Localizing Multi-scale Semantic Patches for Image Classification

Chuanguang Yang[1,2] , **Xiaolong Hu**[1,2] , **Zhulin An**[1*] , **Hui Zhu**[1,2] and **Yongjun Xu**[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{yangchuanguang, huxiaolong18g, anzhulin, zhuhui, xyj}@ict.ac.cn

Abstract

Deep convolutional neural networks (CNN) always non-linearly aggregate the information from the whole input image, which results in the difficult to interpret how relevant regions contribute the final prediction. In this paper, we construct a light-weight AnchorNet combined with our proposed algorithms to localize multi-scale semantic patches, where the contribution of each patch can be determined due to the linearly spatial aggregation before the softmax layer. Visual explanation shows that localized patches can indeed retain the semantics of the original images, while helping us to further analyze the feature extraction of localization branches with various receptive fields. For more practical, we use localized patches for downstream classification tasks across widely applied networks. Experimental results demonstrate that replacing the original images can get a clear inference acceleration with only tiny performance degradation.

1 Introduction

Several concurrent works design some CNN architectures attempting to interpret how relevant patches in the input image contribute the final prediction. BagNet [Brendel and Bethge, 2019] constructs ResNet-like [He *et al.*, 2016] architecture to extract the feature map and implement a spatially linear aggregation (i.e., a simple average) before softmax layer, where each spatial location can be mapped back to a small patch in the input image, thus the contribution of each patch can be determined by activation value. Saccader [Elsayed *et al.*, 2019] follows the BagNet and further introduces a hard attention module to localize the most salient locations, and then estimates the relevance of various image patches. However, the notably common neglection of them is that they only have one kind receptive field (RF) accumulated throughout the CNN, resulting in the interpretability of only single-scale image patches. It is widely known that real objects usually have various scales along with coarse- or fine-grained texture features, which would hinder the effective modeling by only single-scale RF. Moreover, the most relevant patches are unable to be further applied in the downstream classification tasks for more practical inference acceleration, due to the extremely high complexity of upstream localizer.

Inspired by the above works of linear feature mapping for retaining the interpretability, and to further address the regrets, we build a light-weight and multi-scale localizer called AnchorNet, the tail of which are three localization branches with various accumulated RF, i.e. 63×63, 95×95 and 111×111 in this paper, to capture multi-scale patches, where each branch is also equipped with an attention branch to assist the localization of relevant patches by providing a spatial attention map that can also be visualized to display the semantically salient locations. To further make the decisions of the predicted class and which branch to localize the semantic patches is most suitable for the given image, we introduce a simple yet effective mechanism that leverages the softmax distributions generated by three branches to achieve these goals. After that we can capture the semantic patches according to the given class and localization branch by a simple algorithm called LSP (**L**ocalizing **S**emantic **P**atches) we proposed. Extensive experiments on downstream classification demonstrate that AnchorNet (parameters: 1.6M, FLOPs: 0.5G) combined with our algorithms can localize more semantic patches and obtain better performance than state-of-the-art (SOTA) Saccader (parameters: 33.58M, FLOPs: 21.6G) only using an order of magnitude fewer complexity.

By visualizing the localized multi-scale patches across the ImageNet 2012 [Deng *et al.*, 2009] validation set, we observe that the localization branch with wider RF is more prone to localize larger object and coarse-grained global features, while that with narrower RF always localize smaller object and fine-grained local features, which matches our intuitive expectations. To pursue more practical application, we further use multi-scale patches for downstream classification tasks to validate the semantics of them and effectiveness of inference acceleration. Experimental results show that using multiple semantic patches to replace the original images for classification can consistently get clear acceleration for inference with tiny drop of accuracy across widely applied networks, e.g., resulting in about 50% FLOPs reduction for ResNet-50 with only 0.7% top-1 accuracy drop without any modifications of the original model.

In brief, our contribution lies in three folds:

1. We construct a light-weight AnchorNet combined with

*Contact Author

our proposed localization algorithms to adaptively localize multi-scale semantic patches via a linearly interpretable manner.

2. We analyze the characteristics of feature extraction for localization branches with different RF based on the visual explanation, and further interpret the intriguing cases of confusion caused by them.

3. Using multi-scale semantic patches for downstream classification task can get a clear inference acceleration with only tiny performance degradation compared with the original images, which is orthogonal and complementary for popular model-based acceleration methods.

