Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection – Supplementary Material –

Chenchen Zhu Fangyi Chen Uzair Ahmed Zhiqiang Shen Marios Savvides
Carnegie Mellon University
{chenchez, fangyic, uzaira, zhiqians, marioss}@andrew.cmu.edu

S1. Removing Novel Classes from ImageNet

We propose a realistic setting for evaluating the few-shot object detection methods, where novel classes are completely removed from the classification dataset used for training a model to initialize the backbone network in the detector. This can guarantee that the object concept of novel classes will not be encoded in the pretrained model before training the few-shot detector. Because the novel class data is so rare in the real world that pretraining a classifier on it is not realistic.

ImageNet [2] is widely used for pretraining the classification model. It has 1000 classes organized according to the WordNet hierarchy. Each class has over 1000 images for training. We systematically and hierarchically remove novel classes by finding each synset and its corresponding full hyponym (synset of the whole sub-tree starting from that synset) using the ImageNet API [1]. So each novel class may contain multiple ImageNet classes.

For the novel classes in the VOC dataset [3], their corresponding WordNet IDs to be removed are as follows.

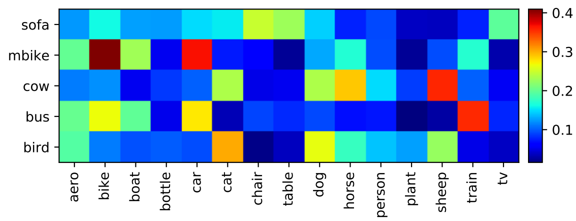
- aeroplane: n02690373, n02692877, n04552348
- bird: n01514668, n01514859, n01518878, n01530575, n01531178, n01532829, n01534433, n01537544, n01558993, n01560419, n01580077, n01582220, n01592084, n01601694, n01608432, n01614925, n01616318, n01622779, n01795545, n01796340, n01797886, n01798484, n01806143, n01806567, n01807496, n01817953, n01818515, n01819313, n01820546, n01824575, n01828970, n01829413, n01833805, n01843065, n01843383, n01847000, n01855032, n01855672, n01860187, n02002556, n02002724, n02006656, n02007558, n02009229, n02009912, n02011460, n02012849, n02013706, n02017213, n02018207, n02018795, n02025239, n02027492, n02028035, n02033041, n02037110, n02051845, n02056570, n02058221
- boat: n02687172, n02951358, n03095699, n03344393, n03447447, n03662601, n03673027, n03873416, n03947888, n04147183, n04273569, n04347754, n04606251, n04612504
- bottle: n02823428, n03062245, n03937543, n03983396, n04522168, n04557648, n04560804, n04579145, n04591713
- bus: n03769881, n04065272, n04146614, n04487081
- cat: n02123045, n02123159, n02123394, n02123597, n02124075, n02125311, n02127052
- cow: n02403003, n02408429, n02410509
- horse: n02389026, n02391049
- motorbike: n03785016, n03791053
- sheep: n02412080, n02415577, n02417914, n02422106, n02422699, n02423022
- sofa: n04344873

For the novel classes in the COCO dataset [7], they are very common in the real world. Removing them from the ImageNet does not make sense as much as removing data-scarce classes. So we suggest for large-scale datasets like COCO, we should follow the long-tail distribution of their class frequency and select the data-scarce classes on the distribution tail to be the novel classes.

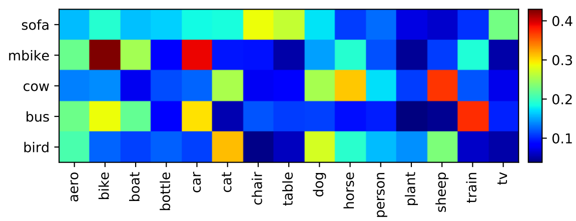
S2. Visualization of Relation Reasoning

Figure S1 visualizes the correlation maps between the semantic embeddings of novel and base classes before and after the relation reasoning, as well as the difference between the two maps. Nearly all the correlations are increased slightly, indicating better knowledge propagation between the two groups of classes. Additionally, it is interesting to see that some novel classes get more correlated than others, e.g. "sofa" with "bottle" and "sofa" with "table", probably because "sofa" can often be seen together

