US 20240045925A1

(19) United States

(12) Patent Application Publication

Savvides et al.

(10) Pub. No.: US 2024/0045925 A1

(43) Pub. Date: Feb. 8, 2024

(54) **SYSTEM AND METHOD FOR IMPROVED FEW-SHOT OBJECT DETECTION USING A DYNAMIC SEMANTIC NETWORK**

(71) Applicant: **CARNEGIE MELLON UNIVERSITY**, Pittsburgh, PA (US)

(72) Inventors: **Marios Savvides**, Pittsburgh, PA (US); **Chenchen Zhu**, Pittsburgh, PA (US); **Fangyi Chen**, Pittsburgh, PA (US); **Uzair Ahmed**, Pittsburgh, PA (US); **Ran Tao**, Pittsburgh, PA (US)

(21) Appl. No.: **18/266,744**

(22) PCT Filed: **Feb. 2, 2022**

(86) PCT No.: **PCT/US2022/014833**

§ 371 (c)(1),
(2) Date: **Jun. 12, 2023**

Related U.S. Application Data

(63) Continuation of application No. 17/408,674, filed on Aug. 23, 2021.

(60) Provisional application No. 63/068,871, filed on Aug. 21, 2020, provisional application No. 63/147,782, filed on Feb. 10, 2021.

Publication Classification

(51) **Int. Cl.**

G06F 18/2136	(2006.01)
G06N 3/04	(2006.01)
G06N 5/02	(2006.01)

(52) **U.S. Cl.**

CPC ***G06F 18/2136*** (2023.01); ***G06N 3/04*** (2013.01); ***G06N 5/02*** (2013.01)

(57) **ABSTRACT**

Disclosed herein is an improved few-shot detector which utilizes a dynamic semantic network which takes as input a language feature and generates trainable parameters for a visual network. The visual network takes a visual feature as input and generates a classification and localization of an object.

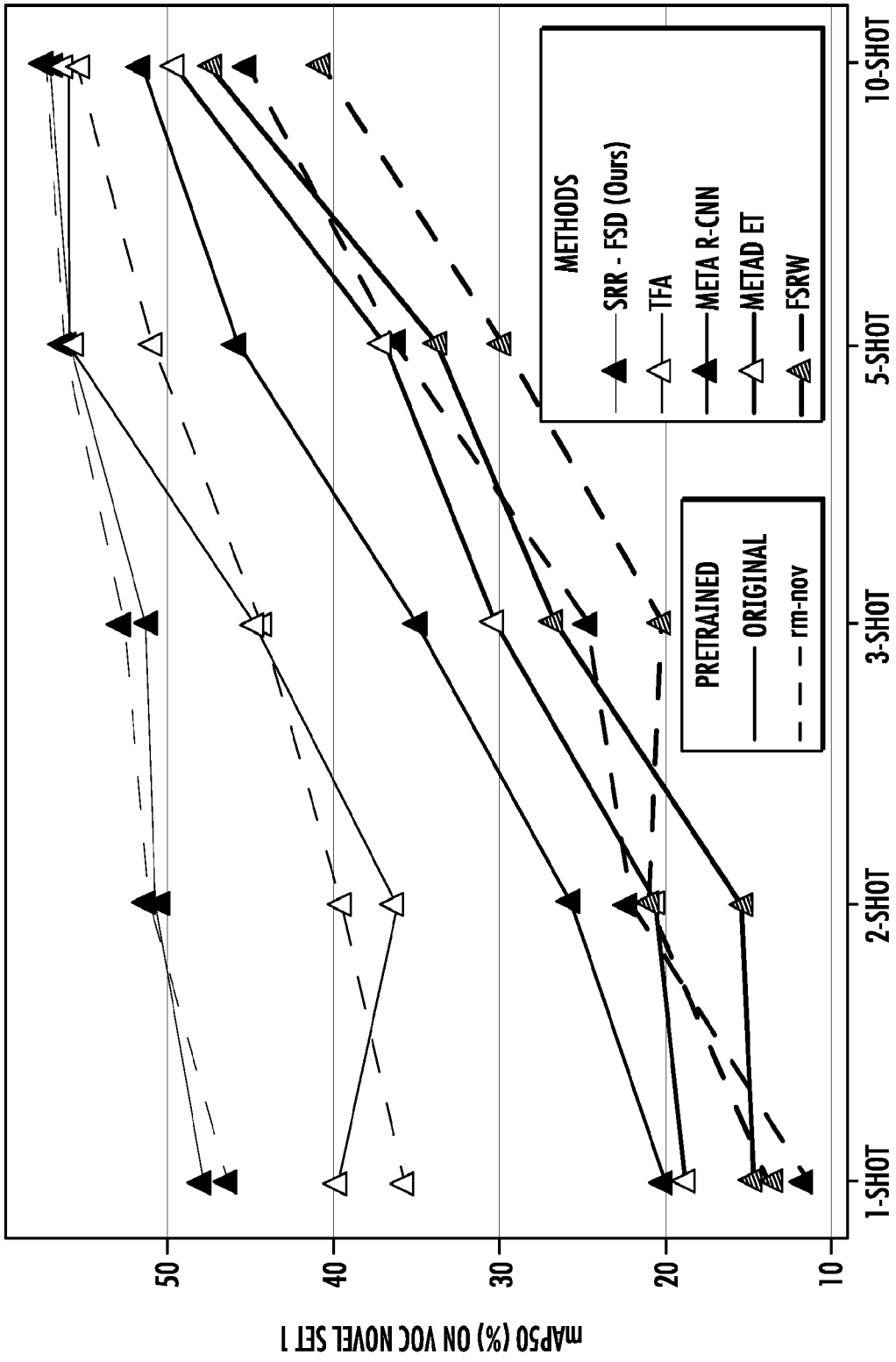
mAP50 (%) ON VOC NOVEL SET 1
50
40
30
20
10
1-SHOT
2-SHOT
3-SHOT
5-SHOT
10-SHOT
METHODS
SRR - FSD (Ours)
TFA
META R-CNN
METAD ET
FSRW
PRETRAINED
ORIGINAL
rm-nov

