Paper ID: 70

BMVC 2022

Re-Attention Transformer for Weakly Supervised Object Localization

Hui Su[1] Yue Ye[1] Zhiwei Chen[1] Mingli Song[2] Lechao Cheng [1*]

[1] Zhejiang Lab, Hangzhou, China

[2] Zhejiang University, Hangzhou, China

之江实验室 ZHEJIANG LAB

浙江大学 ZHEJIANG UNIVERSITY

Motivation

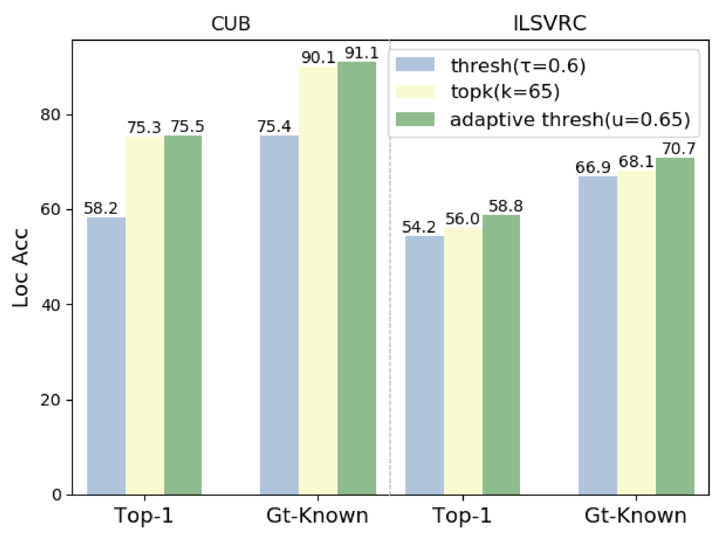
- Weakly supervised object localization is a challenging task which aims to localize objects with coarse annotations such as image categories.
- Existing deep network approaches are mainly based on class activation map, which focuses on highlighting discriminative local region while ignoring the full object.
- Emerging transformer-based techniques constantly put a lot of emphasis on the backdrop that impedes the ability to identify complete objects.

Contributions

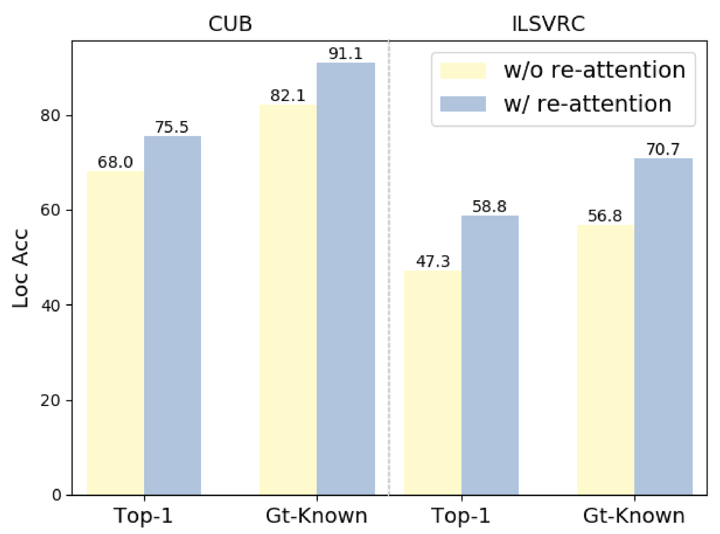
- we propose a re-attention mechanism termed token refinement transformer (TRT) which highlights the precise object of interest.
- we propose an adaptive thresholding strategy based on sampling over cumulative importance that improves the performance significantly in the task of WSOL.
- The experimental results show convincing results of both qualitative and quantitative compared to existing approaches on ILSVRC and CUB-200-2011

