AssocFormer: Association Transformer for Multi-label Classification

Xin Xing[1]
xxi242@g.uky.edu
Chong Peng[2]
pchong1991@163.com
Yu Zhang[1]
y.zhang@uky.edu
Ai-Ling Lin[3]
ai-ling.lin@health.missouri.edu
Nathan Jacobs[4]
jacobsn@wustl.edu

[1] Department of Computer Science, University of Kentucky Lexington, KY, USA
[2] College of Computer Science and Technology, Qingdao University Qingdao, China
[3] Department of Radiology, University of Missouri Columbia, MO, USA
[4] Computer Science & Engineering, Washington University in St. Louis St. Louis, MO, USA

Abstract

The goal of multi-label image classification is to predict a set of labels for a single image. Recent work has shown that explicitly modeling the co-occurrence relationship between classes is critical for achieving good performance on this task. State-of-the-art approaches model this using graph convolutional networks, which are complex and computationally expensive. We propose a novel, efficient association module as an alternative. This is coupled with a transformer-based feature-extraction backbone. The proposed model was evaluated using two standard datasets: MS-COCO and PASCAL VOC. The results show that the proposed model outperforms several strong baseline models.

1 Introduction

Image classification is a fundamental task in computer vision. For a traditional image classification problem where each image contains a single object, the learning task for a given image is to predict the category of the object contained in it. However, real-world images often contain multiple objects and is essentially complex, which makes it challenging for the model to understand the scene. Therefore, multi-label recognition is an essential task in our natural world. Recent works show outstanding improvement on the multi-label recognition task by different model architectures [17, 16], new loss function definition [22], and label association relationship exploration by graph learning [4, 29]. However, these methods increase implementation complexity. Our work achieves the same, or better, accuracy with a significantly simpler implementation. In particular, we adopt the network backbone with a transformer and propose a new module for label correlation computation.

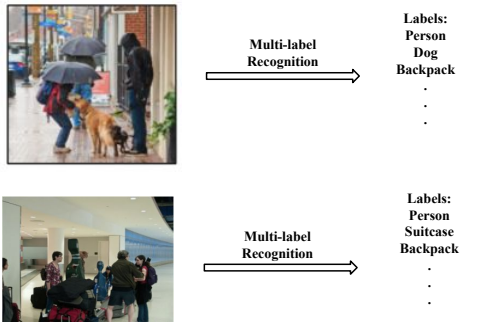

Figure 1: The task of multi-label recognition is predicting the object labels in the input image. We can apply the object association relationship to assist our final prediction; for example, a bag is more likely to be associated with a person than a dog. Therefore, given an image with a dog, the appearance probability of a bag in the same image is low.

The convolutional neural network (CNN) has become a dominant technique in various computer vision tasks, such as classification [15], segmentation [20], and object detection [11]. For the multi-label recognition task, the ResNet, such as ResNet101, has been widely employed in many works [3, 4, 12, 31]. Currently, the TResNet [22] is often considered as an alternative to the ResNet due to its better trade-off between efficiency and performance. However, both of them suffer from a drawback with the convolution operation, which concentrates on the local-area computation and thus omits the global information extraction. For images, especially multi-label images with multi objects in the same scene, the global information is essentially important for classification. It is difficult for the scene understanding by only concentrating on the local regions while neglecting the object dependency relationship. Meanwhile, the vision transformer (ViT) [8, 19, 27], built with the multi-head self-attention module, is capable of the long-range dependency to extract the global information of the whole image with different views. ViT makes up for the flaw of the CNN model and is an ideal backbone for the multi-label recognition task.

The object association has been studied in multi-label classification for recent approaches. The graph convolutional network (GCN) is the main focus in analyzing the label correlation for the final prediction boosting. However, the GCN requires the correlation matrix as prior knowledge, which is challenging in realistic implementation. To address this issue, it is proposed to approximately learn the correlation matrix with the training set [4]. However, with this strategy, it tends to have an overfitting problem for the learned correlation matrix, and thus the learning performance is not guaranteed for the testing set. Moreover, the GCN performs multi-matrix multiplication, which is complex and computationally expensive.

