StarNet: for weakly supervised few-shot detection and explainable classification

Leonid Karlinsky* Joseph Shtok* Amit Alfassy* Moshe Lichtenstein*
Sivan Harary Eli Schwartz Sivan Doveh Prasanna Sattigeri Rogerio Feris
Alexander Bronstein Raja Giryes

Abstract

In this paper, we propose a new few-shot learning method called StarNet, which is an end-to-end trainable non-parametric star-model few-shot classifier. While being meta-trained using only image-level class labels, StarNet learns not only to predict the class labels for each query image of a few-shot task, but also to localize (via a heatmap) what it believes to be the key image regions supporting its prediction, thus effectively detecting the instances of the novel categories. The localization is performed by finding matching regions between all pairs of support and query images of a few-shot task. We evaluate StarNet on multiple few-shot classification benchmarks attaining significant state-of-the-art improvement on the CUB and ImageNetLOC-FS, and smaller improvements on other benchmarks. In addition, we test the proposed approach on the previously unexplored and challenging task of Weakly Supervised Few-Shot Object Detection (WS-FSOD), obtaining significant improvements over the baselines. A more detailed report is located at https://arxiv.org/abs/2003.06798

1. Introduction

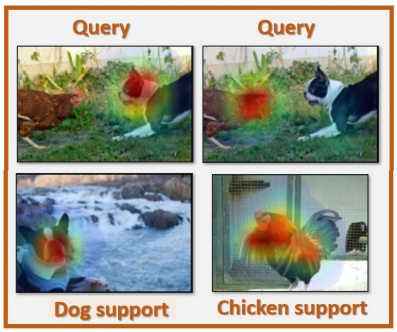

Figure 1. StarNet provides evidence for its predictions by finding large (semantically) matching regions between the query and the support images of a few-shot task, thus providing plausible explanations even for ambiguous cases. Matching regions are drawn as heatmaps for each query (top) and support (bottom) pair.

Recently, great advances have been made in the field of few-shot learning. In this learning regime only a handful of examples for the target classes are available at test time, while the target classes themselves are novel and unseen during training. In most few-shot learning methods (e.g., [17, 8]), the training is performed on a dataset of 'base' classes with a large number of labeled examples for each class. While annotating the samples for target classes is cheap, preparing elaborate annotations (bounding boxes or segmentation masks) for base classes from same visual domain can be quite expensive. This is, however necessary for few-shot learning methods to be effective [2].

Generic weakly-supervised methods, such as the popular GradCAM [15], are able (to some extent) to highlight the pixels responsible for the CNN classifier's prediction. But these are generally much less effective for few-shot classifiers. In this paper, we propose StarNet - a new type of few shot learner that is capable of localizing the novel class instances while being trained for few-shot classification using only image-level class labels. StarNet is comprised of a novel few-shot classifier head, attached to a standard fully-convolutional CNN feature extractor (e.g. ResNet-12). It is trained in meta-learning fashion, where k-shot, n-way training episodes are randomly sampled from the base classes.

StarNet head operates by geometrically matching every pair of support and query images. For each such pair, a non-parametric star-model [13, 7] matching is implemented as an end-to-end differentiable neural network. The StarNet head drives the CNN backbone to learn the optimal features for this matching. A star-model is a probabilistic model where the object parts are generated independently given some latent state variables describing the object (e.g. object center location and size). It allows matching (maximal) partial regions of arbitrary shape between the two images up to arbitrary shift and local deformations accommodating for changes in shape. In StarNet training, the matching regions correspond to instances of same class, present on support and query pairs.

With StarNet, we address a new and challenging task of *Weakly Supervised* Few-Shot Object Detection (WS-FSOD). A few-shot algorithm is trained using only classification data on base classes, and acts as object detector on novel classes, despite no location information is provided in the test few-shot tasks.

In addition, we show that StarNet few-shot learner is effective at few-shot classification, significantly improving the state-of-the-art (SOTA) baselines on the CUB and ImageNetLOC-FSfew-shot benchmarks, and comparing favorably to the SOTA methods on: *mini*ImageNet, CIFAR-FS and FC100. At the same time, StarNet provides plau-

sible explanations for its predictions by highlighting the matching regions it finds for query-support pairs.