Ablations

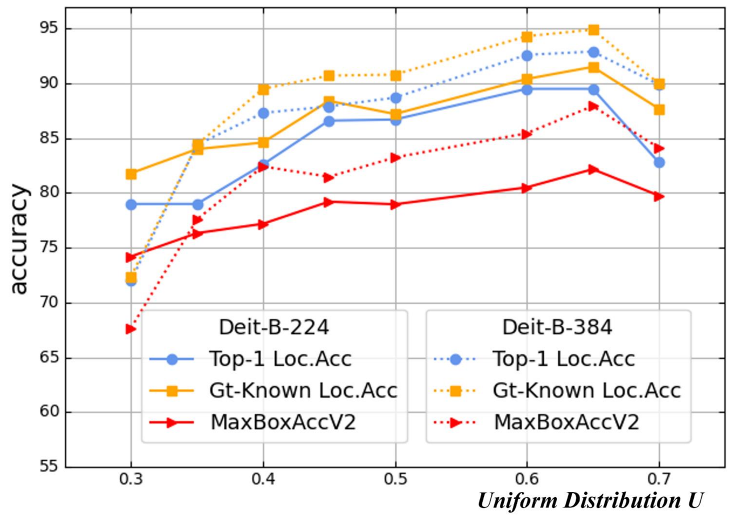
Token Selection Strategies



With or Without Token Re-Attention



Impact Of Uniform Distribution



(I) Input (II) No Re-attention (III) Re-attention

Method

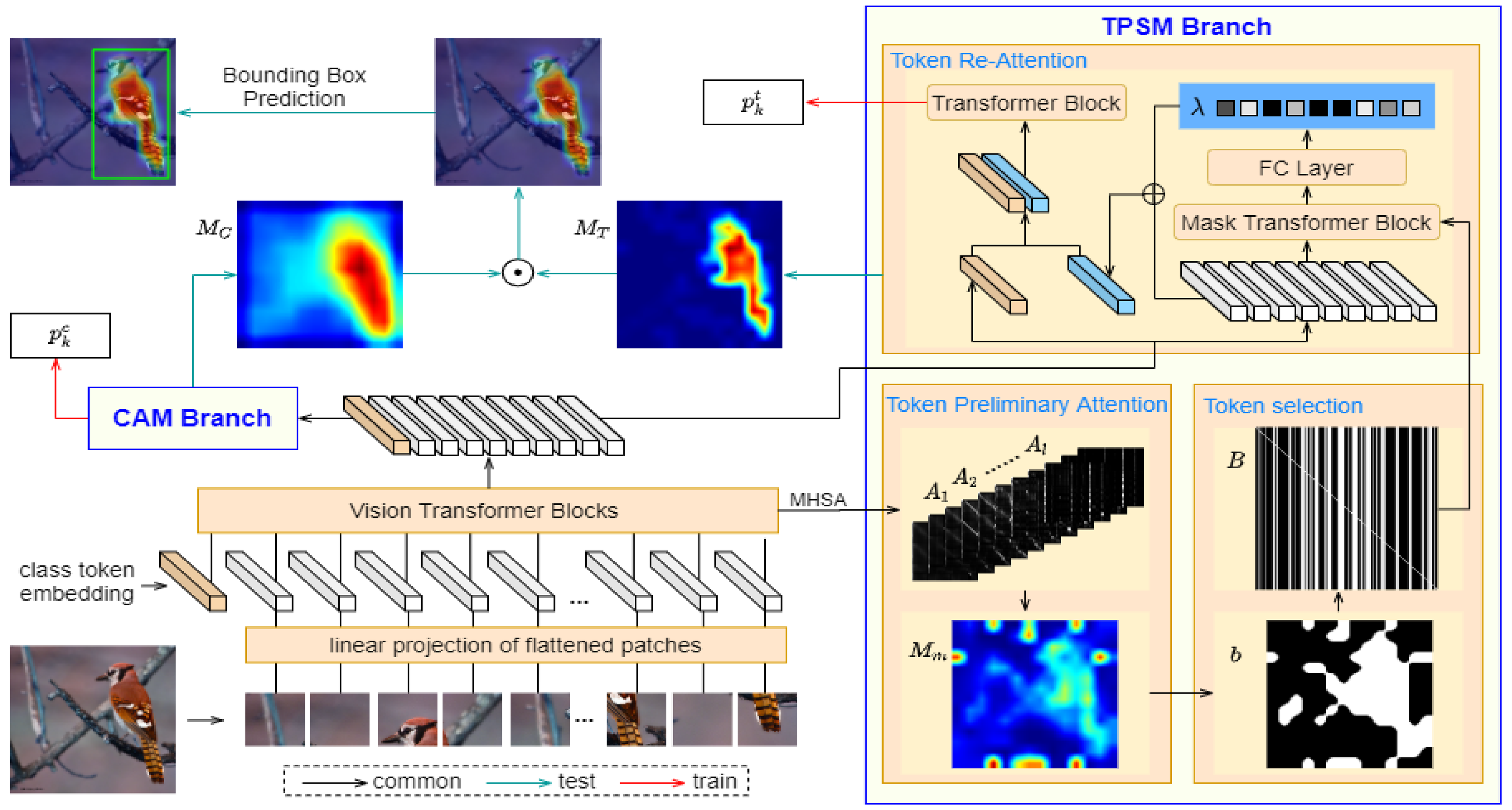


Figure 2: Token Refinement Transformer (TRT) framework. TRT consists of two branches, Token Priority Scoring Module (TPSM) and CAM, respectively. TPSM attempts to generate context-aware features $\mathbf{M}_T$ that contribute most to the target class. CAM is introduced to obtain discriminative features $\mathbf{M}_C$. We finally get the attention map $\mathbf{M}$ by performing element-wise multiplication as $\mathbf{M} = \mathbf{M}_C \odot \mathbf{M}_T$.

Token Preliminary Attention

First, we generate a preliminary attention map by exploiting long-range dependencies of class token and patch tokens over transformer blocks

$$\mathbf{A}_l = softmax\left(\frac{\mathbf{Q}_l \mathbf{K}_l^{\mathbf{T}}}{\sqrt{D}}\right) \qquad \mathbf{m} = \textstyle\sum_{l=1}^{L-1} \mathbf{A}_l[0,1:]$$

Token Selection