FIG. 1

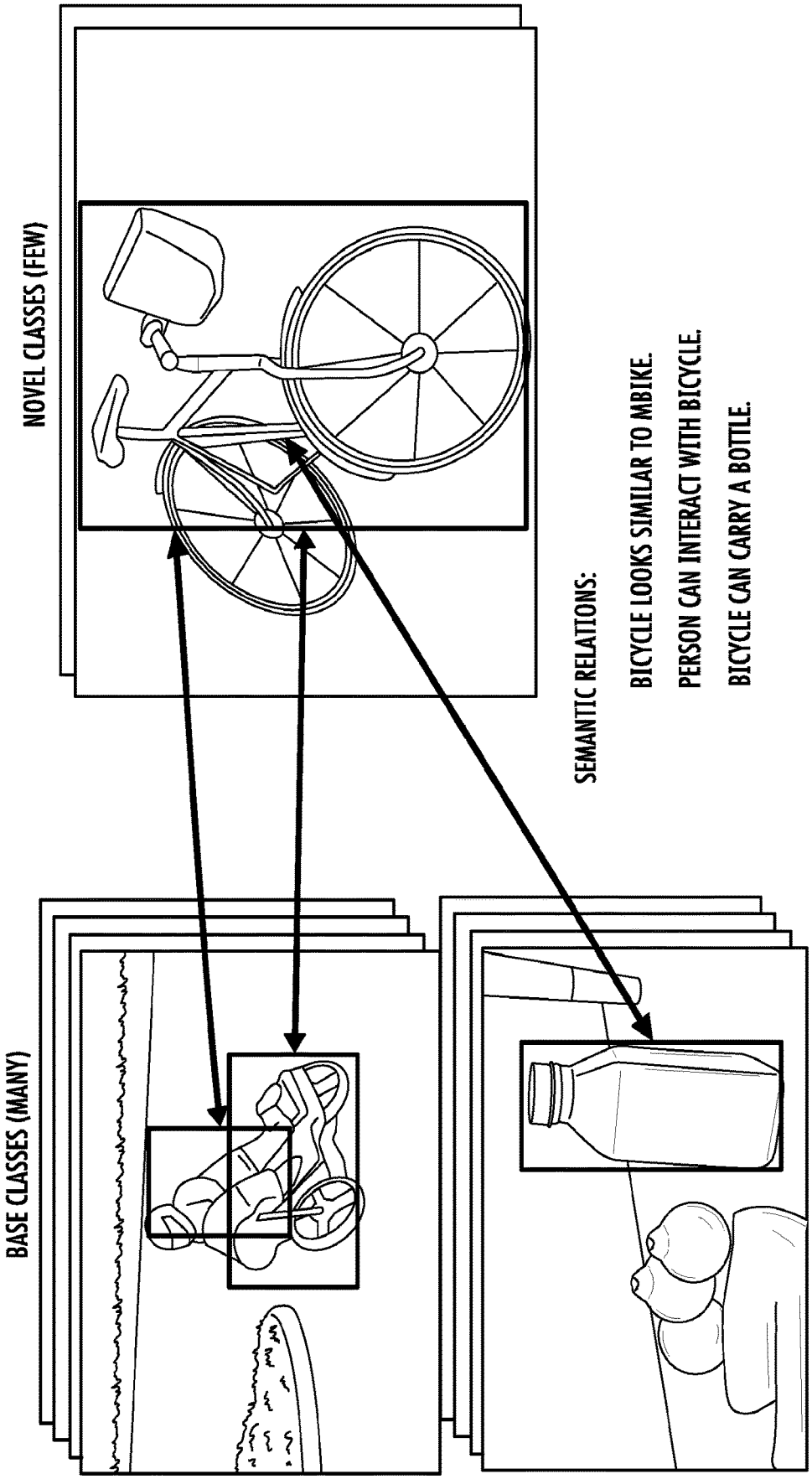
NOVEL CLASSES (FEW)
BASE CLASSES (MANY)
SEMANTIC RELATIONS:
BICYCLE LOOKS SIMILAR TO MBIKE.
PERSON CAN INTERACT WITH BICYCLE.
BICYCLE CAN CARRY A BOTTLE.

