Improving Attention-based Few-shot Object Detection

Yihao Huang
Phillips Academy
yhuang23@andover.edu

Nicholas Zufelt
Phillips Academy
nzufelt@andover.edu

Abstract

Few-shot object detection is a classic computer vision task aimed at developing detection models that can generalize to new classes with few training instances. While existing works have seen tremendous progress in recent years, we observe two critical deficiencies in existing metric-learning based few-shot object detectors. Existing few-shot detection methods often directly extend 1-shot detection methods and do not take into account the differences between support shots. Further, existing methods lose tremendous positional information while aggregating over multiple support shots. To address these deficiencies, we propose a novel Cross-Multiple-Image Spatial Attention module to perform query-informed aggregation of support information across different support shots. We further leveraged 3 dimensional positional embedding to enhance the preservation of locational information in the detector network. Our model can be easily inserted into a Faster RCNN or any other generic two-stage object detectors. The code for our paper can be found at this link: https://github.com/michaelyhuang23/Attention-FS-Det.git.

1 Introduction

In the past years, computer vision has achieved astonishing progress, surpassing humans in the task of image recognition. Despite the progress, these deep learning models require millions of images as well as several GPU-days to train. In contrast, humans can learn from limited data in a short amount of time [16]. The data-hungry nature of traditional deep learning object detection techniques render them impractical in many real-life situations where only limited training data is provided or the model is required to quickly learn to identify previously unseen objects, as is common in robotics [1].

The task of learning from limited number of data samples is termed few-shot learning. Few-shot image classification has seen tremendous progress in the past years. The principle idea is to transfer knowledge learned on a set of base-classes, which have abundant annotations, to quickly learn to classify images of a set of novel classes, which have scarse annotations. In contrast to image classification, few-shot object detection, which involves localizing objects as well as classifying them, has received much less attention. A performant way to solve few-shot object detection is augmenting a conventional object detector, such as the Faster RCNN, with a matching network. The matching network compares novel image features against support instances' features. The results from the comparison is used to classify and localize novel instances in the novel image.

In a recent work, Fan *et al.* proposed to inject support instance information before the Region Proposal Network in a Faster RCNN [6]. Chen *et al.* extended this work, designing a position-aware attention-based matching network, termed Cross-Image Spatial Attention (CISA) [3]. In order to compare a query feature map against support feature maps of multiple support instances, the CISA-based matching networks first generates a query-specific support feature map from all the inputed support feature maps, and then uses this feature map as the basis for comparison. The CISA-based model suffer from two deficiencies. Firstly, the model does not take into account the varying degree to which each support instance is useful for detections on a particular query. Secondly, a large amount of positional information of the support instances is still lost when the support vectors are averaged over all support instances.

To address these deficiencies, we propose an extension of CISA, the Cross-Multiple-Images Spatial Attention (CWISA) module. CWISA leverages a 3 dimensional attention to generate a query-specific support feature map from all support instances. The attention module inherently takes into account the usefulness of each support instance for the detection task on a particular query, thus resolving the first deficiency. The second deficiency is resolved by augmenting the attention module with 3 dimensional positional encoding, which helps preserve the spatial information from the support feature maps. Due to the modularity in its design, a CWISA module can be inserted before the RPN layer and the detector's final classification and regression layer to transform a conventional Faster RCNN into a few-shot detector.

Thus, our main contribution can be summarized as designing a novel CWISA module that has greater theoretical robustness than the existing CISA, in terms of weighting support instances as well as preserving locational information.

2 Related Works

2.1 Object Detection with Deep Learning

The development of deep convolutional neural networks in recent years has enabled deep-learning based object detection systems to achieve high accuracy given sufficient labeled data. Modern object detection approaches fall largely into two paradigms: single-stage detection and two-stage detection.

YOLO [12] is a prominent example of single-stage object detector. It uses a pretrained CNN backbone to extract generic visual informations from an image and then generate bounding boxes as well as their classification results directly. Other examples of single-stage detectors include SSD, YOLOv2, YOLOv3, and its various other derivatives [11, 13, 14].

Faster RCNN, on the other hand, is the paradigm of the two-stage object detection approach. It uses a region proposal network (RPN) to propose regions of interests, which represent the foreground objects in the image. Then, the features within regions of interests are further processed before bounding boxes and classification scores are generated for each of them. [2, 4, 5, 9] are all two-stage detectors and derivatives of Faster RCNN.

