CROSS-LAYER AGGREGATION WITH TRANSFORMERS FOR MULTI-LABEL IMAGE CLASSIFICATION

Weibo Zhang[1,2] *Fuqing Zhu* [1,2] ✉ *Jizhong Han* [1] *Tao Guo* [1] *Songlin Hu* [1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, China

ABSTRACT

Multi-label image classification task aims to predict multiple object labels in a given image and faces the challenge of variable-sized objects. Limited by the size of CNN convolution kernels, existing CNN-based methods have difficulty capturing global dependencies and effectively fusing multiple layers features, which is critical for this task. Recently, transformers have utilized multi-head attention to extract feature with long range dependencies. Inspired by this, this paper proposes a Cross-layer Aggregation with Transformers (CAT) framework, which leverages transformers to capture the long range dependencies of CNN-based features with Long Range Dependencies module and aggregate the features layer by layer with Cross-Layer Fusion module. To make the framework efficient, a multi-head pre-max attention is designed to reduce the computation cost when fusing the high-resolution features of lower-layers. On two widely-used benchmarks (i.e., VOC2007 and MS-COCO), CAT provides a stable improvement over the baseline and produces a competitive performance.

Index Terms— Cross-layer aggregation, transformers, multi-label image classification

1. INTRODUCTION

Multi-label image classification (MLIC) is a fundamental task of computer vision, and significance improvements have been produced by convolutional neural network (CNN) benefiting from the strong feature extraction ability. However, this task still faces some unique challenges (e.g., spatial dependencies and various sizes of object), compared with the conventional single-label image classification.

Recently, a considerable amount of works [1, 2, 3, 4, 5] have been proposed to capture the spatial dependencies and position aware information by designing a variety of spatial and channel attention modules based on CNN. The fixed kernel size in CNN handles local dependencies well, while fails to adapt to the objects of various sizes or formulate the dependencies between two objects with long distance, as illustrated in Fig. 1. In MLIC community [6, 7], multi-layer feature fusion is an effective way for enhancing the image feature representation ability to accommodate the objects of different sizes. In [6], a companion network with fusion module is proposed to aggregate multi-layer feature from the main network. In [7], a multi-layered semantic representation network is designed to utilize label semantics for learning global semantics on different layers, respectively. While the above methods ignore the correlation or dependencies between layers. This paper focuses on how to learn the discriminative features of long range dependencies with cross-layer fusion, simultaneously.

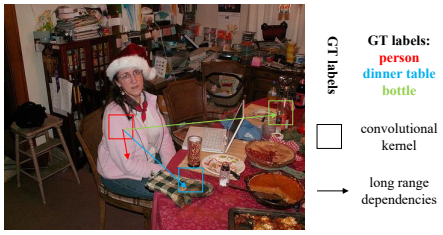

Fig. 1. Fixed CNN kernel size fails to adapt to the objects of various sizes or formulate the dependencies between two objects with long distance.

For long range dependencies modeling, multi-head attention (MHA) in transformers [8] is the first attempt in Natural Language Processing (NLP) community, and gradually improves the performance of various computer vision tasks. By calculating the dependencies between image features at different positions, the limited attention caused by fixed convolution kernel size could be alleviated effectively. In this paper, we make an attempt of introducing MHA into MLIC task. In addition, the image features are transmitted between adjacent layers through convolution layers, the correlation between layers should be considered seriously. Inspired by this, we propose a Cross-layer Aggregation with Transformers (CAT) framework which leverages transformers to capture the spatial dependencies of CNN-based features with Long

✉ Corresponding author: Fuqing Zhu

This research is supported in part by the National Key R&D Program of China under Grant 2020AAA0140000.

