Improving Multi-label Recognition using Class Co-Occurrence Probabilities

Samyak Rawlekar[*1], Shubhang Bhatnagar[*1], VP Srinivasulu[2], Narendra Ahuja[1]

University of Illinois Urbana-Champaign [1]
Vizzhy.com [2]

Problem Definition

Multi-Label Recognition vs Single Label Recognition

Cat Person Dog

A. Multi-Label Recognition (MLR)

Image contains multiple objects
We assign a present/absent label to each class in the image

Dog

Horse

Car

B. Single-Label Recognition (SLR)

Image contains only one object
We assign one label to each image

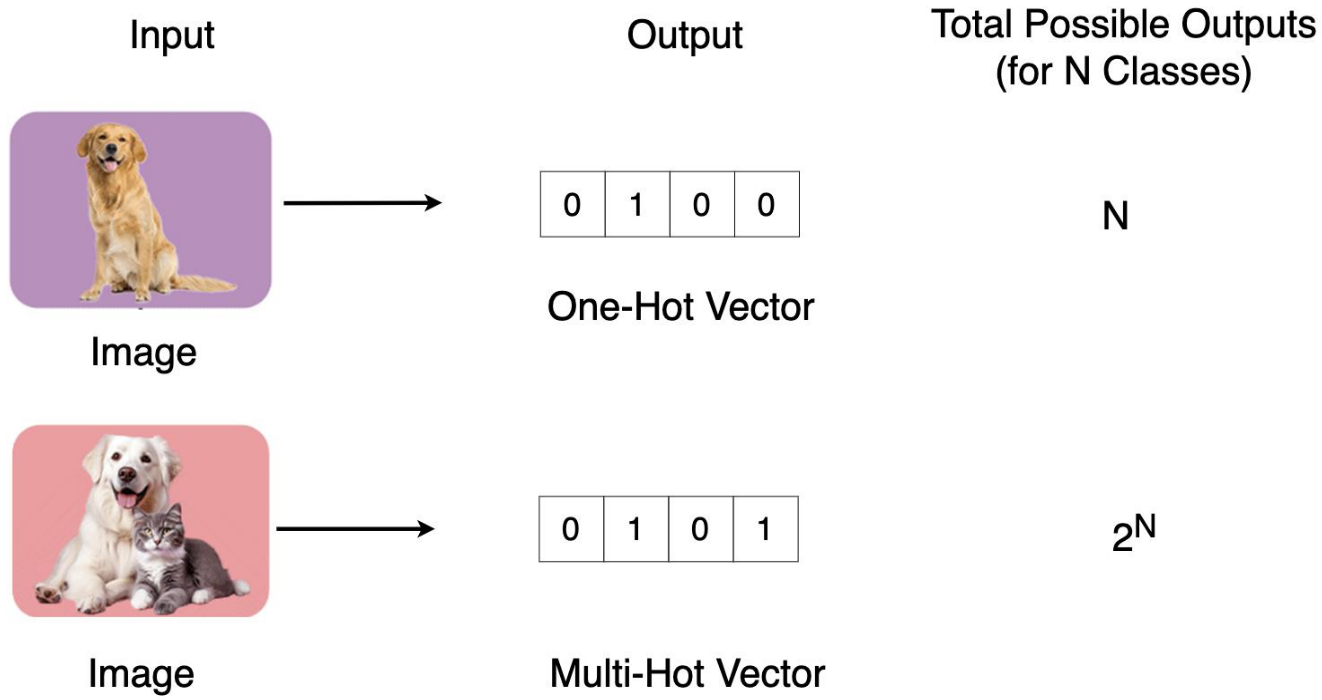
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Challenges of MLR

1. **Expensive Annotation:** Exhaustive annotations needed for each image (N labels vs 1 label)

2. **More Training Data Needed :** Much Larger output space -> Needs much more data to train

3. **Class imbalance:** Some object classes occur more frequently than others in real-world datasets

