A First Look Towards One-Shot Object Detection with SPOT for Data-Efficient Learning

Ria Chakraborty, Madhur Popli, Rachit Lamba, Rishi Verma
Product Assurance, Risk and Security (PARS)
Amazon, India
[riacha, madpopli, lrachit, vermrish]@amazon.com

Abstract

In this work we discuss One-Shot Object Detection, a challenging task of detecting novel objects in a target scene using a single reference image called a query. To address this challenge we introduce SPOT (Surfacing POsitions using Transformers), a novel transformer based end-to-end architecture which uses synergy between the provided query and target images using a learnable *Robust Feature Matching* module to emphasize the features of targets based on visual cues from the query. We curate LocateDS - a large dataset of query-target pairs from open-source logo and annotated product images containing pictograms, which are better candidates for the one-shot detection problem. Initial results on this dataset show that our model performs significantly better than the current state-of-the-art. We also extend SPOT to a novel real-life downstream task of *Intelligent Sample Selection* from a domain with very different distribution.

1 Introduction

The availability of large amount of labeled data [8, 4, 3] for training has been a key enabler for the success of deep learning in Object Detection tasks [10, 5, 1, 9]. While open source data facilitate research and development in academic settings, real-world use cases often need domain specific datasets which are not only costly but challenging to obtain. In many domains like product compliance, where there is dynamicity in the nature of legal requirements, a massive backlog of products is created for reassessments whenever the compliance definitions change. These definitions can range from having certain declarations displayed on the product packaging to mandating the presence of certain regulatory logos and pictograms. Sellers and platforms are then essentially stuck in a situation where they need to wait longer to onboard their offerings because of new or updated compliance requirements and platforms need to onboard a massive human workforce to go through millions of product images to ensure compliance. This end-to-end process is managed manually until human annotators gather enough data and a machine learning model is trained to help with this workload. Therefore data efficient learning becomes critical for success in these areas.

In this paper we target the challenging problem of one-shot object detection to optimise the data labeling process for object detection tasks. We first introduce our one-shot detector SPOT (Surfacing POsitions using Transformers). SPOT is a novel end-to-end trainable model based on transformers [12]. For our problem formulation, we assume that an example of the unseen class (query) would be provided, and the task will be to uncover all the regions in the target image which are visually similar to the query. SPOT emphasizes the features of the target image based on visual cues from the query using *Robust Feature Matching* (RFM) for detecting objects of novel classes. We then demonstrate SPOT's generalization capabilities on previously unseen object categories and beat the state-of-the-art by a significant margin on LocateDS, a curated dataset consisting of open-soure Logo Images, annotated product images containing pictograms and Synthetic Images. We also formulate and

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Figure 1: Qualitative results from SPOT - best viewed when zoomed. All the queried items above are unseen to the model. As can be seen, SPOT is able to locate the queried items in the target images, irrespective of their scale, color and transformation differences.

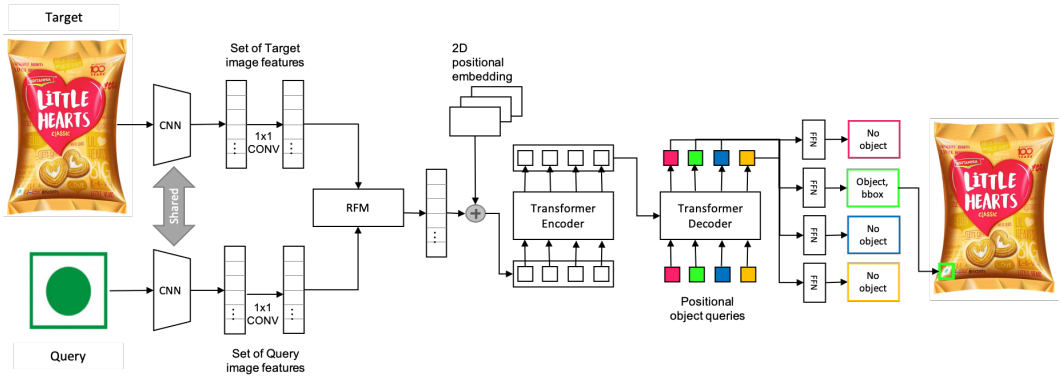

Figure 2: Architecture of our one-shot detector SPOT.

demonstrate SPOT's generalisation capabilities by its extension towards a novel real-life downstream task of *Intelligent Sample Selection*, which essentially optimises the way samples are selected for collecting annotations for object-detection tasks. Some qualitative results from SPOT are presented in Figure 1.

2 Method

We build a one-shot object detection network by extending DETR [2], an end-to-end detection model composed of a backbone (typically a convolutional residual network [6]), followed by a Transformer Encoder-Decoder [12]. DETR streamlines the detection pipeline, effectively removing the need for many hand-designed components like a non-maximum suppression procedure or anchor generation that explicitly encode the prior knowledge about the task.

