FSM: FEATURE SAMPLING MODULE FOR OBJECT DETECTION

Xin Yi, Bo Ma, Jiahao Wu

Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing, China
School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

ABSTRACT

Challenges caused by the acquisition condition of the images, the state of the objects, or the noise in the transmission of the images commonly exist in object detection. In those situations, the features of the objects extracted by CNNs contain certain uncertainty, which increases the difficulty of subsequent classification and regression. Towards enhancing the quality of the features, we propose a Feature Sampling Module (FSM), which learns multiple two-dimensional Gaussian distributions by the sampling network (SN) and applies those Gaussian masks to extract valid information of the features. With this sampling scheme, our method avoids learning the decision boundary from the low-quality features, making the overall model classification performance more robust. To ensure that the SN is capable of sampling the highest quality region, we design a novel sampling loss (SL) to measure the quality of the sampled features. Extensive experimental results validate the effectiveness of our proposed method.

Index Terms— Object detection, Feature sampling, Quality enhancement, Deep learning

1. INTRODUCTION

Object detection is a fundamental problem in computer vision, which can be applied in instance segmentation, scene understanding, pose estimation, image captioning and multiple objects tracking (MOT), to name a few. Recently, with the development of Convolutional Neural Networks (CNNs), learning-based object detection methods have achieved remarkable progress beyond the traditional detection methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. However, object detection in the wild is still tricky mainly due to the various quality of images, which is related to the acquisition condition of the images (such as illumination, weather, camera angle, and resolution), the state of the objects (such as camouflage, occlusion, and pose), or the noise in the transmission.

Under those aforementioned difficulties, a significant feature value shift would appear in the feature space. In addition, the image quality uncertainty (In an occlusion situation, it represents the position and the proportion of the occlusion) would produce the feature uncertainty. Motivated by some previous data uncertainty works in face recognition [13, 14],

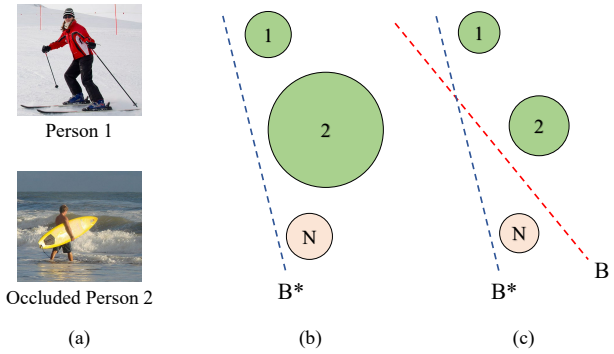

Fig. 1: Geometrical interpretation of the classification in the feature space. (a) Input images with a normal person and an occluded person. (b) Decision boundary B^* learned by previous methods, where the circle 1, 2, and N represent positive sample person 1, positive sample person 2, and negative sample, respectively. The smaller the circle, the higher the quality of the feature. (c) Decision boundary B learned by our method. We sample the features for better classification.

we model the object in the feature space as a Gaussian distribution, where the mean denotes the most likely feature value and the variance denotes the uncertainty of the feature value. Previous detectors apply their classifier to extracted features directly, without considering the quality of the features, which results in learning an undesirable decision boundary. As illustrated in Figure 1(b), due to the uncertainty of $person2$, decision boundary B^* produces a false positive N.

Previous work [15] has exploited the external visual saliency map to construct an effective attention block. The saliency map consists of several partial saliency maps, where each partial saliency map represents a single Gaussian foveating feature over stimuli. Motivated by this, we propose a Feature Sampling Module (FSM) that predicts k Gaussian masks to sample the valid information from the features. We find that in dense detection frameworks, almost all the features of the foreground objects can be generated, and the bottleneck of the performance lies in the classification and regression of the object features. Thus, we append our FSM behind the object feature layers, seeking to sample representative regions of the object by Gaussian masks for quality enhancement. To

ensure these Gaussian masks converge to the optimal or the suboptimal locations, we introduce an additional sampling loss (SL) on the sampled features. By applying this feature sampling, the feature value shift and the feature uncertainty are alleviated. As illustrated in Figure 1(c), a better decision boundary B can be learned after reducing the feature uncertainty.

2. RELATED WORKS

Recently, object detection has improved progressively with the development of CNNs. According to whether to utilize region proposal, the learning-based object detection methods can be divided into two mainstreams, two-stage methods [1, 4, 2, 3, 5] and one-stage methods [8, 9, 10, 11, 12, 6, 7]. Similar to traditional object detection methods, two-stage object detection methods comprise a region proposal stage that pays attention to generating sufficient category-independent candidate region proposals, while one-stage methods achieve object detection without a distinct region proposal stage.