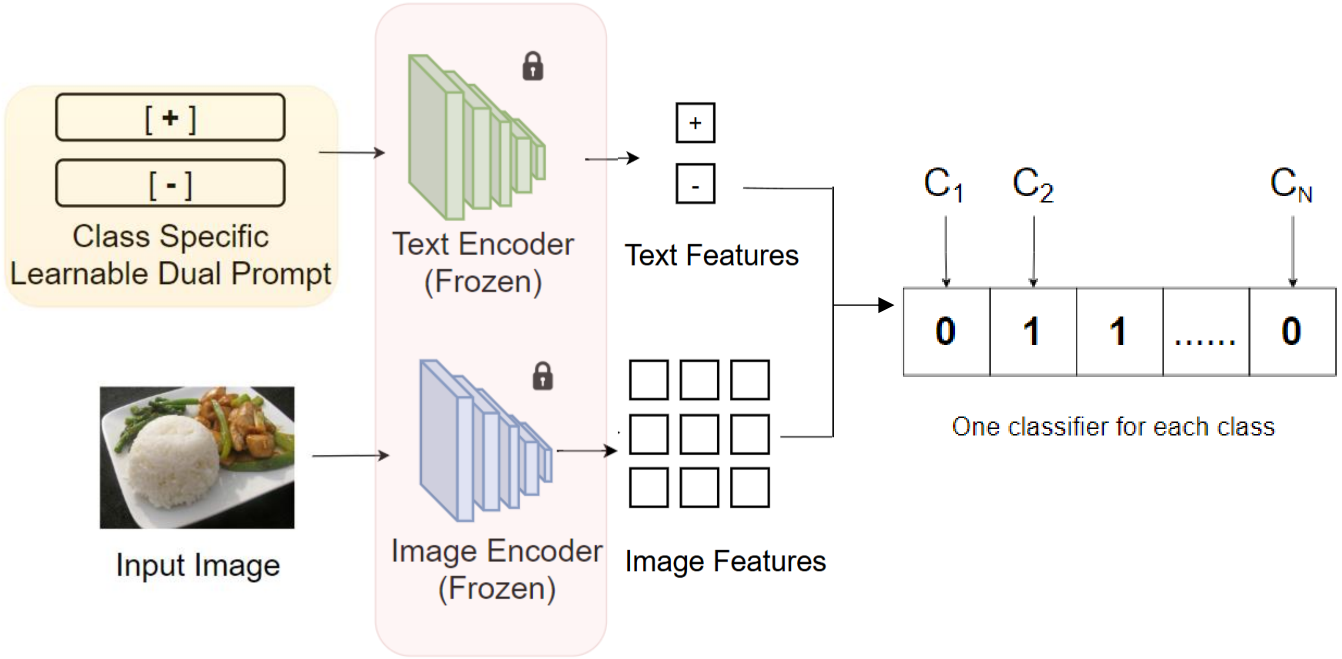


Recent Work in MLR

Vision-Language Models for MLR

To deal with challenges, recent work proposes:

- Adapt information from pretrained vision language models (e.g. CLIP [1]).
- Keep VLM frozen to preserve feature extraction priors
- Using extracted features, learn an independent classifier for each class to detect it's presence /absence
- Classifiers can be in the form of learnable positive/negative text prompts to make use of text priors [2]



[1] Radford et al " Learning transferable visual models from natural language supervision." *ICML* (2021)
[2] Sun et al "Dualcoop: Fast adaptation to multi-label recognition with limited annotations." *NIPS* (2022)

Limitations of Recent MLR Methods

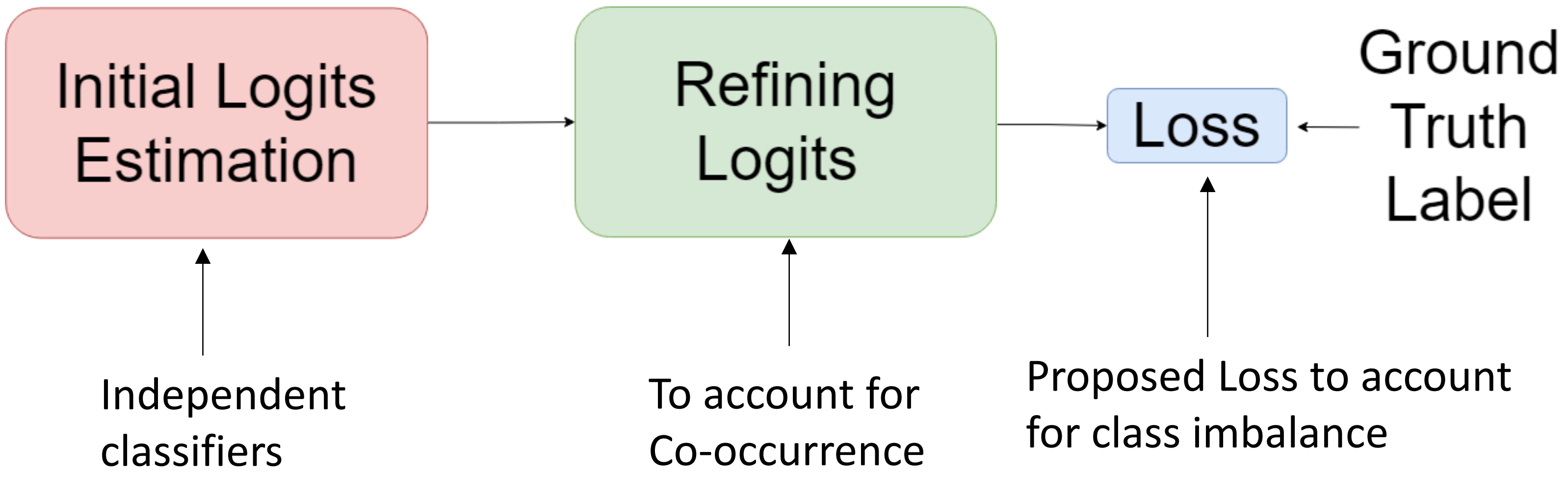
Recent works mitigate the relative paucity of annotations by using VLMs, however they still are limited by:

1. **No Co-occurrence Modeling**
   - **Learn Independent Classifiers**
     - **Ignores occurrence between objects (Crucial in limited data settings)**

2. **Don't Account for Class Imbalance**

Recent methods do not address class imbalance in real world MLR datasets

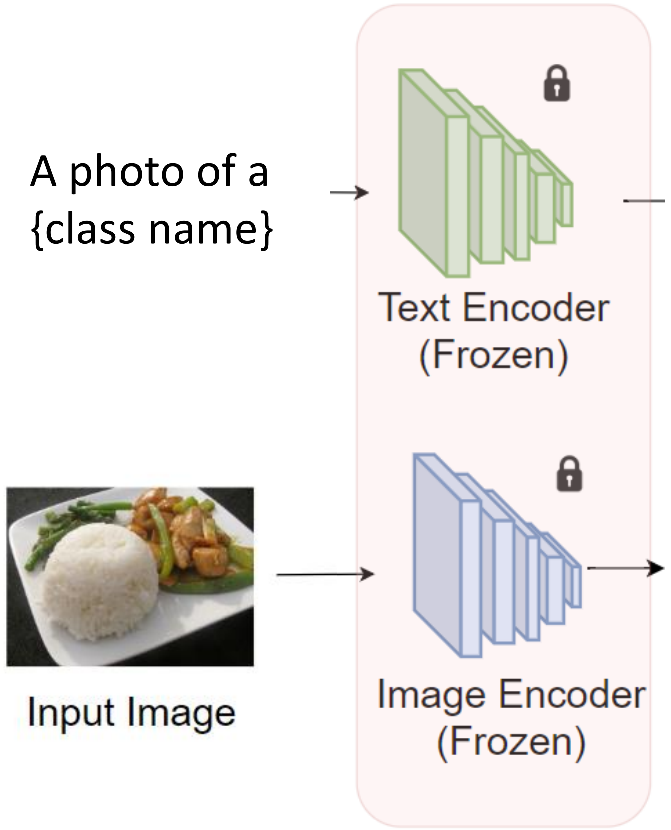
We propose a two-step method:



Method : Initial Logits Estimation

Key Components:

a. CLIP encoders



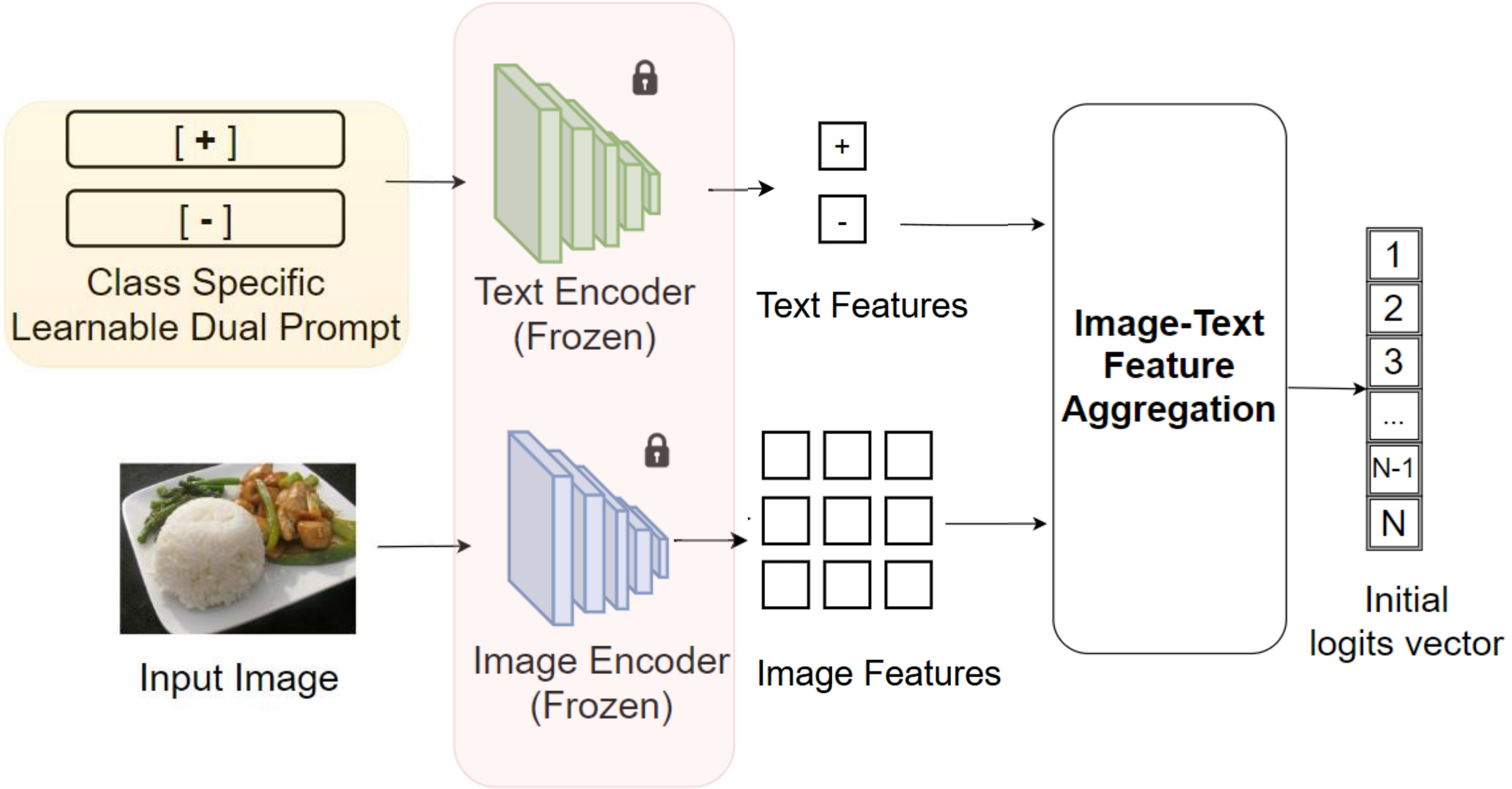
Method : Initial Logits Estimation

Key Components:

a. CLIP encoders

b. Learnable Prompts

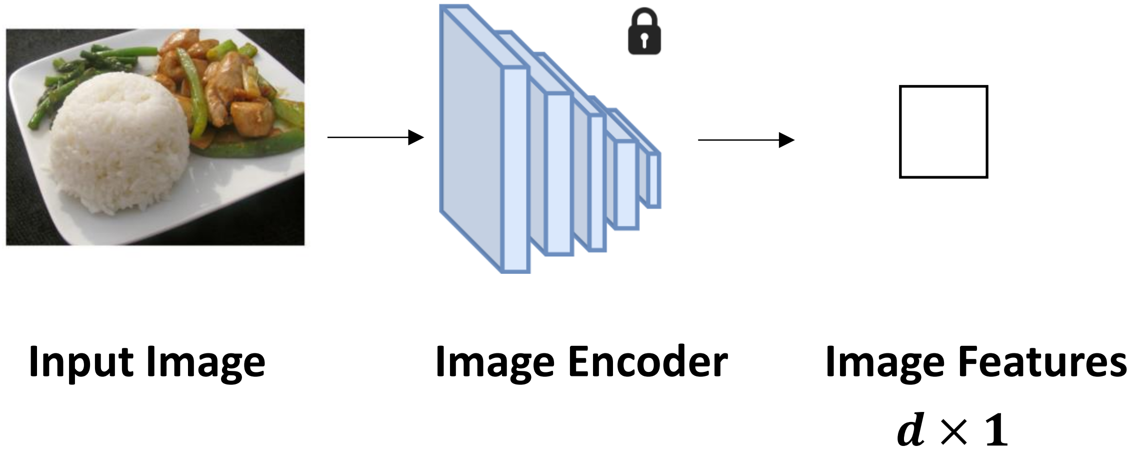
c. Image-Text Feature Aggregation



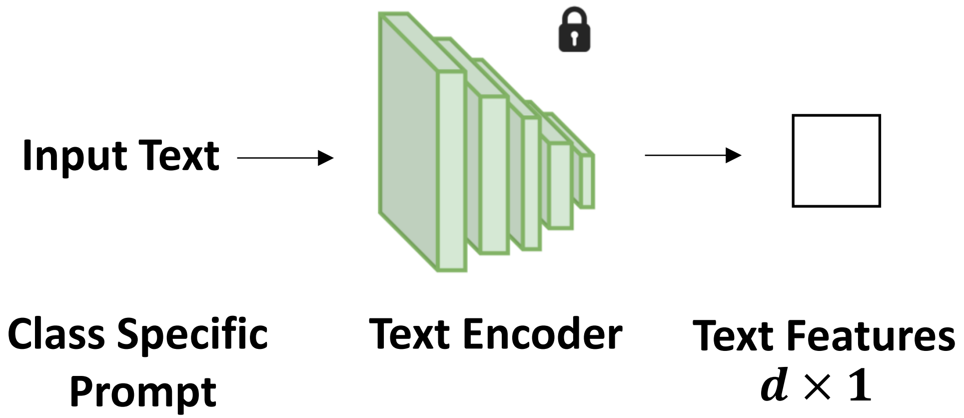
Method

a. CLIP Encoders

Image Encoder



Text Encoder



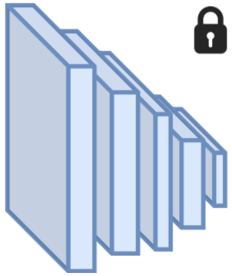
Objects appear in different locations in an image and hence it is crucial to look at features of subimages

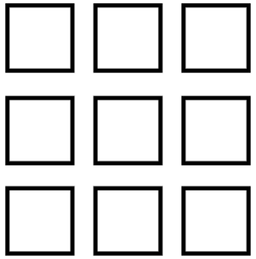
Pooling subimage features mixes the features of multiple objects within an image, which can result in suppression of certain individual object features.

Method

a. CLIP Encoders

Image Encoder





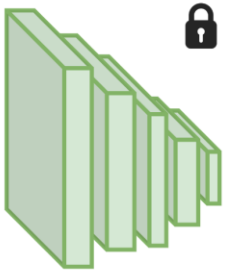
Input Image

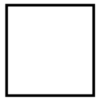
Image Encoder

Image Features
$d \times H \times W$

Text Encoder

Input Text





Class Specific Prompt

Text Encoder

Text Features
$d \times 1$

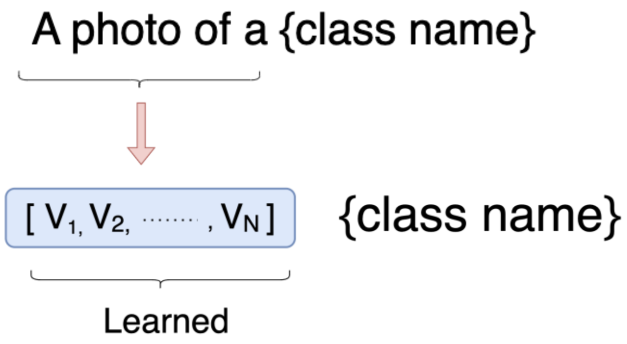
For Image Encoder: Remove the pooling layer and use subimage features.

Method

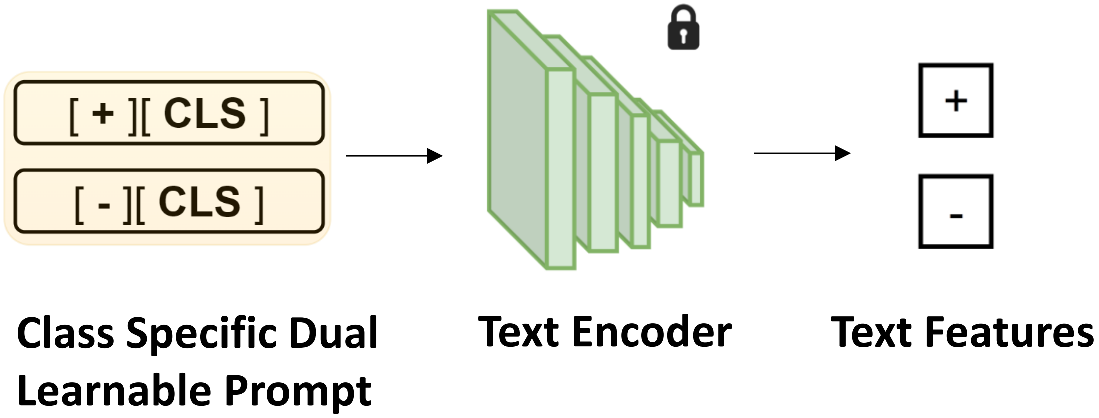
b. Learnable Prompts

Prompt Learning [3]:

- VLMs need an images and texts, we have the image and class names
- We create prompts (text):
  class names ⟶ "A photo of a {class name}"



Prompt Learning for MLR [1]