2 Related Work

2.1 Localizing semantic features

Some previous works aim to interpret the decision of CNN by visualizing the semantic feature heatmap, mainly divided into response-based or gradient-based [Springenberg *et al.*, 2015; Selvaraju *et al.*, 2017; Smilkov *et al.*, 2017] manners. However, visual explanation only displays the semantic region that is unable to implement the downstream classification task due to its irregular shape. Furthermore, attention mechanisms are usually introduced to highlight the spatially semantic locations by providing a spatial attention map [Woo *et al.*, 2018; Fukui *et al.*, 2019; Elsayed *et al.*, 2019]. AnchorNet differs from those prior practices in that we apply multi-branch spatial attention mechanism so as to perform multi-scale semantic localization.

Some seemingly similar but essentially different approaches are region proposal models for object detection [Girshick, 2015; Ren *et al.*, 2015; He *et al.*, 2017a], which typically combine contextual information to infer relevant object regions rather than the local information in the fixed regions, and use ground-truth bounding boxes for training. Unlike these work, our AnchorNet is only supervised by image-level labels and extracts local features in the fixed regions that are strictly spatial alignment to the initial input image. Zhou *et al.* [2016] implement object localization without supervised on any bounding box annotations, which shares the similarity to us of training by image-level labels. However, the information is still gathered from the whole image instead of local regions, hence the contributions of various patches to final prediction would get tangled. Additionally, only one patch to localize the full object would be difficult for downstream classification task due to the dramatically various scale. In contrast, AnchorNet utilizes one or more patches with the same size to cover the object, which are quite advantageous to classification meanwhile obtaining a good performance.

2.2 Inference acceleration

Modern acceleration mothods mainly concentrate on channel pruning [Li *et al.*, 2017; He *et al.*, 2017b; Liu *et al.*, 2019], which aims to remove redundant convolutional filters in the model, or dynamic inference [Huang *et al.*, 2018; Wu *et al.*, 2018], which aims to only use a part of structure of the model conditioned on the input image at inference time.

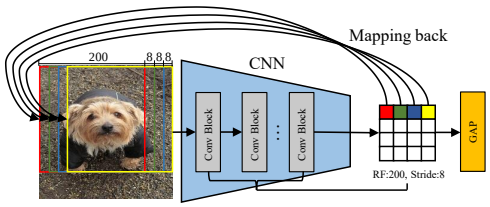

Figure 1: Example of feature mapping. For a input image with 224×224 pixels to the CNN model, which has accumulated 200×200 RF size and 8 strides before GAP, can generate a feature map with $4 \times 4 = 16$ spatial locations. Note that we only depict the spatial demension while omitting channels for brevity and better understanding of spatial mapping rule, which is best viewed in color.

In this work, we localize informative local image patches over the whole image guided by light-weight AnchorNet, then the downstream networks only need to process semantic feature patches which have much smaller size than the original images, thus producing a clear acceleration. Moreover, data-based localization is CNN-agnostic and thus can be regarded as the orthogonal and complementary of model-based pruning or dynamic inference.

3 Methodology

3.1 Review of Feature Mapping

Modern CNNs gradually decrease the spatial resolution for the input image by several convolutional blocks until the global average pooling (GAP) layer. Many hyperparameters in convolutional layer settings, e.g., kernel size, padding or stride can affect resolution size of the output. We set padding 0 across all the convolutional layers in AnchorNet, so each final spatial location of the feature map before GAP layer can be mapped to the input image exactly without the cases of beyond bounds. Given an example for better understanding, assumed that one CNN model receives a image with $H \times W$ pixels as the input, and has accumulated $k \times k$ RF size and s strides before GAP, we will obtain the $[\lfloor (H-k)/s \rfloor + 1] \times [\lfloor (W-k)/s \rfloor + 1]$ spatial locations, where each location can be mapped back to a region with the size of $k \times k$. Figure 1 illustrates the mapping rule.

3.2 AnchorNet

We develop a CNN called AnchorNet which can automatically localize the most suitable semantic features with various patch sizes conditioned on the input image. Figure 2 illustrates the overall architecture of AnchorNet schematically, which contains the following three components:

Head. The input image is firstly processed by a head to extract low-level features, the all details of it is shown in Table 1. We adopt efficient bottleneck unit, SE block [Hu *et al.*, 2018] and the hyperparameter settings following Howard *et al.* [2019], except that we replace most 3×3 convolutions with 1×1 convolutions to restrict accumulated RF throughout the head, and only perform less down-sampling compared with popular networks on ImageNet dataset to retain the higher resolution of feature map for providing more patch mappings to the input image.