2. Related Work

Star Models and **Generalized Hough Transform** techniques were popular classification and detection methods before the advent of CNNs. In these techniques, objects were represented as a collection of parts, independently linked to the object model via Gaussian priors to allow local deformations. Classically, parts were represented using patch descriptors [9], or SVM part detectors in the Deformable Part Model (DPM) [12]. DPM was later extended to CNN based DPM in [3]. Unlike DPM, StarNet is non-parametric, in a sense that parts are not explicit and are not fixed during inference, and unlike all of the aforementioned methods, it is trained with only weak supervision to work in the few-shot setting. In [11] a *non* few-shot classification network is trained through pairwise local feature matching, but unlike in StarNet, no geometrical constraints on the matches are used. **Weakly-supervised object detection** refers to techniques that learn to localize objects despite being trained with only image-level class labels [1, 19]. However, to the best of our knowledge, no prior works have considered this challenging problem in the few-shot setting. **Few-shot with localization and attention** is a recent research direction. Most of these methods rely on bounding box supervision. Using this supervision, several works have extended the popular object detection techniques to the few-shot setting [6]. Recently, several works have further explored the use of geometric relations between intermediate convolutional features in few-shot classification [4, 20]. In CAN [5] attention maps for query and support images are generated by 1×1 convolution applied to a pairwise local feature comparison map. These attention maps are not intended for object localization, so unlike StarNet, geometry of the matches in [5] is not modeled.

3. Single-stage StarNet

Denote by Q and S a pair of query and support images belonging to a k-shot, n-way episode E (a few-shot task) sampled during either meta-training or meta-testing. Let ϕ be a fully convolutional CNN feature extractor, taking a square RGB image input and producing a feature grid tensor of dimensions $r \times r \times f$ (here r is the spatial dimension, and f is the number of channels). Applying ϕ on Q and S computes the query and support grids of feature vectors:

$$\begin{aligned} &\{\phi(Q)_{i,j} \in \mathcal{R}^f \,|\, 1 \le i,j \le r\}, \\ &\{\phi(S)_{l,m} \in \mathcal{R}^f \,|\, 1 \le l,m \le r\} \end{aligned} \tag{1}$$

For brevity we will drop ϕ in further notation and write $Q_{i,j}$ and $S_{l,m}$ instead of $\phi(Q)_{i,j}$ and $\phi(S)_{l,m}$. We first L_2-normalize $Q_{i,j}$ and $S_{l,m}$ for all grid cells, and then compute a tensor D of size $r \times r \times r \times r$ of all pairwise distances between Q and S feature grids cells: $D_{i,j,l,m} = ||Q_{i,j} - S_{l,m}||^2$. In our implementation, the tensor D is efficiently computed for all support-query pairs simultaneously using matrix multiplication with broadcasting. We convert the tensor D into a (same size) tensor of unnormalized probabilities P, where: $P_{i,j,l,m} = e^{-0.5 \cdot D_{i,j,l,m}/\sigma_f^2}$ is the probability that $Q_{i,j}$ matches $S_{l,m}$ in a sense of representing the same part of the same category. Some object part appearances are more rare than others; to accommodate for that, the tensor P is normalized to obtain the tensor R of the same size, where $R_{i,j,l,m} = P_{i,j,l,m}/N_{i,j}$ is the likelihood ratio between 'foreground' match probability $P_{i,j,l,m}$, and the 'background' probability $N_{i,j}$ of 'observing' $Q_{i,j}$ in a random image, approximated as:

$$N_{i,j} = \sum_S \sum_{l,m} P_{i,j,l,m} \tag{2}$$

where $\sum_S$ is computed by matching the same query Q to all of the supports in the episode. Let w be a reference point on S. We set $w = (r/2, r/2)$ to be the center of S feature grid. We compute voting offsets as $o_{l,m} = w - (l,m)$ and the voting target as $t_{i,j,l,m} = (i,j) + o_{l,m}$ being the corresponding location to the reference point w on the query image Q assuming that indeed $Q_{i,j}$ matches $S_{l,m}$. Please note that by construction, $t_{i,j,l,m}$ can be negative, and its values range between $(-r/2, -r/2)$ and $(3r/2, 3r/2)$, thus forming a $2r \times 2r$ hypothesis grid of points in coordinates of Q potentially corresponding to point w on S.

