Few-shot learning for Object Detection

Hieu Trung Le (Eric Le)
Stanfor University
Computer Science Department
leric@stanford.edu

Abstract

Deep neural networks often excels at learning high level features based on a substantial amount of training data with the cost that preparing such high-quality training data is very labor-intensive. Though it often suffers when the data set is small or has poor sample efficiency. This project investigates the benefit of using few-shot learning to train detection model so that it learns to recognize unseen categories object without further training or finetuning.

We will explore the usage of Attention-RPN, Multi-Relation Detector and Contrastive Training strategy as outlined in Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector paper [10], which exploit the similarity between the few shot support set and query set to detect novel objects. We will train the network using the FSOD dataset [10], which is a new dataset which contains 1000 categories of varies objects with high quality annotations. This is also the first dataset specifically designed for few shot object detection. Once the network is trained, we will then apply object detection for unseen classes without further training or fine tuning. We will also demonstrate the effectiveness of our method quantitatively and qualitatively on different datasets.

1. Introduction

Deep convolutional neural networks (CNN) are often known to have huge success in analyzing huge amount of training data set with high precision and able to produce bounding box around object with accuracy. This often comes at a cost of having to produce such high quality training data-set that could cost lots of time and money. Also, deep CNN often are under performing if the training labeled data is scarce. In fact, deep CNNs sometimes fail to generalize over examples that are not well represented. Despite several years of research and evolution, most CNNs show a certain degree of weakness when dealing with high-dimensional and non-stationary environments. The ability of deep CNNs to generalize over novel concepts without abundant labeled data is active area of research and often known as few-shot learning. Few-shot learning refers to the practice of feeding a learning model a very small amount of training data and teaching it to generalize and to be able to recognize novel target object

Few-shot learning is challenging given large variance of illumination, shape, texture in real-world objects. Another challenge of few-shot learning is how to localize an unseen object in a cluttered background of an image. This problem is often caused by the inappropriate low scores of good bounding boxes output from a region proposal network (RPN). In another word, potential bounding boxes can miss unseen new target objects in an image with many objects in the background.

In this paper, we will address the problem of few-shot object detection. Given a support image s_c with a close-up of the target object and a query image q_c which potentially contains objects of the support category c, the task is to find all the target objects belonging to the support category in the query and label them with tight bounding boxes. If the support set contains N categories and K examples for each category, then this will become N-way K-shot problem.

The key to few-shot learning lies in the ability to generalize of the model when presented with novel categories. A high-diversity dataset with a large number of object categories is therefore necessary for training such model that can detect unseen objects. Existing datasets contain very limited categories and they are not designed to be used to train object detection model in few-shot evaluation setting. In this paper, we will use the Few-Shot Object Detection Dataset (FSOD) [10] that specifically designed for few-shot learning.

With designed contrastive training strategy, attention module on RPN and detector, the detection model will exploit matching relationship between object pairs in a weight-shared network to perform online detection on objects of novel categories by producing rectangular bounding box around these objects

2. Related work

General Object Detection Object detection is one of the many classic problems in computer vision. Object detection was usually formulated as a sliding window classification problem using handcrafted features. Dalal et Triggs [8] use Histograms of Oriented Gradient descriptors as feature sets for human detection task. Felzenszwalb et al [24] use discriminative training with partially labeled data for object detection task. Viola et al [28] brings together new algorithms and insights using a Boosted Cascade of Simple Features to construct a framework for robust and extremely rapid object detection. With the rise of deep learning,CNN-based methods become the dominant detection solution. There are 2 general approaches when it comes to CNN object detection: proposal-free detectors and proposal-based detectors. Proposal-free detectors follow a one-stage training strategy and do not generate proposal boxes explicitly. Example are in [18] and [29]. Proposal-based detectors, on the other hand, would first extract class agnostic region proposals of the potential objects in an image and draw boxes around them. These boxes are then further refined and classified to different categories by a specific module. Examples are in [12], [25] and [19]. The proposal-based methods usually perform better than the proposal-free ones and become leading state-of-the-art of detection task. These methods, however, require intensive supervision manner and are hard to expand to novel categories with only few examples. This motivates the need and research interest for few-shot learning in object detection.

