TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization

Wei Gao¹ Fang Wan¹ Xingjia Pan²,³ Zhiliang Peng¹ Qi Tian⁴
Zhenjun Han¹ Bolei Zhou⁵ Qixiang Ye¹*
¹University of Chinese Academy of Sciences ²YouTu Lab, Tencent
³NLPR, Institute of Automation, CAS ⁴Huawei Cloud AI
⁵The Chinese University of Hong Kong
{vasgaowei, xjia.pan}@gmail.com, {wanfang, hanzhj, qxye}@ucas.ac.cn
tian.qi1@huawei.com, bzhou@ie.cuhk.edu.hk

Abstract

Weakly supervised object localization (WSOL) is a challenging problem when given image category labels but requires to learn object localization models. Optimizing a convolutional neural network (CNN) for classification tends to activate local discriminative regions while ignoring complete object extent, causing the partial activation issue. In this paper, we argue that partial activation is caused by the intrinsic characteristics of CNN, where the convolution operations produce local receptive fields and experience difficulty to capture long-range feature dependency among pixels. We introduce the token semantic coupled attention map (TS-CAM) to take full advantage of the self-attention mechanism in visual transformer for long-range dependency extraction. TS-CAM first splits an image into a sequence of patch tokens for spatial embedding, which produce attention maps of long-range visual dependency to avoid partial activation. TS-CAM then re-allocates category-related semantics for patch tokens, enabling each of them to be aware of object categories. TS-CAM finally couples the patch tokens with the semantic-agnostic attention map to achieve semantic-aware localization. Experiments on the ILSVRC/CUB-200-2011 datasets show that TS-CAM outperforms its CNN-CAM counterparts by 7.1%/27.1% for WSOL, achieving state-of-the-art performance. Code is available at https://github.com/vasgaowei/TS-CAM

1. Introduction

Weakly supervised learning refers to methods that utilize training data with incomplete annotations to learn recognition models. Weakly supervised object localization (WSOL) solely requires the image-level annotations indicating the presence or absence of a class of objects in images to learn localization models [23, 24, 28, 46]. WSOL has attracted increasing attention as it can leverage the rich Web images with tags to learn object-level models [46].

As the cornerstone of WSOL [7], the Class Activation Mapping (CAM) [55] utilizes the activation map from the last convolution layer to generate semantic-aware localization maps for object bounding-box estimation. However, CAM suffers from severe underestimation of object regions because the discriminative regions activated through the classification models are often much smaller than objects’ actual extent [2]. Local discriminative regions are capable of minimizing image classification loss, but experience...
difficulty for accurate object localization [46], Fig. 1(b). Much effort has been made to solve this problem by proposing various regularizations [50, 51, 46, 20, 21, 6], divergent activation [33, 46, 48] or adversarial training [8, 21, 43, 50, 48, 33]. However, there is very little work to pay attention to fundamentally solving the inherent defects of CNN’s local representation, Fig. 2(a). Capturing the long-range feature dependency, which can be interpreted as the semantic correlation between features in different spatial locations, is critical for WSOL.

Recently, visual transformer has been introduced to the computer vision area. Visual transformer [10] constructs a sequence of tokens by splitting an input image into patches with positional embedding and applying cascaded transformer blocks to extract visual representation. Thanks to the self-attention mechanism and Multilayer Perceptron (MLP) structure, visual transformers can learn complex spatial transforms and reflect long-range semantic correlations adaptively, which is crucial for localizing full object extent, Fig. 1(d). Nevertheless, visual transformer cannot be directly mitigated to WSOL for the following two reasons: (1) When using patch embeddings, the spatial topology of the input image is destroyed, which hinders the generation of activation maps for object localization. (2) The attention maps of visual transformers are semantic-agnostic (not distinguishable to object classes) and are not competent to semantic-aware localization, Fig. 2(b).

In this study, we propose the token semantic coupled attention map (TS-CAM), making the first attempt for weakly supervised object localization with visual transformer. TS-CAM introduces a semantic coupling structure with two network branches, Fig. 2(c), one performs semantic re-allocation using the patch tokens and the other generates semantic-agnostic attention map upon the class tokens. Semantic re-allocation, with class-patch semantic activation, enables the patch tokens to be aware of object categories. The semantic-agnostic attention map aims to capture long-distance feature dependency between patch tokens by taking the advantages of the cascaded self-attention modules in transformer. TS-CAM finally couples the semantic-aware maps with the semantic-agnostic attention map for object localization, Fig. 2(c).

