模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

General Multi-label Image Classification with Transformers

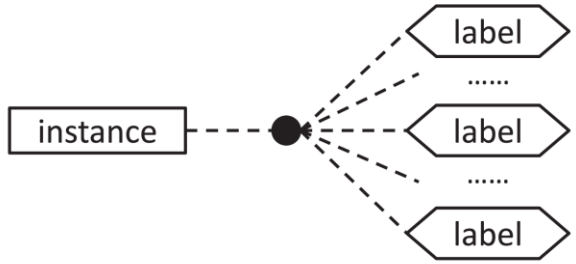
Jack Lanchantin, Tianlu Wang, Vicente Ordonez, Yanjun Qi
University of Virginia

CVPR 2021

Background

multi-class learning VS **multi-label learning**

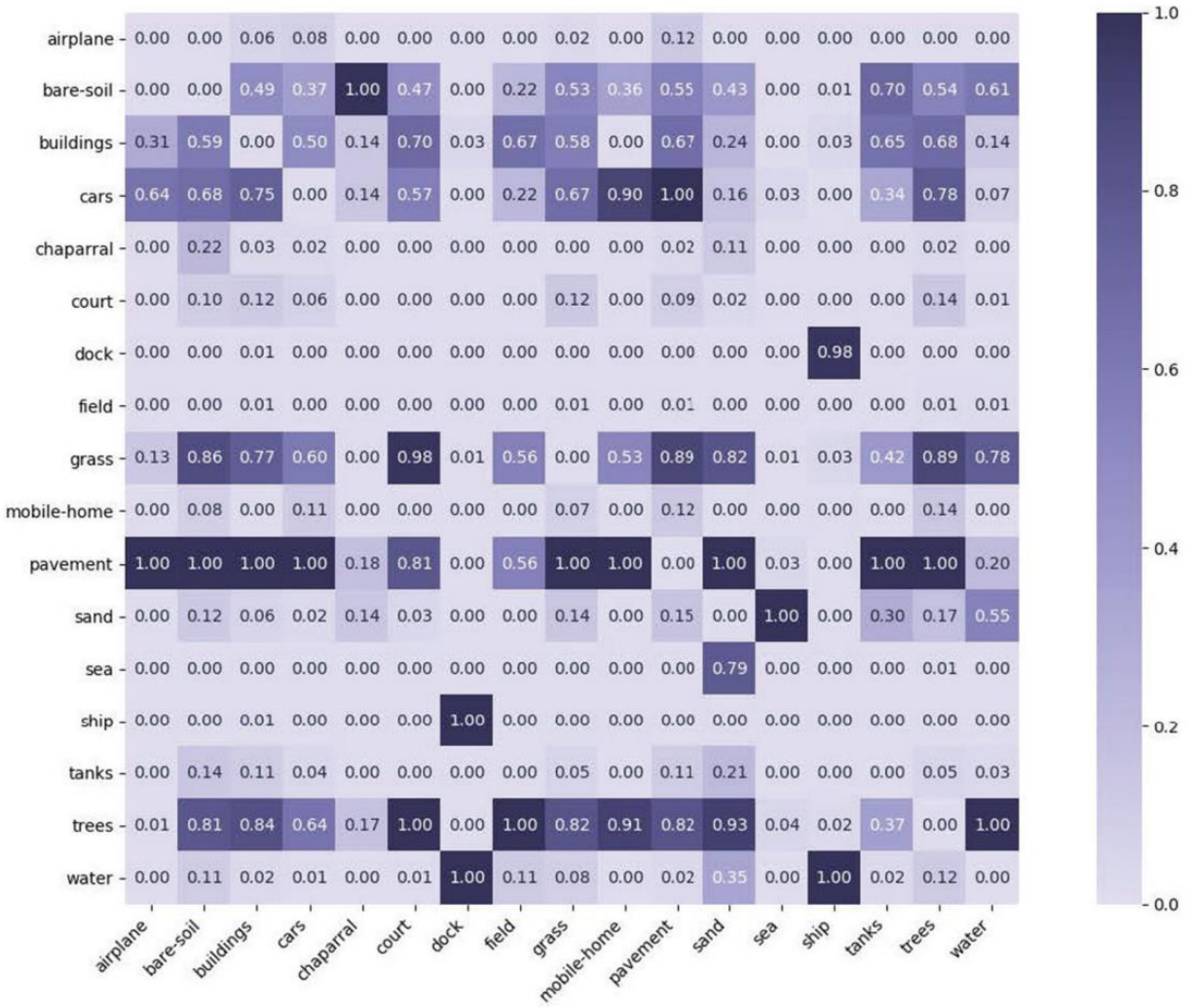
(a) traditional supervised learning



(b) single-instance multi-label learning

Motivation

label co-occurrences



How to tackle the complex dependencies among visual features and labels?

