
	GCN-ML*[9]	-	-	-	-	-	-	94
	KGGR*[7]	-	-	-	-	-	-	93.6
	DualCoOp[41]	81.1	93.3	86.5	83.5	94.1	88.5	94.0
	SCPNet[17]	68.9	91.6	76.8	68.5	93.5	79.1	87.4
	Ours	**81.1**	**94.1**	**87.1**	**83.6**	**94.5**	**88.6**	**94.4**
FoodSeg103 [44]	DualCoOp[41]	44.9	52.7	46.9	59.2	69.2	63.8	49.0
	SCPNet[17]	39.4	54.4	43.2	61.4	67.8	64.4	48.8
	Ours w/o reweigh	44.8	55.0	48.0	58.6	70.2	63.9	51.3
	Ours	**47.1**	**55.1**	**50.8**	**63.7**	**69.9**	**66.7**	**52.9**
UNIMIB [10]	DualCoOp[41]	46.9	54.7	48.4	69.0	79.0	73.7	58.1
	SCPNet[17]	50.5	52.9	49.9	69.6	78.4	73.8	60.0
	Ours w/o reweigh	52.6	59.6	53.8	73.5	83.3	78.1	64.4
	Ours	**66.8**	**65.8**	**64.2**	**80.9**	**86.1**	**83.4**	**72.2**

Table 1: Comparison of results obtained by our method and the state-of-the-art baselines, on four MLR datasets in the low data regime: FoodSeg103, UNIMIB 2016, COCO-small (5% of COCO's training data) and VOC-2007. Our approach achieves the best performance on all metrics: per-class and overall average precisions (CP and OP), recalls (CR and OR), F1 scores (CF1 and OF1), and mean average precision (mAP). * indicates methods that fine-tune the complete backbone network.

UNIMIB 2016: UNIMIB [10] is another multi-label food recognition dataset. It consists of 1027 images with 3616 food instances spanning 73 classes. Similar to FoodSeg103, UNIMIB also follows a long-tail distribution typical of real-world datasets. We comply with the official train-test split.

4.2 Implementation Details

In our experiments, we use CLIP (Contrastive Language-Image Pre-Training) [39] as the VLM. Consistent with recent works that use VLMs for MLR [41,17,1], we select ResNet-101 as the visual encoder and standard transformer within CLIP as the text encoder. Both encoders are kept frozen during our experiments, and we train the GCN and learnable prompts. Following [41,9,17], we resize the images to 448×448 for COCO and VOC datasets and to 224×224 for UNIMIB and FoodSeg103. Similar to previous works [41,17] we apply Cutout [16] and RandAugment [12] to augment training images. We use a 3-layer GCN network for all our experiments. We use SGD for optimizing parameters with an initial learning rate of 0.002, which is reduced by cosine annealing. We train for 50 epochs and use a batch size of 32. We set the loss hyperparameters in Eq. 4 as

$\gamma_- = 3$, $\gamma_+ = 1$ and $\delta = 0.05$. We conduct all experiments on a single RTX A4000 GPU.

4.3 Evaluation Metrics

To evaluate the performance of our approach on the four MLR datasets, we use standard metrics, also used by previous MLR approaches [8,7,9]. The metrics include the commonly used mean average precision (mAP) as well as class and overall precisions (CP and OP), recalls (CR and OR), and F1 scores (CF1 and OF1). mAP is obtained by calculating the mean of individual average precision (AP) values over all classes. For each class, AP is computed as the area under the Precision-Recall curve.

4.4 Results

We compare our approach primarily with DualCoOp [41] and SCPNet [17] which use VLMs like our method, on all four standard benchmarks discussed earlier. As seen in Table 1, our method outperforms DualCoOp by 0.4% and SCPNet by 7.0% mAP on the VOC-2007. On COCO-small, our method outperforms DualCoOp by 2.4% and SCPNet by 3.3% mAP.