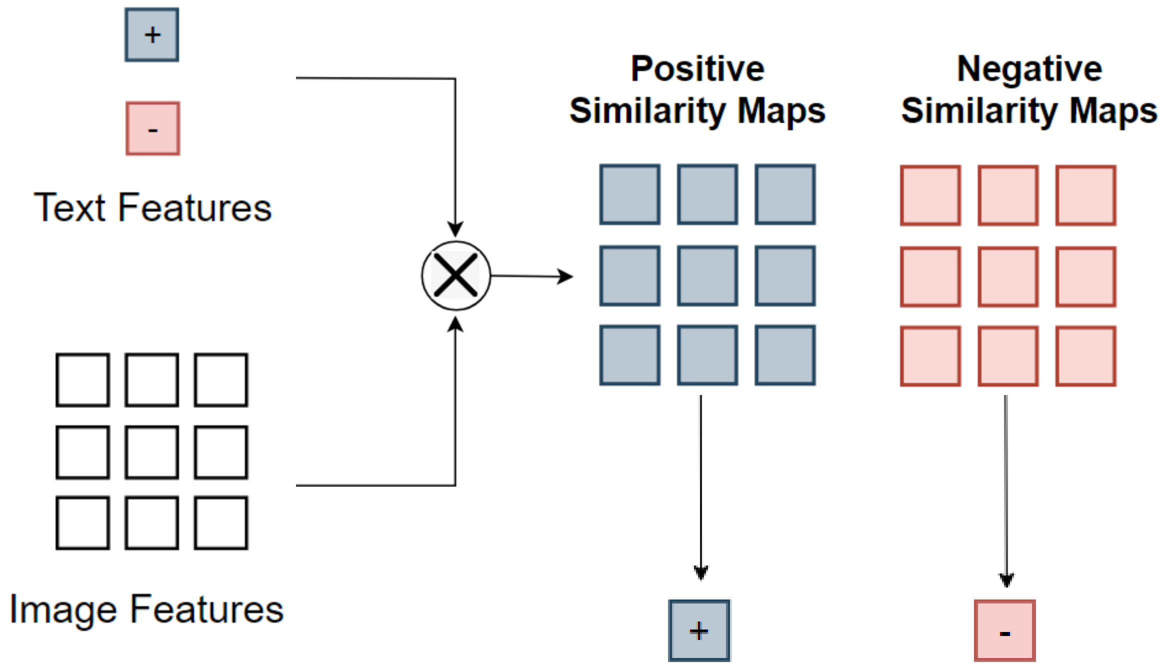


Key Point: We learn two prompts per class: one to detect presence of the class, another to detect its absence

[1] Sun et al. "Dualcoop: Fast adaptation to multi-label recognition with limited annotations." *NIPS* (2022)
[3] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models.", IJCV 2022

Method

c. Image-Text Feature Aggregation



- Obtain the spatial similarity map by the dot product of spatial image and text features
- Aggregate along the spatial regions to obtain initial positive and negative scores
- Compare the positive and negative scores
  The one with higher score is the winner!

× **Product**

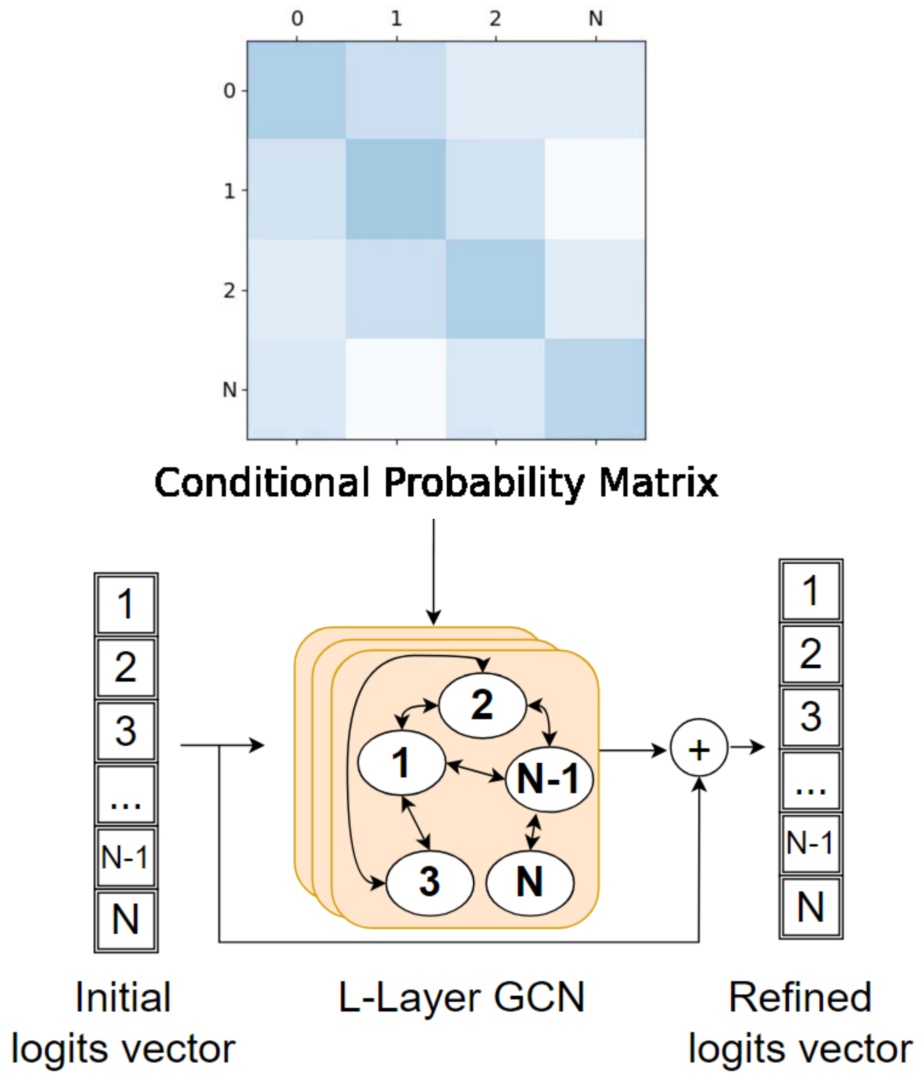
[1] Sun et al. "Dualcoop: Fast adaptation to multi-label recognition with limited annotations." *NIPS* (2022)