Contribution

Framework:

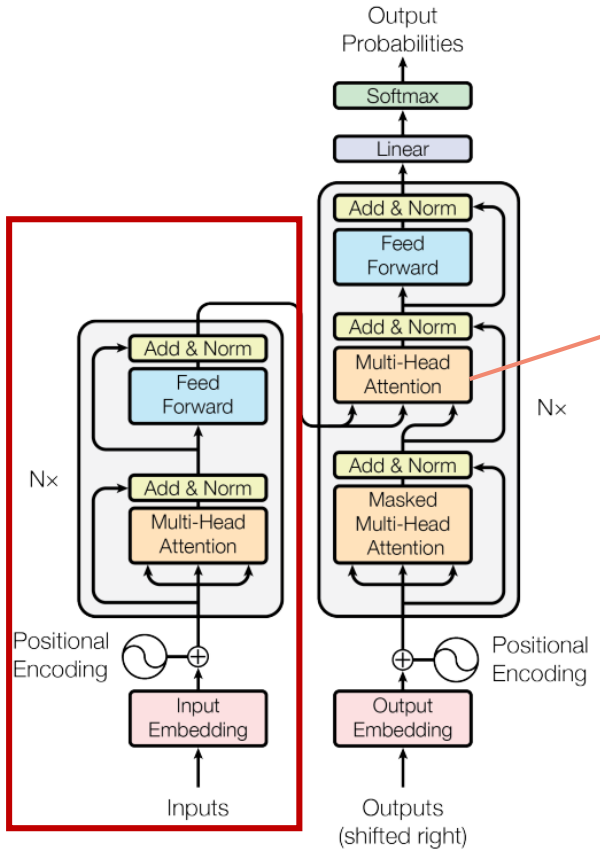
classification Transformer (C-Tran) — a general framework for multi-label image classification that leverages Transformers to exploit the complex dependencies among visual features and labels

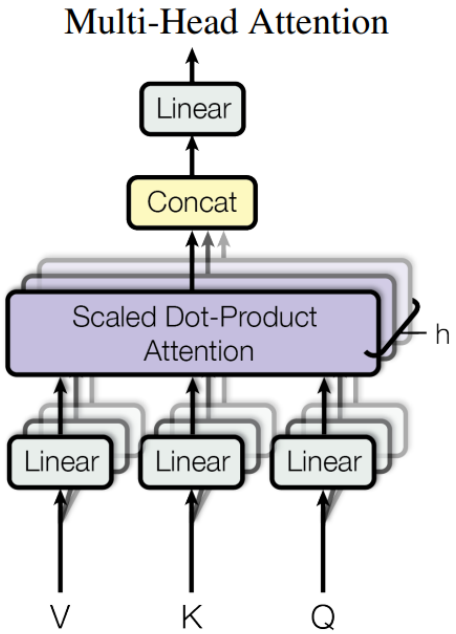
Strategy:

label mask training (LMT) — train the model with an input set of masked labels and visual features from a convolutional neural network