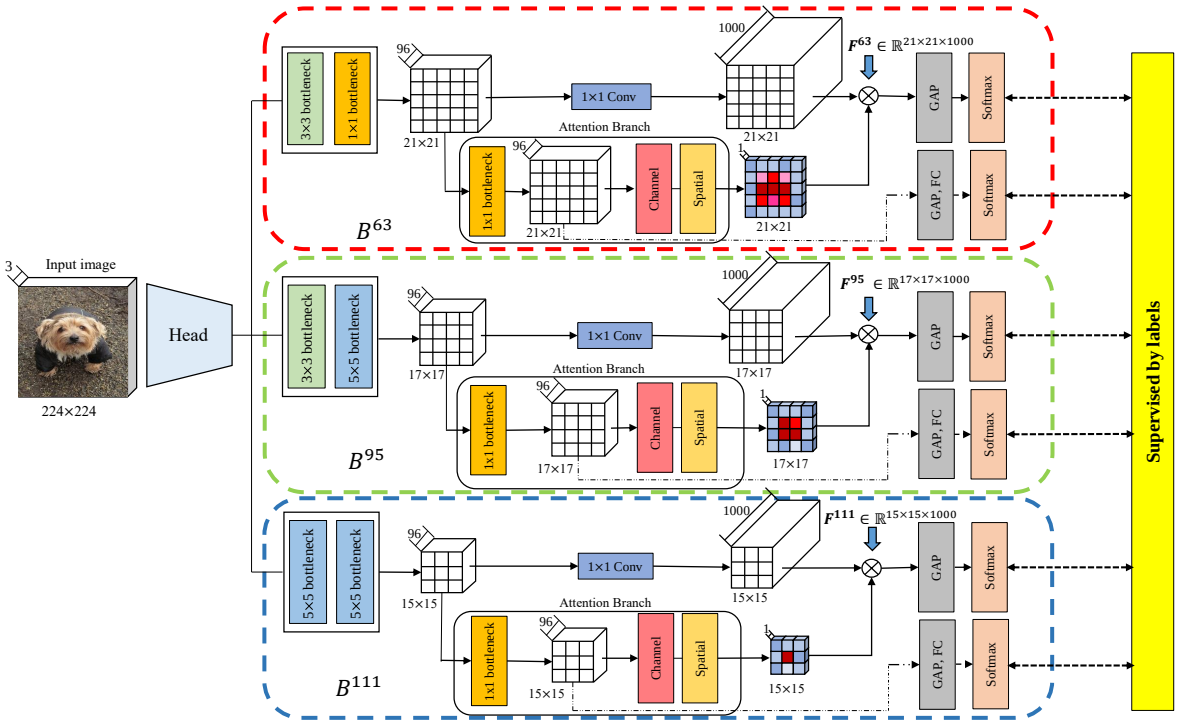

Figure 2: Illustration of the overall architecture of AnchorNet. Three branches associated with various RFs of 63×63, 95×95 and 111×111 after head perform multi-scale patch localization, which we name them as B^{63}, B^{95} and B^{111}, respectively. Each branch is attached with an attention branch, which sequentially includes **channel attention** and **spatial attention**, and the connection to GAP and FC is cancelled at test. Note that we tag the actual spatial size below each feature map, and the number of grids only represents the relative size.

IR	Operator	Exp	Out	SE	NL	s	RF
224^2	conv2d,3×3	-	16	-	HS	2	3^2
111^2	bneck,3×3	16	16	-	RE	2	7^2
55^2	bneck,3×3	72	24	-	RE	2	15^2
27^2	bneck,1×1	88	24	-	RE	1	15^2
27^2	bneck,1×1	96	40	✓	HS	1	15^2
27^2	bneck,1×1	240	40	✓	HS	1	15^2
27^2	bneck,1×1	240	40	✓	HS	1	15^2
27^2	bneck,1×1	120	48	✓	HS	1	15^2
27^2	bneck,3×3	144	48	✓	HS	1	31^2
25^2	bneck,3×3	288	96	✓	HS	1	47^2

Table 1: The head of AnchorNets. IR denotes the input resolution. Exp and Out denote the expansion and output channels, respectively. SE denotes whether there exists a SE block. NL denotes the non-linearity, including h-swish (HS) or ReLU (RE).

