General Multi-label Image Classification with Transformers

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

A. Appendix

A.1. Qualitative Examples

Inference with Partial Labels. In Figure 2, we show qualitative results on COCO-80 demonstrating the use of partial labels. In these examples, we first show the predictions for ResNet-101, as well as C-Tran without using partial labels. The last column shows the C-Tran predictions when using $\epsilon = 25\%$ partial labels (which is 21 labels for COCO-80) as observed, or known prior to inference. For many examples, certain labels cannot be predicted well without using partial labels.

Inference with Extra Labels. In Figure 3, we show qualitative results on CUB-312 demonstrating the use of extra labels. In the CUB-312 dataset, the extra labels are high level concepts of bird species that are not target labels. In these examples, we first show the predictions for C-Tran without using extra labels labels, and the last column shows the C-Tran predictions when using $\epsilon = 54\%$ of the extra labels (which is 60 labels for CUB-312) as observed, or known prior to inference. We can see that many bird species predictions are completely changed after using the extra labels as input to our model.

A.2. Detailed Diagram of C-Tran Settings

Figure 1 shows a detailed diagram of all possible training and inference settings used in our paper, and how C-Tran is used in each setting. By using the same random mask training, we can apply our model to any of the three inference settings.

A.3. Multi-Label Classification Metrics

$$\begin{aligned} \mathrm{OP} &= \frac{\sum_i N_i^c}{\sum_i N_i^p} \\ \mathrm{OR} &= \frac{\sum_i N_i^c}{\sum_i N_i^g} \\ \mathrm{OF1} &= \frac{2 \times \mathrm{OP} \times \mathrm{OR}}{\mathrm{OP} + \mathrm{OR}} \end{aligned} \tag{1}$$

$$\begin{aligned} \mathrm{CP} &= \frac{1}{C} \sum_i \frac{N_i^c}{N_i^p} \\ \mathrm{CR} &= \frac{1}{C} \sum_i \frac{N_i^c}{N_i^g} \\ \mathrm{CF1} &= \frac{2 \times \mathrm{CP} \times \mathrm{CR}}{\mathrm{CP} + \mathrm{CR}} \end{aligned} \tag{2}$$

where C is the number of labels, N_i^c is true positives for the i-th label, N_i^p is the total number of images for which the i-th label is predicted, and N_i^g is the number of ground truth images for the i-th label.

A.4. More Discussions of C-Tran

Connecting to Transformers and BERT. Our proposed method, C-Tran, draws much inspiration from works in natural language processing. The transformer model [7] proposed "self attention" for natural language translation. Self attention allows each word in the target sentence to attend to all other words (both in the source sentence and the target sentence) for translation. [2] introduced BERT for language modeling. BERT uses self attention with masked words to pretrain a language model.

Self attention and BERT are both examples of complete graphs, but on sentences rather than image features and labels. C-Tran uses the same self-attention mechanisms as [7] and [2], but instead of using only the word embeddings from a sentence, we use feature and label embeddings.

In computer vision, [1] used Transformers for object detection. Our method varies in several distinct ways. First, we are primarily interested in using partial evidence for image classification, and our unique state embeddings allow C-Tran to use such evidence. Second, we model image and label features jointly in a Transformer encoder, whereas [1] use an encoder/decoder framework. Our method allows the image features to be updated conditioned on the labels, which is a key characteristic of our model.

Connecting to Graph Based Neural Relational Learning. Another line of recent works employ object localization techniques[11, 9] or attention mechanism[8, 12] to locate semantic meaningful regions and try to identify underlying relations between regions and outputs. However, these