Certain previous works [16, 17, 18] have proposed methods to use the attention mechanism for feature quality improvement. Such as, in CBAM [16], the channel-wise pooling and the spatial-wise convolution are conducted to promote the interaction of the feature information. However, those methods mainly focus on modeling the relationship between different locations of space dimension or channel dimension without directly considering the feature quality assessment. In this work, we explicitly introduce indicators to evaluate the sampled feature directly. Moreover, the computational complexity of the FSM is lower than that of the attention mechanism, which is proportional to the square of the image size.

3. PROPOSED METHOD

3.1. Overall

The whole pipeline of our proposed Feature Sampling Module (FSM) is illustrated in Figure 2. Given an arbitrary image, object feature x is generated by the backbone network. Then, we append an auxiliary FSM on the x, sampling partial information from it for quality enhancement. Finally, we fuse the sampled feature with the original x for the final detection.

The principal of designing this scheme is to mine and sample those more representative and discriminative regions of the object, offsetting the adverse effect in the potential region caused by occlusion, camouflage, or other interference, finally reducing the uncertainty. To this end, the proposed component should sense the global information and seek the local region with more discriminative clues. Thus, we design the sampling network (SN) to perceive then sample high-level semantic features of the object instance by predicting k Gaussian masks. In order to evaluate the quality of the sampled feature and supervise the convergence of the SN, we introduce the sampling loss (SL) to the sampled feature.

3.2. Sampling network

We construct the learnable sampling network (SN) to map the original global features into Gaussian parameters (μ and σ) then transfer those parameters into Gaussian masks. The detail of our SN is illustrated in Figure 2. To be specific, given object feature x with C channels and $W \times H$ spatial resolution, we first downsample the feature into a lower channel dimension by a 1×1 convolution to avoid high complexity. Since the μ of the Gaussian mask is related to the spatial position, we utilize another 1×1 convolution to learn the quality of different spatial positions on the feature map and obtain k values of the μ through the linear layer. In addition, since the σ can be regarded as the scale of the high-quality region, we design the multi-scale pyramid convolutions to extract spatial information and then concatenate them to learn the parameter σ. Finally, to ensure that the μ falls into the range of the spatial resolution of feature ($[0, W]$ and $[0, H]$ in this case) and the σ is no less than 0.1, we apply some post-process on the predicted μ and σ, which can be written as

$$\begin{aligned} \mu &= \frac{1}{2} S_{ratio} * T(\mu) \\ \sigma &= R(\sigma) + 0.1 \end{aligned} \tag{1}$$

where T and R denote the $Tanh$ and $ReLU$ activation functions, respectively. S_{ratio} is the spatial resolution of the feature map. The coordinate origin is set at the feature center.

3.3. Sampling loss

After initialization, different Gaussian masks pay attention to different regions, i.e., different local features are sampled. We hope that these Gaussian masks focus on more representative and discriminative local regions. Assuming that these representative regions should be discriminative enough for a detector to do the classification, e.g., a person's face is strong enough to be distinguished from other categories, we attach the classification loss $\mathcal{L}_c$ to force FSM to learn a better sampling strategy. To be specific, we apply the Cross-Entropy Loss on the sampled object features for supervision, which can be written as

$$\mathcal{L}_c = -\sum_x y(x) log\big(p(x)\big) \tag{2}$$

where $y(x)$ denotes the symbolic function of the label and $p(x)$ denotes the prediction. In each epoch, the generated Gaussian distribution samples a new feature distribution from the original feature. High-quality sampled distribution widens the classification boundary and thus increases the classification confidence, while low-quality sampled distribution decreases the classification confidence. With this supervision,