In this work, we propose a model with a transformer as the backbone and a new association module (AM) for the label-correlation computation. The transformer-based backbone can effectively extract the discriminative feature for further analysis. The AM has a more straightforward architecture than the GCN model and is more efficient without the graph convolution operation. Meanwhile, the AM can dynamically learn the label correlation without any prior or extra information.

We summarize the main contributions of our work as follows:

- We propose to build a simple yet effective end-to-end transformer-based framework

for multi-label classification.

- We propose a new association module to explore label correlation. The module is learnable based on the features and is computation-efficiency compared with the graph convolutional operation.
- We conducted experiments on different benchmark datasets: MS-COCO and PASCAL VOC and obtained superior or comparable performance.

2 Related Work

2.1 Transformer

Recently, the transformer has shown impressive potential in computer vision with its long-dependency ability for global information aggregation. Dosovitskiy *et al*. [8] propose the Vision Transformer (ViT) to bridge the transformer-based model for image classification. By splitting the image into a number of fixed-size patches, the block-stacked transformer encoder can conduct the feature extraction on visual task. Carion *et al*. [1] apply the transformer on the object detection task with a model named DETR. Wu *et al*. [27] propose the CvT by adopting the convolutional layer for the low-level feature extraction and forwarding the feature to the transformer. Liu *et al*. [19] design a new transformer architecture named Swin by patch shifting and merging and demonstrate its generalization on different vision tasks. For the transformer application on multi-label recognition task, Lanchantin *et al*. [16] propose a Classification Transformer (C-Tran) to combine the transformer with inferred label mask together for the final prediction. Cheng *et al*. [5] propose a model named MlTr by merging a cross-attention module on the transformer encoder to improve the performance. Our study mainly leverages the transform as the backbone for multi-label recognition. We use Swin and CvT as the backbone for feature extraction and stack the association module (AM) to boost the final prediction. Compared with C-Tran, we don't need extra or prior information but only the input image. Compared with MlTr, our model has simpler architecture and gains better performance.

2.2 Loss Function

The commonly used loss functions for multi-label recognition are cross entropy and multi-label soft margin loss, which ignore the positive-negative imbalance problem for the labels. Asymmetric loss (ASL), a variant of focal loss [22], is proposed to overcome this issue by asymmetric focusing and asymmetric probability shifting to dynamically operate on the negative and positive samples. ASL shows tremendous improvement and becomes the standard loss on the multi-label recognition tasks. In our implementation, we apply the ASL as the loss function.

2.3 Label Association

Considering that multi-objects appear in the same image, the label association is a good topic to explore in the multi-label recognition study. Therefore, the graph convolutional network (GCN) [14] is adopted to analyze the label correlation. MLGCN [4] proposes to learn a static correlation matrix by counting the label relation of the training set and develop a GCN-based classifier to boost the final prediction. SSGRL [2] proposes to learn the dynamic

correlation matrix based on the extracted feature with the LSTM model. ADDGCN [29] simplifies the correlation matrix learning procedure by applying the class activation module (CAM) [30]. Different from graph learning in exploring the label association relationship, by graph operation, we propose to construct the learnable object association matrix by our end-to-end model directly. Besides, our initial attempt to combine the GCN module with the ViT shows that the whole model is unstable, with either a static graph (MLGCN) or a dynamic graph (ADDGCN). During the training, there is the NaN value on the gradient descent when combining the ViT with GCN module.

3 Approach

In this section, we provide a detailed introduction of our method. First, we summarize the high-level problem in our study. Second, we show the overall network architecture. Then, we introduce detailed information about the transformer and association module basics. Next, we show our fusion operation. Finally, we discuss the loss function used for training.

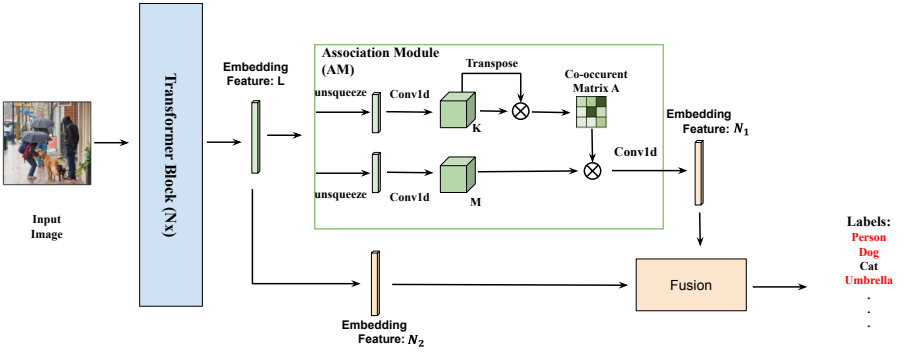

Figure 2: The overall architecture of the proposed model.