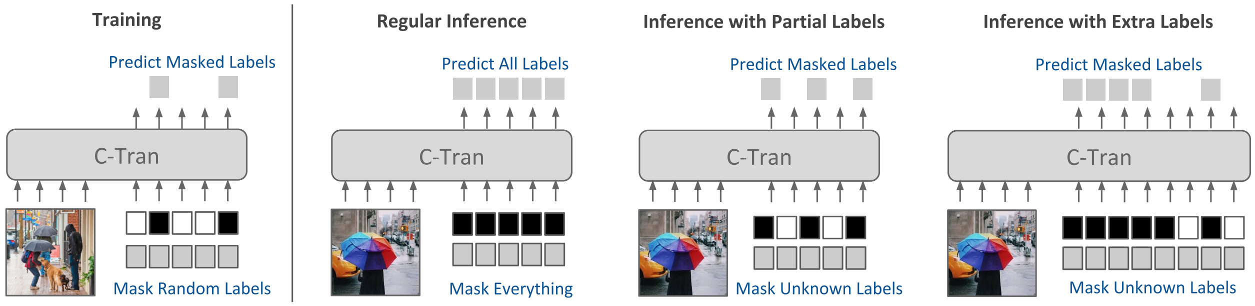


Figure 1. Detailed example of the general training method and three different inference settings where C-Tran can be applied.

Images	True Labels	ResNet-101	C-Tran	C-Tran + partial labels	
 ID:000000362831	fork knife, spoon, bowl, chair, diningtable	**fork,** sandwich, **diningtable,** **spoon,** cup	**fork,** **knife,** **diningtable,** person, cake	spoon=1, trafficlight=0, bench=0, dog=0, ...	**fork,** **knife,** **diningtable,** person, **bowl**
 ID:000000106216	person, car, truck, parkingmeter, horse,	**person,** **car,** **truck,** **horse,** bicycle	**car,** **person,** **truck,** **horse,** bicycle	bicycle=0, motorcycle=0, train=0, boat=0 ...	**car,** **person,** **truck,** **horse,** **parkingmeter**
 ID:000000243213	person, bench, backpack, tennisracket, bottle, chair	**person,** **tennisracket,** **chair,** tie, sportsball	**person,** **tennisracket,** **chair,** sportsball, **bench**	backpack=1 parkingmeter=0, bird=0, zebra=0, ...	**person,** **tennisracket,** **chair,** **bottle,** **bench**
 ID:000000170129	airplane, train	**airplane,** boat, car, truck, person	**airplane**, boat, person, car, bird	car=0, motorcycle=0, bus=0, truck=0, ...	**airplane**, boat, person, bird, **train**
 ID: 000000262896	bottle, spoon, diningtable, cellphone, book	**bottle,** fork, **diningtable,** bowl, **spoon**	fork, **spoon**, bowl, **book,** **diningtable**	diningtable=1, bicycle=0, car=0, truck=0, ...	**spoon**, bowl, **book**, **bottle**, **cellphone**

Figure 2. Qualitative examples of C-Tran + partial labels on the COCO-80 dataset. In the last column, we use $\epsilon = 25\%$ partial labels, some of which are shown. Correctly predicted labels are in bold.

methods either require expensive bounding box annotations or merely get regions of interest roughly due to the lack of label supervision. One recent study by [10] also showed that modeling the associations between image feature regions and labels helps to improve multi-label performance. In our work, C-Tran uses graph attentions and enables each target label to attend differentially to relevant parts of an input image.

For multi-label classfication(MLC), [3] formulate MLC using a label graph and they introduced a conditional de-