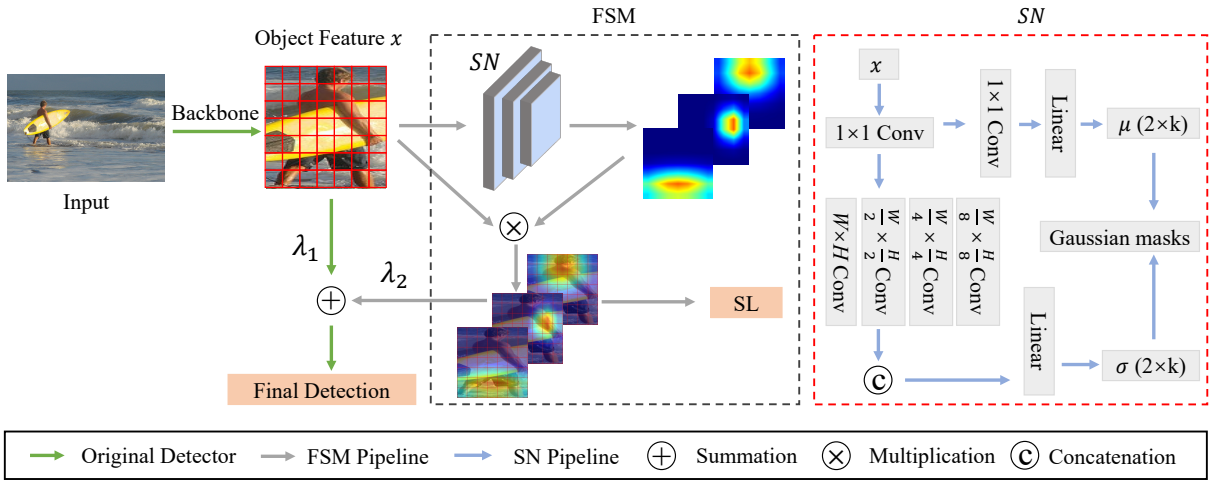

Fig. 2: Illustration of our proposed FSM. We utilize the sampling network (SN) to predict several Gaussian masks for feature sampling. The auxiliary sampling loss (SL) is imposed on the sampled feature to supervise the convergence of the SN. Furthermore, to achieve better-locating performance, we fuse the original feature and the sampled feature for final detection.

the FSM is forced to search for the high-quality region to make the new-attached loss decline.

In order to extract as much high-quality and effective information from different regions as possible, the generated Gaussian distributions need to be sufficiently dispersed. Thus, we propose the distribution loss $\mathcal{L}_d$ to measure the space status. Given k Gaussian masks (G_1 to G_k), we utilize Jensen–Shannon divergence to measure the similarity between two Gaussian masks and sum them. Since $JS(G_i||G_i) = 0$, there are actually $k(k-1)$ effective pairs. Besides, since $JS(G_i||G_j) = JS(G_j||G_i)$, we can take half of the sum as JS_{sum}, and now the number of effective pairs becomes $k(k-1)/2$:

$$JS_{sum} = \frac{1}{2}\sum_{i}^{k}\sum_{j}^{k} JS(G_i||G_j) \tag{3}$$

The value range of JS_{sum} is 0 to $k(k-1)/2$. We introduce $\hat{d}_k$ to measure the distribution similarity as follows,

$$\hat{d}_k = \frac{k(k-1)}{2} - \frac{1}{2}\sum_{i}^{k}\sum_{j}^{k} JS(G_i||G_j) \tag{4}$$

The value range of $\hat{d}_k$ is also 0 to $k(k-1)/2$. We normalize the $\hat{d}_k$ and apply margin m to get the distribution loss $\mathcal{L}_d$:

$$\mathcal{L}_d = \max(0, \frac{2\hat{d}_k}{k(k-1)} - m) \tag{5}$$

Therefore, the final loss can be calculated as

$$\mathcal{L}_s = \mathcal{L}_c + \alpha\mathcal{L}_d \tag{6}$$

3.4. Feature fusion

Although decision boundary is improved by uncertainty reducing, partial global information is sacrificed in those sampled features, which would lead to inaccurate locating. Therefore, to utilize more outline information of the objects, we fuse the sampled object feature with the original one and apply final object detection on the fused feature.

4. EXPERIMENTS

4.1. Datasets and experiments setup

To validate the effectiveness of our proposed method, we conduct experiments on COCO 2017 [19] dataset, which contains *train* set (118k images), *val* set (5k images), and *test-dev* set (20k images). We train some basic and the state-of-the-art object detectors on the *train* set and evaluate the performance on the *test-dev* set. Then, we apply our proposed FSM on those detectors to evaluate the performance enhancement and assess the generalization ability. For a fair comparison, we conduct experiments on the same NVIDIA 2080Ti GPUs.

The extra hyperparameters introduced by our FSM are λ_1, λ_2, margin m, and α. We set λ_1, λ_2, and α as the learnable parameters and manually set margin m to 0.7.

4.2. Results

The results of previous detection methods and our method in COCO 2017 [19] *test-dev* set are listed in Table1. Among those methods, our FSM method on ATSS [22] (ResNet-101) achieves 45.2 mAP without any other tricks. For the basic Faster R-CNN [3] (w/FPN ResNet-101) and SSD513 [6], our FSM method can obtain 2.8 and 2.6 mAP gains, respectively.

Table 1: Statistical evaluation metrics of previous detection methods and our method on COCO 2017 [19] *test-dev* set.