IR	Operator	Exp	Out	SE	NL	s	RF
23^2	bneck,3×3	480	96	✓	HS	1	63^2
21^2	bneck,1×1	576	96	✓	HS	1	63^2
21^2	bneck,1×1	192	96	✓	HS	1	63^2
23^2	bneck,3×3	480	96	✓	HS	1	63^2
21^2	bneck,5×5	576	96	✓	HS	1	95^2
17^2	bneck,1×1	192	96	✓	HS	1	95^2
23^2	bneck,5×5	480	96	✓	HS	1	79^2
19^2	bneck,5×5	576	96	✓	HS	1	111^2
15^2	bneck,1×1	192	96	✓	HS	1	111^2

Table 2: The bottlenecks of localization branches in AnchorNet. The blocks sequentially correspond the localization branches B^{63}, B^{95} and B^{111} in Figure 2, respectively. The final row of each block denotes the bottleneck of attention branch.

Localization Branch.We construct three branches to further localize semantically multi-scale regions along the spatial dimension after head. To this end, bottlenecks with various kernel sizes are intentionally equipped to adjust the accumulated RF sizes of these branches individually. Table 2 elaborates the information of accumulated RF, it means that each spatial location of three feature maps processed by the three localization branches would obtain a mapping patch size of 63×63, 95×95 or 111×111 to the original image. Due to the accumulated stride of all the three branches is $2^3 = 8$, they can map to $21^2 = 441$, $17^2 = 289$, $15^2 = 225$ possible semantic locations for the original 224×224 image, respectively. Before classification, we utilize a linear 1×1 convolution to encode the representations into a 1000-dimensional logits tensor denoted as $\mathbf{F}^j \in \mathbb{R}^{H\times W\times 1000}$ which combined with the spatial attention map by broadcast element-wise multiplication, where $j \in \{63, 95, 111\}$ denotes the given branch, H and W denote the spatial height and width, respectively. And then we apply a global average pooling for $\mathbf{F}^j$ and a softmax layer to obtain the class probability distri-

bution. Compared with the setting of popular fully-connected (FC) layer, it is noteworthy that we just perform a linear average aggregation along the spatial dimension and then attach softmax function that can allows us to pinpoint exactly how various patches contribute the final prediction, this is what we refer to as the concept of ***linear*** in this paper. While FC would facilitate the interaction between patch-wise evidences thus destroying the interpretability of mapping. The outputs of all branches after softmax layer are supervised by the cross-entropy loss with image-level labels.

Attention Branch. To assist feature learning for the localization branch, we further construct an attention branch to emphasize semantic locations by generating a spatial attention map. A bottleneck is applied to produce the feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ for attention localization, where C denote the number of channels. Then a 1×1 convolutional filter compacts $\mathbf{X}$ along the channel dimension to $\tilde{\mathbf{G}} \in \mathbb{R}^{H\times W\times 1}$, and followed by a softmax function to generate spatial weights $\mathbf{G} \in \mathbb{R}^{H\times W\times 1}$:

$$\mathbf{G}_{i,j,1} = \frac{e^{\tilde{\mathbf{G}}_{i,j,1}}}{\sum_{h=1}^{H}\sum_{w=1}^{W} e^{\tilde{\mathbf{G}}_{h,w,1}}} \tag{1}$$

According to normalized spatial weights $\mathbf{G}$, we employ global weighted average pooling to $\mathbf{X}$ and produce a channel attention map $\tilde{\mathbf{C}} \in \mathbb{R}^{1\times 1\times C}$, the c-th channel of $\tilde{\mathbf{C}}$ is as (2), $*$ denotes the broadcast element-wise multiplication here.

$$\tilde{\mathbf{C}}_c = \sum_{h=1}^{H}\sum_{w=1}^{W} \mathbf{X}_{h,w,c} * \mathbf{G}_{h,w,1} \tag{2}$$

Then softmax function normalize the $\tilde{\mathbf{C}}$ to generate the final **channel attention map** $\mathbf{C} \in \mathbb{R}^{1\times 1\times C}$:

$$\mathbf{C}_{1,1,c} = \frac{e^{\tilde{\mathbf{C}}_{1,1,c}}}{\sum_{i=1}^{C} e^{\tilde{\mathbf{C}}_{1,1,i}}} \tag{3}$$

According to normalized channel weights $\mathbf{C}$, we employ weighted shrinking of $\mathbf{X}$ along the channel demension to generate a spatial attention map $\tilde{\mathbf{S}} \in \mathbb{R}^{H\times W\times 1}$:

$$\tilde{\mathbf{S}}_{i,j,1} = \sum_{c=1}^{C} \mathbf{X}_{i,j,c} * \mathbf{C}_{1,1,c} \tag{4}$$