FIG. 2

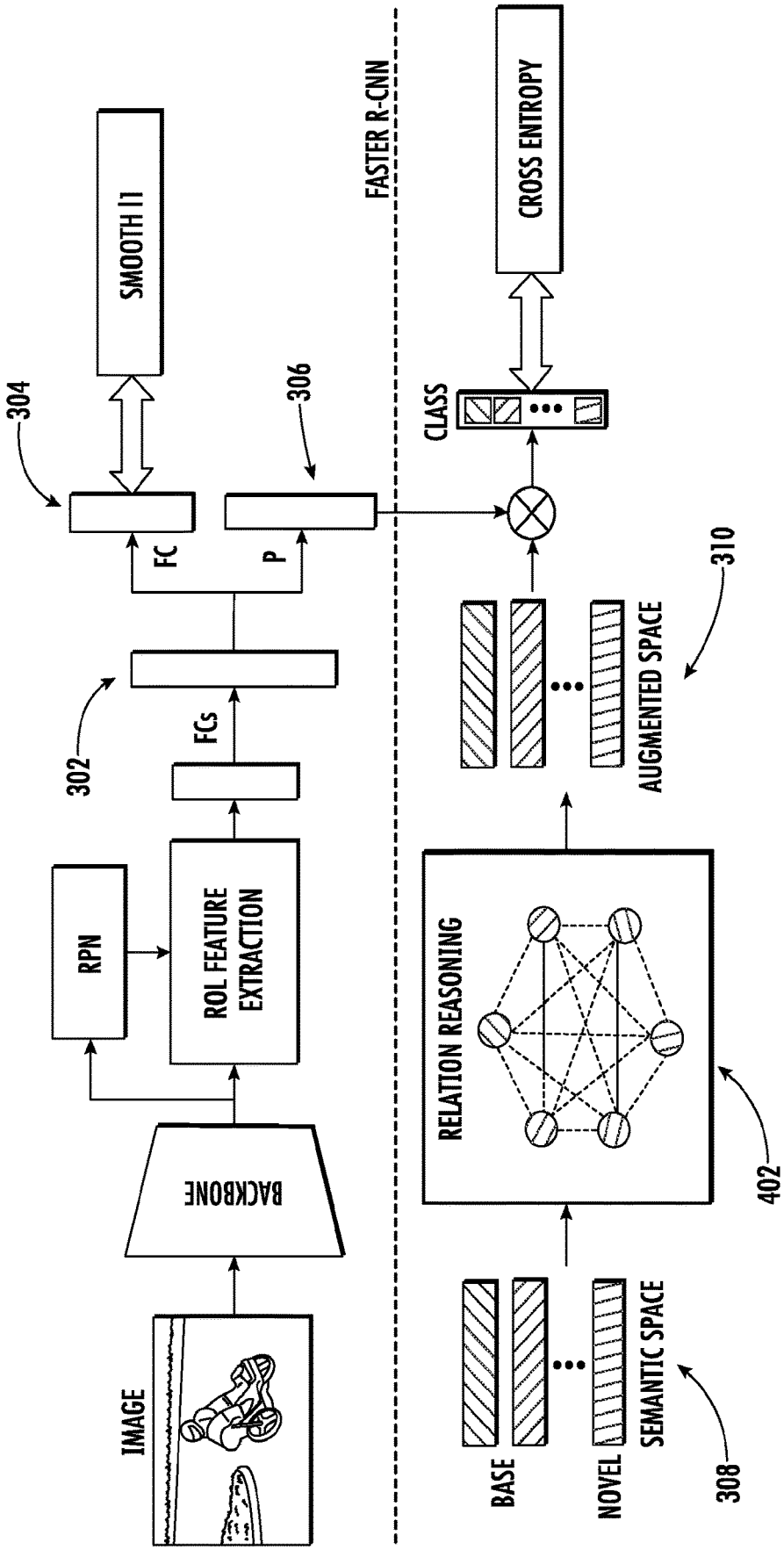
IMAGE
BACKBONE
RPN
ROI FEATURE EXTRACTION
FCs
302
FC
P
304
306
SMOOTH l1
FASTER R-CNN
BASE
NOVEL
SEMANTIC SPACE
308
RELATION REASONING
402
AUGMENTED SPACE
310
CLASS
CROSS ENTROPY

FIG. 3

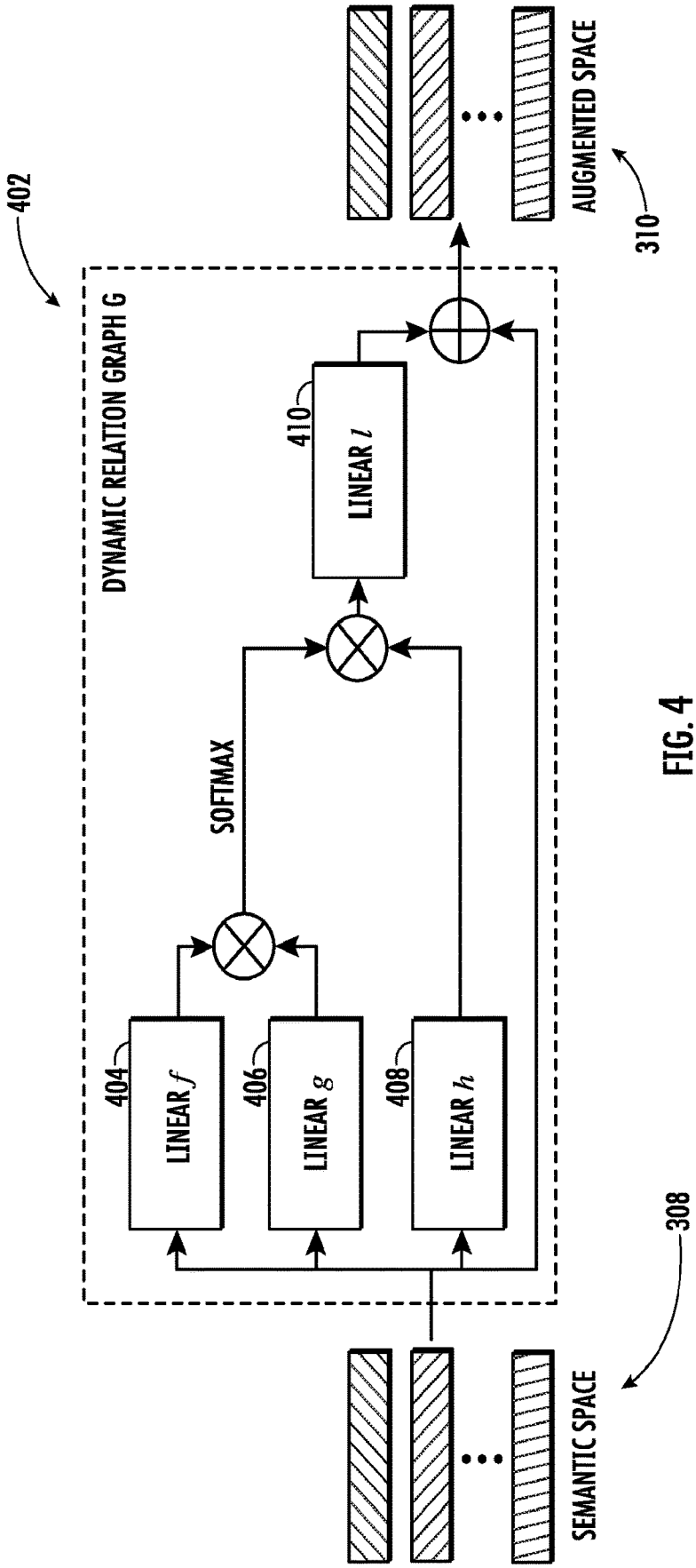
402
DYNAMIC RELATION GRAPH G
404
LINEAR f
406
LINEAR g
408
LINEAR h
SOFTMAX
410
LINEAR l
308
SEMANTIC SPACE
310
AUGMENTED SPACE