The contributions of this work are as follows:

- We propose the token semantic coupled attention map (TS-CAM), as the first solid baseline for WSOL using visual transformer by leveraging the long-range feature dependency.
- We propose the semantic coupling module to combine the semantic-aware tokens with the semantic-agnostic attention map, providing a feasible way to leverage
both semantics and positioning information extracted by visual transformer for object localization.

- TS-CAM achieves a substantial improvement over previous methods on two challenging WSOL benchmarks, fully exploiting the long-range feature dependency in the visual transformer.

2. Related Work

Weakly Supervised Object Localization (WSOL) aims to learn object localizations given solely image-level category labels. A representative study of WSOL is CAM [55], which produces localization maps by aggregating deep feature maps using a class-specific fully connected layer. By removing the last fully connected layer, CAM can also be implemented by fully convolutional networks [15].

Despite the simplicity and effectiveness of CAM-based methods, they suffer identifying small discriminative parts of objects. To improve the activation of CAMs, HaS [33] and CutMix [33] adopted adversarial erasing on input images to drive localization models focusing on extended object parts. ACoL [50] and ADL [8] instead erased feature maps corresponding to discriminative regions and used adversarially trained classifiers to recover missed parts. SPG [51] and F2C [52] increased the quality of localization maps by introducing the constraint of pixel-level correlations into the network. DANet [46] applied a divergent activation to learn complementary visual cues for WSOL. SEM [53] and SPA [22] refined the localization maps by using the point-wise similarity within seed regions. GC-Net [20] took geometric shapes into account and proposed a multi-task loss function for WSOL.

Most of the above methods struggled to expand activation regions by introducing sophisticated spatial regularization techniques to CAM. However, they remain puzzled by the contradiction between image classification and object localization. As observed by the visualization approaches [3, 49], CNNs tend to decompose an object into local semantic elements corresponding local receptive fields. Activating a couple of the semantic elements could bring good classification results. The problem about how to collect global cues from local receptive fields remains.

Weakly Supervised Detection and Segmentation are vision tasks closely related to WSOL. Weakly supervised detection train networks to simultaneously perform image classification and instance localization [41, 39, 27]. Given thousands of region proposals, the learning procedure selects high-scored instances from bags while training detectors. In a similar way, weakly supervised segmentation trains classification networks to estimate pseudo masks which are further used for training the segmentation networks. To generate accurate pseudo masks, [17, 1, 14, 41, 56] resorted to a region growing strategy. Meanwhile, some researchers investigated to directly enhance the feature-level activated regions [18, 44]. Others accumulate CAMs by training with multiple phases [16], exploring boundary constraint [5], leveraging cross-image semantics [35] to refine pseudo masks.

Similar to WSOL, many weakly supervised detection and segmentation approaches are prone to localize object parts instead of full object extent. There is a requirement to explore new classification models to solve the partial activation problem in a systematic way.

Long-Range Feature Dependency. CNNs produce a hierarchical ensemble of local features with different reception fields. Unfortunately, most CNNs [31, 12] are good at extracting local features but experience difficulty to capture global cues.

To alleviate such a limitation, one solution is to utilize pixel similarity and global cues to refine activation maps [41, 42, 52, 53]. Cao et al. [4] found that the global contexts modeled by non-local networks are almost the same for query positions and thereby proposed NLNet [40] with SENet [13] for global context modeling. MST [34] proposed the learnable tree filter to capture the structural property of minimal spanning tree to model long-range dependencies. The other solution is the attention mechanism [40, 26, 54]. The non-local operation [40] was introduced to CNNs in a self-attention manner so that the response at each position is a weighted sum of the features at all (global) positions. SASA [26] verified that self-attention is an effective stand-alone layer for CNNs. Relation Networks [9] proposed to process a set of objects simultaneously through interaction between their features and geometry, allowing modeling the spatial relations between objects. Recent studies introduced a cascaded self-attention mechanism in the transformer model to capture long-range feature dependency [45, 37, 47, 25].

3. Methodology

In this section, we first give the preliminaries for visual transformer. We then introduce the TS-CAM method.