3.1 Problem Formulation

For an input image x with a set of category binary label $y = [y_1, y_2, ..., y_C], y_i \in \{0, 1\}$, the multi-label recognition task is to predict whether each category appears or not in the image. Label $y_i = 1$ means that the $i-th$ category is presented in the image; otherwise, $y_i = 0$. The target of multi-label recognition is to build up a classifier, f, to predict the probability $p = [p_1, p_2, ..., p_C]$ of the appearance of each category of the input image x: $y = f(x)$.

3.2 Architecture

Figure 2 illustrates the overall architecture of our model. We first leverage the transformer as the backbone feature extractor for an input image. The output feature of the backbone is a feature embedding $L \in R^d$. Then, We forward the embedding L to the AM module to calculate the association matrix $A \in R^{CxC}$ and output $N_1 \in R^C$. Depending on the fusion

operation, we can forward the embedding L through a fully connected layer and get another output $N_2 \in R^C$. More details will be discussed in the fusion section.

3.3 Basics of Transformer

Considering the high performance of the transformer in different vision tasks, we leverage the transformer as the backbone in our study. So far, many works study the variants on ViT [8] structure, such as the patch shifting and merging in Swin [19] and convolutional layer hybrid structure in CvT [27]. For simplification, we summarize the transformer basics, which are treated as the standard operation.

Given an image x, the patch embedding is to project the $2D$ image into a $1D$ embedding z_0 by the patch-partitioning operation at the beginning of the transformer:

$$z_0 = PatchEmbedding(x), \tag{1}$$

after the initial patch embedding operation, the consecutive transformer encoder blocks are computed as:

$$\hat{z}^l = MSA(LN(z^{l-1})) + z^{l-1}, \tag{2}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l, \tag{3}$$

where $\hat{z}^l$ and z^l annotate the output feature of the multi-head self-attention (MSA) module and multi-layer perceptron (MLP) module of block l, respectively. LN represents the layer normalization.

3.4 Association Module (AM)

We want to efficiently capture the co-occurrence information with easy implementation. Our association module (AM), motivated by GCN [4, 29], is to aggregate object association information, which is used to boost the final prediction. To dynamically capture the object association, we prefer to learn the association matrix through the extracted feature L of the transformer-based backbone. Since the feature embedding L contains the potential information of each object category, it is the potential to access the object association by appropriate module design. In the implementation, we forward the extracted embedding feature $L \in R^d$ through the AM. We firstly unsqueeze the L and conduct a $1D$ convolution to project $1D$ embedding to $2D$, $\{K, M\} \in R^{C \times d}$:

$$K/M = Conv1D(unsqueeze(L, 1)). \tag{4}$$

We transpose the feature K and conduct the multiplication with K itself attached a sigmoid function to calculate the association matrix $A \in R^{C \times C}$:

$$A = sigmoid(K \times K^t), \tag{5}$$

we finally multiply feature M with association matrix A and apply another Conv1D to get the embedding feature N_1:

$$N_1 = Conv1D(M \times A). \tag{6}$$

It should be noted that there is a stark difference between the self-attention module (SAM) [24] and our model. The purpose of the association module is dynamically learning and aggregating the inter-class correlation through model training. Therefore, association module works on the class domain. Meanwhile, self-attention works on the spatial domain of the input image. A more straightforward view of the difference is by the implementation. Given a dataset contains C classes, we have an input feature $I \in R^{D\times H\times W}$, where D is the channel, H is the height, and W is the width. The attention map $M_{SA} \in R^{N\times N}$ of the self-attention is to calculate the pixel-wise similarity, where $N = H \times W$. However, the association map $A \in R^{C\times C}$, where C is the number of the task classes. As for the computation complexity, it is straightforward that our module needs less than the GCN module, which requires multiple-times matrix multiplication with the same input.