Method : Logits Refinement

Key Components:

a. Conditional Probability Matrix (Information)

b. Graph Convolution Network (GCN) (Enforcer)



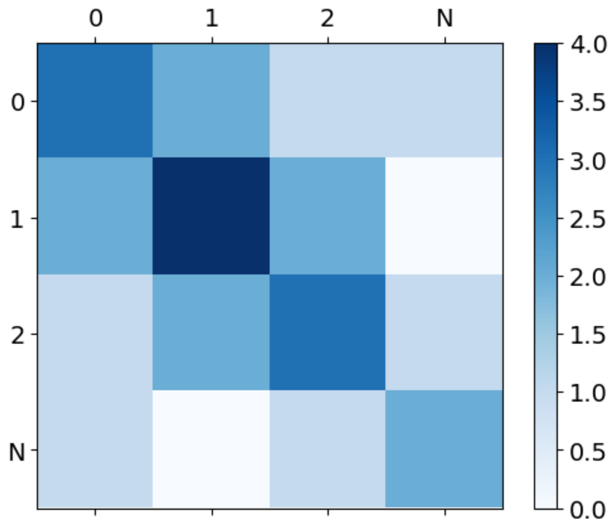
Logits Refinement

Method

a. Conditional Probability Matrix

Co-occurrence Matrix

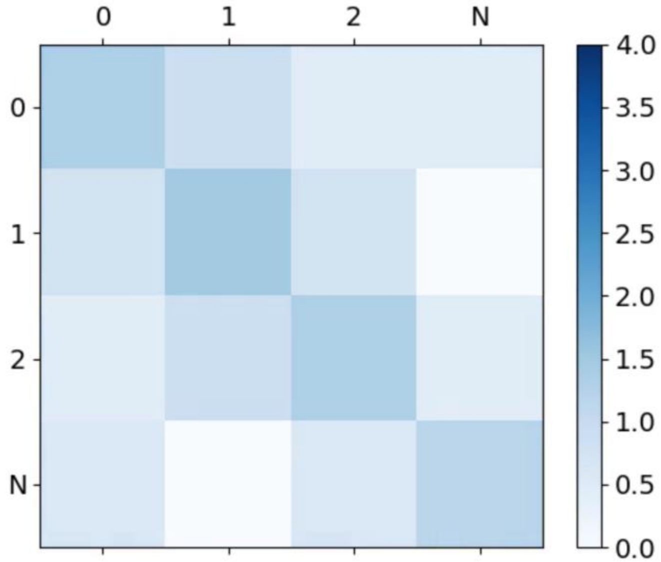
Count pairwise label co-occurrence in the training dataset.



Co-occurrence Matrix

Conditional Probability Matrix

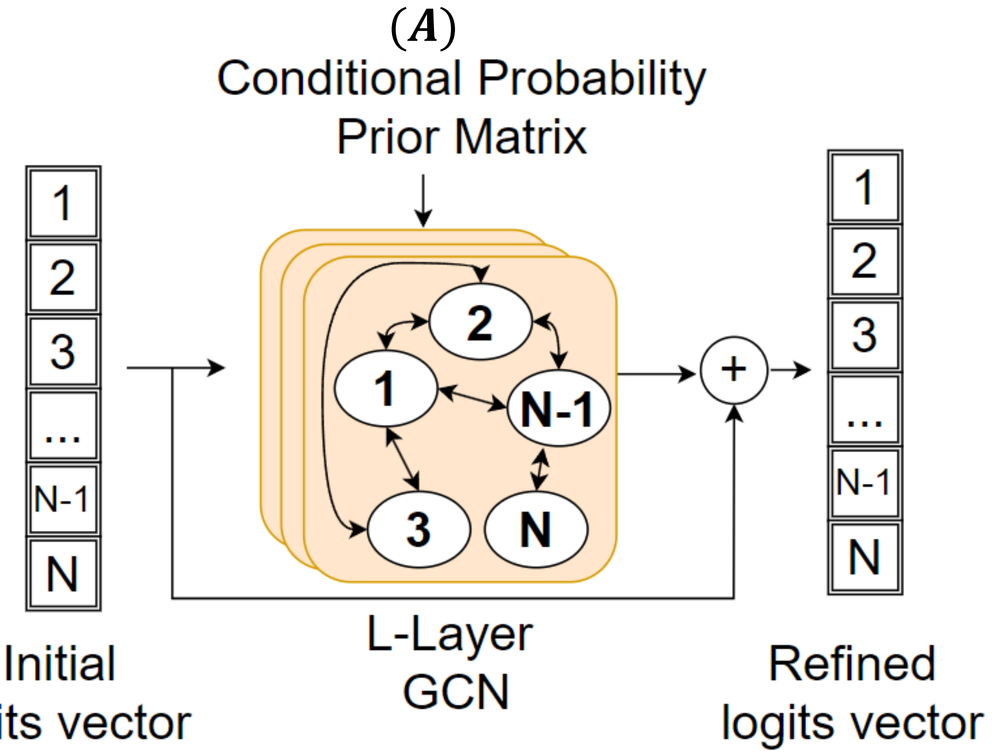
Normalize each row of the co-occurrence matrix by its diagonal entry.



Conditional Probability Matrix

Method

b. Graph Convolution Network



Logits Refinement

Conditional Probability Matrix $(\boldsymbol{A})$ represents the connection weights of the graph which is used to refine the logits.