3.1. Preliminaries

For visual transformer [10], an input image $x$ of $W \times H$ resolution is divided to $w \times h$ patches, where $w = W/P$, $h = H/P$ and $P$ denotes the width/height of a patch. The divided patches are flattened and linearly projected to construct $N = w \times h$ patch tokens $\{t_{n}^{i} \in \mathbb{R}^{1 \times D}, n = 1, 2, ..., N\}$ and a class token $t_{0}^{i} \in \mathbb{R}^{1 \times D}$. Fig. 3. $D$ stands for the dimension of each token embedding. The class token $t_{0}^{i}$ is learnable with random initialization. Each token is added with a learnable position embedding in an element-wised manner. These tokens are fed into $L$ cascaded transformer blocks, each of which consists of a multi-head self-attention layer and a Multilayer Perceptron (MLP) block.
Denote \( t^c_n \) and \( t^c_{L} \) as the \( n \)-th patch token and the class token of the \( L \)-th transformer block, respectively. The last embedded class token \( t^L_L \) is fed to an MLP block to predict the classification probability, as

\[
p = \text{Softmax}(\text{MLP}(t^L_L)),
\]

where \( p \in \mathbb{R}^{1 \times C} \) and \( C \) denotes the number of classes. \( p_c \) denotes the predicted probability to class \( c \). \( \text{MLP}(\cdot) \) denotes the classification function implemented by the MLP block. Denote the ground-truth label for image \( x \) as \( y \in \{1, 2, ..., C\} \), the classification loss function is defined as

\[
\mathcal{L}_T = -\log p_y,
\]

which is used to train the visual transformer.

### 3.2. TS-CAM

We propose the TS-CAM method to generate semantic-aware localization maps upon the trained visual transformer, Fig. 3. In the visual transformer, however, only the class token is semantic-aware while the patch tokens are semantic-agnostic. To fulfill semantic-aware localization, we introduce the semantic re-allocation branch to transfer semantics from the class token to patch tokens and generate semantic-aware maps. Such semantic-aware maps are coupled with the semantic-agnostic attention maps to generate the semantic-aware localization maps.

**Semantic Re-allocation.** Visual transformer uses the class token to predict image categories (semantics), while using semantic-agnostic patch tokens to embed object spatial locations and reflects feature spatial dependency. To generate semantic-aware patch tokens, we propose to re-allocate the semantics from the class token \( t^c_{L} \) to the patch tokens \( \{t^c_1, t^c_2, ..., t^c_N\} \).

As shown in Fig. 3, patch token embeddings of the \( L \)-th visual transformer block are concatenated and transposed as \( t^{L} \in \mathbb{R}^{D \times N} \). They are then reshaped to token feature maps \( t^{L}_{d} \in \mathbb{R}^{D \times W \times H} \), where \( t^{L}_{d}, d \in \{1, 2, ..., D\} \) denotes the \( d \)-th feature map. The semantic aware map \( S_c \) of class \( c \) is calculated by convolution as

\[
S_{c} = \sum_{d} t^{L}_{d} * k_{c,d},
\]

where \( k \in \mathbb{R}^{C \times D \times 3 \times 3} \) denotes the convolution kernel and \( k_{c,d} \) is a \( 3 \times 3 \) kernel map indexed by \( c \) and \( d \). * is the convolution operator. To produce semantic-aware maps, the loss function defined in Eq. 2 is updated to

\[
\mathcal{L}^{\prime}_T = -\log p_y
\]

\[
= -\log \frac{\exp \left( \sum_n S_{n,y}/N \right)}{\sum_c \exp \left( \sum_n S_{n,c}/N \right)},
\]

where \( S_{n,c} \) is the semantic of the \( n \)-th patch token for class \( c \). While optimizing Eq. 2 allocates the semantics to class token \( t^c_{L} \), minimizing Eq. 4 re-allocates the semantics to patch tokens \( \{t^c_1, t^c_2, ..., t^c_N\} \), generating semantic-aware maps for WSOL.

**Semantic-agnostic Attention Map.** To fully exploit the long-range feature dependency of visual transformer, we propose to aggregate the attention vectors of the class token...
to generate the semantic-agnostic attention map. Denoting \( t^l \in \mathbb{R}^{(N+1) \times D} \) as the input of transformer block \( l \), which is calculated by concatenating the embeddings of all tokens (including class token and all patch tokens). In the self-attention operation in the \((l)\)-th transformer block, the embedded tokens \( \tilde{t}^l \) is computed as

\[
\tilde{t}^l = \text{Softmax}\left( (t^l \theta_q^l)(t^l \theta_k^l)^T / \sqrt{D} \right) (t^l \theta_v^l),
\]