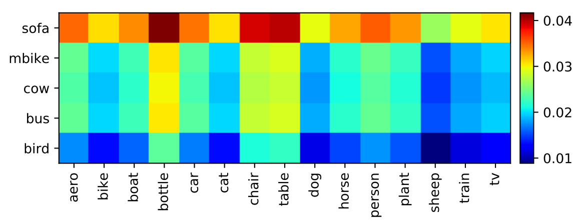
[1] http://image-net.org/download-API



(a) Before relation reasoning



(b) After relation reasoning



(c) Difference between above correlation maps

Figure S1. Correlation of the semantic embeddings before and after the relation reasoning between the base classes and the novel classes on the VOC dataset. The novel classes are from Novel Set 1. The last figure shows how does the correlation change subtly. Some novel classes are getting more correlated with base classes after relation reasoning, e.g. “sofa” with “bottle” and “table”. *Best viewed in color.*

with “bottle” and “table” in the living room but the original semantic embeddings cannot capture these relationships.

S3. Using Other Word Embeddings

In the semantic space projection, we represent the semantic space using word embeddings from the Word2Vec [8]. We could simply set the $\mathbf{W}_e$ to be random vectors. Additionally, there are other language models for obtaining vector representations for words, such as the GloVe [10]. The GloVe is trained with aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. We also explored using word embedding with different dimensions from the GloVe in the semantic space projection step and compared with the results by the Word2Vec. Performance on the VOC Novel Set 1 is reported in Table S1. The Word2Vec can provide better representations than the GloVe of both 300 dimensions and 200 dimensions. The performance of random embeddings is significantly worse than the meaningful Word2Vec and GloVe, which again verifies the importance of semantic information for shot-stable FSOD.

	Novel Set 1				
Word embeddings	shot=1	2	3	5	10
Random-300d	33.2	37.5	43.0	47.0	51.5
Word2Vec-300d [8]	42.8	47.1	49.0	50.8	52.8
GloVe-300d [10]	38.8	44.8	46.6	49.0	54.3
GloVe-200d [10]	39.7	44.6	45.8	49.4	53.0

Table S1. FSOD performance (mAP50) on the VOC Novel Set 1 under different word embeddings in the semantic space projection. All models are using the ResNet-50 network. 300d and 200d mean the numbers of embedding dimension are 300 and 200 respectively. The Word2Vec provides better representations than the GloVe.

	Novel Set 1				
Dimension	shot=1	2	3	5	10
128	40.9	44.6	44.3	48.1	54.1
64	42.0	**47.4**	**48.9**	51.7	54.1
32	**42.4**	**46.8**	**48.1**	**51.9**	**54.7**
16	**44.1**	46.0	47.8	51.7	**54.7**

Table S2. FSOD performance (mAP50) on the VOC Novel Set 1 under different reduced feature dimension in the relation reasoning module. Bold font indicates best or second best results. All models are using the ResNet-50 network.

	Novel Set 1				
Tunable Parameters	shot=1	2	3	5	10
Last layer (TFA [13])	39.8	36.1	44.7	55.7	56.0
+FCs	36.9	34.9	45.3	53.0	55.9
+FCs +RPN	37.2	39.8	44.3	52.7	56.2
+FCs +RPN +Backbone	16.2	19.5	24.8	39.2	44.6

Table S3. FSOD results (mAP50) on the VOC Novel Set 1 with more and more tunable parameters in the finetuning stage. The baseline is TFA [13] which only finetunes the last classification layer in the Faster R-CNN. We gradually unfreeze more previous layers including two fully-connected layers (FCs) after the RoI-pooling, layers in region proposal network (RPN), and layers in the Backbone. This proves that finetuning more parameters does not guarantee better performance in few-shot detection.