FIG. 4

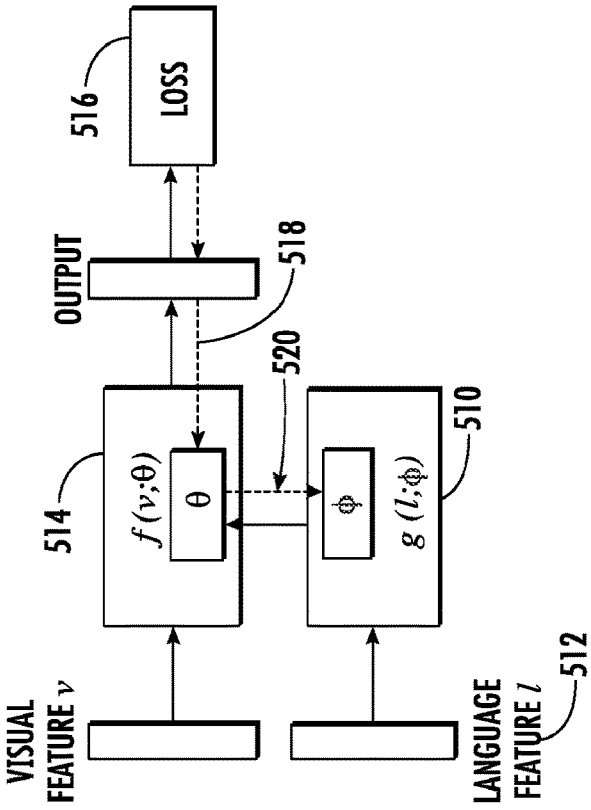
VISUAL FEATURE v
LANGUAGE FEATURE l
512
$f(v;\theta)$
θ
514
ϕ
$g(l;\phi)$
510
520
518
OUTPUT
516
LOSS

FIG. 5B

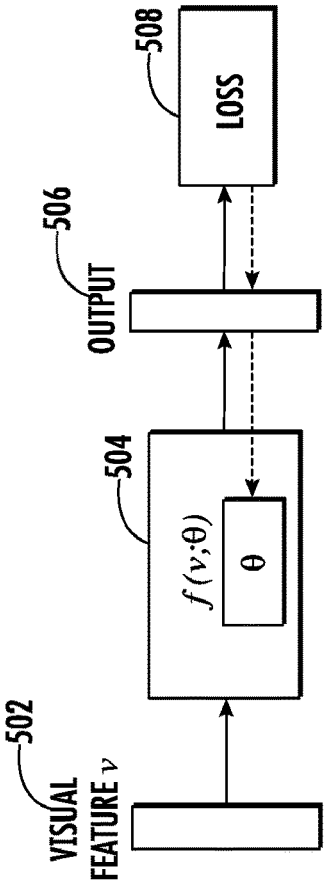
VISUAL FEATURE v
502
$f(v;\theta)$
θ
504
OUTPUT
506
LOSS
508

FIG. 5A

SYSTEM AND METHOD FOR IMPROVED FEW-SHOT OBJECT DETECTION USING A DYNAMIC SEMANTIC NETWORK

RELATED APPLICATIONS

[0001] This application is a continuation-in-part of U.S. patent application Ser. No. 17/408,674, filed Aug. 23, 2021. In addition, this application claims the benefit of U.S. Provisional Patent Application No. 63/147,782, filed Feb. 10, 2021. The contents of these applications are incorporated herein in their entireties.

BACKGROUND

[0002] Deep learning algorithms typically require a large amount of annotated data to achieve superior performance. To acquire enough annotated data, one common way is by collecting abundant samples from the real world and paying annotators to generate ground-truth labels. However, even if all the data samples are well annotated, a problem still exists regarding few-shot learning. Because long-tail distribution is an inherent characteristic of the real world, there always exist some rare cases that have just a few samples available, for example, rare animals, un-common road conditions, etc. In other words, because of the few number of samples available for some classes, the situation is not able to be remedied by simply spending more money on annotation of existing samples.

