AssocFormer: Association Transformer for Multi-label Classification

Xin Xing, Chong Peng, Yu Zhang,
Ai-Ling Lin, Nathan Jacobs

University of Kentucky
Washington University in St. Louis

BMVC 2022

Overview

Recent work has shown that explicitly modeling the co-occurrence relationship between classes is critical for achieving good performance on multi-label classfication task. We propose an end-to-end model by adopting the transformer-based feature-extraction backbone with a novel and efficient association module.

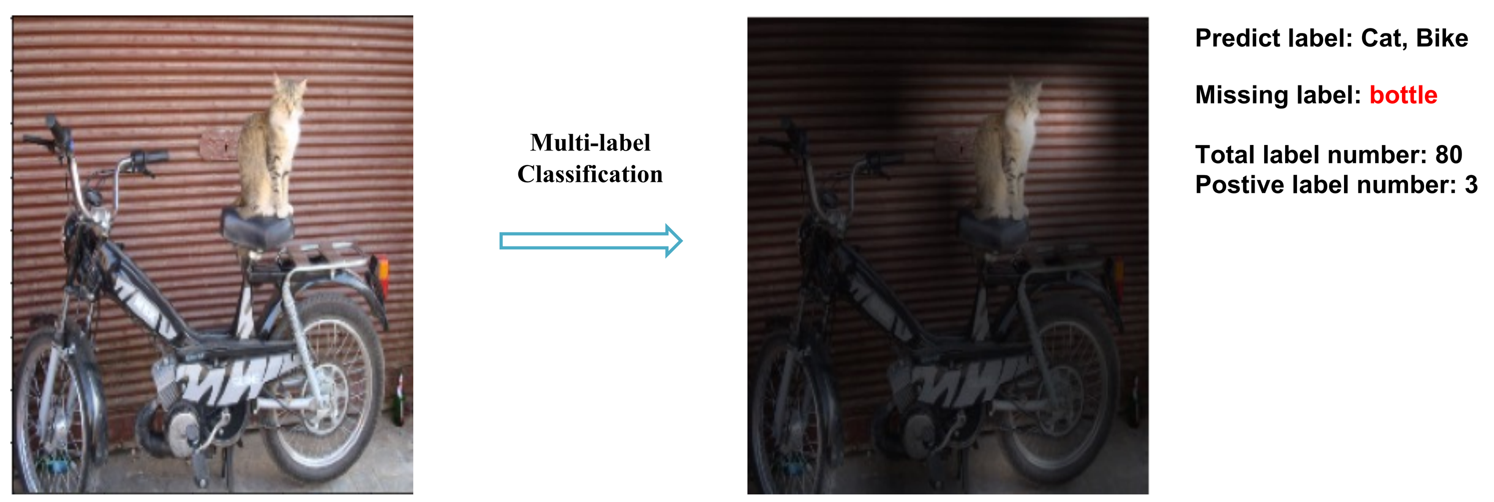
Highlights:

- A simple yet effective end-to-end transformer-based framework for multi-label classification.
- A new association module to explore label correlation. The module is learnable and is computation-efficiency
- Evaluate the proposed model on different benchmark dataset: MS-COCO and PASCAL VOC and obtrain superior or comparable performance.