2.1 SPOT

The architecture of SPOT is shown in figure 2. We removed the Fully Connected layers and the Global Pooling layer from a pre-trained ResNet50, and used the remaining convolutional layers as our feature extractor module. The target and query images are encoded by this same convolutional backbone to produce a sequence of hidden vectors. We then subject these two sets of features through a 1×1 convolution to create the feature maps $f(t) \in \mathbb{R}^{d \times H_t \times W_t}$ and $f(q) \in \mathbb{R}^{d \times H_q \times W_q}$. Here H_t and W_t denote the spatial dimensions of the feature map for target image and d denotes the number of channels. Similary, H_q and W_q are dimensions of the feature map of query image.

Robust Feature Matching The Robust Feature Matching (RFM) module acts as a robust patch-wise similarity encoder. This module has multiple learnable similarity function approximators assessing different aspects of $f(t)$ and $f(q)$ to emphasize those feature vectors in $f(t)$ that bear close similarity

with the feature vectors in $f(q)$. If d is the channel dimension of $f(q)$, then RFM encodes each aspect of similarity as a variant of:

$$g(f(q), f_{i,j}(t)) = f_{i,j}(t)(w \times f(q)) \tag{1}$$

where w can be estimated as: $w = \text{softmax}\left(\frac{f_{i,j}(t)f(q)^\top}{\sqrt{d}}\right)$. i and j are feature accessors for $f(t)$.

These similarity approximator weights were combined in the same way as the Multi-Headed attention mechanism [12]. The output of this layer is a weighted sum of the values in $f(q)$ for each feature value in $f(t)$, where the weight assigned to each value of $f(q)$ is determined by its similarity to the feature channels of $f(t)$. We add 2-D positional embeddings to this emphasized target feature vector to conserve the spatial information. This sequence is fed into the transformer encoder. The transformer decoder takes as input a set of N learned embeddings called object-queries, that can be viewed as slots that the model needs to fill with detected objects. The number of object queries acts as an upper-bound on the number of objects the model can detect simultaneously. All the object queries are fed in parallel to the decoder, which uses cross-attention layers to look at the encoded image and predict the output embeddings for each of the object-queries. The final representation of each object query is independently decoded into box coordinates and class labels using a shared feed-forward layer. For our problem, we have a binary decision for the class label - 0 whenever the predicted patch matches the query, otherwise it is treated as part of the no-object class (encoded as 1 in our case).

Hungarian Algorithm and Losses SPOT infers a fixed-size set of N predictions in a single pass. The training objective is to score predicted objects (class, position, size) with respect to the ground truth. Like DETR, we use the Hungarian Algorithm to find an optimal bipartite matching between predicted and ground truth objects, and then optimize object-specific (bounding box) losses. The matching cost takes into account both the class prediction and similarity of predicted and ground truth boxes. That is, the final loss is a linear combination of negative log-likelihood for class prediction and a box loss. Please see [2] for more information on the losses.

2.2 Training Dataset: LocateDS

One-shot object detection is best suited for two-dimensional objects that have some regularity in the wild. Most works on one-shot object detection alter VOC and COCO datasets. However, the classes represented under these datasets are very broad, e.g., contain both objects and their parts annotated as one class, or the objects look very different because of large viewpoint variations and other 3-D deformations. While a large number of training images help learn these transformations, the same is not possible in the one-shot setting (e.g., detecting a full animal given only a head as the query image). Hence, as part of this work, we curated a new dataset - LocateDS containing query-target image pairs, along with the bounding box annotation(s) from logo and pictogram domains. This dataset has three primary sources: Open-Source, Manual and Synthetic. The Open-source contains images from LogoDet3K [13] and Openlogo [11], adapted for the one-shot problem using the approach outlined in [7]. The Manual source contains manually annotated product images containing pictogram(s) from publicly available product images in E-Commerce portals. For these target images, the queries are the corresponding good quality pictogram images, publicly available in Regulatory websites. We generated Synthetic data by concatenating multiple images together and pasting affine-transformed pictograms on random images. The Synthetic source contains target images with variety of different region(s) of interest together to constrain the model to factor-in the provided query image in order to output the correct region(s). Padding was used to simulate smaller region of interest sizes - a feature lacking in most of the logo datasets, where many times, almost the whole image is a logo. The distribution of different splits were as follows: Train: 89,955, Validation: 56,033 and Test: 25,901 query-target pairs. Number of classes in Train: 2,481, Validation: 630 (90% unseen) and Test: 302 (98% unseen). Please note, the Validation and Test split contains mostly unseen and a few seen classes. Distribution of seen and unseen classes of the Test split is shown in Table 1.