Few-shot learning Few-shot learning is is a challenging problem for traditional machine learning algorithms because it is difficult to learn from just a few training examples. Earlier works such as [21], [5] and [6] attempts few-shot learning by taking advantage of visual concepts and knowledge coming from previously learned categories and learn features such as hand-designed strokes or parts, that can be shared across categories. Others such as [4], [9] and [16] focus on metric learning, metric scaling and metric task conditioning which aimed at manually designing the distance formulation among different categories in few-shot learning. A recent trend, called meta-learning, is a design of a general strategy that can guide the supervised learning within each task, by accumulating knowledge it gains the network an attribute to capture the structure varies across different tasks such as [13], [23] and [7]. Some works such as [30] and [33] exploit local descriptor to reap additional knowledge from the limited data. Gidaris et al [11] and Kim et al[27] introduce Graph Neural Network to model the relationship between different categories. Li et al [17] attempts to traverses across the entire support set and identifies task-relevant features to make metric learning in high-dimensional space more effective.

As we have seen, few-shot learning has achieved much very progress. It, however, mostly focus on the classification task, rarely on other computer vision tasks, such as semantic segmentation or object detection. Few-shot object detection is a task intrinsically different from the general few-shot learning on classification. Dong et al [32] harnesses unlabeled data and iterates between model training and high-confidence sample selection and embedding multiple detection models to optimizes on images without box. This method may be misled by incorrect detection in the weak supervision and requires re-training for a new category and it is out of our scope. Low-Shot Transfer Detector LSTD [14] proposed a novel few-shot object detection framework that can transfer knowledge from one large dataset to another small one, by minimizing the gap of classifying posterior probability between the source domain and the target domain. LSTD, however, strongly depends on source domain and hard to extend to a scenario very different from.

As motivated by Koch et al [13], we now implement a general few-shot object detection deep network that learn the matching on image pairs based on the Faster R-CNN framework

3. Methods

3.1. Problem definition

As mentioned, we will define few-shot object detection task as following: given a support image s_c with a close-up of the target object and a query image q_c which potentially contains objects of the support category c, the task is to find all the target objects belonging to the support category in the query and label them with tight bounding boxes. If the support set contains N categories and K examples for each category, then this will become N-way K-shot problem.

3.2. Deep Attentioned Siamese Framework for few-shot detection

Few-shot recognition is challenging due to the fact that it is hard to capture a common sense from just a few examples. To address this challenge, we will build an intensive attention network that learns a general matching relationship between the support set and queries on both the RPN module and the detector.

This attention network will consist of a siamese framework which has two branches and share weights, whose one branch is for support set and the other is for the query set, where the query branch of the siamese framework is a Faster R-CNN network, which contains two stages of RPN and detector. We will utilize this framework to train the matching relationship between support and query features, in order to enforce the network to learn general knowledge and the ability to capture common sense among the same categories. Based on the framework, we will further introduce

introduce a novel attention regional proposal network (or RPN) and detection with multi-relation modules to prompt an accurate parsing between support and potential boxes in the query.

3.3. Regional Proposal Network (RPN) for few-shot detection

In few-shot detection, regional proposal network (or RPN) helps to produce potentially relevant boxes for facilitating the task of object detection. RPN helps distinguish between objects and non-objects in the background by filtering out negative or irrelevant objects that does not belong to the support category. Without any support image information, the RPN will have to go over every single potential object with high objectness score even though they might not belong in the support category. First we compute the similarity between the support feature map and that of the query via depth-wise. we denote the support features as $X \in t^{S\times S\times C}$ and feature map of the query as $Y \in t^{W\times W\times C}$, the similarity will be defined as:

$$G_{h,w,c} = \sum_{i,j} X_{i,j,c} \cdot Y_{h+i-1,w+j-1,c}$$

where $i, j \in \{1, ..., S\}$. The support features X is used as the kernel to slide on the query feature map, then a depth-wise convolution between them is calculated. The attention map is processed by a 3×3 convolution followed by the objectiveness classification layer and box regression layer

In order to provide support information to RPN, we will implement the attention mechanism to guide the RPN to produce relevant proposals while suppressing proposals in other categories.