[0003] In few-shot object detection (FSOD), there are base classes in which sufficient objects have been annotated with bounding boxes, and novel classes in which very few annotated objects are available. The novel class set does not share common classes with the base class set. Few-shot detectors are expected to learn from limited data in novel classes with the aid of abundant data in base classes and to be able to detect novel objects in a held-out testing set. To achieve this, most prior art few-shot detection methods adopt ideas from meta-learning and metric learning for few-shot recognition and apply them to conventional detection frameworks (e.g., Faster R-CNN, YOLO).

[0004] Although prior art FSOD methods have improved the base-line considerably, data scarcity is still a bottleneck that hurts the detector's generalization from a few samples. In other words, the performance is very sensitive to the number of both explicit and implicit shots and drops drastically as data becomes limited. The explicit shots refer to the available labeled objects from the novel classes. For example, the 1-shot performance of some FSOD methods is less than half of the 5-shot or 10-shot performance, as shown in FIG. **1**.

[0005] FIG. **1** is a graph showing FSOD performance (mAP50) on VOC Novel Set 1 at for different shot numbers. Solid lines (original) indicate the pre-trained model used for initializing the detector backbone trained on the original ImageNet dataset. Dashed lines (rm-nov) indicate classes in Novel Set 1 which have been removed from the ImageNet dataset for the pretrained backbone model. The invention is more stable to the variation of explicit shots (x-axis) and implicit shots (original vs. rm-nov).

[0006] In terms of implicit shots, initializing the backbone network with a model that has been pre-trained on a large-scale image classification dataset is a common practice for training an object detector. However, the classification dataset typically contains many implicit shots of object classes overlapped with the novel classes.

[0007] As such, the detector can have early access to novel classes and encode their knowledge in the parameters of the backbone. Removing those implicit shots from the pre-trained dataset also has a negative impact on the performance, as shown in FIG. **1**. The variation of explicit and implicit shots could potentially lead to system failure when dealing with extreme cases in the real world.

[0008] The reason for shot sensitivity could be due to exclusive dependence on the visual information. Novel objects are learned through images only and the learning is independent between classes. As a result, visual information becomes limited as image data becomes scarce.

SUMMARY OF THE INVENTION

[0009] The key insight in this invention is that the semantic relation between base and novel classes remains constant regardless of the data availability of novel classes. For example, in FIG. **2**, "bicycle" is the novel class. If the prior knowledge that the novel class "bicycle" looks similar to "motorbike", can have an interaction with a "person", and can carry a "bottle" is known, it will be easier to learn the concept "bicycle" rather than relying solely on a few images of a bicycle. Such explicit relation reasoning is even more crucial when visual information is hard to access.

[0010] The disclosed invention introduces semantic relations to few-shot detection. In natural language processing, semantic concepts are represented by word embeddings from language models. Explicit relationships are represented by knowledge graphs.

[0011] The disclosed invention comprises a Semantic Relation Reasoning Few-Shot Detector (SRR-FSD), which incorporates the semantic relation for FSOD. The SRR-FRD learns novel objects from both visual information and the semantic relation in an end-to-end style. Specifically, a semantic space is constructed using word embeddings. Guided by the word embeddings of the classes, the detector is trained to project the objects from the visual space to the semantic space and to align their image representations with the corresponding class embeddings.

[0012] Directly applying these concepts to few-shot detectors leads to non-trivial practical problems (i.e., the domain gap between vision and language, and the heuristic definition of knowledge graph for classes in FSOD datasets). To address these problems, instead of pre-defining a dynamic relation graph based on heuristics, the invention learns a dynamic relation graph driven by the image data. Then, the learned graph is used to perform relation reasoning and to augment the raw embeddings, resulting in a reduced domain gap.

[0013] With the help of the semantic relation reasoning, SRR-FSD demonstrates the shot-stable property in two aspects, as shown in FIG. **1**. In the common few-shot settings (solid lines), SRR-FSD achieves competitive performance at higher shots and significantly better performance at lower shots compared to prior art few-shot detectors. In a more realistic setting (dashed lines) where implicit shots of novel concepts are removed from the classification dataset for the pretrained model, SRR-FSD steadily maintains the performance while some prior art methods have results degraded by a large margin due to the loss of implicit shots.

[0014] The novelty of the invention is the use of semantic relation reasoning for the few-shot detection task. The SRR-FSD achieves stable performance with respect to shot variation and outperforms prior art FSOD methods under several existing settings, especially when the novel class data is extremely limited. Even when implicit shots of novel classes are removed from the classification dataset for the pretrained model, SRR-FSD maintains a steadier performance compared to prior art methods.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] FIG. **1** is a graph showing a comparison of FSOD performance (mAP50) for various few-shot detectors versus the improved few-shot detector if the disclosed herein.

[0016] FIG. **2** is a graphic representation showing how semantic information can be used to learn novel classes.

[0017] FIG. **3** is a block diagram of the semantically-enhanced few-shot object detector disclosed herein.

[0018] FIG. **4** is a block diagram of the relation reasoning module used to provide the semantic component of the improved few-shot detector disclosed herein.

[0019] FIG. **5** is a block diagram showing a second embodiment of the invention in contrast with a block diagram of a conventional approach.

DETAILED DESCRIPTION

[0020] To understand SRR-FSD, it will first be useful to start with an explanation of prior art few-shot object detection. Thereafter, the building of the SRR-FSD comprises integrating semantic relations with the visual information in a Faster R-CNN and allowing it to perform relation reasoning in the semantic space. A two-phase training process is disclosed.