In our work, we model our few shot detector based on the Faster RCNN architecture because of the prevalence of its derivatives and its popularity in prior few-shot object detection works, but it should be noted that our proposed CWISA module works on any object detector.

2.2 Few-shot Image Classification

Most few-shot image classification approaches are meta-learning based, meaning that they seek to transfer meta-level knowledge from base classes to novel classes.

A major line of work focuses on prototypical networks, where query image's feature embedding is compared to prototypical embeddings of different classes to determine the image's class using a nearest neighbor search [17]. Other metric-learning approaches adopt a small neural network to make the comparison between query and support images [18].

Another line of work focuses on inter-class knowledge transfer through weight generation. These methods seek to learn a meta-level weight generator that generates classification weights given from support images [8, 15].

2.3 Few-shot Object Detection

Despite having received much less attention than few-shot image classification, two types of methods for few-shot detection exist: the fine-tuning based methods and the meta-learning based methods. Both methods build upon a traditional object detection model, such as the Faster RCNN.

Pioneered by Wang *et al.* fine-tuning based approaches first train the object detector on image-abundant base classes and then fine-tune the pre-trained model on the limited novel class dataset [20, 7].

The other category of models adopt meta-learning to transfer meta-level knowledge from the base classes to novel classes by learning how to learn on the base classes. A popular line of work is metric learning, in which the model learns to compare a novel image against support instances [21, 22, 6, 23, 3]. In this type of model, a object detector, such as the Faster RCNN, is augmented with a matching network which compares novel image features against support instances' features. Predictions are generated from the results of the comparison.

In recent years, several works [6, 3] focused on injecting the support instance information into the matching network early on, allowing the comparison between query and support to be operated on lower level features that are more generalizable and thus robust in a few-shot setting.

3 Methodology

In this section, we formally define the problem of few-shot object detection and provide a detailed description of our model architecture.

3.1 Problem Definition

Given a query image q and a set of support instances $S_c = \{s_i\}$ of category c, the problem of few-shot object detection aims to find all objects of category c inside q. Each $s_i \in S_c$ contains a bounding box annotation of an object of category c in some support image. The detection task is labeled N-way K-shot when there are N categories and each category's support set S_c contains K instances.

3.2 Cross-Multiple-Image Spatial Attention

Existing matching networks used for few-shot object detection do not fully utilize the informational potential of all K support instances. This is because existing matching models do not directly compare image features from the query image against image features from the support images. Instead, the query image features are compared against reference support features obtained as an average across all support instances. This brute-force averaging approach does not take into account the possibility that some support instances may be more helpful for the detection of a particular query object because it shares more similarities in lighting, color, orientation, or posture with the target query object. These support instances' features should therefore be weighted more highly in the generation of the reference support features. To resolve this deficiency, we propose the Cross-Multiple-Image Spatial Attention (CWISA) module.

The principle idea behind CWISA is to leverage the support feature vectors of every single support instance at every single spatial location in the cropped close-up images of those support instances to produce the reference support features. The reference support features are then compared against the novel image features for the detections of the novel-class object instances. To aggregate support features from all support instances to generate the reference support features, we employ an attention system similar to the QKV Attention module.

The reference support features are generated at each pixel location independently, corresponding to the query image's feature vector at that pixel location. We will now discuss the generation of the support feature vector at one particular pixel location (h, w), which we denote as Y_{hw}.

The query image's feature vector at that pixel location Q_{hw} is used as the query in a QKV Attention framework. The query is then applied across a set of keys from the support instances. Each key is a feature vector of one particular support instance k at one particular pixel location (h, w). We denote it by $S^k_{h'w'}$. The support feature vectors are also used as the values.

The attention map can be formulated as follows

$$A_{hwkh'w'} = \sigma(Q_{hw} \cdot S^k_{h'w'})$$

σ is the categorical cross-entropy activation function and it is applied across all spatial locations as well as across all support instances. Note that each $A_{hwkh'w'}$ is a scalar indicating the similarity or correlation between a particular query feature vector Q_{hw} and a particular support feature vector $S^k_{h'w'}$. Then, the reference support features Y_{hw}, can be determined by

$$Y_{hw} = \sum_{\text{all } k,h',w'} A_{hwkh'w'} S^k_{h'w'}$$

Effectually, Y_{hw} is a weighted aggregation of of all support feature vectors at every pixel location of every support image instance. We can combine the two equations to obtain the full formulation for Y_{hw}

$$Y_{hw} = \sum_{\text{all } k,h',w'} \sigma(Q_{hw} \cdot S^k_{h'w'}) S^k_{h'w'}$$

3.3 Three Dimensional Positional Encoding

To obtain the final reference support features, Y_{hw} is concatenated with a three dimensional positional encoding. The positional encoding is a multi-dimensional extension of the positional encoding introduced for natural language processing [19]. The positional encoding records the k coordinate, h' coordinate, w' coordinate of each $S^k_{h'w'}$ and is concatenated with the feature vector $S^k_{h'w'}$ to form ${S'}^k_{h'w'}$ before it is passed through the attention module.