Our proposed StarNet implements the non-parametric star-model [9, 7] for matching the Q and S feature grids. For every hypothesis location (x,y) on the $2r \times 2r$ hypothesis grid, the star model computes the overall belief $A(x,y)$ for that hypothesis considering independently all the possible matches between support and query feature grids. In probabilistic sense, star-model is a variant of the Naive-Bayes model, and hence to compute $A(x,y)$, should have accumulated *log*-likelihood ratios from the potential matches. However, as in [7], to make the star-model more robust to background noise, in StarNet likelihood ratios are directly accumulated:

$$A(x,y) = \sum_{\substack{\{i,j,l,m\} \; s.t. \\ t_{i,j,l,m} = (x,y)}} R_{i,j,l,m} \tag{3}$$

Following accumulation, the final StarNet posterior for each location hypothesis is computed by convolution of A with

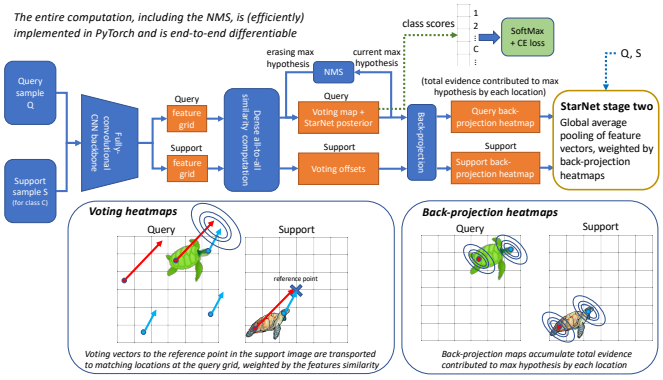

Figure 2. **StarNet overview.** Best viewed in large zoom.

a symmetric Gaussian kernel $G(\sigma_g)$:$V(x,y) = G(\sigma_g) \circledast A(x,y)$. This efficiently accounts for any random relative location shift allowed to occur with the $G(\sigma_g)$ Gaussian prior for any matched pair of $Q_{i,j}$ and $S_{l,m}$.

For the following, we will slightly abuse notation and denote by $V_{Q,S}(x,y)$ the StarNet posterior $V(x,y)$ computed for the pair of query image Q and support image S from the k-shot, n-way episode E. We compute the score (logit) of predicting the category label c for Q as:

$$SC_1(c;Q) = \frac{1}{k} \cdot \sum_{\substack{S \in E \ s.t. \\ C(S)=c}} \max_{x,y} V_{Q,S}(x,y) \qquad (4)$$

where $C(S)$ is the class label of S. The backbone ϕ is trained using Cross Entropy loss between $SC_1(c;Q)$ (after softmax) and the GT class label of Q. By only matching images with the same label, the model learns to match the regions that correspond to instances of the shared class.

4. Back-projection maps

For any pair of query Q and support S, and any hypothesis location $(\hat{x},\hat{y})$ on the $2r \times 2r$ grid, and in particular one with the maximal StarNet posterior value $(\hat{x},\hat{y}) = \arg\max_{x,y} V(x,y)$, we can compute two back-projection heatmaps (one for Q and one for S). These are $r \times r$ matrices in the feature grid coordinates of Q and S respectively, whose entries contain the amount of contribution the corresponding feature grid cell on Q or S gave to the posterior probability $V(\hat{x},\hat{y})$:

$$BP_{Q|S}(i,j) = \sum_{l,m} R_{i,j,l,m} \cdot e^{-0.5 \cdot ||t_{i,j,l,m} - (\hat{x},\hat{y})||^2 / \sigma_g^2}$$

The $BP_{S|Q}(l,m)$ is computed in symmetrical fashion by replacing summation by l,m with summation by i,j. The back-projection heatmaps are highlighting the matching regions on Q and S that correspond to the hypothesis $(\hat{x},\hat{y})$, which for query-support pairs that share some category labels are in most cases the instances of that category.

This back-projection process can be iteratively repeated by suppressing $(\hat{x},\hat{y})$ (and its 3×3 neighborhood) in $V(x,y)$ as part of the Non-Maximal Suppression (NMS) process, also implemented as part of the neural network.

5. Two-stage StarNet

Having computed the $BP_{Q|S}$ and $BP_{S|Q}$ back-projection heatmaps, we follow the best practice of the modern 2-stage CNN detectors, such as FasterRCNN [16], to enhance the StarNet performance with a second stage classifier that benefits from category instances localization produced by StarNet (in $BP_{Q|S}$ and $BP_{S|Q}$). We first normalize each of the $BP_{Q|S}$ and $BP_{S|Q}$ to sum to 1, and then generate the following pooled feature vectors by weighted global average pooling with $BP_{Q|S}$ and $BP_{S|Q}$ weights:

$$F_{Q|S} = \sum_{i,j} BP_{Q|S}(i,j) \cdot Q_{i,j}, \quad F_{S|Q} = \sum_{l,m} BP_{S|Q}(l,m) \cdot S_{l,m}$$