[0021] The conventional object detection problem has a base class set C_b in which there are many instances, and a base dataset D_b with abundant images. D_b consists of a set of annotated images $\{(x_i,y_i)\}$ where x_i is the image and y_i is the annotation of labels from C_b and bounding boxes for objects in x_i. For the few-shot object detection problem, in addition to C_b and D_b, it also has a novel class set C_n and a novel dataset D_n, with $C_b \cap C_n = 0$. In D_n, objects have labels belong to C_n and the number of objects for each class is k for k-shot detection. A few-shot detector is expected to learn from D_b and to quickly generalize to D_n with a small k such that it can detect all objects in a held-out testing set with object classes in $C_b \cup C_n$. It is assumed that all classes in $C_b \cup C_n$ have semantically meaningful names, so the corresponding semantic embeddings can be retrieved.

[0022] A typical few-shot detector has two training phases. The first phase is the base training phase where the detector is trained on D_b similarly to conventional object detectors. Then in the second phase, it is further fine-tuned on $D_b \cup D_n$. To avoid the dominance of objects from D_b, a small subset is sampled from D_b, such that the training set is balanced concerning the number of objects per class. As the total number of classes is increased by the size of C_n in the second phase, more class-specific parameters are inserted in the detector and trained to be responsible for the detection of novel objects. The class-specific parameters are usually in the box classification and localization layers at the very end of the network.

[0023] An overview of SRR-FSD is illustrated in FIG. **3**. A semantic space is built from the word embeddings of all corresponding classes in the dataset and is augmented through a relation reasoning module. Visual features are learned to be projected into the augmented space **310**. ("⊗" is the dot product; "FC" is a fully-connected layer; and "P" is a learnable projection matrix).

[0024] Semantic Space Projection—The SRR-FSD detector disclosed herein is built on top of Faster R-CNN, a prior art two-stage general object detector. In the second-stage of Faster R-CNN, a feature vector **302** is extracted for each region proposal and forwarded to a classification subnet **304** and a regression subnet **306**. In the classification subnet, the feature vector is transformed into a d-dimensional vector $v \in R^d$ through fully-connected layers. Then v is multiplied by a learnable weight matrix $W \in R^{Nxd}$ to output a probability distribution as in Eq. (1).

$$p = \text{softmax}(W_v + b) \tag{1}$$

where:

[0025] N is the number of classes; and

[0026] $b \in R^N$ is a learnable bias vector.

[0027] The probability distribution is used as the classification output of the detector. It represents the object's classification scores by a vector with a length equal to the number of classes.

[0028] To learn objects from both the visual information and the semantic relation, a semantic space **308** is first constructed and the visual feature v is projected into this semantic space. Specifically, the semantic space is represented using a set of d_e-dimensional word embeddings $W_e \in R^{Nxd_e}$ corresponding to the N object classes (including the background classes). The detector is trained to learn a linear projection $P \in R^{d_e xd}$ in the classification subnet such that v is expected to align with its class' word embedding after projection. Mathematically, the prediction of the probability distribution is derived from Eq. (1) as:

$$p = \text{softmax}(W_v P_v + b) \tag{2}$$

[0029] During training, W_e is fixed and the learnable variable is P. A benefit is that the generalization to novel objects involves no new parameters in P. W_e **308** can simply be expanded with embeddings of the novel classes. The b is retained to model the category imbalance in the detection dataset.

[0030] Reducing the Domain Gap Between Vision and Language—W_e **308** encodes the knowledge of semantic concepts from natural language. While it is applicable in zero-shot learning, it will introduce the bias of the domain gap between vision and language to the FSOD task. Unlike zero-shot learning where unseen classes have no support from images, the few-shot detector can rely on both the images and the embeddings to learn novel objects. When there are very few images to rely on, the knowledge from embeddings can guide the detector towards a decent solution. However, when more images are available, the knowledge from the embeddings may be misleading due to the domain gap, resulting in a suboptimal solution. Therefore, there is a need to augment the semantic embeddings to reduce the domain gap. Leveraging the explicit relationship between classes is effective for embedding augmentation, leading to implementation of a dynamic relation graph.

[0031] Relation Reasoning—FIG. **4** shows the network architecture of the relation reasoning module **402** for learning the relation graph G. ("⊗" is the dot product; and "e" is the element-wise addition). The semantic space projection learns to align the concepts from the visual space with the

semantic space **308**. However, each class is still treated independently and there is no knowledge propagation among classes. Therefore, a knowledge graph is introduced to model their relationships. The knowledge graph G is an N×N adjacency matrix representing the connection strength for every neighboring class pairs. G is involved in classification via a graph convolution operation. Mathematically, the updated probability prediction is given as:

$$p=\text{softmax}(GW_vP_v+b) \quad (3)$$

[0032] In zero-shot or few-shot recognition algorithms, knowledge graph G is predefined base on heuristics. It is usually constructed from a database of common sense knowledge rules by sampling a sub-graph through the rule paths such that semantically related classes have strong connections. For example, classes from the ImageNet dataset have a knowledge graph sampled from the WordNet. However, classes in FSOD datasets are not highly semantically related, nor do they form a hierarchical structure like the ImageNet classes. The only applicable heuristics are based on object co-occurrence. Although the statistics of the co-occurrence are straightforward to compute, the co-occurrence is not necessarily equivalent to the semantic relation.

[0033] Instead of predefining a knowledge graph based on heuristics, the disclosed invention learns a dynamic relation graph driven by the data to model the relation reasoning between classes. The data-driven graph is also responsible for reducing the domain gap between vision and language because it is trained with image inputs. Inspired by the concept of the transformer, the dynamic relation graph G is implemented with a self-attention architecture as shown in FIG. **4**.