3.5 The Fusion operation

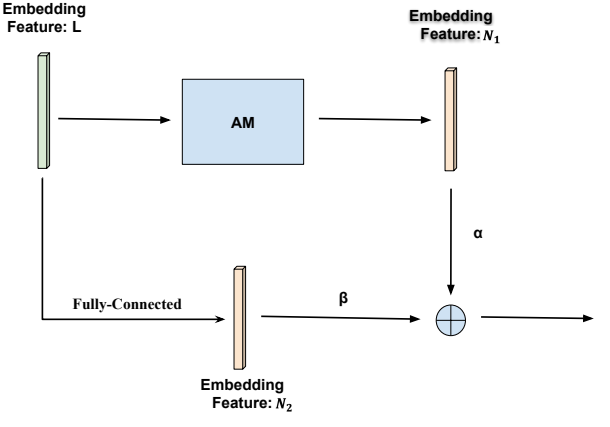

Figure 3: The fusion strategy in our study.

We discuss the potential fusion strategy in this section. Our default way is with our association module only, we also experimented with the model without our association module, and we found it good to combine them together. Illustrated by Figure 3, instead of the output feature N_1, we can simply conduct a fully-connected layer on embedding L to get another output feature $N_2 \in R^C$. Therefore, we can predict the final output $O \in R^C$ by interpolating the N_1 and N_2 as follows:

$$O = \alpha N_1 + \beta N_2, \tag{7}$$

where α and β are the coefficients of the embedding feature N_1 and N_2, respectively. There are different strategies to set values for α and β:

- equal weights, α and β are equal to 0.5: $\alpha = \beta = 0.5$
- single path, we use the embedding N_1 only: $\alpha = 1, \beta = 0$
- learnable path, we set the α as a learnable parameter and let β equal to 1: $\alpha = learnable, \beta = 1$

We set equal weights as our final fusion operation. More details are shown in the supplemental material.

3.6 Loss Function

The cross-entropy loss is the basic loss function for the multi-label recognition problem. However, it neglects the multi-label imbalance problem, where each sample image contains only a sufficiently small number of positive objects from the label set. Therefore, we adopt a new asymmetric loss (ASL) [22], a variant of focal loss, to narrow the imbalance issue. The ASL is defined as:

$$ASL = -\frac{1}{C}\sum_{k=1}^{C} y_k \cdot (1-p_k)^{\gamma+} log(p_k) + (1-y_k)\cdot (p_k)^{\gamma-} log(1-p_k) \tag{8}$$

where the $\gamma+$ and $\gamma-$ values are the asymmetric focusing parameters to decouple the influence of the positive and negative sample. In our implementation, we set $\gamma+ = 0$ and $\gamma- = 2$ as default. Meanwhile, due to the long-tail distribution of the object correlation, we find out that the association matrix $A \in R^{C\times C}$ is easily overfitted. Therefore, we use the L_2 regularization on the association matrix A with a nonnegative balancing parameter θ, which leads to the overall training loss in the following:

$$Loss = ASL + \theta \|A\|_2. \tag{9}$$

4 Evaluation

4.1 Implementation and Metrics

Transformer Encoder. For efficient feature extraction, we adopt pre-trained Swin-L and CvT-w24 as the backbone, respectively. Based on the designation of Swin [19], we set the channel number of the hidden layers in the first stage to be $C = 192$ and the layer number of each Swin block to be $\{2,2,18,2\}$. We adopt the pre-trained Swin model from "timm" [26]. For the CvT-w24 model structure parameters, the hidden embedding size of each block on CvT-w24 is $\{192,768,1024\}$, the layer number of each CvT-w24 block is $\{2,2,20\}$. The pre-trained CvT-w24 model is accessed from CvT [27] official implementation.

Training. Our models are implemented using PyTorch. We train the model with 40 epochs using Adam [13] optimizer and 1-cycle policy [23]. The maximal learning rate is $1e-4$. The batch size is 16. For augmentation, we use Cutout [7] with a factor of 0.5 and RandAugment [6]. For simplification, we normalize the image size to $0 \sim 255$. For regularization, we use True-wight-decay [21] with value $1e-4$. We follow the ASL [22] implementation to use exponential moving average (EMA) to model parameters with a decay of 0.9997. The reported best performance is either the original model or the EMA model. The mixed precision is used to speed up the training processing.