Eq. 5 implies that the embedding \( \tilde{t}^l_\omega \) of class token of the semantic-attention operation is calculated by multiplying its attention vector \( A^l_\omega \) with the embedding \( t^l \) in the \((l)\)-th transformer block. \( A^l_\omega \) is therefore able to "see" all patch tokens, where \( A^l_\omega \) implies how much attention is paid on each token. When Eq. 4 is optimized, the attention vector \( A^l_\omega \) is driven to focus on object regions (e.g., long-range features of semantic correction) for image classification. The final attention vector \( A_s \) is defined as

\[
A_s = \frac{1}{L} \sum_l A^l_\omega,
\]

Semantic-Attention Coupling. As the attention vector \( A_s \) is semantic-agnostic, we use an element-wise multiplication to couple it with the semantic-aware maps to obtain the semantic-coupled attention map \( M_c \) for each class \( c \), Fig. 3.

The coupling procedure is formulated as

\[
M_c = \Gamma^{w \times h}(A_s) \otimes S_c,
\]

where \( \otimes \) denotes an element-wise multiplication and addition operations. \( \Gamma^{w \times h}(\cdot) \) denotes the reshape function which converts the attention vector (\( \mathbb{R}^{1 \times N} \)) to the attention map (\( \mathbb{R}^{w \times h} \)). \( M_c \) is up-sampled to a semantic-aware localization map, which is used for object bounding box prediction with a thresholding approach [51].
In Table 2, we compare TS-CAM with its CNN counterparts (CAM) and the SOTAs on the localization accuracy by using tight bounding boxes on the ILSVRC. TS-CAM respectively outperforms CAM on the VGG16 [32] by 10.6% and 9.4% in terms of Top-1 Loc. Acc and Top-5 Loc. Acc. Compared with SOTAs with the VGG16 backbone [32], TS-CAM outperforms by ~6% and ~4% in terms of Top-1 Loc. Acc and Top-5 Loc. Acc. Compared with $I^2C$, TS-CAM achieves performance gains of 6.0% Top-1 Loc. Acc and 5.8% Top-5 Loc. Acc, which are significant margins to the challenging problem. Compared with SOTAs on the well-designed Inception V3 [36], TS-CAM also achieves the best performance. Specifically, TS-CAM achieves performance gains of 7.1% and 6.1% in terms of Top-1 and Top-5 Loc. Acc compared with CAM. Compared with $I^2C$ which leverages pixel-level similarities across different objects to prompt the consistency of object features within the same categories, TS-CAM achieves comparable results with a cleaner and simpler pipeline. The right half of Fig. 4 illustrates examples of localization maps on ILSVRC.

CAM [55] tends to activate local discriminative regions and cannot retain the object structure well. Due to the lack of category semantics, TransAttention activates almost the salient objects within images (e.g., the building in the first image and the branch in the second image). TS-CAM takes the advantage of self-attention mechanism in visual transformer and thus activates the full extent of objects.

In Fig 5, we compare localization accuracy among variant CAM methods under different IoUs on CUB-200-2011 [38]. TS-CAM outperforms the CAM [55] and RCAM [53] under each IoU by large margins. In addition, TS-CAM achieves larger gains as IoU threshold increases, which indicates that the localization maps of our method cover the object extent accurately.

**Parameter Complexity.** Under similar parameter complexity and computational cost overhead, TS-CAM (with 25.1M parameters) achieves a cleaner and simpler pipeline.
parameters and 5.29G MACs) respectively outperforms VGG16-CAM (with 19.6M parameters and 16.3G MACs) by 27.1% (71.3% vs. 44.2%) and GoogleNet-CAM (with 16.7M parameters and 13.5G MACs) by 27.2% (71.3% vs. 41.1%) in Table 3.

**Error Analysis.** To further reveal the effect of TS-CAM, we categorize the localization errors into five as in [22]: classification error (Cls), multi-instance error (M-Ins), localization part error (Part), localization more error (More), and others (OT). **Part** indicates that the predicted bounding box only cover the parts of object, and IoU is less than a certain threshold. **More** indicates that the predicted bounding box is larger than the ground truth bounding box by a large margin. Each metric calculates the percentage of images belonging to corresponding error in the validation/test set. Table 4 lists localization error statistics of M-Ins, Part, and More. TS-CAM effectively reduces the M-Ins, Part, and More errors on both benchmarks, which indicates more accurate localization maps. For CUB-200-2011, TS-CAM significantly reduces both Part and More-type errors by ~ 17% and ~ 3% compared with CAM [55] on the basis of well-designed Inception V3.