[0034] The original word embeddings W_e **308** are transformed by three linear layers: f **404**, g **406** and h **408**, and a self-attention matrix is computed from the outputs of f and g. The self-attention matrix is multiplied with the output of h followed by another linear layer l **410**. A residual connection adds the output of l **410** with the original W_e **308**. Another advantage of learning the dynamic relation graph G is that it can easily adapt to new classes. Because the graph is not fixed and is generated on the fly from the word embeddings **308**, it is not necessary to redefine a new dynamic relation graph G and retrain the detector. Corresponding embeddings for new classes can simply be inserted and the detector fine-tuned.

[0035] Decoupled Fine-Tuning—In the second fine-tuning phase, only the last few layers of SRR-FSD are unfrozen. For the classification subnet, the parameters in the relation reasoning module and the projection matrix P are fine-tuned. For the localization subnet, it is not dependent on the word embeddings but shares features with the classification subnet. The learning of localization on novel objects can interfere with the classification subnet via the shared features, leading to many false positives. Decoupling the shared fully-connected layers between the two subnets can effectively make each subnet learn better features for its task. In other words, the classification subnet and the localization subnet have individual fully-connected layers and they are fine-tuned independently.

[0036] In one embodiment, SRR-FSD is implemented based on Faster R-CNN with ResNet-101 and a Feature Pyramid Network as the backbone using the MMDetection framework. All models are trained with Stochastic Gradient Descent (SGD) and a batch size of 16. For the word embeddings, the L2-normalized 300-dimensional Word2Vec vectors from the language model trained on large unannotated texts like Wikipedia are used. In the relation reasoning module, we reduce the dimension of word embed-dings to 32 which is empirically selected. In the first base training phase, we set the learning rate, the momentum, and the weight decay to 0.02, 0.9, and 0.0001, respectively. In the second fine-tuning phase, the learning rate is reduced to 0.001 unless otherwise mentioned. The input image is sampled by first randomly choosing between the base set and the novel set with a 50% probability and then randomly selecting an image from the chosen set.

[0037] The training of the few-shot detector usually involves initializing the backbone network with a model pretrained on large-scale object classification datasets such as ImageNet. The set of object classes in ImageNet (i.e., C_0) is highly overlapped with the novel class set C_n in the existing settings. This means that the pretrained model can get early access to large amounts of object samples, (i.e., implicit shots), from novel classes and encode their knowledge in the parameters before it is further trained for the detection task. Even the pretrained model is optimized for the recognition task. The extracted features still have a big impact on the detection of novel objects, as shown in FIG. **1**. However, some rare classes may have highly limited or valuable data in the real world such that pretraining a classification network on it is not realistic.

[0038] Therefore, a more realistic setting for FSOD, which extends the existing settings may be used. In addition to $C_b \cap C_n=0$, an additional requirement is that $C_0 \cap C_n=0$. To achieve this, the novel classes are systematically and hierarchically removed from C_0. For each class in C_n, its corresponding synset is found in ImageNet and its full hyponym (the synset of the whole subtree starting from that synset) is obtained using the ImageNet API. The images of this synset and its full hyponym are removed from the pretrained dataset. The classification model is trained on a dataset with no novel objects.

[0039] Semantic Space Projection Guides Shot-Stable Learning—The baseline Faster R-CNN can already achieve satisfying results at 5-shot and 10-shot. However, at 1-shot and 2-shot, performance starts to degrade due to exclusive dependence on images. The semantic space projection, on the other hand, makes the learning more stable to the variation of shot numbers. The space projection guided by the semantic embeddings is learned well enough in the base training phase so it can be quickly adapted to novel classes with a few instances. A major boost occurs at lower shot conditions compared to the baseline. However, the raw semantic embeddings limit the performance at higher shot conditions. The performance at 5-shot and 10-shot drops below the baseline. This verifies the domain gap between vision and language. At lower shots, there is not much visual information to rely on, so the language information can guide the detector to a decent solution. But, when more images are available, the visual information becomes more precise than the language information starts to be misleading. Therefore, the word embeddings are refined to reduce the domain gap.

[0040] Relation Reasoning Promotes Adaptive Knowledge Propagation—The relation reasoning module **402** explicitly learns dynamic relation graph G that builds direct connections between base classes and novel classes. The detector can learn the novel objects using the knowledge of

base objects besides the visual information. Additionally, the relation reasoning module **402** also functions as a refinement to the raw word embeddings with a data-driven relation graph. Because relation graph G is updated with image inputs, the refinement tends to adapt the word embeddings for the vision domain. Applying relation reasoning improves the detection accuracy of novel objects under different shot conditions.

[0041] Decoupled Fine-Tuning (DF) Reduces False Positives—Most of the false positives are due to misclassification into similar categories. With DF, the classification subnet can be trained independently from the localization subnet to learn better features specifically for classification.

[0042] Second Embodiment Using Dynamic Semantic Networks—In a second embodiment of the invention, a dynamic semantic network is used to simultaneously tackle few-shot classification and few-shot localization in a unified and coherent way. The dynamic semantic network leverages cross-domain knowledge about model parameter generation from semantic concepts in natural language. There exists a generic and class-agnostic transformation from the semantic concepts to models for detecting the corresponding visual objects. Such a transformation could be effectively modeled by the dynamic semantic network. Specifically, the second embodiment automatically learns the transformation with a deep dynamic network conditioned on the semantic word embeddings. The method also learns a dynamic relation graph to allow explicit knowledge propagation between the semantic concepts. The method greatly facilitates object detection in the small sample size regime on a broad range of object detection benchmarks.