Metrics. We follow the conventional setting of the previous works [5, 16, 29] on model evaluation. The threshold of the image label is 0.5. We report the mean average precision (mAP), which is the most important metric in the task of multi-label recognition.

4.2 Experiment

4.2.1 MS-COCO

Microsoft COCO [17] is a widely used benchmark for multi-label image recognition. It contains 82,801 training images and 40,504 validation images. There are 80 categorized

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

Table 1: Results of the MS-COCO dataset under different state-of-the-art models. Take the mAP as the main reference, our models outperform the previous works. All metrics are in %. The 22k denotes the model pre-trained with ImagNet-22k.

objects in this dataset, with an average of 2.9 object labels per image. Since this data set lacks of test set, the validation images are often used for evaluation.

Quantitative results are reported in Table 1. We compare our model with 10 baseline models, including SRN [31], ResNet101 [12], CADM [3], ML-GCN [4], KSSNet [18], SSGRL [2], C-Tran [16], ADD-GCN [29], ASL [22], and MlTr-l [5]. The proposed model CvT-24w-AM achieves the performance with $mAP = 90.1$, outperforming the previous approaches. Another transformer backbone model Swin-L-AM achieves the $mAP = 89.8$, which beats most of the baselines. To further validate the significance of adopting the AM in our model, we perform ablation study as follows. We remove the AM from our model and only leverage the transformer backbone to obtain two models as baselines, namely Swin-L and CvT-24w, respectively. Experimental results show that the performances of both Swin-L and CvT-24 are significantly improved by integrating the AM, which confirms the effectiveness and significance of adopting the AM in our model.

4.2.2 PASCAL VOC

Model	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [31]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
Fev+Lv [31]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
MCAR [31]	99.7	99.0	98.5	98.2	85.4	96.9	97.4	98.9	83.7	95.5	88.8	99.1	98.2	95.1	99.1	84.8	97.1	87.8	98.3	94.8	94.8
ML-GCN [4]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
SSGRL [2]	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	98.8	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	93.4
ADD-GCN [29]	99.8	99.0	98.4	99.0	86.7	98.1	98.5	98.3	85.8	98.3	88.9	98.8	99.0	97.4	99.2	88.3	98.7	90.7	99.5	97.0	96.0
Swin-L-AM (22k)(Ours)	99.8	92.7	99.4	99.3	90.3	98.2	94.0	97.2	89.7	99.4	97.1	86.5	99.3	96.5	89.8	93.1	99.1	94.2	99.6	88.1	96.0
CvT-24w-AM (22k)(Ours)	99.8	98.4	99.6	99.2	96.0	98.7	95.8	99.2	94.0	99.4	96.8	98.2	99.5	99.2	92.2	95.5	99.5	96.7	99.8	97.4	96.2

Table 2: Results of the VOC2007 dataset under different state-of-the-art models. We report the mAP of different methods. For each category, we show the accuracy per class. All metrics are in %. The 22k denotes the model pre-trained with ImagNet-22k.

PASCAL Visual Object Classes Challenge (VOC2007) [9] is another widely used datasets

used for multi-label recognition. VOC2007 contains 5,011 images as the train-val set and 4,952 images as the test set. The number of the class is 20, with 2.5 categories per image. In the literature, some works adopt some extra datasets in the training stage for pre-training, such as the MS-COCO dataset. In our experiments, for fair comparison we compare our model with the baseline methods that are pre-trained with ImageNet-1k or ImageNet-22k only and show the results in Table 2. Compared with baselines, our proposed models achieve superior performance with mAP = 96.0 of the Swin-L-AM model and mAP = 96.2 of the CvT-24w-AM model, which achieves state-of-the-art performance. Besides, we conduct the experiment of the baselines without AM module (backbone Swin-L and CvT-24w only), and get $mAP_{Swin-L} = 95.8$ and $mAP_{CvT-24w} = 96.0$, respectively.

In terms of each category accuracy, both of our models outperform other baselines in the tiny objects, such as 'bottle' and 'plant'. This contribution is jointly by the transformer-based backbone and association module. In a multi-labeling classification task, the small object is one of the bottlenecks of previous convolutional neural network (CNN) based models. The transformer is built upon the self-attention module, which efficiently captures the feature of pixel granularity. Meanwhile, the association module can boost the performance by the object correlation (e.g. a person with a bottle).

