南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

General Multi-label Image Classification with Transformers

Jack Lanchantin, Tianlu Wang, Vicente Ordonez, Yanjun Qi
University of Virginia
{jjl5sw,tianlu,vicente,yq2h}@virginia.edu

CVPR 2021

Existing methods for Multi-Label Classification:

1.Binary Relevance

Train a classifier for each label, and then use all the classifiers to make predictions on the samples.

2.Classifier Chain

When predicting the current label each time, consider not only the feature data but also the previous label.

3.Label Powerset

Treat the label set of a sample as a single category label.

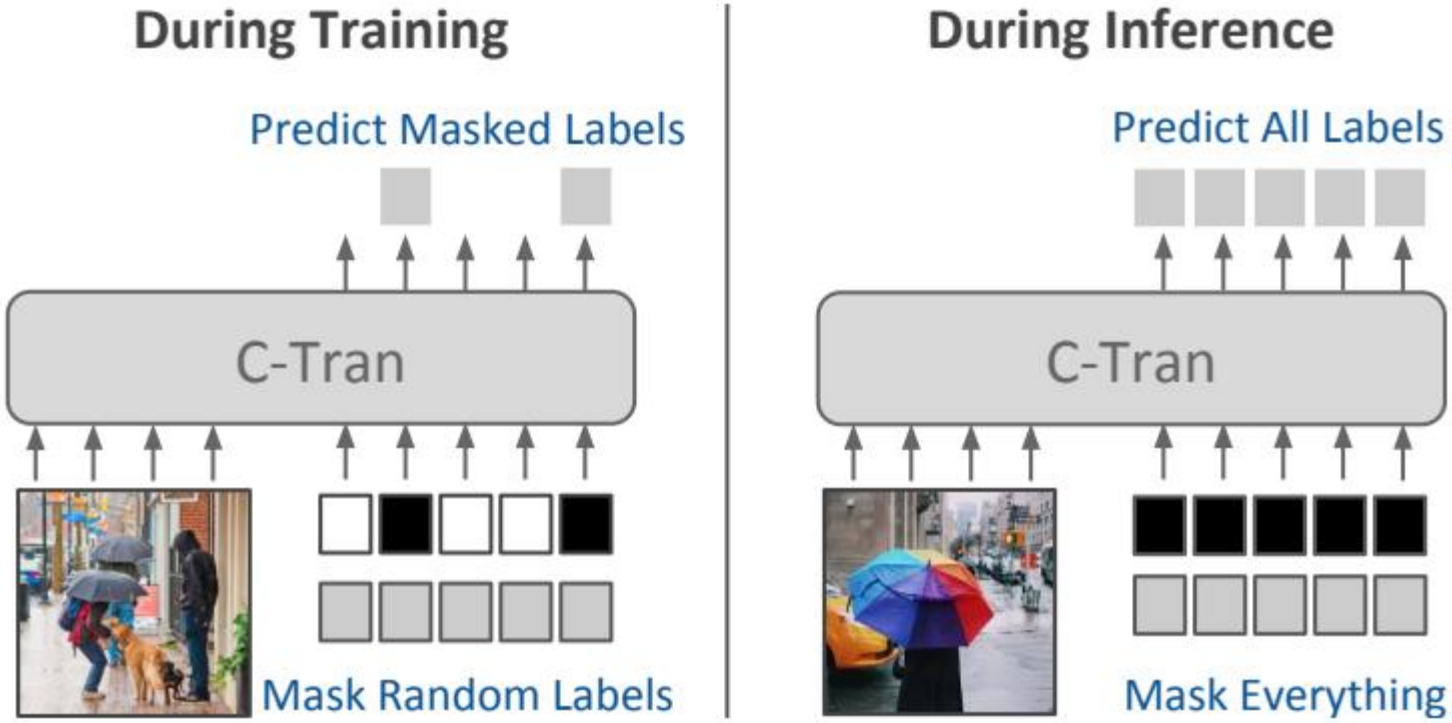
Motivation

In the multi-label setting, the output set of labels has some structure that reflects the structure of the world.
For example, surfboard is unlikely to co-occur with grass, while fork is more likely to appear next to a plate.