Then, an adaptive thresholding strategy is introduced to screen out patch tokens with high response in preliminary attention map

$$F(x) = \mathbf{P}_r(\mathbf{m} < x) = \mathbf{P}_r(\mathbb{T}(U) < x) = \mathbf{P}_r(U < \mathbb{T}^{-1}(x)) = \mathbb{T}^{-1}(x)$$

We sort values in m from high to low and calculate cumulative attention distribution with function F. We set fixed u as threshold of cumulative attention distribution , adaptive threshold r' is obtained based on T.

Token Re-Attention

Finally, we perform re-attention operation on the selected tokens to capture more effective global relationships.

$$\mathbf{B} = \mathbf{J} \otimes \mathbf{b} + \mathbf{J} \otimes (\mathbf{J}^{\mathbf{T}} - \mathbf{b}) \odot \mathbf{I}_N \qquad \mathbf{S} = \frac{\mathbf{Q}_{L-1}\mathbf{K}_{L-1}^{\mathbf{T}}}{\sqrt{D}} \qquad \mathbf{A}_{ij}^r = \frac{\exp(\mathbf{S}_{ij}) * \mathbf{B}_{ij}}{\sum_{k=1}^{N} \exp(\mathbf{S}_{ik}) * \mathbf{B}_{ik}}$$

$$r = \frac{\sum_{k=1}^{N} \mathbf{m}_k * \mathbf{b}_k}{\sum_{k=1}^{N} \lambda_k} \qquad \mathbf{m}' = \mathbf{m} \odot (\mathbf{J}^{\mathbf{T}} - \mathbf{b}) + \lambda * r$$