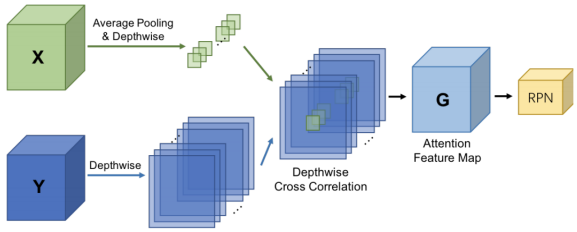

Figure 1. The attention RPN module

3.4. Multi-Relation Detector

The RPN module is then followed by a multi relation detector that is able to re-scoring proposals and class recognition. In another words, the detector will have strong discriminative ability to distinguish different object categories. The detector module will consist of three attention modules

- the patch-relation head to learn a deep non-linear metric for patch matching via one-to-many pixel relationship
- the global-relation head: to learn a deep embedding for global matching by using global representation to match images
- the local-correlation head: to learn the pixel-wise and depth-wise cross correlation between support and query proposals by capturing pixel-to-pixel matching relationship

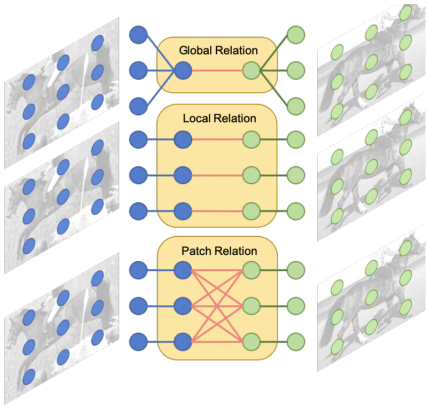

Figure 2. The multi relation detector with the global-relation head, the local-correlation head and the patch-relation head

We first concatenate the support and query proposal feature maps in depth. Then the combined feature map are fed into the patch-relation module, whose structure is shown in Table 1. All the convolution and pooling layers in this module have 0 padding to reduce the feature map from 7 x 7 to 1 x 1 which is used as inputs for the binary classification and regression heads

Type	Filter Shape	Stride/Padding
Avg Pool	3x3x4096	stride = 1/padding = 0
Conv	1x1x512	stride = 1/padding = 0
Conv	3x3x512	stride = 1/padding = 0
Conv	1x1x2048	stride = 1/padding = 0
Avg Pool	3x3x2048	stride = 1/padding = 0

Table 1: Architecture of patch-relation CNN

The global-relation head extends the patch relation to model the global-embedding relation between the support and query proposals. Given a concatenated feature of support and its query proposal, we average pooling the feature to a vector with a size of 1 x 1 x 2C. We then use an multilayer perceptron with two fully connected (FC) layers followed by ReLU and a final FC layer to generate matching scores.

The local-correlation head then computes the pixel-wise and depth-wise similarity between object ROI feature and the proposal feature by performing dot product on feature

pair on the same depth. We first use a weight-shared 1 x 1 x C convolution to process support and query features individually. The depth-wise similarity feature of size 1 x 1 x C is then calculated. Finally, a successive FC layer is used to generate matching scores.

All three relation heads contain different attributes and can well handle the complex attributes where the patch-relation head can generate flexible embedding that be able to match intraclass variances, global-relation head handle stable and general matching, and local-relation patch do matching on parts by capturing pixel-to-pixel matching relationship

The three matching modules will complement each other to produce higher performance to distinguish between different object categories.

Instead of following naive training strategy which is to match the same category objects by constructing a training pair (q_c, s_c) where the query image q_c and support image s_c are both in the same c-th category object, our model instead will not only be able to match objects from the same category but also distinguish objects in different categories. The 2-way contrastive training is used to achieve this objective.

