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, *InstituteofInformationEngineering, ChineseAcademyofSciences, China*

2, *SchoolofCyberSecurity, UniversityofChineseAcademyofSciences, China*

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.

Motivation

1. Label objects in the image are distributed in different positions, and there may be a certain spatial distance relationship. However, **the existing CNN-based methods are limited by the size of the convolution kernel, and it is difficult to capture long-range spatial dependencies.**
2. Label objects in the image are **variable-sized objects**, existing multi-scale fusion generally adopts a simple fusion method, and **rarely considers the relationship between the features of different layers**.

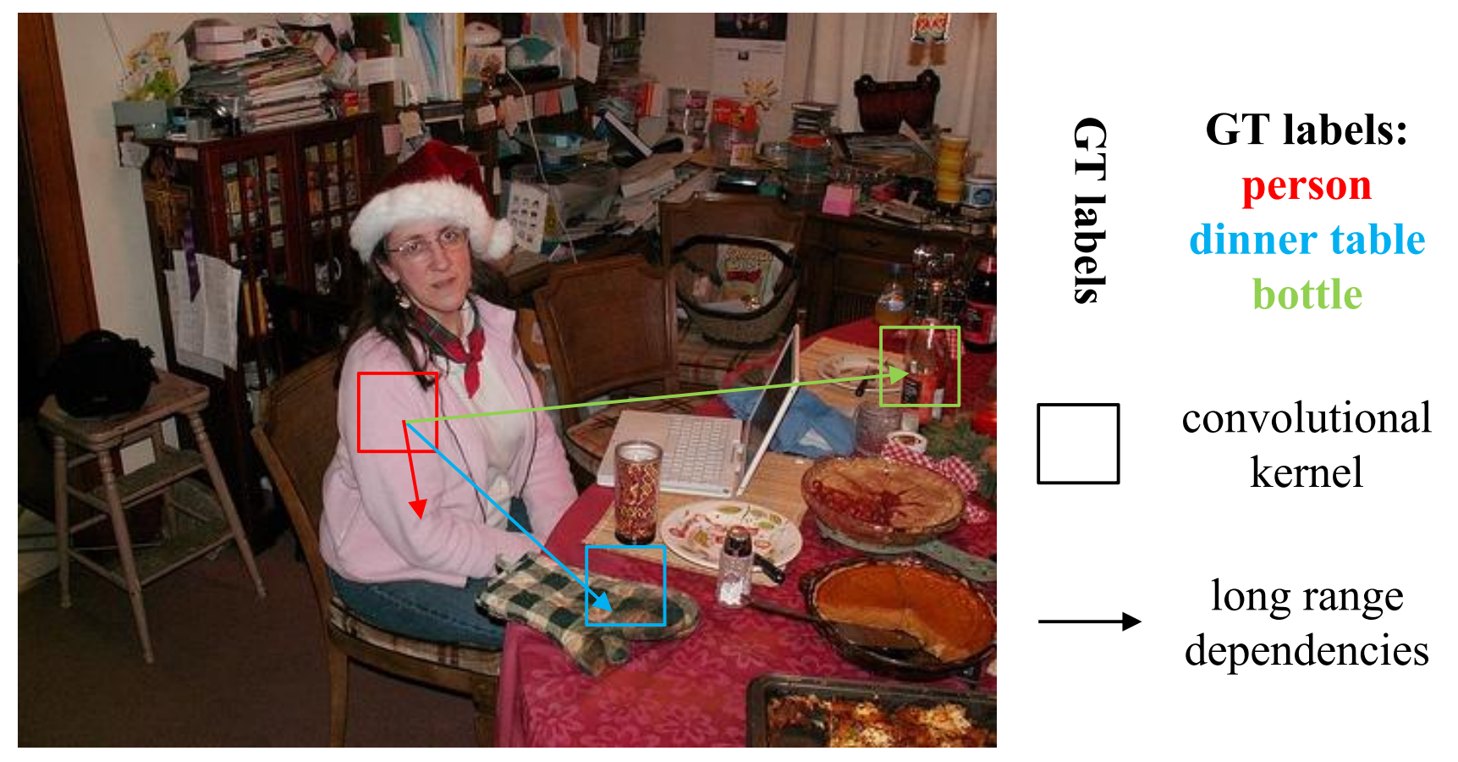

Figure 1: Fixed CNN kernel size fails to adapt to the objects of various sizes or formulate the dependencies between two objects with long distance.

Contributions

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

Methodology

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) 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.

