Generating Features with Increased Crop-related Diversity for Few-Shot Object Detection Supplementary Material

Jingyi Xu
Stony Brook University
jingyixu@cs.stonybrook.edu

Hieu Le
EPFL
minh.le@epfl.ch

Dimitris Samaras
Stony Brook University
samaras@cs.stonybrook.edu

1. Overview

In this document, we provide additional experiments and analyses. In particular:

- Section 2 provides visualizations of the detection results for inaccurate bounding boxes.
- Section 3 provides the results of using different numbers of additional samples to fine-tine the model.
- Section 4 provides additional visualizations of the detection results.
- Section 5 shows the impact of the mapping function on the final results.
- Section 6 provides the details of how we generate additional training data.

2. Detection Results for Inaccurate Bounding Boxes

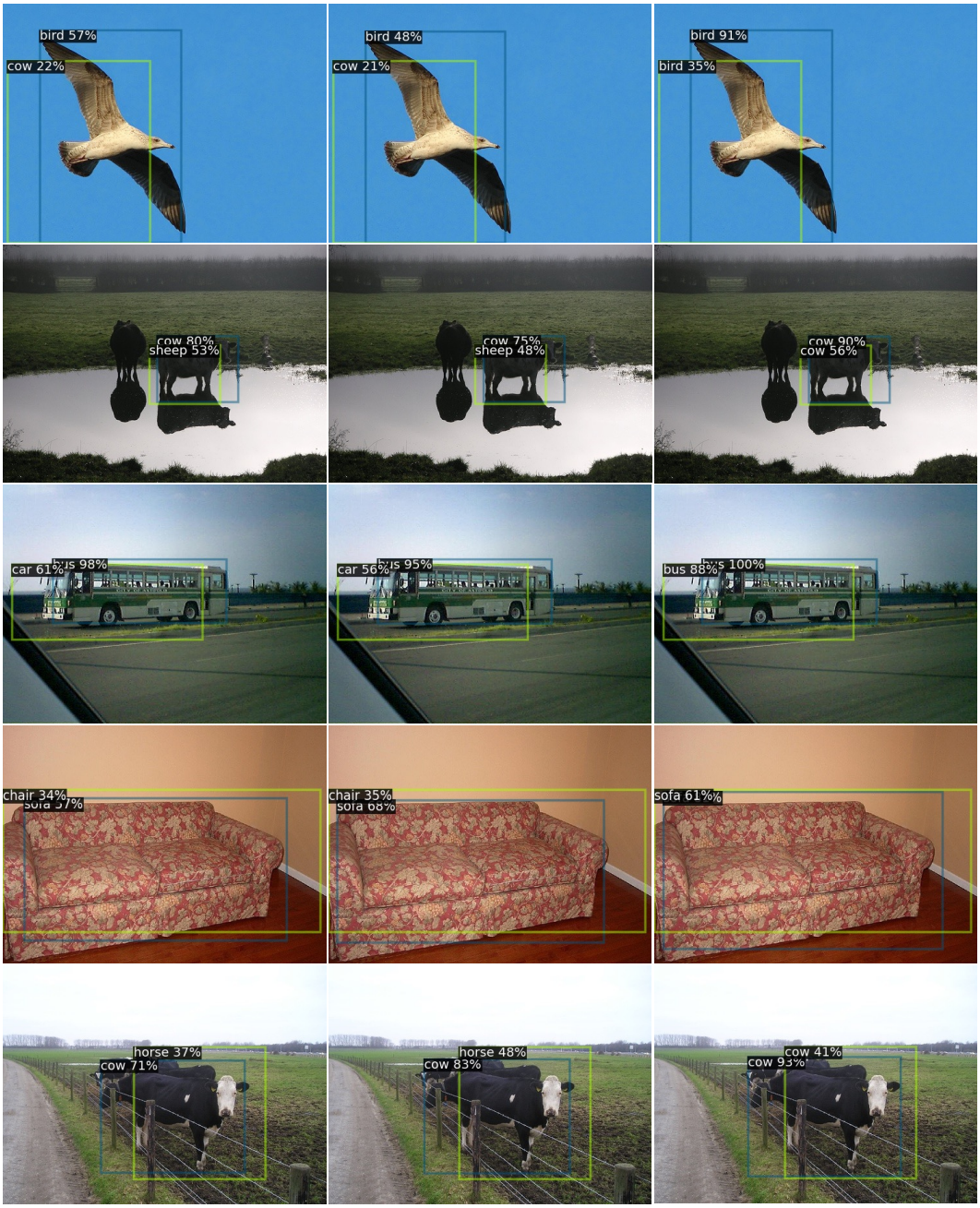
In this section, we provide qualitative visualizations of the detected objects of the 1-shot model on PASCAL VOC Novel Split 1. As shown in Figure 1, for each input image, the blue box is the original prediction result from the object detector. We then randomly create an augmented bounding box based on the ground-truth bounding box and input the augmented box to the classifier of the object detector. The prediction result on the augmented box is denoted as the yellow box. For the examples shown in the figure, the baseline DeFRCN model [2] and the model trained with features from a vanilla VAE predict the class labels correctly on the original input boxes while both fail on the augmented boxes. By contrast, the model trained with features from Norm-VAE can classify both the original box and the augmented box correctly. As can be seen, crop-related variation is crucial for object detection and our method can enhance the object detector's robustness against the variation successfully.