The aim of 2-way contrastive training is match the same category while distinguishing different categories. First we randomly choose one query image q_c, one support image s_c with the same c-th category object and one other support image s_n containing a different n-th category object. We will combine to form the training triplet (q_c, s_c, s_n), where $c \neq n$ Only the c-th category objects in the query image are labeled as foreground while all other are treated as background. During training, the model learns to match every proposal generated by the attention RPN in the query image with the object in the support image. Therefore, we have trained the model to learn not only match the same category objects between (q_c, s_c) but also distinguish objects in different categories between (q_c, s_n)

For the baseline, we will use the Faster RCNN (Ren et al. 2015) [26] and SSD (Liu et al. 2016) [22]. Faster RCNN is a popular region-proposal architecture, where object proposals are firstly generated from region proposal network (RPN) and then fed into Fast RCNN. SSD is a widely used one-stage detection architecture, where the multi-layer design of bounding box regression can efficiently localize objects with various sizes. We will also compare the performance against LSTD [15] and RepMet[20]

4. Dataset and Features

We will use the FSOD: A Highly-Diverse Few-Shot Object Detection Dataset. The key to few-shot learning is the generality ability of the model on novel categories. Therefore, a high-diversity dataset with a large number of object categories is necessary to train a model so that it can be generalized enough to detect unseen objects in unknown categories and also to provide a convincing evaluation

The dataset is build from existing massive supervised object detection datasets as mentioned in Krizhevsky et al [1] and Kuznetsova et al [2]. This raw dataset cannot be used directly for training because 1) the label system of different datasets are inconsistent (some objects with the same semantic use different words) 2) large amount of annotations are less than satisfactory due to the inaccurate and missing labeling 3) the train/test split contains the same categories, while for few-shot dataset we want the train/test sets contain different categories

4.1. Data Preprocessing

The data is first pre-processed by merging the leaf labels in their original label trees, group those of same semantics to one category, and remove some semantics that does not belong to any leaf categories. The images with bad labeling quality and those with boxes of improper size (smaller than 0.05 %) are removed. The dataset is then split into the training set and test set whose categories has no overlap. The test set which contains 200 categories is then split by choosing those with the largest distance with existing training categories. The remaining categories are merged into the training set that in total contains 800 categories.

4.2. Data Analysis

This dataset is specifically designed for few-shot learning and for evaluating the generality of a model on novel categories, which contains 1000 categories with 800/200 split for training and test set. This has around 66,000 images and 182,000 bounding boxes in total. The dataset contains 83 parent semantics which are split into 1000 leaf categories with very clear category split for training and testing, where 531 categories come from ImageNet Dataset and 469 from Open Image Dataset. The dataset contains objects with large variance on box size and aspect ratios. The test set also contains a large number of boxes of categories not included in the label system. This makes it challenging for our few-shot detection model during testing. The detailed statistic is shown in Table 2

	Train	Test
No. Class	800	200
No. Image	52350	14152
No. Box	147489	35102
Avg No. Box / Img	2.82	2.48
Min No. Img / Cls	22	30
Max No. Img / Cls	208	199
Avg No. Img / Cls	75.65	74.31
Box Size	[6, 6828]	[13, 4605]
Box Area Ratio	[0.0009, 1]	[0.0009, 1]
Box W/H Ratio	[0.0216, 89]	[0.0199, 51.5]

Table 2: Dataset Summary

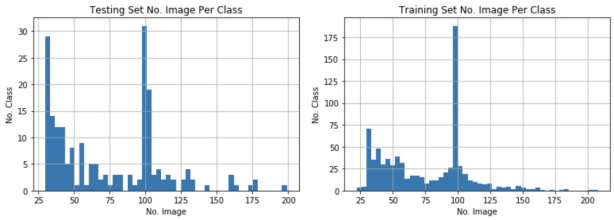

Figure 3. Image distribution by class

The dataset is built from existing large-scale object detection datasets for supervised learning. During training, the shorter side of the query image is resized to 600 pixels. We will also cap the longer side of the image to be 1000. The support image is cropped to be 16-pixel image context, zero-padded and then resized to a square image of 320 x 320

5. Experiments/ Results/Discussion

First we will train our model on Few shot Object Detection (FSOD) Dataset training set (800 categories) and evaluate on the test set (200 categories) using average precision (AP) metrics. Average precision (AP) is a popular metric in measuring the accuracy of object detectors. Average precision computes the average precision value for recall value over 0 to 1.