method	ImageNetLOC-FS 1-shot	ImageNetLOC-FS 5-shot	CUB 1-shot	CUB 5-shot	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot
Baseline++ [2]	-	-	69.55	85.17	53.97	76.16
ProtoNet$^{(2)}$ [17]	-	-	72.03	87.42	54.16	74.65
CTM [10]	-	-	-	-	62.05	78.63
Δ-encoder [14]	-	-	69.80	82.60	59.90	69.70
CAN [5]	57.1$^{(3)}$	73.9	75.01$^{(3)}$	86.8	**63.85**	79.44
MetaOpt$^{(1)}$ [8]	57.7$^{(3)}$	74.8	72.75$^{(3)}$	85.83	61.77 (62.64)	77.9 (78.63)
StarNet (ours)	**63.0**	**78.0**	**79.58**	**89.5**	62.33	**79.6**

Table 1. Few-shot classification results, in % (accuracy). $^{(1)}$As we base our implementation on MetaOpt public code, for fair comparison MetaOpt results are reported as they are reproduced using their public code using their reported optimal hyper-parameters; MetaOpt paper reported results are in parenthesis. $^{(2)}$Results according to [2]. $^{(3)}$using official code.

Here the feature grids $Q_{i,j}$ and $S_{l,m}$ can be computed using ϕ as above, or using a separate CNN backbone trained jointly with the first stage network. Our second stage is a variant of the Prototypical Network (PN) classifier [17]. We compute the prototype for class c and the query Q embedding to be compared to it as:

$$F^P_{c|Q} = \frac{1}{k} \cdot \sum_{\substack{S \in E \ s.t. \\ C(S)=c}} F_{S|Q}, \qquad F^P_{Q|c} = \frac{1}{k} \cdot \sum_{\substack{S \in E \ s.t. \\ C(S)=c}} F_{Q|S}$$

Note that as opposed to PN, our query embedding and class prototypes are jointly query and class dependent. For a given query Q, we have a different query embedding $F^P_{Q|c}$ and a different class prototype $F^P_{c|Q}$ for each class-query pair. Finally, the score (logit, input to SoftMax) of the second stage classifier for assigning label c to the query Q is computed as:

$$SC_2(c;Q) = -||F^P_{Q|c} - F^P_{c|Q}||_2 \qquad (5)$$

In our experiments, whenever 2-stage StarNet is used, we compute its final score as a geometric mean of the first and the second stage classifiers ($sm = softmax$):

$$SC(c;Q) = \sqrt{sm(SC_1(c;Q)) \cdot sm(SC_2(c;Q))} \qquad (6)$$

6. Experiments

In all of experiments, only the class labels were used for training, and the object bounding boxes were used only in weakly-supervised few-shot detection tests. For each dataset we used the standard train / validation / test splits of classes and samples.

6.1. Few-shot classification

We evaluated the StartNet few-shot classification performance on datasets *mini*ImageNet, *tiered*ImageNet, CIFAR-FS, FC100, CUB, ImageNetLOC-FS [6]. Standard protocol [8], was used for this evaluation, with 1000 random 5-way episodes. The results are summarized in Table 1.

For fair comparison, we only list results that were obtained: (1) without using validation data for training; (2)

without transductive setting; (3) using standard 84×84 (or 32×32) resolution; (4) using plain ResNet backbones (we use ResNet-12); and (5) not using additional semantic information.

StarNet main performance gains, of over 4.5% and 5% above SOTA baselines in 1-shot setting, are observed in the CUB and ImageNetLOC-FS respectively. This is expected, as StarNet is optimized to classify the objects through their localization, and hence has an advantage for benchmarks with images that are less cropped around the objects. Interestingly, we observe these gains also above the SOTA attention based method of [5] on these benchmarks.

6.2. Weakly-Supervised Few-Shot Object Detection

For WS-FSOD experiments, we have used ImageNetLOC-FS [6] and CUB datasets. The bounding box annotations were used only for evaluation. The ImageNetLOC-FS is a diverse dataset with over 200 test categories, on which the fully-supervised few-shot object detector RepMet [6] serves as a natural performance upper bound for StarNet. For the CUB, the same split as for the few-shot classification experiments was used. Since, to the best of our knowledge, StarNet is the first method proposed for WS-FSOD, we compare its performance to several baselines, with results summarized in Table 2. The first two are based on MetaOpt classifier [8] combined with GradCAM or SelectiveSearch for localization. PCL [18] is a recent WSOD method, and CAN [5] is a SOTA attention based few-shot method with some localization ability. The same training split was used for all methods. For StarNet, MetaOpt+GradCAM, and CAN, the bounding boxes were obtained from the predicted heatmaps using the CAM algorithm from [21] (as in most WSOD works). StarNet results are higher by a large margin then results obtained by all the compared baselines and is almost on par with the fully supervised few-shot RepMet detector for $IoU \geq 0.3$