Preknowledge

Transformer



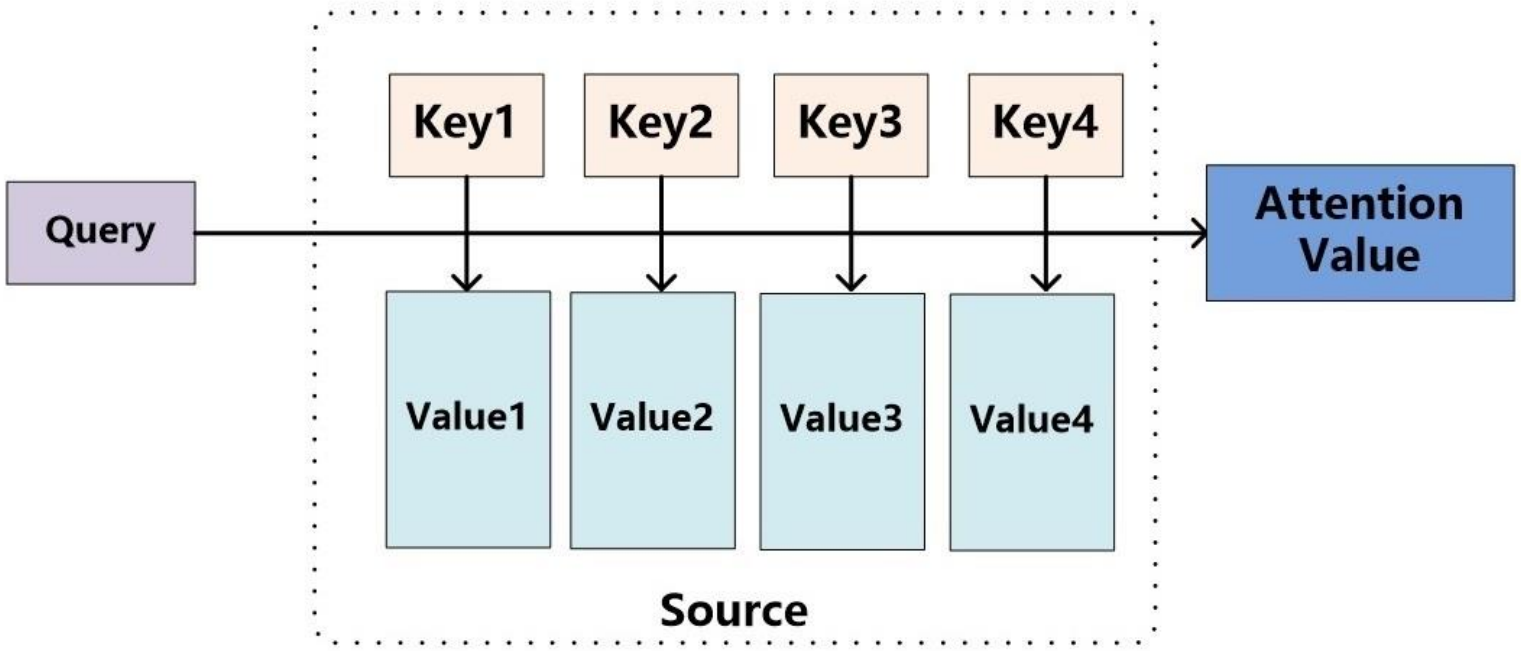


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

left: encoder right: decoder

Attention

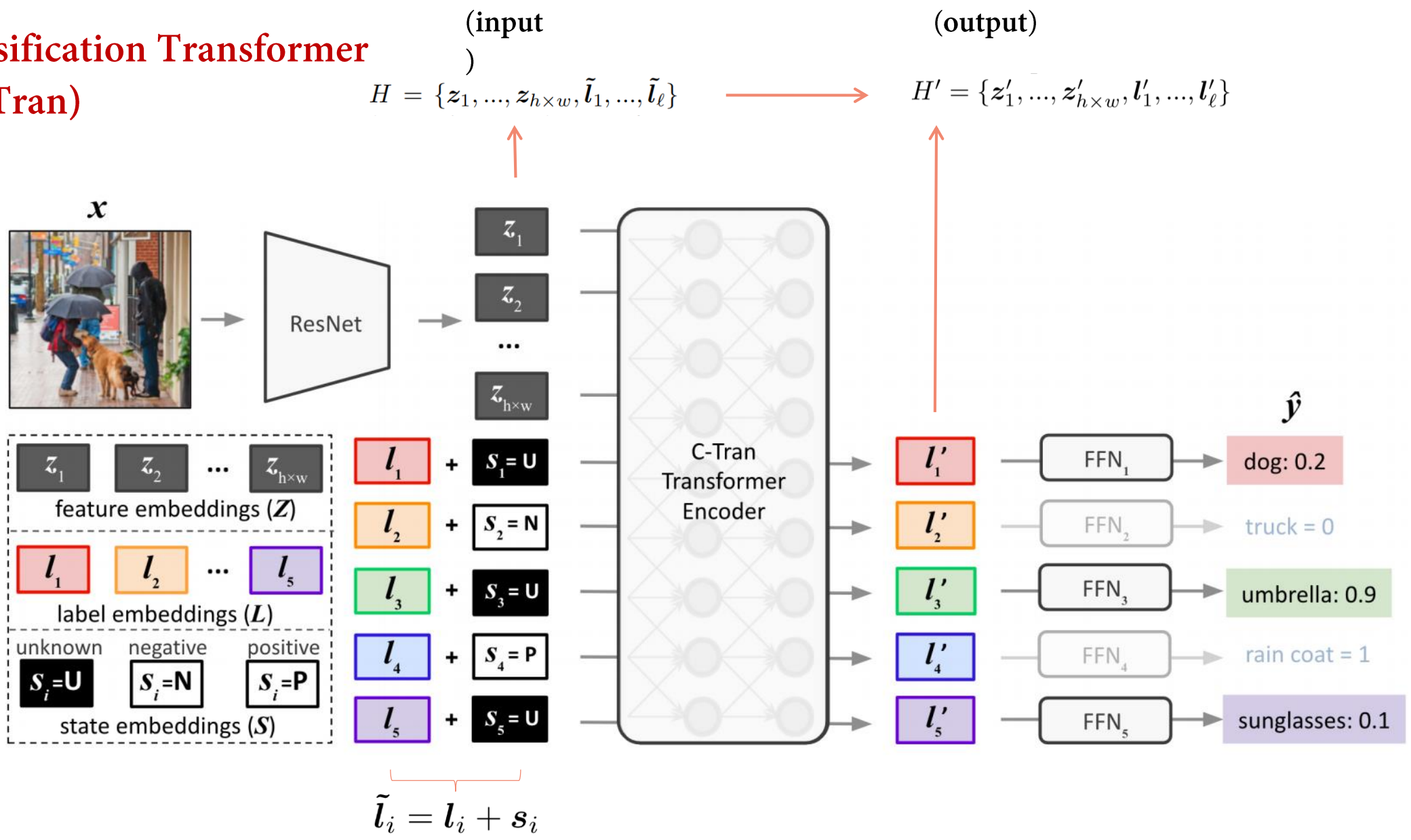