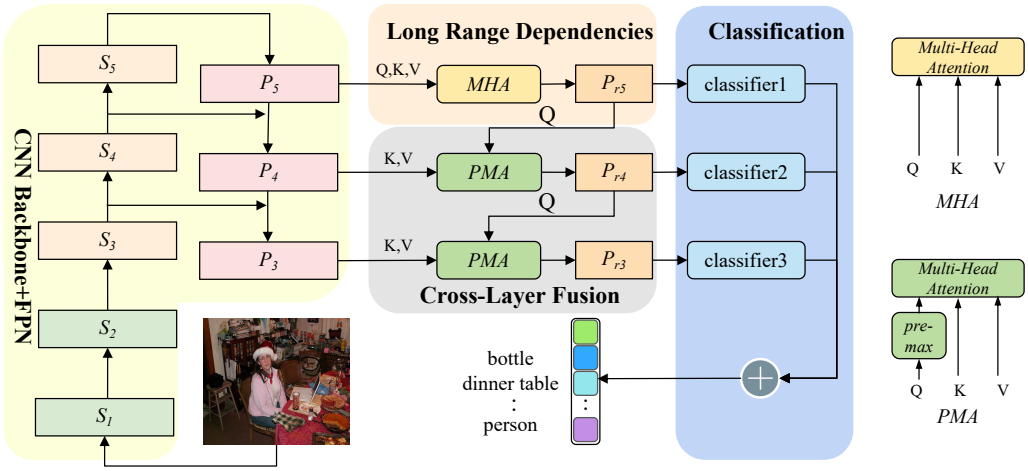

Fig. 2. The overall framework of the proposed CAT for multi-label image classification.

Range Dependencies module and aggregate the features layer by layer with Cross-Layer Fusion module. First, multi-layer image features with strong semantics are obtained leveraging a CNN backbone to extract features from multiple layers, which are then fed into Feature Pyramid Network (FPN) [9]. Subsequently, the features with long range dependencies could be obtained by applying a vanilla MHA to the highest-layer features. Finally, the cross-layer fusion is conducted to aggregate the multi-layer features layer by layer in a top-down manner for classification. Besides, to fuse the high-resolution features of lower-layers efficiently, we design a multi-head pre-max attention in cross-layer fusion to replace the vanilla MHA in transformers. In this way, the computational complexity is linearly discounted. The contributions of this paper are summarized as follows:

- A Cross-layer Aggregation with Transformers (CAT) framework is proposed to learn the discriminative features of long range dependencies with cross-layer fusion.
- A multi-head pre-max attention is designed to reduce the computation cost when fusing the high-resolution features of lower-layers.
- Experimental results on two widely-used benchmarks (i.e., VOC2007 [10] and MS-COCO [11]) demonstrate that CAT provides a stable improvement over the baseline and produces a competitive performance.

2. METHODOLOGY

2.1. Overview

In this paper, a Cross-layer Aggregation with Transformers (CAT) framework, which consists of a Long Range Dependencies (LRD) module and a Cross-Layer Fusion (CLF) module, is proposed as illustrated in Fig. 2. First, multi-layer image features with strong semantics are obtained by leveraging a pre-trained CNN (ResNeXt50-swsl [12]) to extract multiple layers features followed by FPN. Subsequently, the features with long range dependencies are obtained by applying a vanilla MHA to the higest-layer features. Finally, the cross-layer fusion is conducted to aggregate the multi-layer features layer by layer in a top-down manner for classification, where a multi-head pre-max attention is designed to make the framework efficient.

2.2. Problem Definition

Given an image in the train set D denoted as x, multi-label image classification task aims predict a L-dim binary vector $y = [y^1, y^2, ..., y^l, ..., y^L]$. 0 or 1 is assigned to y^l for the absence or presence of the l-th label.

2.3. Framework

Multi-Layer Image Features. In this paper, ResNeXt50-swsl [12] is adopted as the backbone which outputs five stages features by each stage's last convolution block, denoted as $\{S_{1-5}\}$. Given an input image $x \in R^{3\times H\times W}$, where H and W are the origin height and width of the image, we first extract features of the last three stages $\{S|S_i\}$, where $i \in \{3, 4, 5\}$ denotes the i-th stage. We do not include S_1 and S_2 due to the large memory accounts. At this time, the S_4, S_5 are semantically weak and direct feature aggregation may harm representational capacity for classification. Thus, semantically strong features of all layers $\{P_3, P_4, P_5\}$ could be obtained by passing S to FPN which utilizes lateral connections to strength the semantics of S_4 and S_5. Specifically, the resolution of each P_i is $\frac{H}{32}\times\frac{W}{32}$, $\frac{H}{16}\times\frac{W}{16}$ and $\frac{H}{8}\times\frac{W}{8}$, respectively. The channel number of each P_i is 1024.