On the FoodSeg103 dataset, our approach significantly improves upon DualCoOp by 3.9% and SCPNet by 4.1% mAP. In the UNIMIB dataset, our method achieves substantial performance gains of 14.1% over DualCoOp and 12.2% mAP over SCPNet.

Furthermore, for VOC, we extend our comparison to approaches that do not use VLMs and instead rely on complete fine-tuning[8,7,9]. These approaches also use a ResNet-101 backbone similar to our visual encoder, but the backbone is initialized with weights pre-trained on ImageNet instead [15]. Our method also outperforms these methods. More detailed comparison can be found in Table. 1.

4.5 Impact of the Strength of Conditional Probability on Performance

In this section, we determine the impact of the strength of conditional probability of a pair of classes on MLR performance on the COCO-small dataset. Specifically, we observe how the improvement in average precision of a class of objects (ΔAP) brought by our method varies with the average conditional probability of the class paired with the top three other classes it co-occurs the most with. Note that we choose to average the top three values of conditional probabilities of the class because the COCO dataset typically contains an average of around three objects per image. The improvement in (ΔAP) for a class = AP achieved by logits after refinement - AP achieved by the raw VLM logits before refinement.

We visualize the variation in ΔAP with the avg. conditional probability in Figure 2, where we observe an increasing trend of ΔAP with increase in the average of the top-3 conditional probabilities for a given class. Note that for

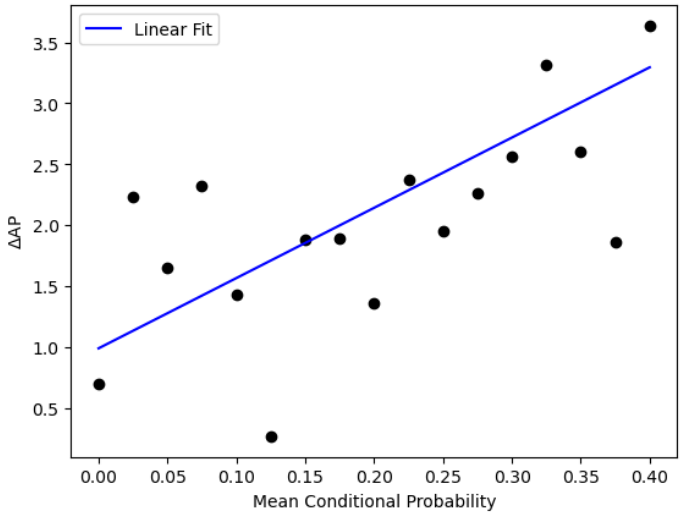

Fig. 2: Improvement in average precision (ΔAP) of a class obtained by refining VLM-based initial logits to incorporate the information provided by conditional probabilities, shown as a function of the mean conditional probability of most co-occurring three classes.

ease of visualization of the scatterplot, we use bins of size 0.02 to group together classes having similar average conditional probabilities. The points represent the average ΔAP value of all classes within the respective bin.

This implies that classes having stronger conditional probabilities with other classes benefit more from our approach of refining logits using conditional probabilities, as is intuitively expected.

4.6 Performance on Classes that are Difficult to Recognize

In this section, we empirically explore the impact of our approach on classes that are difficult to recognize when using image features exclusively. For concreteness, we focus our analysis on the 10 classes in FoodSeg103 and UNIMIB datasets on which the previous state-of-the-art approach (DualCoop[41]) performs (in terms of CF1) the worst.

Table 2 compares the performance of our method on these classes with the previous state-of-the-art DualCoOp[41]. We see that our method significantly improves the performance of DualCoOp[41], which relies solely on VLMs without modeling any conditional probabilities. Specifically, we observe a growth of 22.3% in CP, 33.9% in CR and 34.8% for CF1 on UNIMIB2016. On FoodSeg103, our method significantly improves CP, CR and CF1, by 15.6%, 7.2% and 11.89%, respectively. This underscores the importance of the information obtained by modeling joint class probabilities in recognizing classes of objects that are difficult to recognize from image features alone.