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



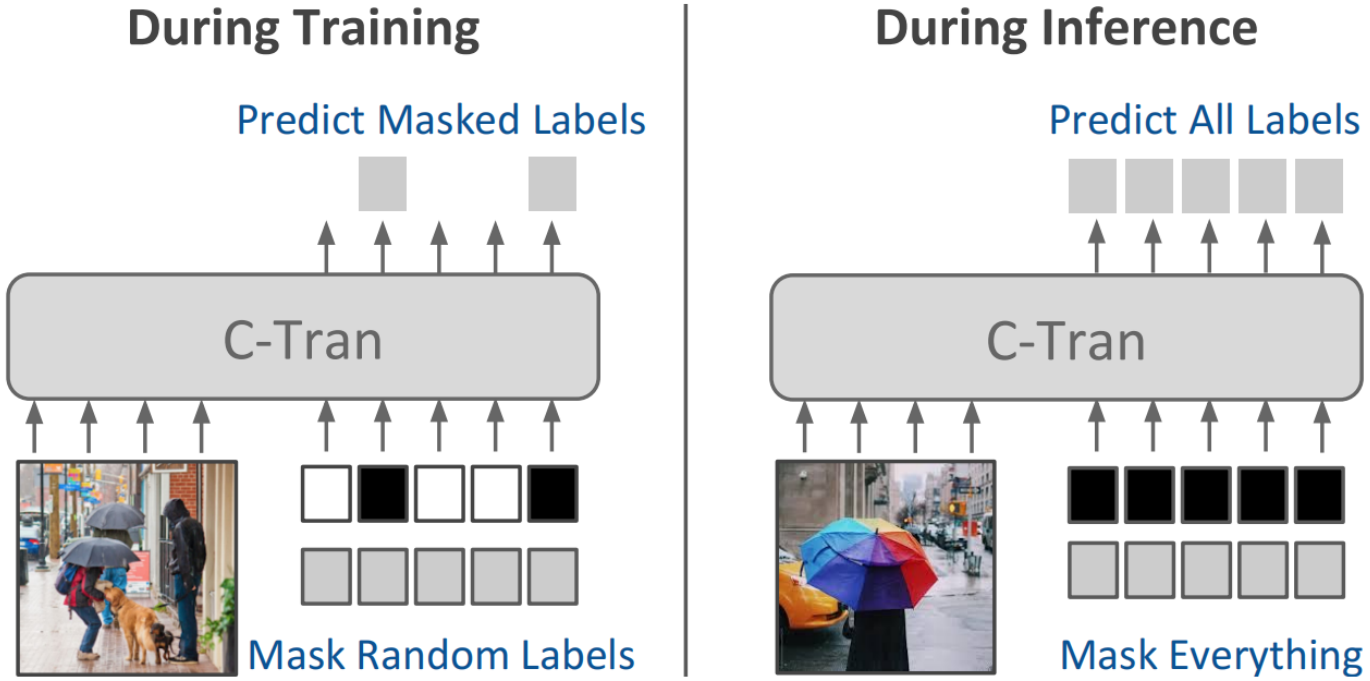
the Attention is the weighted sum of the Values
the weight of Value is the similarity between Query and Key