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

H^{l-1} is the Input to layer l

W^l is the weights for layer l

ρ is the non-linearity

Key Point: We refine logits using a GCN that enforces co-occurrence

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

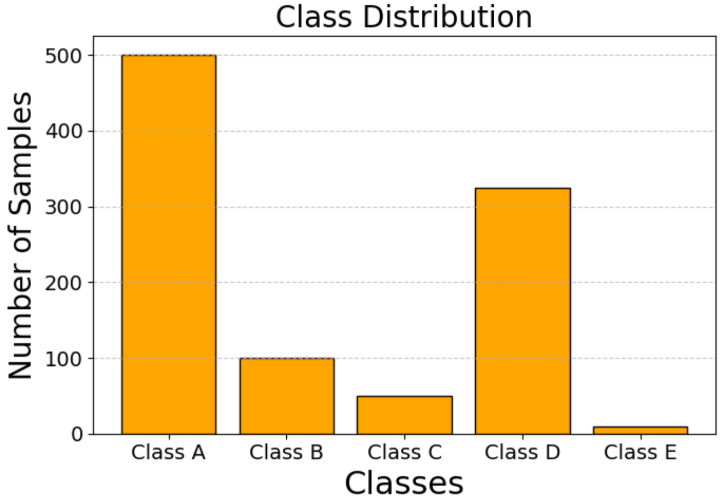
Training : Tackling Imbalance (RASL)

Imbalance in MLR:

a. Image level Imbalance

- 3 positive labels (person, dog, bench)
- 77 Negative Labels

b. Dataset level Imbalance



- Class imbalance in the dataset

We use ASL for image level imbalance, but for imbalance in the whole dataset we:

$$L_{RASL} = -\frac{1}{N}\sum_{i=1}^{D}\sum_{j=1}^{N}(\alpha_j)\cdot\left[\,(y_i^j)\cdot\left(1-p_i^j\right)^{\gamma^+}\cdot\log\left(p_i^j\right)\;+\;(1-y_i^j)\cdot\left(p_i^j\right)^{\gamma^-}\cdot\log\left(1-p_i^j\right)\right]$$

$$\alpha_j = \frac{\sum_{j=1} a_{jj}}{a_{jj}}$$

Results

Tested MLR performance on

- MS-COCO 2014 – small: 4k images (sampled 5% of the total data)
- PASCAL VOC 2007: 4k images
- FoodSeg103: 5k images
- UNIMIB-2016: 700 images

Using the standard MLR metrics

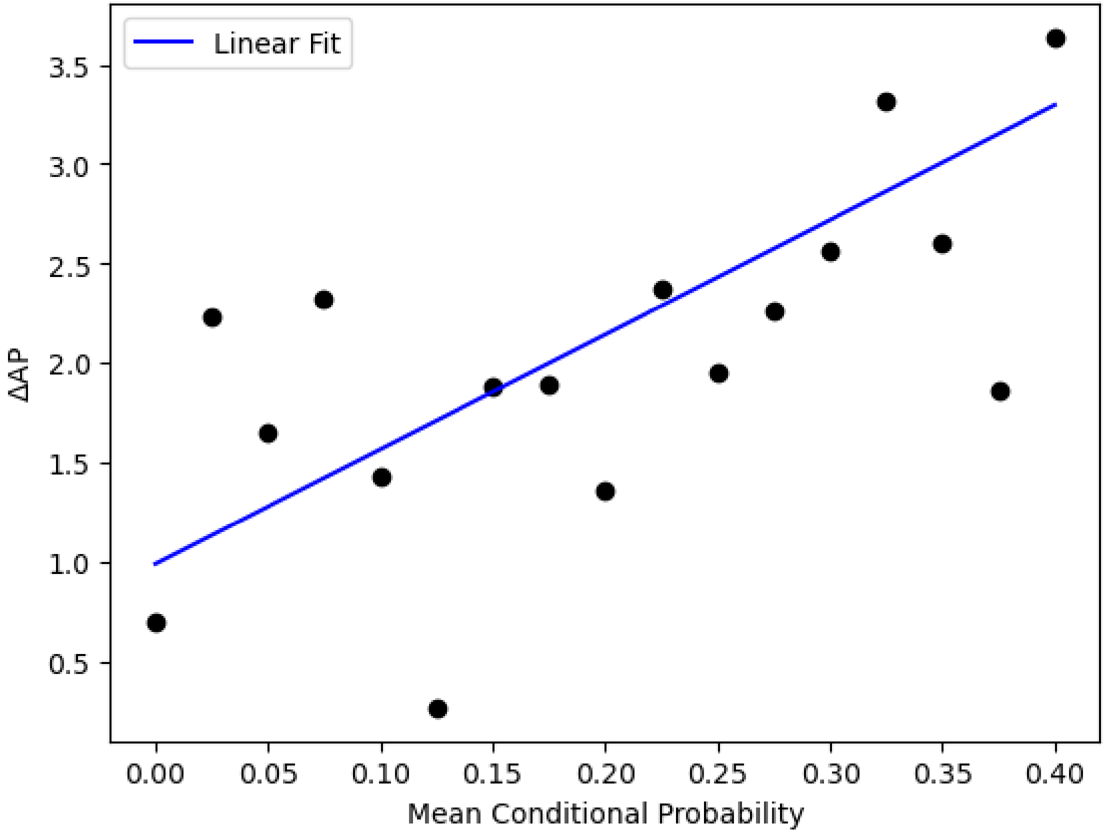
- Precision
- Recall
- F1 - score
- Mean Average Precision (mAP)