After applying softmax function to $\tilde{\mathbf{S}}$, we output the final **spatial attention map** $\mathbf{S} \in \mathbb{R}^{H\times W\times 1}$ that we need:

$$\mathbf{S}_{i,j,1} = \frac{e^{\tilde{\mathbf{S}}_{i,j,1}}}{\sum_{h=1}^{H}\sum_{w=1}^{W} e^{\tilde{\mathbf{S}}_{h,w,1}}} \tag{5}$$

It is noteworthy that we introduce a FC layer and softmax function behind the attention features $\mathbf{X}$ to implement a additionally direct supervision by cross-entropy loss with image-level labels beside the main localization branches, which can be more easier to learn discriminative features and facilitate the attention localization. Note that the connectivity from attention branch to GAP is occluded at inference stage.

Algorithm 1 Localizing Semantic Patches (LSP)

Input: input image $\mathbf{I}$, heatmap $\mathbf{M}^\theta \in \mathbb{R}^{H\times W}$
Parameter: patch size $\theta \times \theta$, maximum number of selected patches K^θ, IoU threshold T, percentage of coverage $\mathcal{P}^\theta$
Output: collection of localized image patches $\mathcal{S}$

$coordinates$=Reverse_Sort(Flatten($\mathbf{M}^\theta$))
$\mathcal{S}$={Map($\mathbf{I}$, θ, $coordinates[1]$)} # A patch in $\mathbf{I}$ mapped by the coordinate in $\mathbf{M}^\theta$ with the maximum activation
for $i = 2 : H \times W \times \mathcal{P}^\theta$ **do**
  if $\forall s \in \mathcal{S}$,$IoU$(Map($\mathbf{I}$, θ, $coordinate[i]$)), s)$<T$ **then**
    $\mathcal{S}$ = $\mathcal{S}\cup${Map($\mathbf{I}$, θ, $coordinate[i]$)}
  end if
  if len($\mathcal{S}$) $== K^\theta$ **then**
    return $\mathcal{S}$
  end if
end for
return $\mathcal{S}$

3.3 Localizing Multi-scale Semantic Patches

Given a input image x to AnchorNet, localization branch B^j can predict the class probability distribution $[B_1^j(x), B_2^j(x), ..., B_{1000}^j(x)]$, where $B_y^j(x)$ denotes the probability of the class y, and $y \in \{1, 2, ..., 1000\}$, $j \in \{63, 95, 111\}$. Then we make a simple decision for individual branch as following:

$$Y^j = \arg\max_y B_y^j(x),\ P^j = \max_y B_y^j(x) \tag{6}$$

Where Y^j and P^j denote the predicted class and its probability by branch B^j, respectively. Then we can implement the systematic decisions of final class γ and branch B^θ for patch localization according to (7) and (8) as following:

$$\gamma = \begin{cases} y, if\ \exists y, \sum_j (y == Y^j) \geqslant 2 \\ Y^{\arg\max_j P^j}, otherwise \end{cases} \tag{7}$$

$$B^\theta = B^{\arg\max_j [(P^j == B_\gamma^j(x))\cdot P^j]} \tag{8}$$

Given the branch B^θ, each channel of logits tensor $\mathbf{F}^\theta$ corresponds the specific class activation map, which emphasizes class-specific semantical regions. Given the predicted class label γ, the heatmap $\mathbf{M}^\theta \in \mathbb{R}^{H\times W}$ can be obtained that is equal to $\mathbf{F}^\theta_{:,:,\gamma}$, which represents the interpretable contribution of each mapped patch for predicted class γ. Instead of simply selecting top K patches with maximum activations, we perform LSP as Algorithm 1 to ensure the localized patches that are not only semantic but also partly separated to cover more information. First, we flatten the $\mathbf{M}^\theta$ to a list including $H \times W$ 2-demensional coordinates $[(h, w)]_{h=1,2,...,H;w=1,2,...,W}$, and sort them from maximum to minimum according to their corresponding activation values. Then, we straightforward map the first coordinate point which has the maximum activation to the corresponding patch, mapping rule is as mentioned in section 3.1, and put it in the collection $\mathcal{S}$. Next, we visit each point sequentially from front to back, the mapped patch with the size of $\theta \times \theta$ of which can be put in the $\mathcal{S}$ only if the