Method

classification Transformer (C-Tran)



label mask training (LMT)



the black is masked(unknown), the white is not(positive or negative).

Method

C-Tran

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

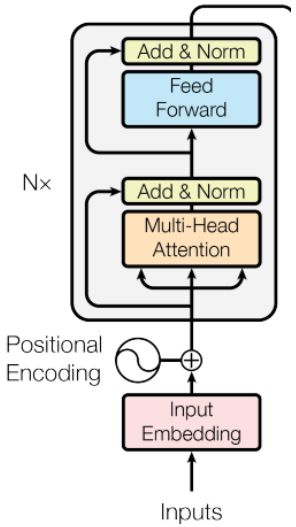
(input $H=\{\boldsymbol{z}_1,...,\boldsymbol{z}_{h\times w},\tilde{\boldsymbol{l}}_1,...,\tilde{\boldsymbol{l}}_\ell\}$)

$\boldsymbol{h}_i \in H \qquad \boldsymbol{h}_j \in H$

$$\alpha_{ij}=\text{softmax}\big((\mathbf{W}^q\boldsymbol{h}_i)^\top(\mathbf{W}^k\boldsymbol{h}_j)/\sqrt{d}\big),$$

$$\bar{\boldsymbol{h}}_i=\sum_{j=1}^{M}\alpha_{ij}\mathbf{W}^v\boldsymbol{h}_j,$$

$$\boldsymbol{h}'_i=\text{ReLU}(\bar{\boldsymbol{h}}_i\mathbf{W}^r+\boldsymbol{b}_1)\mathbf{W}^o+\boldsymbol{b}_2,$$



(output) $H'=\{\boldsymbol{z}'_1,...,\boldsymbol{z}'_{h\times w},\boldsymbol{l}'_1,...,\boldsymbol{l}'_\ell\}$

Method

C-Tran

$$\text{(output)}\ H' = \{z'_1, ..., z'_{h\times w}, l'_1, ..., l'_\ell\}$$

- z'_i : feature embedings
- l'_i : label embedings

$$\hat{y}_i = \text{FFN}_i(l'_i) = \sigma\left((\mathbf{w}_i^c \cdot l'_i) + b_i\right)$$

- σ(·) : simoid function
- FFN(·) : feedforward network

Loss function:

$$L = \sum_{n=1}^{N_{tr}} \mathbb{E}_{p(\mathbf{y}_k)}\{\text{CE}(\hat{\mathbf{y}}_u^{(n)}, \mathbf{y}_u^{(n)})|\mathbf{y}_k\}$$

- y_k : known labels
- y_u : unknown labels

Experiment

COCO-80:

	All							Top 3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [50]	61.2	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention [52]	-	-	-	-	-	-	-	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [6]	-	-	-	-	-	-	-	79.1	58.7	67.4	84.0	63.0	72.0
ML-ZSL [31]	-	-	-	-	-	-	-	74.1	64.5	69.0	-	-	-
SRN [58]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet101 [21]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
Multi-Evidence [19]	-	80.4	70.2	74.9	85.2	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
ML-GCN [10]	83.0	85.1	72.0	78.0	85.8	75.4	80.3	89.2	64.1	74.6	90.5	66.5	76.7
SSGRL [8]	83.8	**89.9**	68.5	76.8	**91.3**	70.8	79.7	**91.9**	62.5	72.7	**93.8**	64.1	76.2
KGGR [7]	84.3	85.6	72.7	78.6	87.1	75.6	80.9	89.4	64.6	75.0	91.3	66.6	77.0
C-Tran	**85.1**	86.3	**74.3**	**79.9**	87.7	**76.5**	**81.7**	90.1	**65.7**	**76.0**	92.1	**71.4**	**77.6**