Experiments

Performance Comparision

Methods	Backbone	*Loc.Acc*		
		Top-1	Top-5	Gt-Known
CAM[53]	VGG16	44.2	52.2	56.0
SPG[51]	VGG16	48.9	57.2	58.9
SLT-Net[18]	VGG16	67.8	-	87.6
CAM[53]	InceptionV3	41.1	50.7	55.1
SPG[51]	InceptionV3	46.7	57.2	-
I2C[52]	InceptionV3	66.0	68.3	72.6
SLT-Net[18]	InceptionV3	66.1	-	86.5
TS-CAM[16]	Deit-B	75.8	84.1	86.6
TS-CAM*[16]	Deit-B-384	77.8	88.6	90.8
TRT(Ours)	Deit-B	76.5	88.0	91.1
TRT(Ours)	Deit-B-384	80.5	91.7	94.1

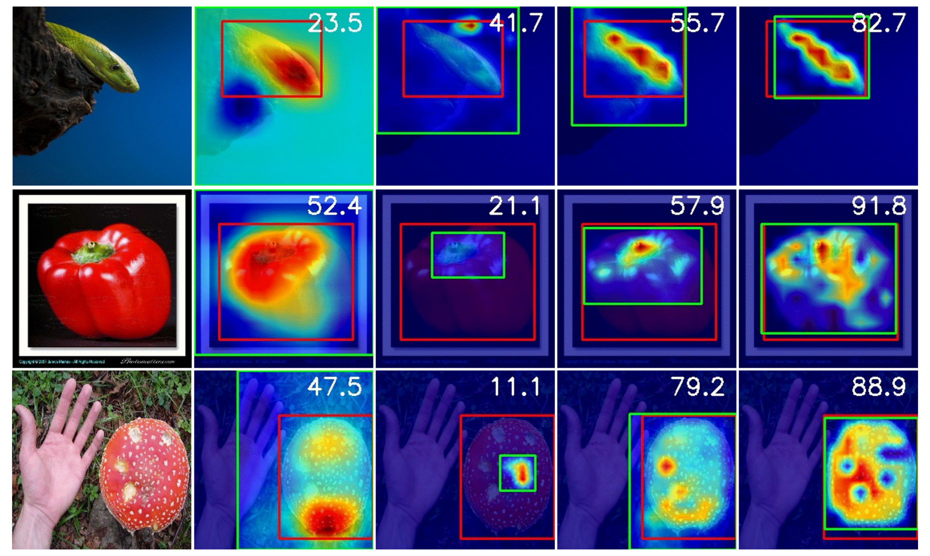
Table 1: Experimental results on CUB-200-2011 for metrics of *Loc.Acc*.

Methods	Backbone	MaxBoxAccV2
CAM[53]	VGG	63.70
ADL[7]	VGG	66.30
VITOL[19]	Deit-B	73.17
TS-CAM*[16]	Deit-B	76.74
TRT(Ours)	Deit-B	82.08

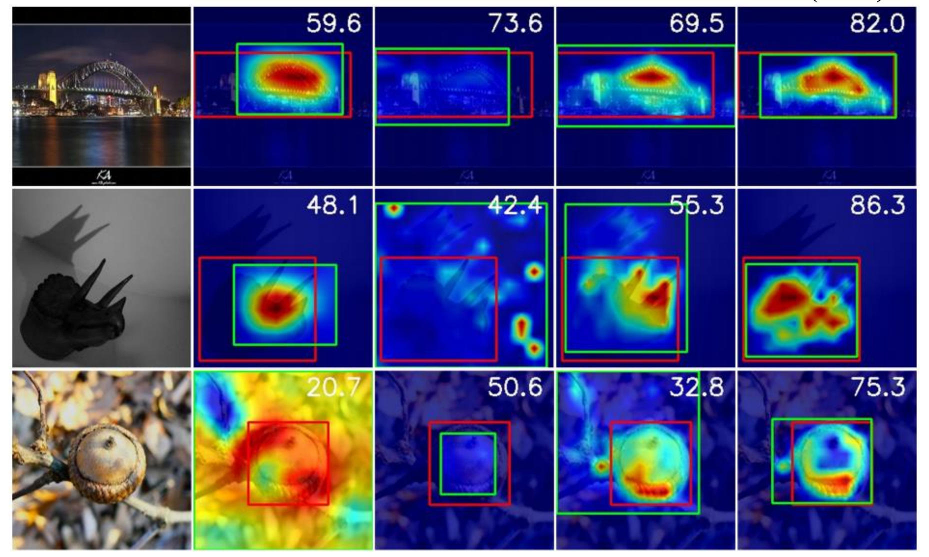
Table 2: Experimental results on CUB-200-2011 for MaxBoxAccV2