We follow the N-way K-shot evaluation protocol proposed in RepMet to evaluate our relation heads and other components. The K-way N-shot evaluation helps evaluate our approach using standard object detection evaluation. For each image in the test set, a test episode is constructed by the test image, N random support images containing its category k, and N support images for each of other K 1 categories, where the K 1 categories are randomly selected. By testing in one episode, we in fact perform a K-way N-shot evaluation, where each query is detected by 5 supports of its category and the other four non-matching categories

We will conduct the ablation study of the proposed multi-relation detector under the naive 1-way 1-shot training strategy and 5-way 5-shot evaluation on the FSOD dataset. The metrics we will use to evaluate is AP with the thresholds of 0.5 (AP_{50}) and 0.75 (AP_{75}).We use the same evaluation setting and metrics for all ablation studies on the FSOD dataset.

The model is end-to-end trained based on 4 Tesla GPUs using SGD with a weight decay of 0.0002 and momentum of 0.8. The learning rate is 0.001 for the first 50000 iterations, and 0.0001 for the later 10000 iterations. During our training, we find that having more training iterations will impact the performance negatively. We think may be too many training iterations led to over-fitting of model on the training set.

5.1. Relation head Ablation Study

We follow the N-way K-shot evaluation protocol proposed in RepMet [20] to evaluate which relation heads are important. Table 3 shows the ablation study of the proposed multi-relation detector under the naive 1-way 1-shot training strategy and 5-way 5-shot evaluation on the FSOD dataset.

Training	Global Rel	Local Rel	Patch Rel	AP_{50}	AP_{75}
1-w-1-s	✓			45.2	32.0
1-w-1-s		✓		51.2	33.2
1-w-1-s			✓	42.3	32.1
1-w-1-s	✓		✓	49.6	34.7
1-w-1-s		✓	✓	52.5	37.8
1-w-1-s	✓	✓		54.2	38.4
1-w-1-s	✓	✓	✓	57.3	40.1
5-w-5-s	✓			63.1	52.0
5-w-5-s		✓		67.4	55.2
5-w-5-s			✓	60.3	32.1
5-w-5-s	✓		✓	65.6	54.7
5-w-5-s		✓	✓	66.7	57.2
5-w-5-s	✓	✓		69.5	58.1
5-w-5-s	✓	✓	✓	71.7	56.1

Table 3: Experimental results for different relation head combinations in the 1-way 1-shot and 5-way 5-shot training strategy

For individual heads, the local-relation head performs best on both AP_{50} and AP_{75} evaluations. The patch-relation head performs worse than other relation heads. This could be because patch-relation head tries to model more complicated relationship between images. We believe that the complicated relation head makes the model difficult to learn. When combining any two types of relation head, we obtain better performance than that of individual head. By combining all relation heads, we obtain the full multi-relation detector and achieve the best performance, showing that the three proposed relation heads are complementary to each other for better differentiation of targets from non-matching objects. All the following experiments thus adopt the full multi-relation detector

5.2. Attention RPN vs regular RPN

We will examine the effect of using attention RPN and those with the regular RPN in different training strategies. The attention RPN produces 3.7%/3.1% gain in the 1-way 1-shot training strategy, 3.0%/2.3% gain in the 2-way 5-shot training strategy and 6.5%/4.3% gain in the 3-way 5-shot training strategy on the AP_{50}/AP_{75} evaluation. The model with attention RPN consistently performs better than the regular RPN on both AP_{50} and AP_{75} evaluation. These results confirm that the attention RPN generates better pro-

posals and benefits the final detection prediction

Training Strategy	Attenion RPN	AP50	AP75
1-way 1-shot	no	53.2	37.1
1-way 1-shot	yes	56.9	40.2
2-way 1-shot	no	62.3	41.7
2-way 5-shot	no	65.2	44.2
2-way 5-shot	yes	68.2	46.5
3-way 1-shot	no	57.8	37.5
3-way 2-shot	no	58.2	36.9
3-way 5-shot	no	59.1	38.5
3-way 2-shot	yes	61.2	41.9
3-way 5-shot	yes	63.8	40.2
3-way 5-shot	yes	65.6	42.8

Table 4: Experimental results for training strategy and attention RPN

5.3. ImageNet dataset

We will compare our results with those of LSTD [15] and RepMet[20] on the challenging ImageNet based 50-way 5-shot detection scenario. Table shows experimental results on ImageNet Detection dataset for 50 novel categories with 5 supports using LSTD [15], RepMet[20] and our proposed method.