Results: Comparison with SOTA

Dataset	Method	CP	CR	CF1	mAP
COCO-small	DualCoOp	53.3	73.5	59.8	70.2
	SCPNet	51.9	70.3	59.7	69.3
	Ours	**54.1**	**74.3**	**62.6**	**72.6**
VOC	DualCoOp	81.1	93.3	86.5	94.0
	SCPNet	68.9	91.6	76.8	87.4
	Ours	**81.1**	**94.1**	**87.1**	**94.4**
FoodSeg103	DualCoOp	44.9	52.7	46.9	49.0
	SCPNet	39.4	54.4	43.2	48.8
	Ours w/o RASL	44.8	55.0	48.0	51.3
	Ours	**47.1**	**55.1**	**50.8**	**52.9**
UNIMIB	DualCoOp	46.9	54.7	48.4	58.1
	SCPNet	50.5	52.9	49.9	60.0
	Ours w/o RASL	52.6	59.6	53.8	64.4
	Ours	**66.8**	**65.8**	**64.2**	**72.2**

- **We outperform SOTA approaches across all metrics on four MLR datasets.**
- **Datasets in very low data regime and strong co-occurrence (FoodSeg103 and UNIMIB) benefit more from RASL.**

Results: Impact of Conditional Probability



- ΔAP is the change in AP value for a class before and after enforcing conditional probability.
- Mean conditional probability is the average of conditional probability of the top-3 classes that commonly occur with the chosen class.

As the strength of conditional probability (co-occurrence) increases, performance improves on the COCO dataset.

Results: Performance on Classes that are Difficult to Recognize using Visual Features

	UNIMIB			FoodSeg103		
Methods	CP	CR	CF1	CP	CR	CF1
DualCoOp	25.4	26.2	24.3	13.7	19.7	16.5
SCPNet	30.5	34.8	32.5	12.9	21.1	16.0
Ours w/o reweigh	41.9	57.5	44.9	14.8	22.5	18.7
Ours	**57.6**	**60.0**	**59.1**	**28.7**	**26.9**	**28.4**

Performance comparison of the 10 classes with the lowest F1 scores shows

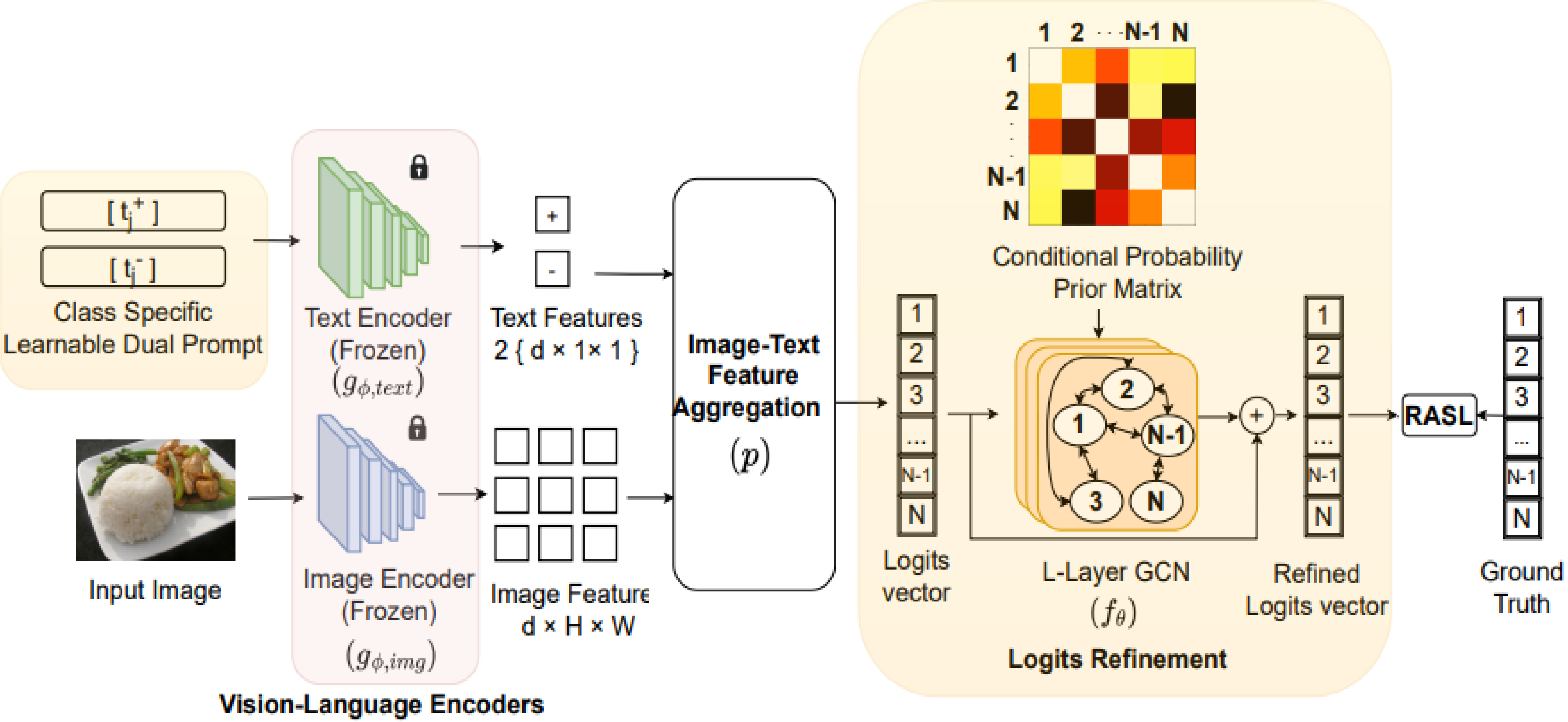
- Our approach significantly enhances MLR performance on these challenging classes by leveraging information from class conditional probabilities.

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Conclusion

- Previous methods overlook valuable co-occurrence information by detecting object labels independently
- We use CLIP for initial object logits and refine them with a graph convolution network (GCN) to enforce label correlations
- Re-weighted Asymmetric Loss (RASL) tackles imbalance
- Surpass all SOTA MLR methods on four benchmark datasets
- Limitations: Our method provides lesser benefit over independent classifiers when objects rarely co-occur (weaker co-occurrence)

Questions ?



Project Page