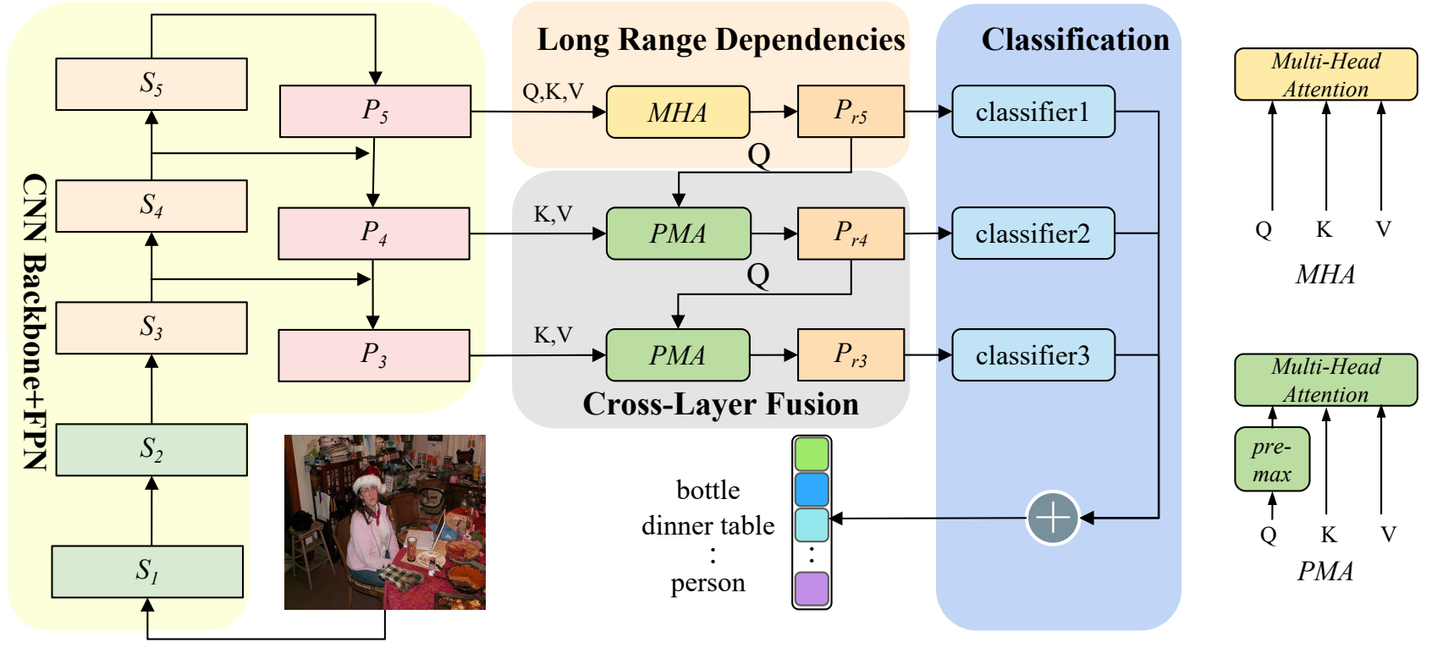

Figure 2: The overall framework of the proposed CAT for multi-label image classification.

Image Encoder

Multi-layer image features $\{P_3, P_4, P_5\}$ are obtained by extracting features with multiple layers of the neural network and passing through the Feature Pyramid Network.

$$P5 = conv_1(S5), \tag{1}$$

$$P4 = S4 + conv_2(P5), \tag{2}$$

$$P3 = S3 + conv_2(P4), \tag{3}$$

Long Range Dependencies

This paper utilizes vanilla multi-head self-attention mechanism in Transformers to obtain long-range dependencies between the current position and other positions.

$$P_{r5} = FFN(P_5MHA(P_5)), \tag{4}$$

Cross Layer Fusion

The features of the upper layer are used as query, and the features of the next layer are used as key and value for fusion between layers. A multi-head pre-max attention is designed in cross-layer fusion to replace the vanilla multi-head self-attention in Transformers.

$$P_{r4} = CLA(P_{r5}, P_4), \tag{5}$$

$$P_{r3} = CLA(P_{r4}, P_3), \tag{6}$$

$$CLA = FFN(PMA(I_1, I_2, I_2)), \tag{7}$$

Classification

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 $\hat{y}$.

$$\hat{y}_1 = mlp_1(P_{r5}) \tag{8}$$

$$\hat{y}_2 = mlp_2(P_{r4}) \tag{9}$$

$$\hat{y}_3 = mlp_3(P_{r3}) \tag{10}$$

$$\hat{y} = \alpha\hat{y}_1 + \beta\hat{y}_2 + \gamma\hat{y}_3 \tag{11}$$

$$\mathcal{L}_{loss} = -\frac{1}{L}\sum(y^l * \log(\hat{y}^l)) + (1 - y^l) * \log(1 - \hat{y}^l)), \tag{12}$$

Experiments

We conduct experiments on two of the most commonly used multi-label image classification datasets, namely, Pascal VOC2007 and MS-COCO datasets. And to verify the effectiveness of each module in the proposed CAT, we conducted ablation experiments. We also visualize of the class activation map for each category with the baseline and CAT.