Method	Training Dataset	Finetuning	AP_{50}	AP_{75}
LSTD	COCO	✓Imagenet	36.3	20.4
RepMet	COCO	✓Imagenet	38.2	21.2
Ours	COCO	✓Imagenet	41.1	22.2
Ours	FSOD [10]		44.6	28.8
Ours	FSOD [10]	✓Imagenet	46.9	32.1

Table 5: experimental results on ImageNet Detection dataset

Our approach produces 2.9% performance gain compared to the state-ofthe-art (SOTA) on the AP_{50} evaluation

To demonstrate the generalization ability of our approach, we apply our model trained on FSOD dataset on the test set and we obtain 44.6% on the AP_{50} evaluation. Here our model is trained on FSOD dataset can be directly applied on the test set without fine-tuning to achieve SOTA performance. Although the difference in performance gain between our model trained on FSOD dataset and the fine-tuned SOTA model on the MS COCO dataset is not that significant, our model surpasses the fine-tuned model by 6.6% on the AP_{75} evaluation, which shows that our proposed FSOD dataset significantly benefits few-shot object detection. With further fine-tuning our FSOD trained model on the test set, our model achieves the best performance

5.4. MS COCO dataset

We will compare our model with other few-shot learning method such as Feature Reweighting [3] and Meta R-CNN [31] on MS COCO minival set. We use the same evaluation protocol used in the paper

Method	Training Dataset	Finetune	AP_{50}	AP_{75}
Feature Reweighting	COCO	✓COCO	12.6	5.2
Meta R-CNN	COCO	✓COCO	18.2	6.9
Ours	COCO	✓COCO	20.3	12.4
Ours	FSOD [10]		32.7	16.7
Ours	FSOD [10]	✓COCO	34.4	20.1

Table 6: experimental results on MS COCO Detection dataset

Our fine-tuned model with the same MS COCO training dataset outperforms Meta R-CNN by 2.1%/5.5% and Feature Reweighing by 7.7%/7.2%on AP_{50}/AP_{75} metrics. This demonstrates the strong learning and generalization ability of our model. In the few-shot scenario, learning general matching relationship is more than just simply to learn category-specific embeddings. Our model trained on FSOD achieves more significant improvement of 14.1%/7.7% on AP_{50}/AP_{75} metrics

5.5. Novel Category Detection

We will now work with real world task. Given a massive number of images in a photo album or TV drama series without any labels, the task is to annotate a novel target object in the given massive collection without knowing which images contain the target object, which can be in different sizes and locations if present. Following this setting, we perform the evaluation as follows: We mix all test images of FSOD dataset, and for each object category, we pick 5 images that contain the target object to perform this novel category object detection in the entire test set. Our method can be applied to detect object in novel categories without any further retraining or fine-tuning. We compare our model with LSTD [14] and Faster R-CNN [25] and the result is below

Method	(Pre)-Training Dataset	Finetune on FSOD	AP_{50}	AP_{75}
Faster RCNN	no	✓	10.3	5.3
Faster RCNN	FSOD [10]	✓	21.5	11.3
LSTD	FSOD [10]	✓	25.1	11.2
Ours	FSOD [10]	✓	28.6	19.8

Table 7: experimental results on FSOD dataset

Our method outperforms LSTD by 4.5%/8.6% and its backbone Faster R-CNN by 7.1%/8.5% on all 200 testing categories on AP_{50}/AP_{75} metrics. Without pre-training on our dataset, the performance of Faster R-CNN significantly drops.

6. Conclusion/Future Work

In this paper, we implement a novel few-shot object detection network with Attention-RPN, Multi-Relation Detectors and Contrastive Training strategy. We trained the model on the new FSOD [10] which contains 1000 categories of various objects with high-quality annotations. Through various experiement and study, we have shown that the model trained on FSOD can detect objects of novel categories requiring no pre-training or further network adaptation. The model has been validated by quantitative and qualitative results on different datasets.