[0043] The second embodiment leverages cross-domain knowledge about model parameter generation from semantic concepts in natural language. In conventional approaches, FSOD learns exclusively from visual information, as illustrated in FIG. **5**(*a*). A visual feature v **502** is input to a network f **504** parameterized by trainable θ to predict the output y **506**. A loss function **508** then compares the output with the ground truth and computes the gradient, which is back-propagated to update θ. Mathematically, the output is represented as follows:

$$y=f(v;\theta) \tag{4}$$

[0044] In the second embodiment of the invention, instead of training θ, θ is generated from a dynamic semantic network g **510**, parameterized by trainable ϕ, as shown in FIG. **5**(*b*). The network g **510** takes in a language feature l **512** (e.g., a word embedding) representing the semantic representation of a class and outputs a class-specific parameter for the visual network. Specifically:

$$y=f(v;g(l;\phi)) \tag{5}$$

[0045] The visual network f **514** receives gradients **518** from loss function **516** and computes partial derivatives using chain rules, and then backpropagates the gradients **520** to the dynamic semantic network **510**, where they are used to update the trainable parameters ϕ of dynamic semantic network **510**.

[0046] This approach can be applied to both the classification and the localization subnets in the few-shot detector, thus simultaneously tackling few-shot classification and few-shot localization in a unified and coherent manner. With the cross-domain knowledge from natural language, this approach provides great performance in the small sample size regime of few-shot detection.

[0047] The dynamic semantic network g of the second embodiment is a generalization of the relation graph G of the previous embodiment. The relation graph G is implemented with a self-attention module, wherein the attention map is computed using the projections of the same input vector. The dynamic semantic network g, on the other hand, can be any network with learnable parameters. Note that relation graph G is only generating weights for classification, but dynamic network g generates any class-specific weights for both classification and localization in the detection network.

[0048] In conclusion, disclosed herein is semantic relation reasoning for few-shot object detection. The invention explicitly integrates semantic relations between base and novel classes with the available visual information, which assists in improved learning of the novel classes, especially when the novel class data is extremely limited. The semantic relation reasoning is applied to the standard two-stage Faster R-CNN and demonstrates robust few-shot performance against the variation of shot numbers. Compared to prior-art methods, SRR-FSD achieves state-of-the-art results on several few-shot detection settings, as well as a more realistic setting where novel concepts encoded in the pretrained backbone model are eliminated. The key components of SRR-FSD (i.e., semantic space projection and relation reasoning), can be straightforwardly applied to the classification subnet of other few-shot detectors.

[0049] As would be realized by one of skill in the art, the disclosed methods described herein can be implemented by a system comprising a processor and memory, storing software that, when executed by the processor, performs the functions comprising the method.

[0050] As would further be realized by one of skill in the art, many variations on implementations discussed herein which fall within the scope of the invention are possible. Moreover, it is to be understood that the features of the various embodiments described herein were not mutually exclusive and can exist in various combinations and permutations, even if such combinations or permutations were not made express herein, without departing from the spirit and scope of the invention. Accordingly, specific methods and systems disclosed herein are not to be taken as limitations on the invention but as an illustration thereof. The scope of the invention is defined by the claims which follow.

1. A few-shot object detector comprising:

a visual network with trainable parameters producing an output based on an input of a visual feature; and

a dynamic semantic network with trainable parameters producing the trainable parameters for the visual network based on input of a language feature.

2. The few-shot detector of claim **1** wherein the language feature input to the dynamic semantic network represents a language representation of a class for which the visual network is trained to detect.

3. The few shot detector of claim **2** further comprising a loss function generating a gradient for backpropagating to the visual network.

4. The few shot detector of claim **3** wherein the visual network backpropagates the gradients to the dynamic semantic network.

5. The few shot detector of claim **4** wherein the visual network computes partial derivatives of the gradients using chain rules before backpropagating the gradients to the dynamic semantic network.

6. The few shot detector of claim **5** wherein the dynamic semantic network uses the gradient received from the visual network to update the trainable parameters of the dynamic semantic network.

7. The few shot detector of claim **6** wherein the visual network comprises a classification sub-network and a localization sub-network.

8. The few shot detector of claim **7** wherein the dynamic semantic network updates trainable parameters of both the classification sub-network and the localization sub-network.

9. The few shot detector of claim **1** wherein the visual network is trained on a dataset comprising may instances of base class objects and few instances of novel class objects.

10. A system comprising:

a processor; and

memory, storing software that, when executed by the processor, implements the few-shot detector of claim **1**.

11. A method comprising:

training a visual network with trainable parameters to produce an output based on an input of a visual feature; and

training a dynamic semantic network with trainable parameters to produce the trainable parameters for the visual network based on an input of a language feature.

12. The method of claim **11** wherein the language feature input to the dynamic semantic network represents a language representation of a class for which the visual network is trained to detect.

13. The method of claim **12** wherein the visual network:

receives backpropagated gradients from a loss function.

14. The method of claim **13** wherein the visual network:

backpropagates the gradients to the dynamic semantic network.

15. The method of claim **14** wherein the visual network:

computes partial derivatives of the gradients using chain rules before backpropagating the gradients to the dynamic semantic network.

16. The method of claim **15** wherein the dynamic semantic network:

uses the gradients received from the visual network to update the trainable parameters of the dynamic semantic network.

17. The method of claim **16** wherein the visual network comprises a classification sub-network and a localization sub-network.

18. The method of claim **7** wherein the dynamic semantic network:

generates the trainable parameters of both the classification sub-network and the localization sub-network.

19. The method of claim **11** wherein the visual network is trained on a dataset comprising may instances of base class objects and few instances of novel class objects.

20. A system comprising:

a processor; and

memory, storing software that, when executed by the processor, performs the method of claim **11**.

* * * * *