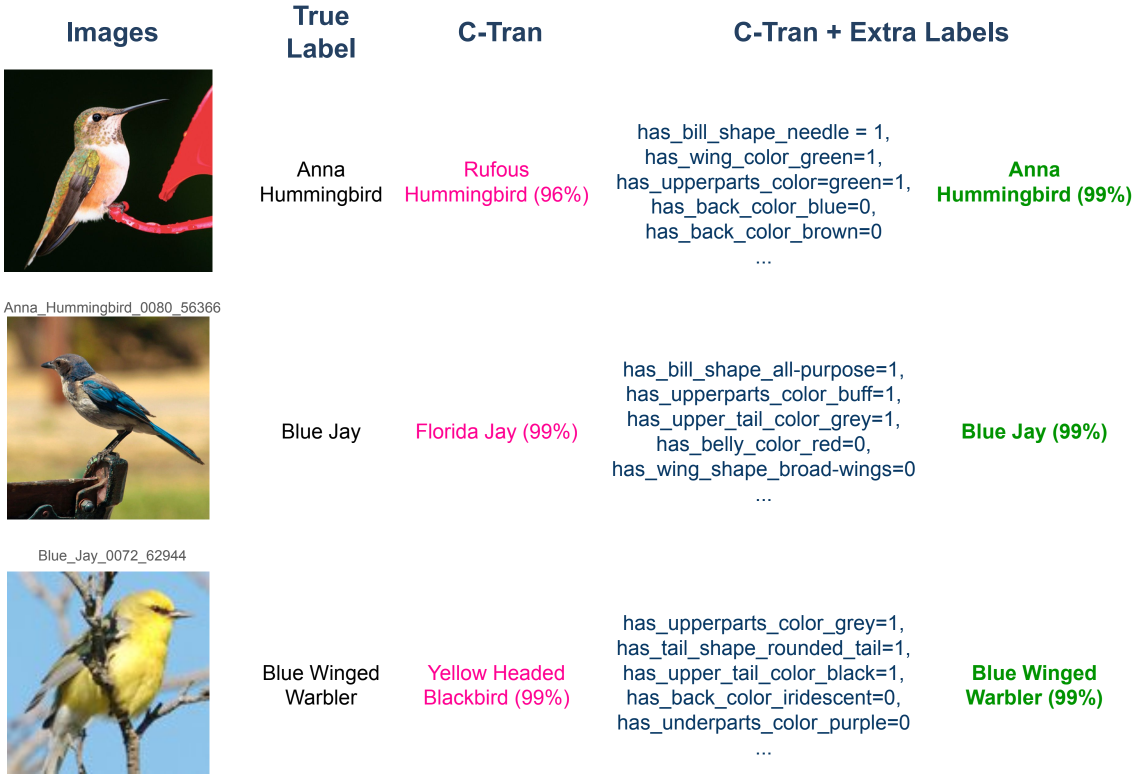

Images
True Label
C-Tran
C-Tran + Extra Labels
Anna Hummingbird
Rufous Hummingbird (96%)
has_bill_shape_needle = 1,
has_wing_color_green=1,
has_upperparts_color=green=1,
has_back_color_blue=0,
has_back_color_brown=0
...
Anna Hummingbird (99%)
Anna_Hummingbird_0080_56366
Blue Jay
Florida Jay (99%)
has_bill_shape_all-purpose=1,
has_upperparts_color_buff=1,
has_upper_tail_color_grey=1,
has_belly_color_red=0,
has_wing_shape_broad-wings=0
...
Blue Jay (99%)
Blue_Jay_0072_62944
Blue Winged Warbler
Yellow Headed Blackbird (99%)
has_upperparts_color_grey=1,
has_tail_shape_rounded_tail=1,
has_upper_tail_color_black=1,
has_back_color_iridescent=0,
has_underparts_color_purple=0
...
Blue Winged Warbler (99%)

Figure 3. Qualitative examples of C-Tran + extra labels on the CUB-312 dataset. In the last column, we use $\epsilon = 54\%$ extra labels, some of which are shown.

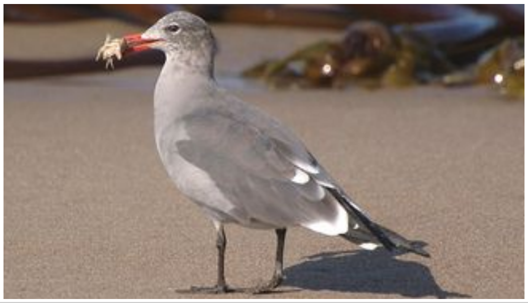
pendency SVM where they first trained separate classifiers for each label given the input and all other true labels and used Gibbs sampling to find the optimal label set. The main drawback is that this method requires separate classifiers for each label. [6] proposes a method to label the pairwise edges of randomly generated label graphs, and requires some chosen aggregation method over all random graphs. The authors introduce the idea that variation in the graph structure shifts the inductive bias of the base learners. One recent study [5] used graph neural networks for multi-label classification on sequential inputs. The proposed method models the label-to-label dependencies using GNNs, however, does not represent input features and labels in one coherent graph. A key aspect of C-Tran is that the Transformer encoder can be viewed as a fully connected graph which is able to learn any relationships between features and labels. The Transformer attention mechanism can be regarded as a form of graph ensemble learning [4]. Above all, previous methods using graphs to model label dependencies do not allow for partial evidence information to be included in the prediction.