Comparison with SOTA baselines

Methods	mAP
HCP	90.9
CNN-RNN	84.0
RLSD	88.5
ML-GCN	94.0
MSRN	94.9
ADD-GCN	96.0
ASL	95.8
ResNeXt50-swsl	94.4
CAT(ours)	**96.2**

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

Methods	All						
	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN	61.2	-	-	-	-	-	-
SRN	77.1	81.6	65.4	71.2	82.7	69.9	75.8
ML-GCN	83.0	85.1	72.0	78.0	85.8	75.4	80.3
DecoupleNet	82.2	83.1	71.6	76.3	84.7	74.8	79.5
MSRN	83.4	86.5	71.5	78.3	86.1	75.5	80.4
ADD-GCN	85.2	84.7	75.9	80.1	84.9	79.4	82.0
ASL	86.5	87.2	**76.4**	81.4	88.2	79.2	81.8
MGTN	87.0	86.1	77.9	81.8	87.7	79.4	83.4
ResNeXt50-swsl	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.

CAT outperforms other baseline models, especially on mAP. Among them, the AP on the Pascal VOC2007 dataset, please refer to the results in the paper.

Ablation Study

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.

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.

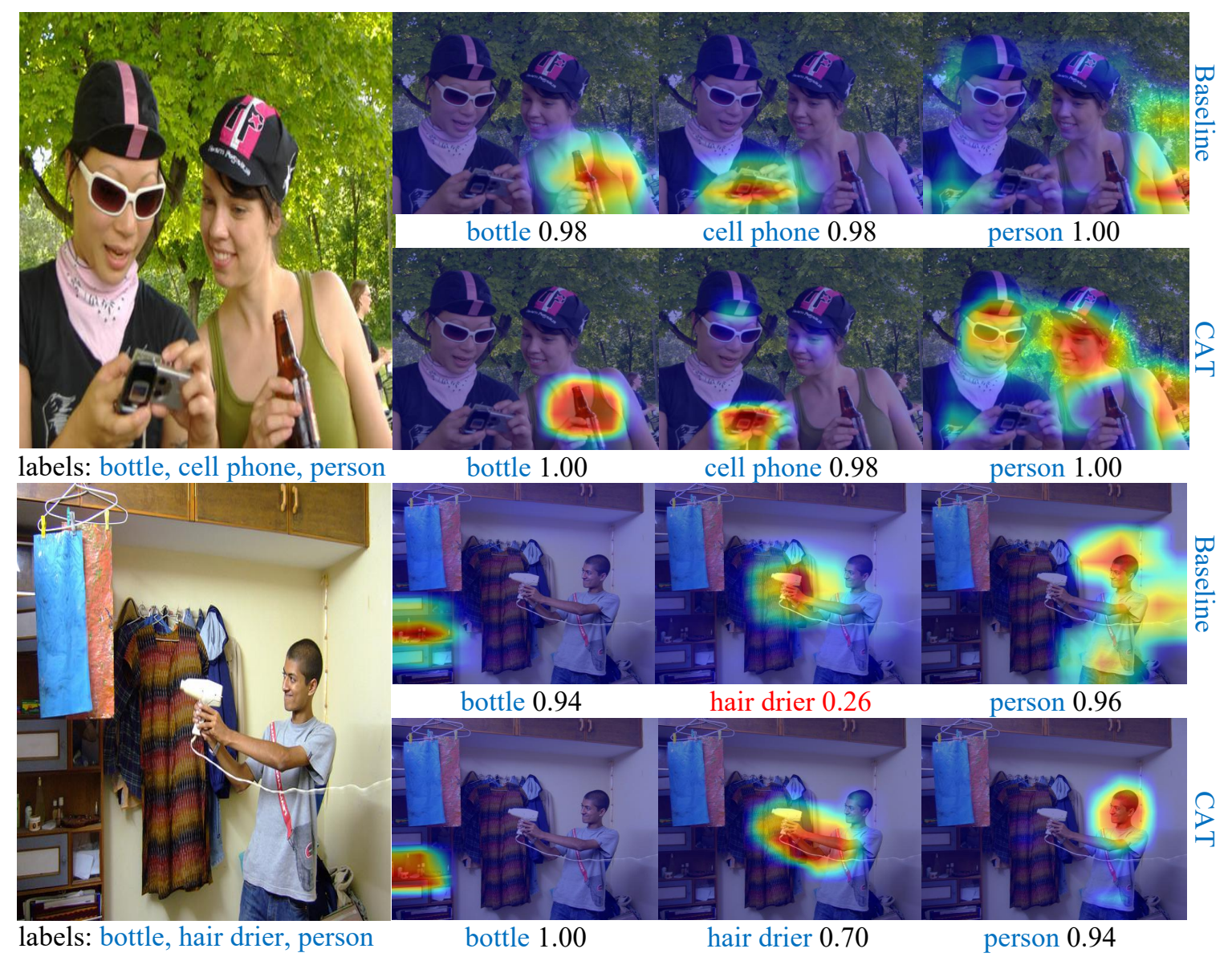

Figure 3: Visualization of class activation map for each label with probability value. The red font represents missing labels.

Compared with the baseline, CAT learns more discriminative features for each category, as shown in the Figure **??**. 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.

Conclusions

- 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 widelyused benchmarks demonstrate that CAT provides a stable improvement over the baseline and produces a competitive performance.