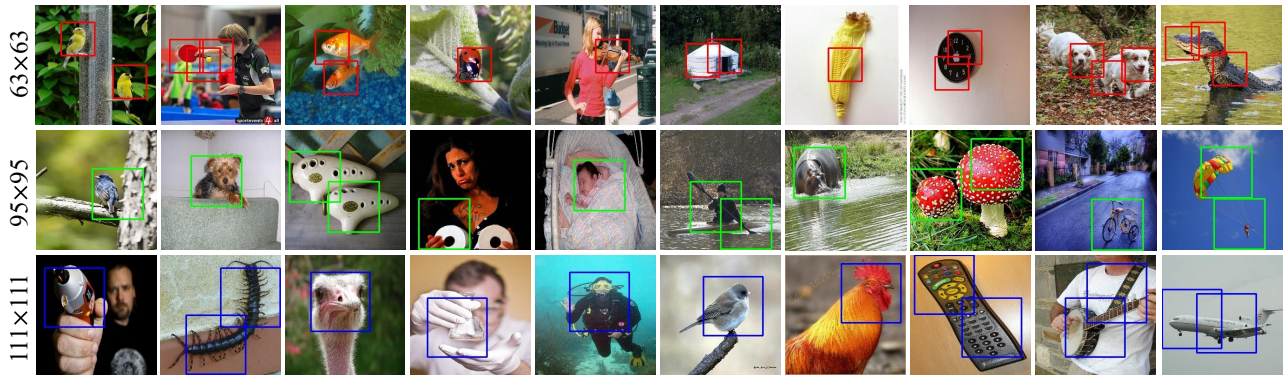

Figure 3: Examples of the localized semantic image patches. The first, second and third row denote the results localized by B^{63}, B^{95}, B^{111}, respectively. Note that each input image is assigned to one of three branches for localization according to its object property.

point meets the following conditions: the IoU of this patch and any patches in $\mathcal{S}$ is less than the threshold T. Where $\theta = \{63, 95, 111\}$, and IoU is a quite practical indicator to quantify the intersection between two patches A and B:

$$IoU = |A \cap B| \div |A \cup B| \tag{9}$$

Where $|\cdot|$ calculate the pixel number of the region. That means that localized patches can be controlled to be separated and semantic concurrently by introducing the IoU mechanism. When the number of patches in $\mathcal{S}$ achieves the upper limitation K, the final collection of patches can be obtained.

4 Experiments

4.1 Dataset and Settings

We experiment AnchorNet on ImageNet 2012 dataset [Deng *et al.*, 2009] to validate the effectiveness of localizing multi-scale semantic patches. ImageNet is a large-scale dataset for image recognition, which contains 1.2 million training images and 50k validation images with 1000 classes, and each image is resized to 224×224 pixels at test time on the validation set. For training, standard data augmentation is employed following He *et al.* [2016], and we use synchronous SGD with a momentum of 0.9 , batch size 256 and weight decay 10^{-4} for 100 epochs. The learning rate starts at 0.1 and decayed by a factor of 10 every 30 epochs.

About LSP algorithm, we set $K^{63} = 5$, $K^{95} = 3$, $K^{111} = 2$, $T = 0.3$, $\mathcal{P}^{63} = 0.05$, $\mathcal{P}^{95} = 0.04$, $\mathcal{P}^{111} = 0.03$ for visualizing localized semantic patches and analyzing characteristics among branches with various RFs. Further, we use slacker settings: $K^{63} = 24$, $K^{95} = 10$, $K^{111} = 8$, $T = 0.8$, $\mathcal{P}^{63} = \mathcal{P}^{95} = \mathcal{P}^{111} = 0.3$, for producing more semantic patches so as to fine-tune the downstream models and enhance its robustness for recognition tasks.

4.2 Localized Multi-scale Semantic Patches

As can be observed in Figure 3, we can obviously conclude that the branch with wider RF size may be more prone to localize larger object and coarse-grained global features, such as scuba diver, remote control and airliner, which occupy the most part in images, are captured by B^{111}. While the branch with relatively narrower RF always localize smaller object and fine-grained local features, e.g., B^{63} can not only captures the miniature object such as ladybug, fish and violin, but also identify the local texture features of large objects, such as corn and crocodile. Here we consider whether the size of object is large or not that is relative to the size of the corresponding image.