Introduction

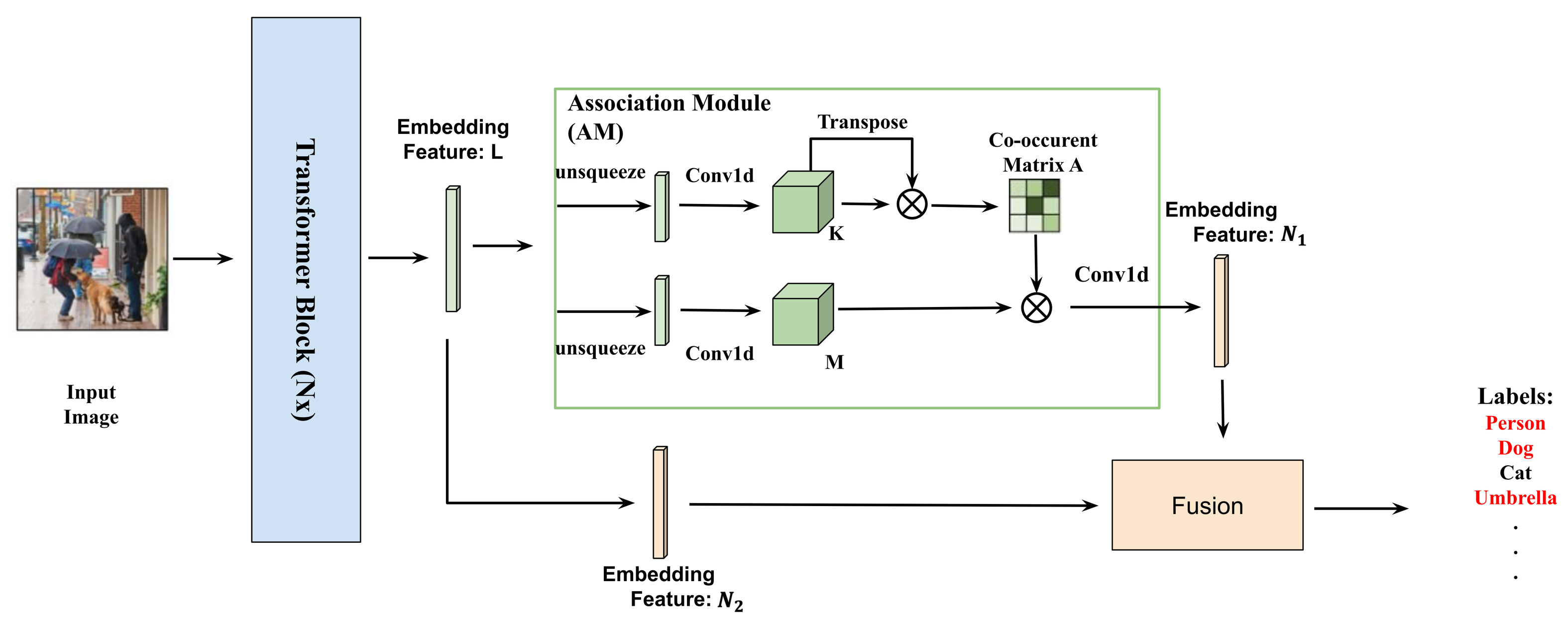
Background: The goal of Multi-Label Classification (MLC) is to predict a set of labels for a single image.

Chanllenge: 1). tiny object dection and 2). positive and negative label imbalance.



Motivation: The missing tiny object dection is one of the main chanllenge of MCL. We propose to overcome this issue by adopting the Vision Transforemr (ViT) as backbone with label association information to boost the final predcition.

Approach



Architecture Overview: We leaverage the transformer as the backbone feature extractor to get the extracted feature $L \in R^d$. We forward L through the Association Module (AM) to calculate the association matrix $A \in R^{C\times C}$ and output $N_1 \in R^c$. Meanwile, we get another output $N_2 \in R^C$ by forwarding L through a fully connected layer. The final prediction is the fusion operation of the N_1 and N_2.

Operation of AM:

We first unsqueeze the feature L and conduct a 1D covolution to project 1D embedding to 2D, $\{K, M\} \in R^{C\times d}$:

$$K, M = Conv1D(unsqueeze(L, 1))$$

We transpose the feature K and conduct multiplication with K itself attached a sigmoid function to calcaulate the association matrix A:

$$A = sigmoid(K\times K^T)$$

We finally multily feagure M with association matrix A and apply another Conv1D to get the output N_1:

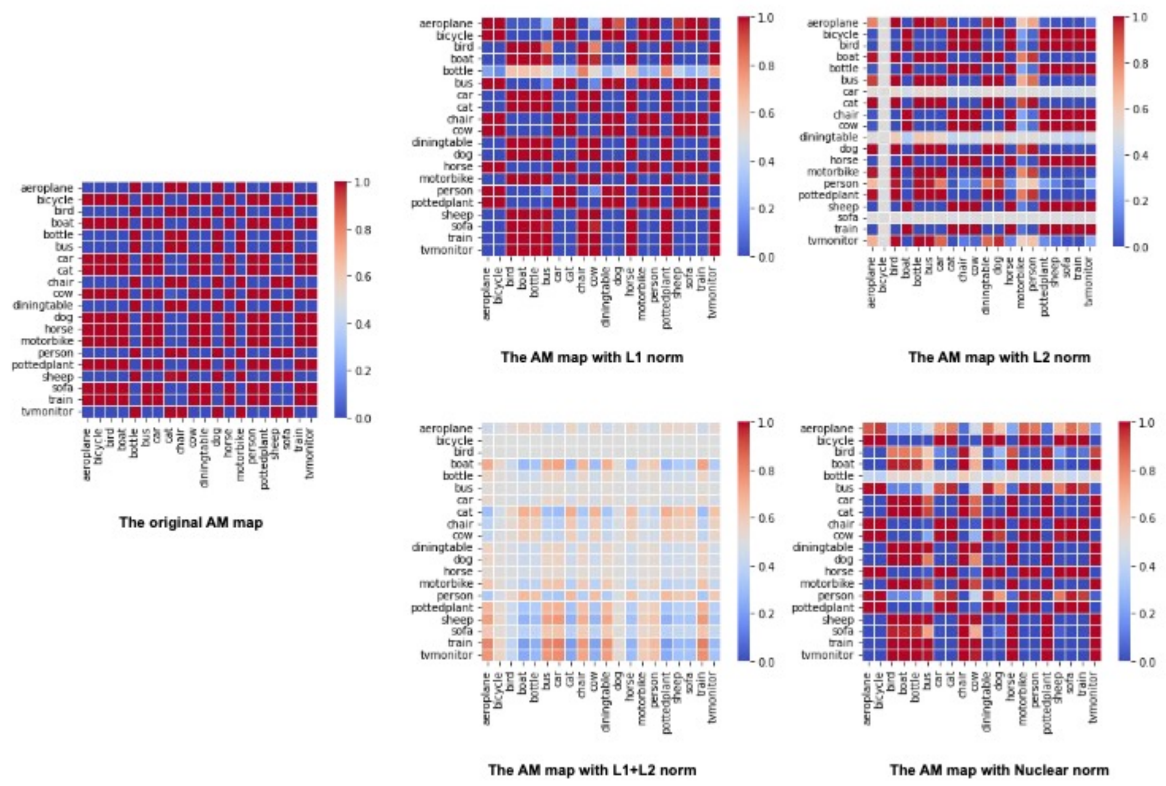
$$N_1 = Conv1D(M\times A)$$

Data & Metrics: We evaluate the model with two public datasts: MS-COCO and PASCAL VOC. To evaluate the performance of our model, we use mean average precision (mAP) as metric.

Experiments

Model	Resolution	mAP
SRN [31]	224 × 224	77.1
ResNet101 [12]	224 × 224	78.3
CADM [3]	448 × 448	82.3
ML-GCN [4]	448 × 448	83.0
KSSNet [18]	448 × 448	83.7
SSGRL [2]	576 × 576	83.8
C-Tran [16]	576 × 576	85.1
ADD-GCN [29]	576 × 576	85.2
ASL(22k) [22]	448 × 448	88.4
MlTr-l(22k) [5]	384 × 384	88.5
Swin-L(22k) [19]	384 × 384	89.2
Swin-L-AM(22k)(Ours)	384 × 384	89.8
CvT-24w(22k) [27]	384 × 384	88.9
CvT-24w-AM (22k)(Ours)	384 × 384	90.1

Results on MSCOCO: The proposed model outperforms the baseline models. Meanwhile, the proposed models with AM gain better performance than the vision transformer only baselines.



Visualization of different AM with different regularization terms on the PASCAL VOC.

Conclusion

We proposed AssocFormer, which combines a transformer backbone with a light-weight association module, for the task of multi-label image classification. This approach outperforms prior work on two standard public benchmark datasets, while simultaneously being simpler to implement.