Table 1. Results of *regular inference* on COCO-80 dataset. The threshold is set to 0.5 to compute precision, recall and F1 scores (%). Our method consistently outperforms previous methods across multiple metrics under the settings of all and top-3 predicted labels. Best results are shown in bold. "-" denotes that the metric was not reported.

Experiment

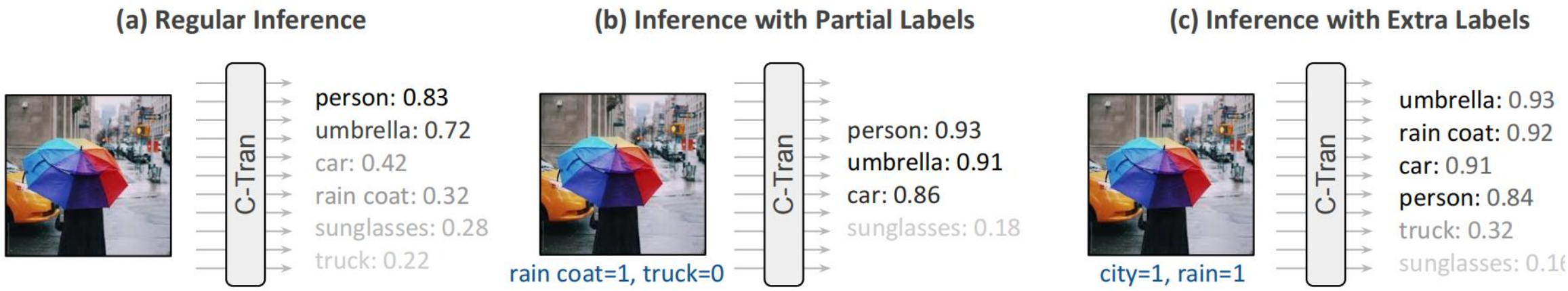
VG-500:

	All							Top 3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
ResNet101[21]	30.9	39.1	25.6	31.0	61.4	35.9	45.4	39.2	11.7	18.0	75.1	16.3	26.8
ML-GCN [10]	32.6	42.8	20.2	27.5	66.9	31.5	42.8	39.4	10.6	16.8	77.1	16.4	27.1
SSGRL [8]	36.6	-	-	-	-	-	-	-	-	-	-	-	-
KGGR [7]	37.4	47.4	24.7	32.5	**66.9**	36.5	47.2	48.7	12.1	19.4	78.6	17.1	28.1
C-Tran	**38.4**	**49.8**	**27.2**	**35.2**	**66.9**	**39.2**	**49.5**	**51.1**	**12.5**	**20.1**	**80.2**	**17.5**	**28.7**

Table 2. Results of *regular inference* on VG-500 dataset. All metrics and setups are the same as Table 1. Our method achieves notable improvement over previous methods.

Problem Setup

three multi-label image classification scenarios as follows:



Note that : All these settings are during testing, and we assume that all labels are available during training.

Inference with partial labels:

	COCO-80				VG-500				NEWS-500				COCO-1000			
Partial Labels Known (ϵ)	0%	25%	50%	75%	0%	25%	50%	75%	0%	25%	50%	75%	0%	25%	50%	75%
Feedbackprop [51]	80.1	80.6	80.8	80.9	29.6	30.1	30.8	31.6	14.7	21.1	23.7	25.9	29.2	30.1	31.5	33.0
C-Tran	**85.1**	**85.2**	**85.6**	**86.0**	**38.4**	**39.3**	**40.4**	**41.5**	**18.1**	**29.7**	**35.5**	**39.4**	**34.3**	**35.9**	**37.4**	**39.1**

Table 3. Results of *inference with partial labels* on four multi-label image classification datasets. Mean average precision score (%) is reported. Across four simulated settings where different amounts of partial labels are available (ϵ), our method significantly outperforms the competing method. With more partial labels available, we achieve larger improvement.