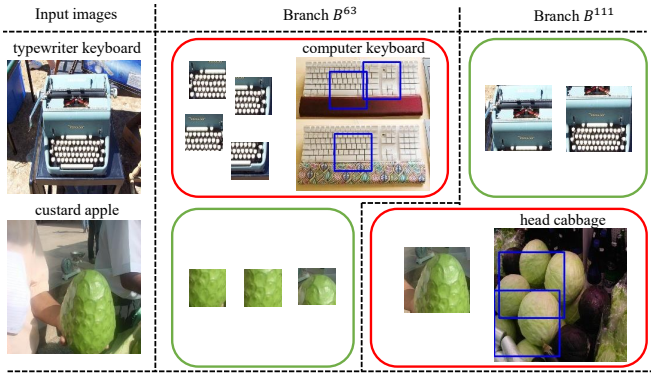

Figure 4: Misclassified cases of branch B^{63} and B^{111}. For each branch, we show the produced patches with the most class evidences in a box, where the green and red margins indicate correct and incorrect predictions, respectively. In the cases of red box, we further show the representative image of misclassified category and localize the most relevant patches for classification by the corresponding branch.

Another intriguing case is the misclassification that may take place in both B^{63} and B^{111} due to their characteristics of feature extraction, as illustrated in Figure 4. Combined with the above discussion, we further consider that although narrow RF can capture local features, it may ignore the more informative global features, e.g., B^{63} concentrates on local keyboard yet omits the global typewriter, leading to confusion with computer keyboard. In contrast, wide RF prefers localizing coarse-grained features but ignore local fine-grained features, e.g., the outline and color of custard apple are interpreted as evidences for head cabbage by B^{111}, which omits the different texture information between them.

Model	Scale	FLOPs (G)		Top-1 (%)		Top-5 (%)	
ResNet-50	224^2	4.1		72.6		91.0	
	63^2	1.9		70.0		88.4	
	95^2	2.2	**2.1**	72.1	**71.9**	89.9	**89.7**
	111^2	2.2		73.4		90.5	
ResNeXt-50	224^2	4.3		75.7		92.8	
	63^2	2.0		72.8		89.8	
	95^2	2.3	**2.2**	74.3	**74.8**	91.0	**91.1**
	111^2	2.4		77.0		92.5	
DenseNet-169	224^2	3.4		74.8		92.5	
	63^2	1.6		71.9		89.6	
	95^2	1.8	**1.7**	73.5	**73.6**	91.1	**91.0**
	111^2	1.8		75.3		92.0	
HCGNet-C	224^2	7.1		78.6		94.3	
	63^2	3.3		74.8		91.1	
	95^2	3.8	**3.7**	77.3	**76.9**	92.3	**92.4**
	111^2	3.9		78.4		93.7	

Table 3: Comprehensive performance of multi-scale semantic patches for classification. FLOPs denotes the average number of floating point operations for processing one validation image, which refers to the initial image if 224×224 scale, otherwise the all corresponding generated patches. We report top-1 and top-5 accuracy to measure the performance of classification. Each bold entry denotes the overall result by weighted average of three branches.

Model	63^2-R	63^2-L	95^2-R	95^2-L	111^2-R	111^2-L
ResNet-50	15.6	**38.2**	37.4	**54.9**	45.1	**60.2**
ResNeXt-50	17.3	**41.0**	40.5	**57.9**	48.4	**63.0**
DenseNet-169	17.9	**38.8**	42.8	**56.6**	48.9	**59.9**
HCGNet-B	12.8	**27.8**	40.9	**54.2**	51.5	**61.8**

Table 4: Comparison of top-1 accuracy obtained by rescaling (-R) and localizing (-L) across the CNN models.

4.3 Using Semantic Patches for Classification

We further conduct downstream classification according to localized patches so as to verify their representations for semantics of the original images. As shown in Table 3, we utilize pre-trained ResNet-50, ResNeXt-50 [Xie *et al.*, 2017], DenseNet-169 [Huang *et al.*, 2017] and HCGNet-C [Yang *et al.*, 2019] fine-tuned on training patches to implement classification tasks. The results are evaluated on ImageNet validation set, where each image is localized by one of three branches. Across all 50K validation images, where 15050, 18047, 16903 images are localized by B^{63}, B^{95} and B^{111} corresponding with 5.6, 2.9 and 2.1 patches for an image on average, respectively. Since one image may generate multiple relevant patches, we implement the final decision by adding the softmax distributions of them, and determine the class with maximum probability. Table 3 shows that without any changes of SOTA models, using multiple semantic patches instead of the original images can achieve about $2\times$ acceleration with tiny drop of top-1 accuracy, varying from 0.7% on ResNet-50 as minimum to 1.7% on HCGNet-C as maximum.

To further demonstrate the performance of remarkable acceleration and good accuracy is attributed to localizing but can not be obtained by simple scale reduction from the original images, we make a comparison between them and evaluated on the same networks without any extra trainings. Each 224×224 image is only localized one semantic patch with maximum activation for the correspond scale decided by AnchorNet, meanwhile it is also performed simply rescaling as the counterpart. Table 4 shows that rescaling consistently incurs significant accuracy drop compared with localizing, indicating the semantic patches are indeed effective.