Benchmark	Methods	FSM	BackBone	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
COCO	SSD513 [6]	×	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
	SSD513	✓	ResNet-101	33.8	53.2	35.8	11.8	36.4	52.6
	Faster R-CNN w/FPN [3]	×	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
	Faster R-CNN w/FPN	✓	ResNet-101	39.0	62.3	42.2	20.1	41.6	51.5
	RetinaNet [7]	×	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
	RetinaNet	✓	ResNet-101	40.6	60.6	43.7	23.3	43.6	52.0
	Cascade R-CNN [20]	×	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
	Cascade R-CNN	✓	ResNet-101	44.1	63.8	48.0	24.9	47.2	57.6
	FCOS [21]	×	ResNet-101	41.5	60.7	45.0	24.4	44.8	51.6
	FCOS	✓	ResNet-101	43.3	62.4	46.9	25.4	46.5	53.7
	ATSS [22]	×	ResNet-101	43.6	62.1	47.4	26.1	47.0	53.6
	ATSS	✓	ResNet-101	45.2	63.9	48.5	27.0	48.7	55.6

Table 2: Ablation study of FSM on COCO 2017 [19] *test-dev* set. The baseline is ResNet-101-FPN Faster R-CNN, and all the experiments are conducted in 3 Gaussian masks.

SN	$\mathcal{L}_c$	$\mathcal{L}_d$	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
			36.2	59.1	39.0	18.2	39.0	48.2
✓	✓		37.1	60.2	39.8	19.0	40.1	49.3
✓		✓	38.5	61.5	41.6	19.9	41.0	51.1
✓	✓	✓	39.0	62.3	42.2	20.1	41.6	51.5

In addition, we integrate our FSM into more recent methods such as RetinaNet [7], Cascade R-CNN [20], FCOS [21], and ATSS [22]. The mAP gains in the ResNet-101 backbone are 1.5, 1.3, 1.8, and 1.6, respectively.

We also calculate the increment of different scale, where the AP_S is 1.6, 1.9, 1.5, 1.2, 1.0, 0.9 and AP_L is 2.8, 3.3, 1.8, 2.4, 2.1, 2.0. This experimental result demonstrates that our FSM has less feature enhancement and performance improvement in small objects. We assume the reason is that the valid information of the small object occupies fewer pixels of the image (feature), inevitably increasing the sampling difficulties and making the sampling network tend to be overfitting.

4.3. Ablation study

To demonstrate the effectiveness of different components in our FSM, we conduct an ablation study by using COCO [19] dataset. In Table 2, we select ResNet-101 Faster R-CNN [3] as the baseline and gradually integrate our modules into the baseline. The results show that the ablation study of adding the SN, $\mathcal{L}_c$, and $\mathcal{L}_d$, mAP is increased from 36.2 to 39.0. If we only add SN and $\mathcal{L}_d$, mAP is increased from 36.2 to 38.5. The reason is that the final detection losses can also be backpropagated to optimize the parameters of the SN, but $\mathcal{L}_c$ attached to the sampled feature is more effective than the common final detection losses in guiding the convergence of Gaussian masks.

Table 3: Ablation study on the number of Gaussian masks. "$0^\dagger$" denotes base ResNet-101 Faster R-CNN. We only evaluate the results with less than 6 masks since more masks are not time-friendly and unnecessary.

Mask Nums	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$0^\dagger$	36.2	59.1	39.0	18.2	39.0	48.2
1	36.9	59.6	39.8	18.5	39.6	49.3
2	38.1	61.4	40.8	19.0	40.3	50.7
3	39.0	62.3	42.2	20.1	41.6	51.5
4	38.7	61.9	42.0	19.8	41.4	51.2
5	38.8	61.8	42.3	20.0	41.1	51.6

The effect of mask number on the final detection results is shown in Table 3. FSM with three Gaussian masks achieves the best performance because three two-dimensional Gaussian distributions are enough to fit complex planar distributions, i.e., high-quality information at different locations can be all extracted by three Gaussians. More Gaussian masks can also achieve this effect, but there would be higher risks of Gaussian centers overlap and overfitting. In addition, more Gaussian masks are not time-friendly.

5. CONCLUSION

In this work, we propose a Feature Sampling Module (FSM) for object detection, which samples valid information from the feature to enhance the quality and reduce the uncertainty. With this scheme, the detector can learn a better decision boundary. Extensive experimental results validate the effectiveness and the generalization of our proposed method.

6. REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

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

[3] 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*, 2015, pp. 91–99.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[5] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun, "Light-head r-cnn: In defense of two-stage object detector," *arXiv preprint arXiv:1711.07264*, 2017.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[9] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[10] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[12] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[13] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei, "Data uncertainty learning in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5710–5719.

[14] Yichun Shi and Anil K Jain, "Probabilistic face embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6902–6911.

[15] Abraham Montoya Obeso, Jenny Benois-Pineau, Mireya Saraí García Vázquez, and Alejandro Álvaro Ramírez Acosta, "Visual vs internal attention mechanisms in deep neural networks for image classification and object detection," *Pattern Recognition*, vol. 123, pp. 108411, 2022.

[16] 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)*, 2018, pp. 3–19.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[18] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[19] 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 *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] Zhaowei Cai and Nuno Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[21] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[22] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.