Long Range Dependencies (LRD). Capturing long range dependencies is a process of feature optimization by connecting the features of any position with all other positions in the feature map. In this paper, we leverage a two-layer transformer encoders to capture the spatial long range dependencies of the last stage features. There are two considerations: i) the resolution of the last stage features is relatively small within an acceptable range of computation cost, while the resolution of lower-layer features is two times of the upper-layer feature, which may cause large amounts of computational complexity. ii) The lack of lower-level long range dependencies can be mitigated by Cross-Layer Fusion as described in the following sections. MHA can obtain independent attention attending to parts of the sequence differently in parallel. Specifically, we adopt P_5 as the *query* (Q), *key* (K) and *value* (V), and the refined feature P_{r5} with long range dependencies is formulated as follow

$$MHA(P_5, P_5, P_5) = [head_1, .head_h, .head_H]W_o, \quad (1)$$

where W_o is a linear projection. $[\cdot]$ represents concatenation alongs the channel. H is the number of heads and set to 8 by default.

$$head_h = softmax((W_h^q P_5)^T (W_h^k P_5)/\sqrt{d}), \quad (2)$$

where W_h^q and W_h^k are two linear projection functions, T denote matrix transposition and $d = C/H$.

$$P_{r5} = FFN(P_5 MHA(P_5)), \quad (3)$$

where FFN consists of two linear transformations with a ReLU activation. In this way, P_{r5} could integrate the local advantages of CNN and the long range dependencies advantages of transformer to perceive information relevant to classification in images. P_{r5} has two functions: the first is to be directly fed into a independent classifier, and the second is to aggregate features layer by layer as the initial *query* feature.

Cross-Layer Fusion (CLF). Cross-layer fusion is proposed to aggregate multi-layer features in a top-down manner. The output features of each layer are fed into a independent classifier for classification and also sent as *query* to the lower-layer as *key* and *value*. Since CAT needs to deal with high-resolution features of lower-level, we propose a multi-head pre-max attention (PMA) to replace the vanilla MHA to make the aggregation efficient. The PMA is denoted as follow

$$PMA(P_{i+1}, P_i, P_i) = [head_1, .head_h, .head_H]W_t, \quad (4)$$

$$head_h = softmax((W_h^{i+1} PM(P_{i+1}))^T (W_h^i P_i)/\sqrt{d}), \quad (5)$$

where PM denotes the global maximization operation on the query. Given the resolution of P_{i+1} and P_i is $c \times h \times w$ and $c \times 2h \times 2w$ respectively, the corresponding computation complexity of MHA and PMA is

$$\Omega(MHA) = 8(hw)^2 c, \quad (6)$$

$$\Omega(PMA) = 8hwc \quad (7)$$

where the computation complexity is $h \times w$ times lower than vanilla MHA. Through cross-layer fusion, the model can aggregate multiple layers of effective semantic features to fit variable-sized objects, so as to facilitate label prediction.

Classifier and Loss. The output features of each CLA layers are fed into three independent fully connected layers respectively and then sum up to get the final prediction $\hat{y}$. The binary entropy loss function $\mathcal{L}_{loss}$ is adopted to optimize the model.

$$\mathcal{L}_{loss} = -\frac{1}{L}\sum(y^l * \log(\hat{y}^l)) + (1 - y^l) * \log(1 - \hat{y}^l)), \quad (8)$$

where y^l and $\hat{y}^l$ denote the ground-truth label and predicted label at *l*-th position, respectively.