Methods	Backbone	Loc.Acc		
		Top-1	Top-5	Gt-Known
CAM[53]	VGG16	42.8	54.9	59.0
ADL[7]	VGG16	44.9	-	-
SLT-Net[18]	VGG16	51.2	62.4	67.2
CAM[53]	InceptionV3	46.3	58.2	62.7
ADL[7]	InceptionV3	48.7	-	-
SLT-Net[18]	InceptionV3	55.7	65.4	67.6
TS-CAM*[16]	Deit-B	47.8	60.0	64.4
LCTR*[6]	Deit-B	53.4	63.9	67.1
TRT(ours)	Deit-B	58.8	68.3	70.7

Table 3: Experimental results on ILSVRC for metrics of Loc.Acc.

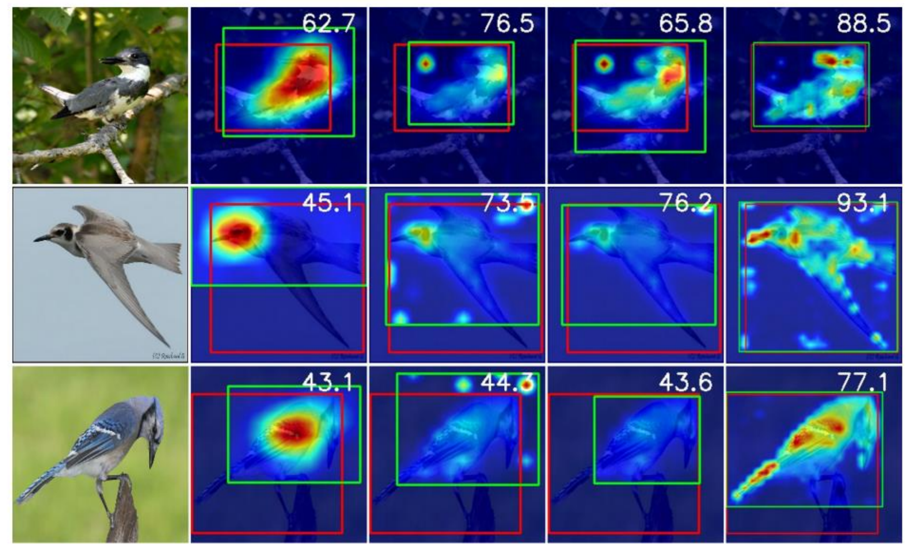


(I) Input (II) CAM (III) Transformer (IV) Transformer-CAM (V) TRT (Ours)



(I) Input (II) CAM (III) Transformer (IV) Transformer-CAM (V) TRT (Ours)

ILSVRC



(I) Input (II) CAM (III) Transformer (IV) Transformer-CAM (V) TRT (Ours)

CUB-200-2011

- Table 1 and Table 2 showcase the competitive results of our proposed TRT framework on the CUB-200-2011.
- Table 3 demonstrated that TRT is superior to both existing CAM-based and transformer-based approaches on ILSVRC.
- Pictures on the right show visualization of localization maps on CUB-200-2011 and ILSVRC datasets. Red means ground truth and green means predicted bounding box.

Code: https://github.com/su-hui-zz/ReAttentionTransformer.