Method

C-Tran (Classification Transformer) :

A Transformer encoder is trained to reconstruct a set of target labels given an input set of masked label embeddings and a set of features obtained from a convolutional neural network. C-Tran uses a label mask training objective that allows us to represent the state of the labels as positive, negative, or unknown. At test time, C-Tran is able to predict a set of target labels using only input visual features by masking all the input labels as unknown.



C-Tran (Classification Transformer) :

C-Tran can also be used at test time with partial or extra label annotations by setting the state of some of the labels as either positive or negative instead of masking them out as unknown.

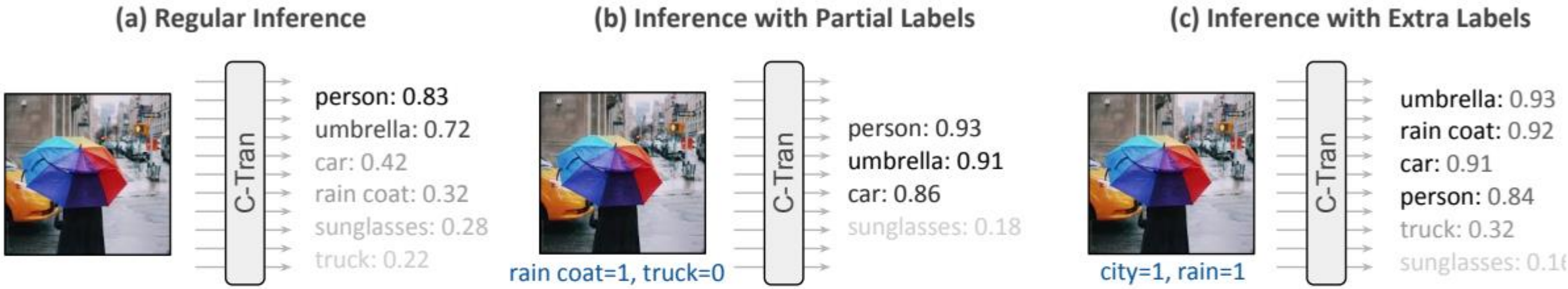


Figure 2. Different inference settings for general multi-label image classification: (a) Standard multi-label classification takes only image features as input. All labels are unknown $\mathbf{y}_u$.; (b) Classification under partial labels takes as input image features as well as a subset of the target labels that are known. The labels *rain coat* and *truck* are known labels $\mathbf{y}_k$, and all others are unknown labels $\mathbf{y}_u$; (c) Classification under extra labels takes as input image features and some related extra information. The labels *city* and *rain* are known extra labels $\mathbf{y}_k^e$, and all others are unknown target labels $\mathbf{y}_u^t$.