3. EXPERIMENTS

3.1. Evaluation Metrics

This paper follows the same evaluation metrics in [13] for fair comparison, i.e., overall precision (OP), overall recall (OR), overall F1-score (OF1), per-class precision (CP), per-class recall (denoted as CR), per-class F1-score (CF1) and mean Average Precision (mAP) for MS-COCO and Average Precision (AP) for VOC2007.

3.2. Implement Details

We implement the experiments with pytorch [1] and a NVIDIA Tesla V100 GPU with 32G memory. ResNeXt50-swsl [12] is adopted as the backbone. During training, we choose the multi-crop augmentation strategy as the same in [13]. Input images are first resized to 512×512 and randomly cropped to 448×448. SGD is adopted as the optimizer with momentum 0.9. Weight decay is 10^{-4}. The learning rate is 0.01, which decays by a factor 10 every 30 epochs with an equal interval strategy. The batch size is 16. Training ends after 60 epochs. During testing, images are resized to 448×448 directly. We set a strong baselines by using ResNeXt50-swsl [12], which provides an excellent performance for image classification.

3.3. Experimental Results

Performance on VOC2007. VOC2007 is a well-know dataset, including totally $9,963$ images of 20 categories. We use the official train/test split, where $5,011$ images are used for training, and $4,952$ for testing. To verify the effectiveness of our framework, we also compare CAT with other state-of-the-art approaches, such as HCP [14], CNN-RNN [15], RLSD [16], ML-GCN [13], MSRN [7], ADD-GCN [17] and ASL [18]. AP and mAP on VOC2007 are shown in the Table 1. Compared to the baseline, mAP is improved by

[1] https://pytorch.org

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
HCP [14]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
CNN-RNN [15]	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
RLSD [16]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
ML-GCN [13]	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
MSRN	100.0	98.8	98.9	99.1	81.6	95.5	98.0	98.2	84.4	96.6	87.5	98.6	98.6	97.2	99.1	87.0	97.6	86.5	99.4	94.4	94.9
ADD-GCN [17]	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
ASL [18]	99.9	98.4	98.9	98.7	86.8	98.2	98.7	98.5	83.1	98.3	89.5	98.8	99.2	98.6	99.3	89.5	99.4	86.8	99.6	95.2	95.8
ResNeXt50-swsl [12]	99.7	96.6	99.2	98.4	83.4	95.8	97.4	98.5	83.5	97.2	87.4	99.0	99.0	96.6	98.8	83.2	97.5	84.2	99.3	93.5	94.4
CAT(ours)	99.8	98.8	99.2	98.3	86.2	97.2	98.2	99.1	87.3	98.1	91.4	99.5	99.0	98.1	99.3	89.0	99.2	89.4	99.5	97.1	96.2

Table 1. Quantitative results (%) on the VOC2007 dataset.

Methods	All						
	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN [15]	61.2	-	-	-	-	-	-
SRN [1]	77.1	81.6	65.4	71.2	82.7	69.9	75.8
ML-GCN [13]	83.0	85.1	72.0	78.0	85.8	75.4	80.3
DecoupleNet[2]	82.2	83.1	71.6	76.3	84.7	74.8	79.5
MSRN [7]	83.4	86.5	71.5	78.3	86.1	75.5	80.4
ADD-GCN [17]	85.2	84.7	75.9	80.1	84.9	79.4	82.0
ASL [18]	86.5	87.2	**76.4**	81.4	88.2	79.2	81.8
MGTN ([19])	87.0	86.1	77.9	81.8	87.7	79.4	83.4
ResNeXt50-swsl [12]	83.3	84.8	73.1	78.5	86.5	76.3	81.1
CAT(ours)	**87.4**	**88.5**	76.1	**81.9**	**88.6**	**79.4**	**83.7**

Table 2. Quantitative results (%) on the MS-COCO dataset.

Methods	MS-COCO	VOC2007
CAT(ours)	**87.4**	**96.2**
CAT w/o LRD	86.6	95.6
CAT w/o PMA	OOM	OOM
CAT w/o CLF	86.8	95.8
Backbone	83.3	94.4
Backbone+FPN	84.6	95.6

Table 3. Ablation study (%) on MS-COCO and VOC2007, respectively. OOM represents out-of-memory.