3 Initial Results

We compare our method with the current state-of-the-art model for one-shot object detection, CoAE [7]. The comparison results are summarised in Table 1. The results show that SPOT outperforms existing state-of-the-art method by a significant margin for both seen and unseen classes. SPOT's

Table 1: Comparison with existing state-of-the-art one-shot object detection method (CoAE) on LocateDS for both seen and unseen logo related classes in terms of Average Precision (%) at IoU=0.5. Our model's performance is shown under method SPOT.

	Unseen		Seen				
Method	**LogoDet3K - 297 classes**	**Stitched**	**Bottega Veneta**	**Prada**	**Under Armour**	**Maxwell House**	**Thomson Reuters**
SPOT	**81.10**	**46.79**	**86.25**	**100.00**	**80.00**	**85.90**	**73.97**
CoAE	64.62	29.09	66.25	94.23	43.64	84.62	17.81
Counts	13,667	11,896	80	52	55	78	73

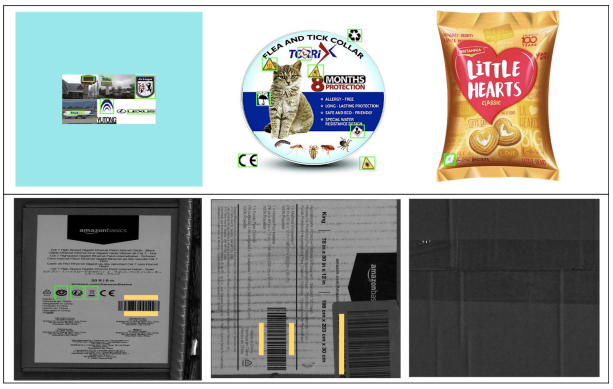

Figure 3: First row shows example images SPOT was exposed to during Training. Images in second row are examples of the domain SPOT was tested as an Intelligent Sample Selector. In this domain, the first image would be relevant since it contains pictograms of interest. The remaining two images would be irrelevant. Relevant regions marked in green. Confidential contents redacted in Yellow.

ability to generalise can be attributed to the RFM module's feature emphasizing capabilities and the end-to-end detection framework which enables it to be robust to scale and other transformation variations and free from prior encodings to a great extent.

3.1 Extending SPOT for Intelligent Sample Selection

In many real-life problems, only a handful of the total images contain region(s) of interest. As a result, randomly drawn samples for an annotation task ends up inflating the sample size with a significant portion of irrelevant images. This in turn, increases the amount of time taken towards completing the task. To demonstrate SPOT's generalisability, we simulated a system which needs to distinguish between a set of 12 pictograms from images of shipping boxes captured at E-commerce warehouses. Images containing pictograms were marked relevant and images without relevant pictograms were marked irrelevant. We then extended SPOT to cut down a portion of these irrelevant images. For this test, we collected a total of 5,256 randomly sampled annotated images where 1,022 contain pictograms. Example images from training and testing domains are shown in Figure 3.

Training and Inference We derived negative samples from LocateDS and trained SPOT on the same. For each negative category, we provided query instances not present in the target and empty bounding boxes. We then used each shipping box image as target and all the relevant pictogram images from Regulatory websites as the queries and flagged those target images for which the predicted score did not exceed the set threshold for any of these queries. Please note both these pictograms and images were unseen for SPOT.

Results and Future Work Under this extremely different distribution, SPOT was able to reduce the original dataset size by about 42% by removing irrelevant images with 97.1% precision. That is, SPOT was able to cut down 2,221 images from total sample of 5,256, resulting in reduced sample size of 3,035 samples. Within these 3,035 samples, 93.6% of the relevant images were covered. Being a one-shot detector, SPOT can naturally be extended to also auto-generate the bounding box annotations around novel regions of interest - an approach we are currently exploring to further optimise data labeling tasks.

Acknowledgments and Disclosure of Funding

We would like to thank Santosh Sahu, Pranesh Bhimarao Kaveri, Vesselin Diev and Mingwei Shen for their support and constructive feedback on this work. Special thanks to Shantanu Rai, Varun Nagaraj Rao and Gwang Lee for helping to compile the pictogram datasets.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:cs.CV/2004.10934

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. arXiv:cs.CV/2005.12872

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (Jan. 2015), 98–136.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:cs.CV/1311.2524

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:cs.CV/1512.03385

[7] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. 2019. One-Shot Object Detection with Co-Attention and Co-Excitation. arXiv:cs.CV/1911.12529

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:cs.CV/1405.0312

[9] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. arXiv:cs.CV/1612.08242

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:cs.CV/1506.01497

[11] Hang Su, Xiatian Zhu, and Shaogang Gong. [n.d.]. Open Logo Detection Challenge. In *British Machine Vision Conference*.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:cs.CL/1706.03762

[13] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. 2020. LogoDet-3K: A Large-Scale Image Dataset for Logo Detection. arXiv:cs.CV/2008.05359