	UNIMIB			FoodSeg103		
	DualCoOp	Ours w/o reweigh	Ours	DualCoOp	Ours w/o reweigh	Ours
CP	25.4	41.9	**57.6**	13.7	14.8	**28.7**
CR	26.2	57.5	**60.0**	19.7	22.5	**26.9**
CF1	24.3	44.9	**59.1**	16.5	18.7	**28.4**

Table 2: A comparison of the average performance of our approach with the previous state-of-the-art VLM-based method DualCoOp[41] on classes that are difficult to recognize using only visual features (having 10 lowest CF1 values on the FoodSeg103[44] and UNIMIB[10]). Our approach significantly improves MLR performance on such classes due to its use of information derived from class conditional probabilities.

4.7 Effect of Loss Reweighing

In this subsection, we investigate the effects of our loss reweighing strategy for UNIMIB2016 and FoodSeg103, which exhibit significant class imbalance. We analyze its impact on performance on (1) All classes as a whole and (2) Classes that are difficult to recognize using only visual features, and hence have a greater reliance on information obtained from conditional probabilities.

(1) **All classes as a whole:** As observed in Table 1, loss re-weighing improves the performance of our method by 1.6% and 7.8% in mAP on FoodSeg103 and UNIMIB2016, respectively.

This demonstrates the utility of using loss re-weighing when training a method modeling the joint probability distribution of classes as opposed.

(2) **Classes that are difficult to recognize using only visual features:** As seen in Table 2, loss re-weighing also significantly boosts the performance of our method on these classes. It improves CP, CR and CF1 by 13.9%, 4.4% and 9.7%, respectively on the FoodSeg103 dataset and by 15.7%, 4.5% and 14.2%, respectively on the UNIMIB2016 dataset. Note that these increases are higher than corresponding gains observed for all classes in the dataset, signifying that loss re-weighing delivers larger benefits to classes more reliant on conditional probabilities for recognition (as opposed to those that derive more information from initial logits estimated from logits)

5 Conclusions

In this paper, we present a novel two-stage framework for multi-label recognition when only a small number of annotated images are available. Our framework builds on recent methods that make use of VLMs to counter this paucity of labeled data but overlook information derived from co-occurence of object pairs. Our framework refines the logit predictions made by VLMs adapted for multi-label recognition by leveraging known conditional probabilities of class pairs derived from the training data distribution. Specifically, we use a graph convolutional network to enrich the logits predicted by the VLM with information from

conditional probabilities of classes. Our method outperforms all state-of-the-art approaches on 4 MLR benchmarks: COCO-14-small, VOC 2007, FoodSeg103 and UNIMIB2016 in a low data regime, demonstrating the utility of modeling class co-occurrence in such cases.

6 Limitations

(1) If the independent classifiers learned by state-of-the-art approaches (relying on only visual information, not modeling the conditional probability of class pairs) are strong, and characterized by a high average precision (AP) of each class, our method would yield lower improvements. However, in practice, many MLR datasets are not very large, with independent classifiers learned from them being relatively weak. MLR on such datasets is likely to benefit significantly from our method.
(2) As shown in Figure. 2, the advantage provided by our method is higher when the conditional probability of pairs of classes co-occurring in an image is higher. For images that consist of objects which are rarely found together, our method provides very little added benefit over independent classifiers.

7 Acknowledgement

The support of the Office of Naval Research under grant N00014-20-1-2444, of USDA National Institute of Food and Agriculture under grant 2020-67021-32799/1024178 and Vizzhy.com are gratefully acknowledged.