3. Number of Generated Samples

In our main experiment, we generate 30 samples per class and use them together with the original few-shot samples to fine-tune the object detector. In this section, we investigate the impact of the number of the generated samples. Table 1 shows the AP50 on PASCAL VOC Novel Split 1 with different numbers of generated features under 1-shot, 2-shot and 3-shot settings. As the number of generated samples increases, the performance gradually improves and then plateaus and drops slightly (less than 0.5% decrease in performance).

4. Visualization of the Detection Results on PASCAL VOC dataset

We show a few visualization results of DeFRCN [2] and our proposed method in Figure 2. As can be seen from the figure, the model trained with additional features performs better than DeFRCN. For instance, in the third row, DeFRCN fails to recognize both the two instances of the "*bird*" class while both Vanilla-VAE and Norm-VAE recognize them. It can be seen that with additional data from Norm-VAE, the FSOD model can recognize objects that are undetected with the model trained with just the original training data. The Norm-VAE model is generally more robust in recognizing objects. It works well even when the objects are cropped (2nd row) or small (two bottom rows).



DeFRCN [2] Vanilla-VAE Norm-VAE

Figure 1. **Qualitative visualizations of the detected objects on PASCAL Novel Split 1**. "Vanilla-VAE" denotes the model trained with features generated from a vanilla VAE and "Norm-VAE" denotes the model trained with features generated from Norm-VAE. The blue box is the detector's prediction on the original image and the yellow box is the prediction on the augmented box. Our proposed Norm-VAE can generate features that enhance the model's robustness against crop-related variation.

5. Mapping Function Analyses

We use a simple pre-defined linear function $g(x) = w \times x + b$ to map from an IoU score x to the new norm of a latent code. Here we only consider proposals with IoU scores ranging from 0.5 to 1. Proposals with lower IoU scores are noisy since they contain mostly background areas. With our VAE architecture and the training data, we observe that the norms of

# Generated Features	0	5	10	15	20	25	30	35
1-shot	56.3	60.5	61.6	61.8	62.0	61.9	**62.1**	62.0
2-shot	60.3	62.0	63.7	63.6	63.6	64.1	**64.9**	64.5
3-shot	62.0	65.6	67.0	67.2	67.2	**67.8**	**67.8**	67.3

Table 1. **Impact of the number of the generated samples under PASCAL VOC Novel Split 1**. As the number of generated samples increases, the performance gradually improves and then saturates and drops slightly.

DeFRCN [2] Vanilla-VAE Norm-VAE

Figure 2. **Visualization of the detection results on PASCAL VOC dataset.** The FSOD model trained with additional features performs better than DeFRCN. It works well even when the objects are partially cropped (2nd row) or small (two bottom rows). The detection score threshold is 0.5. Please view in magnification for cases with small objects.

the original latent codes are ranged approximately from $\sqrt{512}$ to $5\sqrt{512}$. We would like the rescaled norms to be in the same range and, at the same time, the latent code of an easy proposal has a small norm and the latent code of a hard proposal has a large norm. Thus, we set the parameters of $g(x)$ such that $g(0.5) = 5\sqrt{512}$ and $g(1) = \sqrt{512}$.

We also conduct experiments with different ranges and the results are shown in Table 2. Note that here $\sqrt{512}$ is a scaling constant that corresponds to the number of dimensions ($N = 512$) of the latent space. As can be seen from the table, we observe better performance when the IoU score inversely correlates with the latent norm. In this case, a proposal with a low IoU score (i.e., hard case) has a higher latent norm and is placed further away from the origin. A possible reason is that features of hard instances often exhibit higher variance. Thus, it is more optimal to use latent codes with larger norms to represent them [1].

	$g(1)$	$g(0.5)$	AP50
Inverse Correlation	$1\times\sqrt{512}$	$2\times\sqrt{512}$	61.6
	$1\times\sqrt{512}$	$5\times\sqrt{512}$	**62.1**
	$1\times\sqrt{512}$	$10\times\sqrt{512}$	61.8
Correlation	$2\times\sqrt{512}$	$1\times\sqrt{512}$	60.6
	$5\times\sqrt{512}$	$1\times\sqrt{512}$	61.3
	$10\times\sqrt{512}$	$1\times\sqrt{512}$	60.6

Table 2. **Performance with different configurations of the mapping function**. We conduct experiments using different coefficients for function $g(\cdot)$, which defines the value range of the new norm of the latent code.

6. Details on Generating Augmented Training Data

We extract the image features from image crops from the base classes and use them to train a feature generator to generate features for the novel classes. Specifically, we apply the RoI head feature extractor on the ground-truth bounding box b_i from the base classes to get the RoI feature f_i. To enrich the diversity of the RoI feature, we randomly create N additional augmented bounding boxes by randomly moving the starting point and the ending point of the original box, annotated as $\{b_i^1, b_i^2, ...b_i^N\}$. These augmented bounding boxes overlap the ground-truth bounding box differently and have different IoU scores. With a set of augmented bounding boxes $\{b_i^1, b_i^2, ...b_i^N\}$, we extract the corresponding RoI features $\{f_i^1, f_i^2, ...f_i^N\}$ and use them to train our VAE model.

References

[1] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14220–14229, 2021. 4

[2] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8661–8670, 2021. 1, 2, 3