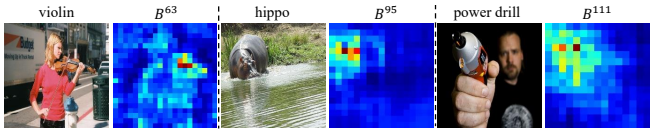

Figure 5: Visualization of spatial attention maps, which are generated by attention branches of B^{63}, B^{95}, B^{111} from left to right conditioned on the corresponding input images, respectively.

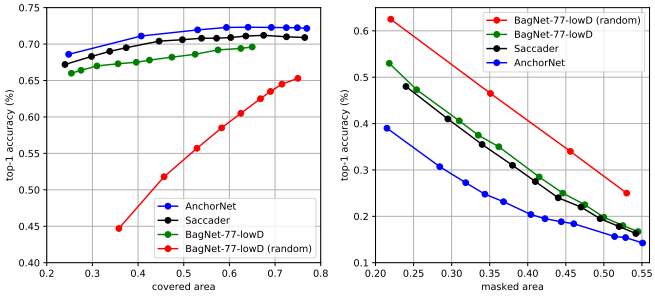

Figure 6: Relationships between accuracy and covered area, or masked area by localized patches using various localizers. Accuracy is evaluated on downstream pre-trained ResNet-50.

4.4 Analysis of AnchorNet

What have attention branches learned?

Attention branches are introduced to assist semantic feature localization, we further visualize the generated spatial attention maps from various branches, which are depicted in Figure 5. All heatmaps show that all attention branches can not only attend to the informative locations, but also adaptively capture various scale objects based on their individual RFs.

AnchorNet localizes relevant patches for classification.

From Figure 6, it can be observed that all localizers generally lead to better accuracy as the covered area increases. Moreover, using relevant patches localized by AnchorNet can achieve the best performance compared with other localizers under the same coverage, which proves the superiority of AnchorNet for downstream classification task is not simply attributed to wider image coverage. We further investigate the importance of localized patches by using them to mask the original images (i.e., set the pixels to 0) and then perform a classification on resulting images. Figure 6 show that masking by AnchorNet leads to more significant drop in performance than other localizers. Based on the above analysis, we think that AnchorNet outperforms other SOTA localizers, largely due to the multi-scale localization capability.

5 Conclusion

We construct a AnchorNet combined with our LSP algorithm to adaptively localize multi-scale semantic patches, meanwhile retaining the interpretability by linearly spatial information aggreation. Compared with previous SOTA localizers, AnchorNet is more feasible for downstream classification and can obtain a better performance due to its capability of light-weight and multi-scale feature extraction. We hope our AnchorNet may inspire the future study of interpretable semantic feature localization and application.

References

[Brendel and Bethge, 2019] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[Elsayed *et al.*, 2019] Gamaleldin Elsayed, Simon Kornblith, and Quoc V Le. Saccader: Improving accuracy of hard attention models for vision. In *Advances in Neural Information Processing Systems*, pages 700–712, 2019.

[Fukui *et al.*, 2019] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019.

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

[He *et al.*, 2016] 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.

[He *et al.*, 2017a] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[He *et al.*, 2017b] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1389–1397, 2017.

[Howard *et al.*, 2019] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[Huang *et al.*, 2018] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, and Kilian Q Laurens van der Maaten. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018.

[Li *et al.*, 2017] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017.

[Liu *et al.*, 2019] Zechun Liu, Haoyuan Mu, Xiangyu Zhang, Zichao Guo, Xin Yang, Tim Kwang-Ting Cheng, and Jian Sun. Metapruning: Meta learning for automatic neural network channel pruning. *arXiv preprint arXiv:1903.10258*, 2019.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[Smilkov *et al.*, 2017] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[Springenberg *et al.*, 2015] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations*, 2015.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[Wu *et al.*, 2018] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8817–8826, 2018.

[Xie *et al.*, 2017] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[Yang *et al.*, 2019] Chuanguang Yang, Zhulin An, Hui Zhu, Xiaolong Hu, Kun Zhang, Kaiqiang Xu, Chao Li, and Yongjun Xu. Gated convolutional networks with hybrid connectivity for image classification. *arXiv preprint arXiv:1908.09699*, 2019.

[Zhou *et al.*, 2016] 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.