7. Conclusions

We have introduced StarNet, a model for few-shot classification which naturally includes object localization in images. StarNet is capable of providing plausible explanations for its predictions by highlighting the corresponding image regions on query and matched support images. Moreover, our method allows to approach, for the first time (to the best of our knowledge) the task of *weakly-supervised* few-shot object detection, with reasonable accuracy.

		1-shot		5-shot	
dataset	method	IoU>0.3	IoU>0.5	IoU>0.3	IoU>0.5
Imagenet LOC-FS	RepMet (upper bound)	59.5	56.9	70.7	68.8
	MetaOpt+GC	32.4	13.8	51.9	22.1
	MetaOpt+SS	16.1	4.9	27.4	10.2
	PCL [18]	25.4	9.2	37.5	11.3
	CAN [5]	23.2	10.3	38.2	12.7
	StarNet (ours)	**50.0**	**26.4**	**63.6**	**34.9**
CUB	MetaOpt+GC	53.3	12.0	72.8	14.4
	MetaOpt+SS	19.4	6.0	26.2	6.4
	PCL [18]	29.1	11.4	41.1	14.7
	CAN [5]	60.7	19.3	74.8	26.0
	StarNet (ours)	**77.1**	**27.2**	**86.1**	**32.7**

Table 2. Average precision (AP, %) of weakly supervised 5-way few-shot detection and comparison to baselines on the ImagenetLOC-FS and CUB datasets. GC = GradCAM, SS = SelectiveSearch. RepMet is a fully-supervised upper bound.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly Supervised Deep Detection Networks. CVPR, pages 2846–2854, 2016. 2

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A Closer Look At Few-Shot Classification. In ICLR, 2019. 1, 3

[3] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. CVPR, pages 437–446, 2015. 2

[4] Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao. Collect and Select : Semantic Alignment Metric Learning for Few-Shot Learning. ICCV, pages 8460–8469, 2019. 2

[5] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross Attention Network for Few-shot Classification. NeurIPS, 10 2019. 2, 3, 4

[6] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M. Bronstein. RepMet: Representative-based metric learning for classification and one-shot object detection. CVPR, pages 5197–5206, 2019. 2, 3, 4

[7] Leonid Karlinsky, Joseph Shtok, Yochay Tzur, and Asaf Tzadok. Fine-grained recognition of thousands of object categories with single-example training. CVPR, pages 965–974, 2017. 1, 2

[8] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-Learning with Differentiable Convex Optimization. In CVPR, 2019. 1, 3, 4

[9] Bastian Leibe, Ales Leonardis, and Bernt Schiele. An Implicit Shape Model for Combined Object Categorization and Segmentation. (May):508–524, 2006. 2

[10] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding Task-Relevant Features for Few-Shot Learning by Category Traversal. 1, 2019. 3

[11] Tsung-Yu Lin, Aruni Roychowdhury, and Subhransu Maji. Bilinear CNNs for Fine-grained Visual Recognition. TPAMI, 2017. 2

[12] David McAllester Pedro F. Felzenszwalb, Ross B. Girshick and Deva Ramanan. Object Detection with Discriminatively Trained Part Based Models. PAMI, 32(9):1627–1645, 2010. 2

[13] E. Sali and S. Ullman. Combining Class-Specific Fragments for Object Classification. (1994):1–21, 2013. 1

[14] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Delta-Encoder: an Effective Sample Synthesis Method for Few-Shot Object Recognition. NeurIPS, 2018. 3

[15] 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. ECCV, pages 618–626, 2017. 1

[16] Ross Girshick Jian Sun Shaoqing Ren, Kaiming He. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Neural Information Processing Systems (NIPS), 2015. 3

[17] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In NIPS, 2017. 1, 3

[18] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. PAMI, 42(1):176–191, 2020. 4

[19] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly Supervised Region Proposal Network and Object Detection. In ECCV, 2018. 2

[20] Davis Wertheimer and Bharath Hariharan. Few-Shot Learning with Localization in Realistic Settings. 2019. 2

[21] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. Lecture Notes in Computer Science, pages 610–625, 2018. 4