4.3 Association Map Visualization

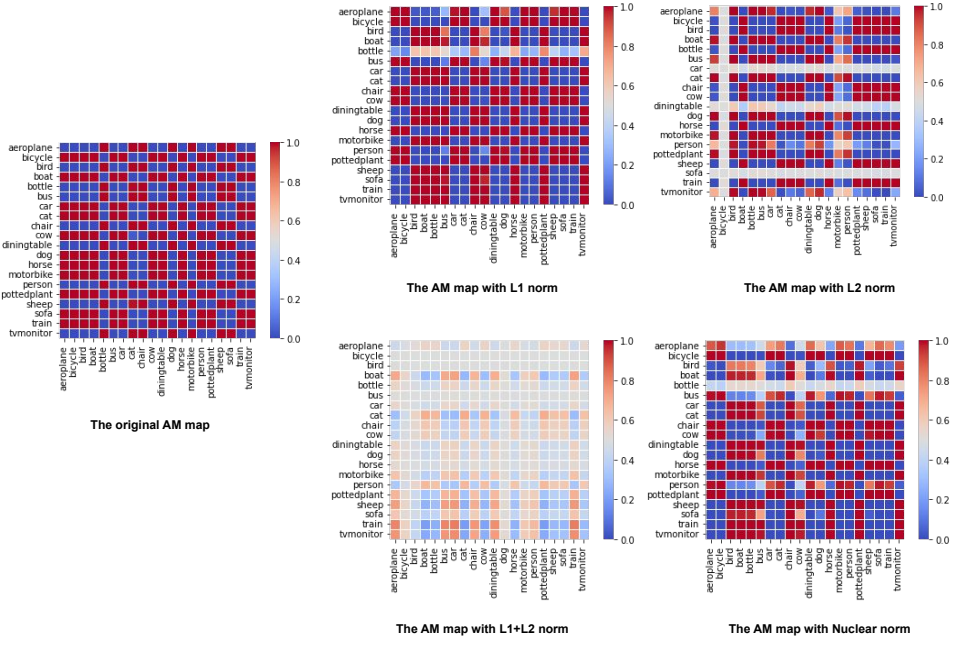

Figure 4: The visual illustration of association matrices that are learned with different regularization terms on the VOC2007 dataset.

norm	mAP
L_1	95.4
L_2	96.0
$L_1 + L_2$	92.2
Nuclear norm	95.8

Table 3: The results of the loss function with different norm strategies on the benchmark VOC2007.

In our model, we adopt the L_2 regularization for the association map of the object. In this section, we conduct experiments with the same coefficient $\theta = 0.1$ to show it is more proper to adopt L_2 regularization than the other regularizations, such as L_1 regularization, elastic-net $(L_1 + L_2)$ regularization, and the nuclear norm regularization. We visually show the object AM in the cases of adopting different regularizations in Figure 4. From the figure, it is observed that the AM map memorizes the most object association without any regularization term, which implies a long-tail distribution issue similar as the MLGCN [4]. By the regularization constraint, the AM map efficiently captures the object association with significantly reduced overfitting effects. It is seen that both the L_2 and the nuclear norms have stronger constraining effects than the L_1 norm. We quantitatively show the best performance of our model with different regularization terms in Table 3, where it is found that the L_2 norm promotes the learning performance better than the others. This observations confirms the properness of adopting the L_2 norm in our model.

5 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 matches or outperforms prior work on two standard public benchmark datasets, while simultaneously being simpler to implement. We found that the form of regularization on the association matrix was critical for achieving strong quantitative performance. We conjecture that this is due to the long-tail distribution of class-pair co-occurrences (with many pairs co-occurring infrequently, but a few that commonly co-occur).

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.

[2] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019.

[3] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019.

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[5] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Nian Shi, and Honglin Liu. Mltr: Multi-label classification with transformer. *arXiv preprint arXiv:2106.06195*, 2021.

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[10] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing*, 30:5920–5932, 2021.

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[16] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 700–708, 2018.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[23] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[26] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[27] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[28] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016.

[29] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European conference on computer vision*, pages 649–665. Springer, 2020.

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[31] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5513–5522, 2017.