Inference with extra labels:

Extra Label Groups Known (ϵ)	0%	36%	54%	71%
Standard [25]	82.7	82.7	82.7	82.7
Multi-task [25]	**83.8**	83.8	83.8	83.8
ConceptBottleneck [25]	80.1	87.0	93.0	97.5
C-Tran	**83.8**	**90.0**	**97.0**	**98.0**

Table 4. Results of *inference with extra labels* on CUB-312 dataset. We report the accuracy score (%) for the 200 multi-class target labels. We achieve similar or greater accuracy than the baselines across all amounts of known extra label groups.

Ablation:

Partial Labels Known (ϵ)	COCO-80		VG-500		NEWS-500		COCO-1000	
	0%	50%	0%	50%	0%	50%	0%	50%
C-Tran (no image)	3.60	21.7	2.70	24.6	6.50	33.3	1.50	27.8
C-Tran (no LMT)	84.8	85.0	38.3	38.8	16.9	17.1	33.1	34.0
C-Tran	**85.1**	**85.6**	**38.4**	**40.4**	**18.1**	**35.5**	**34.3**	**37.4**

Table 5. C-Tran component ablation results. Mean average precision score (%) is reported. Our proposed Label Mask Training technique (LMT) improves the performance, especially when partial labels are available.

Experiment

LMT's ablation:

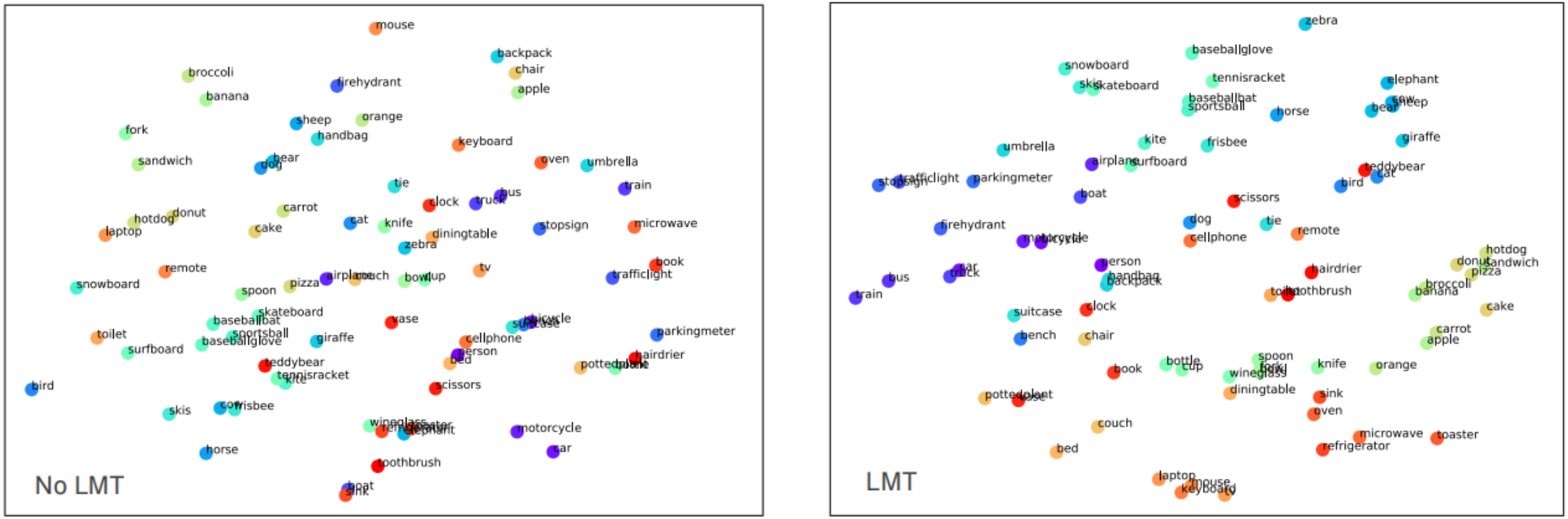


Figure 4. Comparison of the learned label embeddings for COCO-80 using t-SNE. The left figure shows the embedding projections without using label mask training (LMT), and the right shows with LMT. Labels are colored using the COCO object categorization. We can see that using label mask training produces much semantically stronger label representations.

Thanks