S4. Reduced Dimension in Relation Reasoning

In the relation reasoning module, the dimension of word embeddings is reduced by linear layers before computing the attention map, which saves computational time. We empirically test different dimensions and select the one with the best performance, i.e. when the dimension is 32. But other choices are just slightly worse. Table S2 reports the results on VOC dataset under different dimensions. All the experiments are following the same setting as in the main paper. The only exception is that we use ResNet-50 [4] to

Shot	Method	Novel Set 1						Novel Set 2						Novel Set 3					
		bird	bus	cow	mbike	sofa	mean	aero	bottle	cow	horse	sofa	mean	boat	cat	mbike	sheep	sofa	mean
1	FSRW	13.5	10.6	31.5	13.8	4.3	14.8	11.8	**9.1**	15.6	23.7	18.2	15.7	10.8	44.0	17.8	18.1	5.3	19.2
	Meta R-CNN	6.1	32.8	15.0	35.4	0.2	19.9	23.9	0.8	23.6	3.1	0.7	10.4	0.6	31.1	28.9	11.0	0.1	14.3
	MPSR	33.5	41.2	57.6	54.5	21.6	41.7	21.2	**9.1**	36.0	30.9	25.1	24.4	14.9	47.8	57.7	34.7	**22.8**	35.6
	SRR-FSD (Ours)	**38.1**	**53.8**	**58.7**	**64.1**	**24.4**	**47.8**	**27.9**	4.6	**50.5**	**53.9**	**25.5**	**32.5**	**16.2**	**57.2**	**62.9**	**48.3**	16.0	**40.1**
2	FSRW	21.2	12.0	16.8	17.9	9.6	15.5	28.6	0.9	27.6	0.0	19.5	15.3	5.3	46.4	18.4	26.1	12.4	21.7
	Meta R-CNN	17.2	34.4	43.8	31.8	0.4	25.5	12.4	0.1	44.4	50.1	0.1	19.4	10.6	24.0	36.2	19.2	0.8	18.2
	MPSR	**38.2**	28.6	56.5	57.3	32.0	42.5	**36.5**	**9.1**	45.1	21.6	34.2	29.3	**17.9**	49.6	59.2	**49.2**	32.9	**41.8**
	SRR-FSD (Ours)	35.8	**57.7**	**59.3**	**61.8**	**38.0**	**50.5**	34.4	5.7	**57.1**	**44.0**	**35.5**	**35.3**	15.5	**51.4**	**62.6**	44.4	**33.7**	41.5
3	FSRW	26.1	19.1	40.7	20.4	27.1	26.7	29.4	4.6	34.9	6.8	37.9	22.7	11.2	39.8	20.9	23.7	33.0	25.7
	Meta R-CNN	30.1	44.6	50.8	38.8	10.7	35.0	25.2	0.1	50.7	**53.2**	18.8	29.6	16.3	39.7	32.6	38.8	10.3	27.5
	MPSR	35.1	**60.6**	56.6	61.5	**43.4**	**51.4**	**49.2**	9.1	47.1	46.3	**44.3**	**39.2**	14.4	**60.6**	57.1	37.2	**42.3**	42.3
	SRR-FSD (Ours)	**35.2**	55.6	**61.3**	**62.9**	41.5	51.3	42.3	**11.5**	**57.0**	43.6	41.2	39.1	**23.1**	50.6	**60.0**	**49.3**	38.6	**44.3**
5	FSRW	31.5	21.1	39.8	40.0	37.0	33.9	33.1	9.4	38.4	25.4	44.0	30.1	14.2	57.3	50.8	38.9	41.6	40.6
	Meta R-CNN	35.8	47.9	54.9	55.8	34.0	45.7	28.5	0.3	50.4	56.7	38.0	34.8	16.6	45.8	53.9	41.5	48.1	41.2
	MPSR	39.7	**65.5**	55.1	**68.5**	**47.4**	55.2	**47.8**	10.4	45.2	47.5	**48.8**	39.9	**20.9**	**56.6**	**68.1**	48.4	**45.8**	**48.0**
	SRR-FSD (Ours)	**46.1**	58.6	**64.6**	63.5	43.2	**55.2**	44.2	**12.3**	**56.5**	51.3	39.8	**40.8**	20.4	55.5	65.4	**51.9**	41.3	46.9
10	FSRW	30.0	62.7	43.2	60.6	39.6	47.2	43.2	13.9	41.5	58.1	39.2	39.2	20.1	51.8	55.6	42.4	36.6	41.3
	Meta R-CNN	**52.5**	55.9	52.7	54.6	41.6	51.5	**52.8**	3.0	52.1	**70.0**	49.2	45.4	13.9	**72.6**	58.3	47.8	47.6	48.1
	MPSR	48.3	**73.7**	**68.2**	**70.8**	**48.2**	**61.8**	51.8	16.7	53.1	66.4	**51.2**	**47.8**	**24.4**	55.8	**67.5**	**50.4**	**50.5**	**49.7**
	SRR-FSD (Ours)	45.0	67.4	63.1	65.2	43.3	56.8	46.2	**18.4**	**54.0**	59.1	41.4	43.8	17.1	55.1	67.4	47.5	44.7	46.4