References

1. Abdelfattah, R., Guo, Q., Li, X., Wang, X., Wang, S.: Cdul: Clip-driven unsupervised learning for multi-label image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1348–1357 (2023)
2. Anthimopoulos, M.M., Gianola, L., Scarnato, L., Diem, P., Mougiakakou, S.G.: A food recognition system for diabetic patients based on an optimized bag-of-features model. IEEE journal of biomedical and health informatics **18**(4), 1261–1271 (2014)
3. Bhatnagar, S., Ahuja, N.: Piecewise-linear manifolds for deep metric learning. In: Conference on Parsimony and Learning. pp. 269–281. PMLR (2024)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
5. Chang, W.C., Jiang, D., Yu, H.F., Teo, C.H., Zhang, J., Zhong, K., Kolluri, K., Hu, Q., Shandilya, N., Ievgrafov, V., et al.: Extreme multi-label learning for semantic matching in product search. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 2643–2651 (2021)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

7. Chen, T., Lin, L., Chen, R., Hui, X., Wu, H.: Knowledge-guided multi-label few-shot learning for general image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1371–1384 (2020)
8. Chen, T., Xu, M., Hui, X., Wu, H., Lin, L.: Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 522–531 (2019)
9. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5177–5186 (2019)
10. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments, and results. IEEE journal of biomedical and health informatics **21**(3), 588–598 (2016)
11. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 933–942 (2021)
12. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
13. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
14. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
16. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
17. Ding, Z., Wang, A., Chen, H., Zhang, Q., Liu, P., Bao, Y., Yan, W., Han, J.: Exploring structured semantic prior for multi label recognition with incomplete labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3398–3407 (2023)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
19. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Computational intelligence **20**(1), 18–36 (2004)
20. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
21. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
22. Huang, H., Rawlekar, S., Chopra, S., Deniz, C.M.: Radiology reports improve visual representations learned from radiographs. In: Medical Imaging with Deep Learning. pp. 1385–1405. PMLR (2024)

23. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`, if you use this software, please cite it as below.
24. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
25. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: International Conference on Learning Representations (2020)
26. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. arXiv preprint arXiv:2310.09291 (2023)
27. Khan, S., Hayat, M., Zamir, S.W., Shen, J., Shao, L.: Striking the right balance with uncertainty. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 103–112 (2019)
28. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
30. Liu, F., Xiang, T., Hospedales, T.M., Yang, W., Sun, C.: Semantic regularisation for recurrent image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2872–2880 (2017)
31. Liu, W., Tsang, I.: On the optimality of classifier chain for multi-label classification. Advances in Neural Information Processing Systems **28** (2015)
32. Liu, W., Wang, H., Shen, X., Tsang, I.W.: The emerging trends of multi-label learning. IEEE transactions on pattern analysis and machine intelligence **44**(11), 7955–7974 (2021)
33. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539–550 (2008)
34. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
35. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
36. Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.P.: Im2calories: towards an automated mobile vision food diary. In: Proceedings of the IEEE international conference on computer vision. pp. 1233–1241 (2015)
37. Misra, I., Lawrence Zitnick, C., Mitchell, M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2930–2939 (2016)
38. Park, S., Lim, J., Jeon, Y., Choi, J.Y.: Influence-balanced loss for imbalanced visual classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 735–744 (2021)

39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
40. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)
41. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems **35**, 30569–30582 (2022)
42. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294 (2016)
43. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7959–7971 (2022)
44. Wu, X., Fu, X., Liu, Y., Lim, E.P., Hoi, S.C., Sun, Q.: A large-scale benchmark for food image segmentation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 506–515 (2021)
45. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
46. Yang, J., Price, B., Cohen, S., Yang, M.H.: Context driven scene parsing with attention to rare classes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3294–3301 (2014)
47. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. AI Open **5**, 30–38 (2024)
48. Yazici, V.O., Gonzalez-Garcia, A., Ramisa, A., Twardowski, B., Weijer, J.v.d.: Orderless recurrent models for multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13440–13449 (2020)
49. Zhang, H., Li, F., Ahuja, N.: Open-nerf: Towards open vocabulary nerf decomposition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3456–3465 (2024)
50. Zhang, H., Li, F., Qi, L., Yang, M.H., Ahuja, N.: Csl: Class-agnostic structure-constrained learning for segmentation including the unseen. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7078–7086 (2024)
51. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European conference on computer vision. pp. 493–510. Springer (2022)
52. Zhang, Z., Pfister, T.: Learning fast sample re-weighting without reward data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 725–734 (2021)
53. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)