### 4.3. Ablation Study

**Attention and activation.** Using Deit-S as the backbone, we conduct ablation studies to verify the components in TS-CAM. Specifically, **TransAttention** solely uses semantic-agnostic attention map ($A_s$) for object localization, while **TransCAM** solely uses semantic-aware maps ($S_s$). $A_s$ and $S_s$ are respectively generated by Eq.6 and Eq.3, and are illustrated in Fig. 3. In Table 5, we evaluate the performance of the TS-CAM on CUB-200-2011 and observed significant improvements over TS-CAM components. Specifically, TS-CAM obtains gains of 12.4%, 14.1%, and 14.8% in terms of Top-1, Top-5 and Gt-Known Loc. Acc compared with **TransAttention**. Using the semantic-aware map, **TransCAM** struggles from distinguishing an object from the background due to the destruction of topology. Taking advantage of both modules, TS-CAM generates semantic-aware localization maps by coupling semantic-agnostic attention from transformer and token semantics from the classifier.

Table 6 shows the results on ILSVRC validation set with different layers are summed on CUB-200-2011 [38] test set.

<table>
<thead>
<tr>
<th>$L$</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Gt-Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>65.2</td>
<td>75.8</td>
<td>79.0</td>
</tr>
<tr>
<td>9</td>
<td>68.5</td>
<td>80.0</td>
<td>83.6</td>
</tr>
<tr>
<td>10</td>
<td>70.2</td>
<td>81.2</td>
<td>85.5</td>
</tr>
<tr>
<td>11</td>
<td>71.2</td>
<td>83.7</td>
<td>87.7</td>
</tr>
<tr>
<td>12</td>
<td>71.3</td>
<td>83.8</td>
<td>87.7</td>
</tr>
</tbody>
</table>

Table 7. Ablation results of TS-CAM when attention maps ($A_l^L$) from different layers are summed on CUB-200-2011 [38] test set.
generate structure-preserving activation maps and thus cannot differentiate the objects from background. It obtains a significant performance degradation compared with TS-CAM and CAM. TransAttention achieves 10.4% and 12.4% performance degradation compared TS-CAM in terms of Top-1 and Top-5 Loc. Acc. The class-agnostic features puzzled TransAttention toward false localization, Fig. 2(b).

Why summarize all attention maps? In Fig. 7, we visualize attention maps $A_l^*$ from all layers and semantic-agnostic attention map $A_*$. As $\{A_1^*, ..., A_L^*\}$ are complementary, we summarize them for full object extent localization. Ablation study in Table 7 on CUB-200-2011 dataset demonstrates that summing all $A_l^*$ achieves the highest localization accuracy.

Why attention instead of activation? The reasons are two folds: (1) Visual transformer leverages embeddings of a low-resolution class token for image classification and can not produce high-resolution CAM. (2) The semantic-aware maps in TS-CAM, which are calculated by re-allocating the semantics from the class token to patch tokens, fail to discriminatively activate the object regions. By visualizing the similarities among patch token embeddings $\{t_1^L, ..., t_N^L\}$ in Fig. 8 right, we observed that the patch token embeddings are similar with each other, which implies that the activation results (semantic-aware activation maps) generated by these embeddings experience difficulty to discriminate objects from their backgrounds, Fig. 6(c).

5. Conclusion

We proposed the token semantic coupled attention map (TS-CAM) for weakly supervised object localization. TS-CAM takes full advantage of the cascaded self-attention mechanism in the visual transformer for long-range feature dependency extraction and object extent localization. To solve the semantic agnostic issue of the patch tokens, we proposed to re-allocate category-related semantics for patch tokens, enabling each of them to be aware of object categories. We proposed the semantic coupling strategy to fuse the patch tokens with the semantic-agnostic attention map to achieve semantic-aware localization results. Experiments on the ILSVRC/CUB-200-2011 datasets show that TS-CAM significantly improved the WSOL performance, in striking contrast with its CNN counterpart (CAM). As the first and solid baseline with transformer, TS-CAM provides a fresh insight to the challenging WSOL problem.

Acknowledgment. This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61836012, 61771447 and 62006216, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000, and the China National Postdoctoral Program for Innovative Talents under Grant BX20190323.
References


[16] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *IEEE ICCV*, pages 2070–2079, 2019. 3