Method

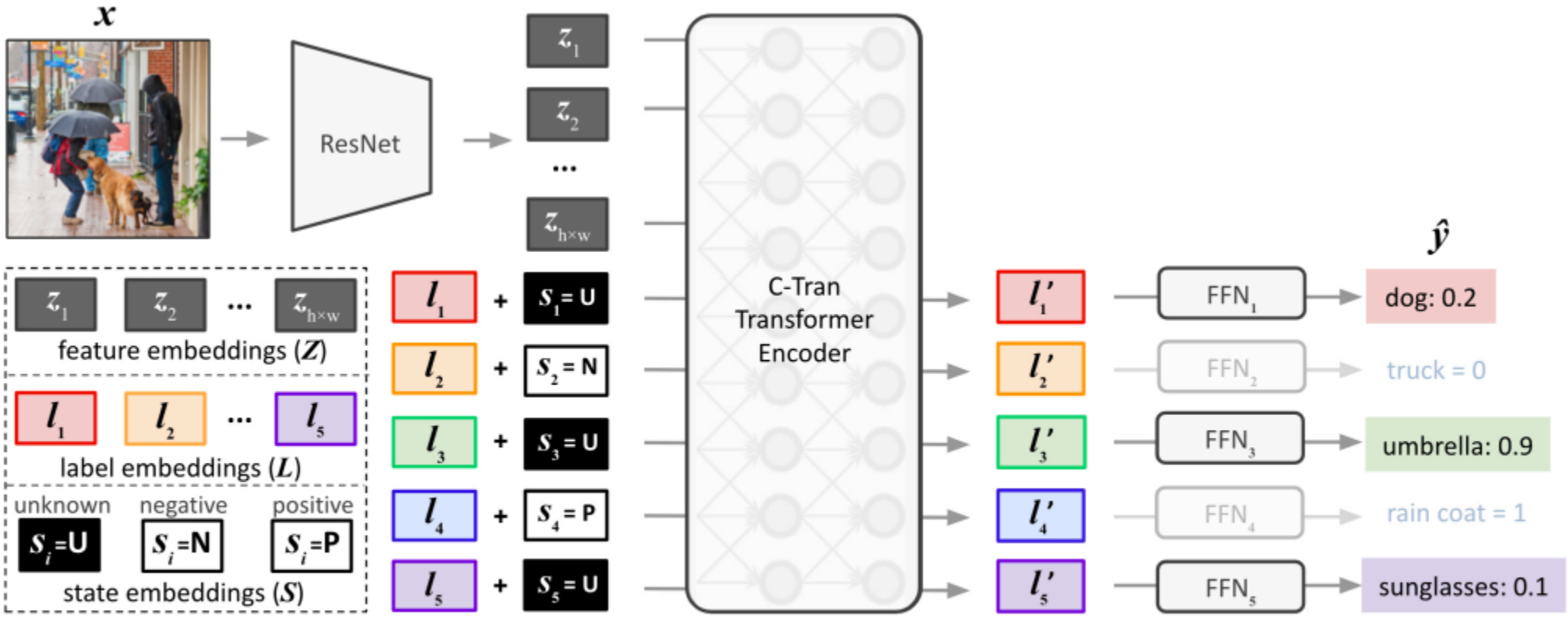


Figure 3. C-Tran architecture and illustration of label mask training for general multi-label image classification. In this training image, the labels *person*, *umbrella*, and *sunglasses* were randomly masked out and used as the unknown labels, $\mathbf{y}_u$. The labels *rain coat* and *truck* are used as the known labels, $\mathbf{y}_k$. Each unknown label is added the unknown state embedding U, and each known label is added its corresponding state embedding: negative (N) , or positive (P). The loss function is only computed on the unknown label predictions $\hat{\mathbf{y}}_u$.

Method

Image Feature Embeddings: $x \in \mathbb{R}^{H \times W \times 3} \longrightarrow z \in \mathbb{R}^{h \times w \times d}$

Label Embeddings L: $\boldsymbol{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, ..., \boldsymbol{l}_\ell\}, \boldsymbol{l}_i \in \mathbb{R}^d$

Label embeddings are learned from an embedding layer of size d × l

Adding Label Knowledge via State Embeddings S: In traditional architectures, there is no way to encode partially known or extra labels as input to the model. To address this drawback, we propose a technique to easily incorporate such information. Given label embedding li, we simply add a "state" embedding vector, si $\in \mathbb{R}^d$:

$$\tilde{\boldsymbol{l}}_i = \boldsymbol{l}_i + \boldsymbol{s}_i$$

where the si takes on one of three possible states: unknown (U), negative (N), or positive (P). The state embeddings are retrieved from a learned embedding layer of size d × 3, where the unknown state vector (U) is fixed with all zeros.

Method

Modeling Feature and Label Interactions with a Transformer Encoder

Let H = $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_{h\times w}, \tilde{\boldsymbol{l}}_1, ..., \tilde{\boldsymbol{l}}_\ell\}$ be the set of embeddings that are input to the Transformer encoder.

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

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

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

We denote the final output of the Transformer encoder after L layers as

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

Lastly, after feature and label dependencies are modeled via the Transformer encoder, a classifier makes the final label predictions. Feedforward network (FFNi)

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

Label Mask Training (LMT)

During training, we randomly mask a certain amount of labels, and use the ground truth of the other labels (via state embeddings) to predict the masked labels.
Given that there are ℓ possible labels, the number of "unknown" labels for a particular sample. n is chosen at random between 0.25 ℓ and ℓ .

Essentially, our label mask training pipeline tries to minimize the following loss approximately:

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

Regular Inference

	All							Top 3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [48]	61.2	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention [50]	-	-	-	-	-	-	-	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 [32]	-	-	-	-	-	-	-	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 [20]	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 [16]	-	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 [9]	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.

Regular Inference

	All							Top 3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
ResNet101[20]	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 [9]	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.

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 [49]	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.

Ablation Study

Partial Labels	COCO-80		VG-500		NEWS-500		COCO-1000	
Known (ϵ)	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.

Thanks