Table S4. AP50 performance of each novel class on the few-shot VOC dataset. Bold font indicates the best result in the group. Our SRR-FSD trained with visual information and semantic relation demonstrates shot-stable performance.

reduce the computational cost of tuning hyperparameters.

S5. Finetuning More Parameters

Similar to TFA [13], we have a finetuning stage to make the detector generalized to novel classes. For the classification subnet, we finetune the parameters in the relation reasoning module and the projection matrix while all the parameters in previous layers are frozen. Some may argue that the improvement of our SRR-FSD over the baseline is due to more parameters finetuned in the relation reasoning module compared to the Faster R-CNN [11] baseline. But we show that finetuning more parameters does not necessarily lead to better results in Table S3. We take the TFA model which is essentially a Faster R-CNN finetuned with only the last layer trainable and gradually unfreeze the previous layers. It turns out more parameters involved in finetuning do not change the results substantially and that too many parameters will lead to severe overfitting.

S6. Complete Results on VOC

In Table S4, we present the complete results on the VOC [3] dataset as in FSRW [5] and Meta R-CNN [16]. We also include the very recent MPSR [15] for comparison. MPSR develops an auxiliary branch to generate multi-scale positive samples as object pyramids and to refine the prediction at various scales. Note that MPSR improves its baseline by a considerable margin but its research direction is *orthogonal and complimentary* to ours because it is still exclusively dependent on visual information. Therefore, our approach combining visual information and semantic relation reasoning can achieve superior performance at extremely low shot (e.g. 1, 2) conditions.

S7. Interpretation of the Dynamic Relation Graph

In the relation reasoning module, we propose to learn a *dynamic* relation graph driven by the data, which is conceptually different from the predefined fixed knowledge graphs used in [14, 1, 9]. We implement the dynamic graph with the self-attention architecture [12]. Although it is in the form of a feedforward network, it can also be interpreted as a computation related to the knowledge graph. If we denote the transformations in the linear layers f, g, h, l as $\mathbf{T}_f$, $\mathbf{T}_g$, $\mathbf{T}_h$, $\mathbf{T}_l$ respectively, we can formulate the relation reasoning in Eq. (S1)

$$\mathbf{W}'_e = \delta(\mathbf{W}_e\mathbf{T}_f\mathbf{T}_g^T\mathbf{W}_e^T)\mathbf{W}_e\mathbf{T}_h\mathbf{T}_l + \mathbf{W}_e \qquad \text{(S1)}$$

where $\mathbf{W}'_e$ is the matrix of augmented word embeddings after the relation reasoning which will be used as the weights to compute classification scores and δ is the softmax function operated on the last dimension of the input matrix. The item $\delta(\mathbf{W}_e\mathbf{T}_f\mathbf{T}_g^T\mathbf{W}_e^T)$ can be interpreted as a $N \times N$ dynamic knowledge graph in which the learnable parameters are $\mathbf{T}_f$ and $\mathbf{T}_g$. And it is involved in the computation of the classification scores via the graph convolution operation [6], which connects the N word embeddings in $\mathbf{W}_e$ to allow knowledge propagation among them. The item $\mathbf{T}_h\mathbf{T}_l$ can be viewed as a learnable transformation applied to each embedding independently.

References

[1] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 3

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8420–8429, 2019. 3

[6] T.N. Kipf and M. Welling. Semi-supervised classification with graph convolutional network. In *International Conference on Learning Representations (ICLR)*, 2017. 4

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2

[9] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 441–449, 2019. 3

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[13] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 2, 3

[14] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018. 3

[15] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European conference on computer vision*. Springer, 2020. 3

[16] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9577–9586, 2019. 3