In other words, using $\oplus$ to denote the concatenation operation, the modified formulation for Y_{hw} can be written as

$$Y_{hw} = \sum_{\text{all } k,h',w'} \sigma(Q_{hw} \cdot S^k_{h'w'}) \left(S^k_{h'w'} \oplus P_k \oplus P_{h'} \oplus P_{w'}\right)$$

3.4 Model Architecture

Following the structure pioneered by [3], we concatenate the output standard support features Y_{hw} with the query features Q_{hw}. The concatenated result is passed through a single layer CNN before being passed into the Region Proposal Network of our Faster RCNN. Alternatively, the concatenated result is passed through a MLP network and then consumed by the final classifcation and regression layers. For Faster RCNN models that use Feature Pyramid Network (FPN) [10], the CWISA module is inserted at every level of the feature pyramid.

During a forward pass, the query image, along with all support images, are passed through the shared backbone of the Faster RCNN. Then, support instance regions are cropped from the support images based upon each instance's bounding box. A CWISA module is applied between the query features and the support features of all support instances. The generated standard support features are concatenated with the query features and passed into a CNN followed by the RPN to generate object proposals for the Faster RCNN. During training, only object proposals of the target category are considered foregrounds. The rest is considered part of the background. RoI pooling is performed on the original Q_{hw} features. Before the final classification and localization layers, another CWISA module is applied on pooled query features and the support features of support instance regions pooled by a similar scale. The final concatenated result is used to predict categories and generate bounding boxes.

4 Conclusion

In this work, we pointed out the deficiencies of existing metric-learning based few-shot object detectors with regard to the preservation of locational information and the inefficacy of the brute-force aggregation-based approach to combine features of different support shots. We designed a novel attention-based feature aggregator that aggregates features at all pixel position of all support shots in a manner that is informed by the query image features. We proposed to use a three dimensional positional embedding to enhance the preservation of positional information in the aggregation process. Finally, we implemented our model with the Detectron2 framework, which can be found at this GitHub link: https://github.com/michaelyhuang23/Attention-FS-Det.git.

References

[1] A. Ayub and A. R. Wagner. Tell me what this is: Few-shot incremental object learning by a robot. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8344–8350, 2020. 1

[2] Z. Cai and N. Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *CoRR*, abs/1906.09756, 2019. 2

[3] T.-I. Chen, Y.-C. Liu, H.-T. Su, Y.-C. Chang, Y.-H. Lin, J.-F. Yeh, W.-C. Chen, and W. H. Hsu. Dual-awareness attention for few-shot object detection. 2021. 1, 2, 3, 4

[4] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016. 2

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks, 2017. 2

[6] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai. Few-shot object detection with attention-rpn and multi-relation detector, 2019. 1, 2, 3

[7] Z. Fan, Y. Ma, Z. Li, and J. Sun. Generalized few-shot object detection without forgetting, 2021. 2

[8] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. 2

[9] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Light-head r-cnn: In defense of two-stage object detector, 2017. 2

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection, 2016. 4

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015. 2

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection, 2015. 2

[13] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. 2

[14] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 2

[15] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell. Meta-learning with latent embedding optimization, 2018. 2

[16] L. B. Smith, S. S. Jones, B. Landau, L. Gershkoff-Stowe, and L. Samuelson. Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1):13–19, 2002. 1

[17] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning, 2017. 2

[18] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning, 2017. 2

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. 4

[20] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu. Frustratingly simple few-shot object detection, 2020. 2

[21] Y.-X. Wang, D. Ramanan, and M. Hebert. Meta-learning to detect rare objects. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9924–9933, 2019. 2

[22] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin. Meta r-cnn : Towards general solver for instance-level few-shot learning, 2019. 2

[23] G. Zhang, Z. Luo, K. Cui, and S. Lu. Meta-detr: Image-level few-shot object detection with inter-class correlation exploitation, 2021. 2