A.5. Label Mask Training

In Algorithm 1, we detail the label mask training (LMT) procedure. For each training sample, we select a random amount of labels to be used as "known" input labels to the model. The loss function is then computed on all unknown labels.

Algorithm 1: C-Tran Label Mask Training Procedure

```
loss = 0
for sample $(\mathbf{x}, \mathbf{y})$ in batch do
    label_idxs = range(1, $\ell$);
    $n$ = randint(0.25$\ell$,$\ell$);
    unk_idxs = sample(label_idxs, $n$);
    $\mathbf{y}_u = \{y_i$ for i in unk_idxs$\}$;
    $\mathbf{y}_k = \{y_j$ for j in in label_idxs excluding unk_idxs$\}$;
    $\hat{\mathbf{y}}_u = f(\mathbf{x}, \mathbf{y}_k; \theta)$;
    for label index j in $\mathbf{y}_u$ do
        loss += $-(y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j))$
    end
end
Run backprop;
Update $\theta$;
```



Attribute Value	Class Prediction
has_yellow_underparts=1	Heermann_Gull
has_yellow_underparts=0	Glaucous_windged_Gull

Figure 4. Counterfactual example. The ground truth is *Heermann_Gull*. If we incorporate the "has yellow underparts" attribute as input to the model, it correctly predicts the bird class.

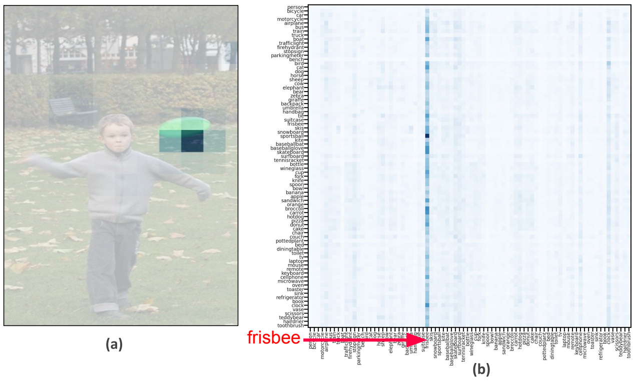


Figure 5. **(top)** Frisbee-to-image attention. The frisbee label attends to the frisbee in the image. **(bottom)** Label-to-label attention. Most labels attend to the frisbee label.

A.6. Counterfactual Testing

Counterfactual testing, as introduced by Koh et al., is performed to answer the question "If I know that some label is true, how does it change the prediction of other labels?". Fig. 4 shows a counterfactual example on the CUB-312 dataset. Here we show the bird class prediction of our model contingent on *has_yellow_underparts* being true (=1) or false (=0). In other words, this allows the user to answer the question "What kind of bird would this be if it has (or doesn't have) yellow underparts?".

A.7. Attention Weight Analysis

In Fig. 5 we demonstrate attention weight analysis from the third layer of our model on the COCO-80 dataset. Fig. 5 (a), shows label-to-image attention for the "frisbee" label. The frisbee label attends to the frisbee object in the image. Fig. 5 (b) shows label-to-label attention. Most labels attend to the frisbee label. We found that sometimes the predictions can be worse if the model relies too much on partial labels, but overall the performance is improved.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[3] Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1300, 2011. 2

[4] Kazuyuki Hara, Daisuke Saitoh, and Hayaru Shouno. Analysis of dropout learning regarded as ensemble learning. In *International Conference on Artificial Neural Networks*. Springer, 2016. 3

[5] Jack Lanchantin, Arshdeep Sekhon, and Yanjun Qi. Neural message passing for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 138–163. Springer, 2019. 3

[6] Hongyu Su and Juho Rousu. Multilabel classification through random graph ensembles. In *Asian Conference on Machine Learning*, pages 404–418, 2013. 3

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017. 1

[8] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[9] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, Sep. 2016. 1

[10] Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, and Yao Lu. Correlative multi-label multi-instance image annotation. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 651–658, USA, 2011. IEEE Computer Society. 2

[11] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In Lourdes Agapito, Tamara Berg, Jana Kosecka, and Lihi Zelnik-Manor, editors, *Proceedings - 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 280–288, United States of America, 2016. IEEE. 1

[12] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2027–2036, 2017. 1