+**1.8%** (from 94.4% to 96.2%). We achieve 96.2% which is comparable to the state-of-the-art method. The comparison with other methods also demonstrates the superiority of CAT.

Performance on MS-COCO. MS-COCO is another widely-used dataset which includes 122,218 images with 80 labels, 82,787 images for training and 40,504 validation images for testing. Besides the baseline, we also set a numerous of methods, including CNN-RNN [15], SRN [1], ML-GCN [13], DecoupleNet [2], MSRN [7], ADD-GCN [17], ASL[18] and MGTN[19]. As shown in Table 2, CAT consistently outperforms the baseline and other state-of-the-art methods in terms of OF1, CF1, and mAP, as well as some other less important metrics. Specifically, CAT improves the mAP of baseline by a large margin +**4.1%** (from 83.3% to 87.4%), which shows the superiority of CAT.

Fig. 3. Visualization of class activation map for each label with probability value. The red font represents missing labels.

3.4. Ablation Study and Visualization

To verify the effectiveness of each module in the proposed CAT, we conducted ablation experiments. We also visualize of the class activation map [20] for each category with the baseline and CAT. The contribution of LRD, CLA module can be observed with the comparison results shown in Table 3. PMA is proved efficient with CAT w/o PMA. Compared with the baseline, CAT learns more discriminative features for each category, as shown in the Figure 3. Specifically, CAT can locate both the *persons* in the first image clearly, and predict the *hair drier* correctly which is missed by the baseline in the second images.

3.5. Conclusion

In this paper, we propose a Cross-Layer Aggregation with Transformers (CAT) framework to capture the long range dependencies and fuse the multi-layer feature via cross-layer fusion. In addition, we design a multi-head pre-max attention to linearly reduce the computation cost for training the framework efficiently. Experimental results on two widely-used benchmarks demonstrate that CAT provides a stable improvement over the baseline and produces a competitive performance.

4. REFERENCES

[1] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. CVPR*, 2017, pp. 5513–5522.

[2] Luchen Liu, Sheng Guo, Weilin Huang, and Matthew R Scott, "Decoupling category-wise independence and relevance with self-attention for multi-label image classification," in *Proc. ICASSP*, 2019, pp. 1682–1686.

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

[4] Yan Luo, Ming Jiang, and Qi Zhao, "Visual attention in multi-label image classification," in *Proc.CVPRW*. IEEE Computer Society, 2019, pp. 820–827.

[5] Zheng Yan, Weiwei Liu, Shiping Wen, and Yin Yang, "Multi-label image classification by feature attention network," *IEEE Access*, vol. 7, pp. 98005–98013, 2019.

[6] Y. Niu, Z. Lu, J. R. Wen, T. Xiang, and S. F. Chang, "Multi-modal multi-scale deep learning for large-scale image annotation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1720–1731, 2019.

[7] Xiwen Qu, Hao Che, Jun Huang, Linchuan Xu, and Xiao Zheng, "Multi-layered semantic representation network for multi-label image classification," *arXiv preprint arXiv:2106.11596*, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proc.CVPR*, 2017, pp. 2117–2125.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[11] 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 *Proc. ECCV*, 2014, pp. 740–755.

[12] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.

[13] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. CVPR*, 2019, pp. 5177–5186.

[14] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1901–1907, 2015.

[15] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proc. CVPR*, 2016, pp. 2285–2294.

[16] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu, "Multilabel image classification with regional latent semantic dependencies," *IEEE Transaction on Multimedia*, vol. 20, no. 10, pp. 2801–2813, 2018.

[17] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Proc. ECCV*, 2020, pp. 649–665.

[18] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proc. ICCV*, 2021.

[19] Hoang D Nguyen, Xuan-Son Vu, and Duc-Trong Le, "Modular graph transformer networks for multi-label image classification," in *Proc. AAAI*, 2021, pp. 9092–9100.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921–2929.