For future work, researcher can make further study into different methods and models with the usage of large-scale FSOD dataset. It would be interesting to see how this work can be extended to make few shot learning better and more powerful

7. Contributions & Acknowledgements

This paper make use of this public code repo `https://github.com/fanq15/FSOD-code` which makes available the Pytorch Implementation of FSOD [10]

References

[1] I. S. Alex Krizhevsky and G. E. Hinton. Imagenet classification with deep convolutional neural networks: A low-shot transfer detector for object detection. 2012.

[2] N. A. J. U. I. K. J. P.-T. S. K. S. P. M. M. T. D. Alina Kuznetsova, Hassan Rom and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. 2018.

[3] X. W. F. Y. J. F. Bingyi Kang, Zhuang Liu and T. Darrell. Few-shot object detection via feature reweighting. 2019.

[4] P. R. L. Boris Oreshkin and A. L.- coste. Tadam: Task dependent adaptive metric for improved few-shot learning. 2018.

[5] J. G. Brenden Lake, Ruslan Salakhutdinov and J. Tenenbaum. One shot learning of simple visual concepts. 2011.

[6] R. R. S. Brenden M Lake and J. Tenenbaum. One-shot learning by inverting a compositional causal process. 2013.

[7] P. A. Chelsea Finn and S. Levine. Modelagnostic meta-learning for fast adaptation of deep networks. 2017.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. 2005.

[9] R. Z. Eleni Triantafillou and R. Urtasun. Few-shot learning through an information retrieval lens. 2017.

[10] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020.

[11] S. Gidaris and N. Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. 2019.

[12] R. Girshick. Fast r-cnn. 2015.

[13] R. Z. Gregory Koch and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. 2015.

[14] G. W. Hao Chen, Yali Wang and Y. Qiao. Lstd: A low-shot transfer detector for object detection. 2018.

[15] G. W. Y. Q. Hao Chen, Yali Wang. Lstd: A low-shot transfer detector for object detection. 2018.

[16] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. 2017.

[17] S. D. M. Z. Hongyang Li, David Eigen and X. Wang. Finding task-relevant features for fewshot learning by category traversal. 2019.

[18] R. G. Joseph Redmon, Santosh Divvala and A. Farhadi. You only look once: Unified, real-time object detection. 2016.

[19] P. D. Kaiming He, Georgia Gkioxari and R. G.- shick. Mask r-cnn. 2017.

[20] S. H. E. S. A. A. R. F. Leonid Karlinsky, Joseph Shtok. Repmet: Representative-based metric learning for classification and few-shot object detection. 2018.

[21] R. F. Li Fei-Fei and P. Perona. One-shot learning of object categories. 2006.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[23] T. Munkhdalai and H. Yu. Meta networks. 2017.

[24] D. M. Pedro F. Felzenszwalb, Ross B. Girshick and D. Ramanan. Object detection with discriminatively trained part based models. 2010.

[25] R. G. Shaoqing Ren, Kaiming He and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015.

[26] R. G. J. S. Shaoqing Ren, Kaiming He. Faster R-CNN: Towards real-time object detection with region proposal networks. 2015.

[27] T. K. Sungwoong Kim Chang D. Yoo Jongmin Kim. Edge-labeling graph neural network for few-shot learning. 2019.

[28] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. 2001.

[29] D. E. C. S. S. R. C.-Y. F. Wei Liu, Dragomir Anguelov and A. C. Berg. Ssd: Single shot multibox detector. 2016.

[30] J. X. J. H. G. Y. Wenbin Li, Lei Wang and J. Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. 2019.

[31] A. X. X. W. X. L. Xiaopeng Yan, Ziliang Chen and L. Lin. Meta r-cnn : Towards general solver for instance-level low-shot learning. 2019.

[32] F. M. Y. Y. Xuanyi Dong, Liang Zheng and D. Meng. Few-example object detection with model communication. 2018.

[33] S. P. Yann Lifchitz, Yannis Avrithis and A. Bursuc. Dense classification and implanting for few-shot learning. 2019.