Multi modal representation learning and cross-modal semantic matching

Wang, X.

3

FINE-GRAINED LABEL LEARNING IN OBJECT DETECTION WITH WEAK SUPERVISION OF CAPTIONS

This chapter is based on the following publication:

Wang, X., Du, Y., Verberne, S. Verbeek, F.J. Fine-Grained Label Learning in Object Detectionwith Weak Supervision of Captions. Multimedia Tools and Applications. (under review)

CHAPTER SUMMARY

This chapter addresses RQ2 and RQ3.

RQ2: How to utilize additional knowledge base to measure semantic matching?
RQ3: To what extent can curriculum learning measure the distribution of visual complexity and improve weak supervision for semantic matching?

This chapter addresses the task of fine-grained label learning in object detection with the weak supervision of auxiliary information attached to images. Most of the recent work focused on the label prediction for objects in the same category space as in training data under the supervised learning framework and cannot be expanded to the learning of more fine-grained categories that have not been defined in training sets. In this chapter, we propose a new approach, called label inference curriculum network (LICN), to fine-grained label learning by incorporating the coarse category labels and captions provided in public datasets. First, we build a semantic label map based on embedding techniques and a knowledge base to describe the correspondence between coarse labels and fine-grained label proposals; second, we introduce the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning. To evaluate the performance of fine-grained label learning, we construct multiple datasets based on widely-used public datasets. Experimental results demonstrate the effectiveness of our approach in the task of fine-grained label learning.

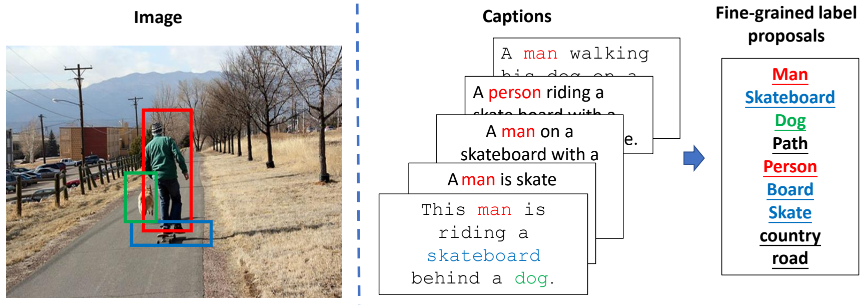


Figure 3.1: An illustration of the image-caption pair. For an image, the location of objects (bounding boxes), the corresponding coarse labels, and the attached captions are provided in the datasets for training. In general, the captions consist of a set of fine-grained label proposals for the objects in the image.

Visual object detection and classification is a fundamental problem in computer vision research and has a wide range of applications, such as face perception, autonomous vehicles and pedestrian detection. Since the renaissance of deep neural networks, object detection has been revolutionized by a series of groundbreaking works, including Faster-RCNN [30], Mask-RCNN [57] and YOLO [58].

Despite these achievements, most deep learning methods have an important limitation: they are trained with exhaustive and clean human annotations. These annotations are expensive as they require human to mark the label and the bounding boxes. Furthermore, labels provided by different annotators are possibly inconsistent. An alternative approach is to relax this requirement of exhaustively labeled data and to use web sources of annotated data, such as social media services like Flickr and Twitter, which have user-generated image tags or captions [59][60]. These data can be seen as natural annotations of the images, providing weak supervision of the collected data, which is a cheap way to increase the scale of datasets near-infinitely.

Weakly supervised object detection (WSOD) is training an object detection model without explicit bounding box annotations. The classic WSOD problem formulation [61][62] treats all object labels per image as a bag of proposals (image-level supervision), and learns to assign instance-level semantics to these proposals using multiple instance learning (MIL). The state-of-the-art model for weakly supervised object detection has reached 43.1% Mean Average Precision [63] on the Pascal VOC 2007 test set. However, there has a strong critical assumption of WSOD is that the image-level labels should be precise, indicating at least one proposal object in the image associated with one label in the image-level labels. This is always the case, especially not in real-world problems and real-world supervision.

A challenge of user-generated labeling (tags or captions) is that these annotations have a lot of noisy labels: Past work has shown that weakly supervised learning algorithms can use these noisy labels [64][65]. However, captions lack information on minor objects or information that may be deemed unimportant, a phenomenon known as reporting bias [66][67]. For example, Fig.3.1 illustrates image captions that describe the same object (marked by a red bounding box) in the image but using different words (person and man) than the predefined category label (person). It is noteworthy that the word "man" is more fine-grained than "person" in describing this object. Also, references to objects may be ambiguous, for example in cases where there are multiple persons in the image.

In this chapter, we focus on a new problem called fine-grained label learning that is different from the traditional WSOD problems. Suppose we have a set of data that is paired image and captions, as shown in Fig.3.1, where the location and coarse labels are provided as ground truth in training sets. In this chapter, we aim to detection objects and learn the fine-grained labels under the joint supervision of the coarse label for an object and the captions for an image. The problem has the following two characteristics. First, the fine-grained labels need to be learned from captions, and thus the supervision of captions is considerably weak, noisy and ambiguous as analyzed above. Second, the uncertainty of noise and ambiguity in the supervision of captions results in different difficulties in the learning process for different examples, and thus the order of training data may affect the learning performance.

To address the problem, this chapter formulates the task of fine-grained label learning with the joint supervision of coarse labels and captions and proposes a novel approach called label inference curriculum network (LICN).

First, we build a semantic mapping that provide a correspondence between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and a knowledge base. Furthermore, we design a curriculum learning process for the Faster R-CNN backbone, where a term called the complexity of samples (CoS) is defined to determine the order of training data in the curriculum learning process.

In summary, our contributions are four-fold. First, we introduce and formulate the problem of fine-grained label learning based on the joint supervision of the coarse category labels and captions. Second, we build a semantic mapping between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and a knowledge base. Third, we propose a novel approach called LICN and design the weakly supervised curriculum learning process for improving the learning performance, where the complexity of samples (CoS) is defined to determine the order of training data in the curriculum learning process. Finally, we construct the datasets consisting of both coarse and fine-grained labels based on MS COCO and Visual Genome for the evaluation of our approach, and the experimental results demonstrate the effectiveness of our approach.

The rest of this chapter is organized as follows. Section 3.1 presents a brief overview of related work. Section 3.2 formulates the problem of fine-grained label learning and introduces our approach in details. Section 3.3 provides the experimental results and analysis, and Section 3.4 concludes the chapter.

3.1. RELATED WORK

The task of weakly supervised object detection involves the correlation of different media and the information distribution from images to the corresponding captions. Therefore, we review the related work in terms of lexico semantic analysis and weakly supervised entity localization.

The task in this chapter has some differences from the following related problems:

Learning from Text: Ye et al. [63] harvest detection models from free-form text and use a label inference module to amplify signals in the free-formed texts to supervise the learning of a multiple instance detection network. Fang et al. [65] use multiple instance learning to train visual detectors for words that commonly occur in captions. Most learning from text model does not use the same semantic word in text as new category.

Weakly Supervised Object Detection and Segmentation(WSOD) [68, 69, 70]: In general, this task aims to detect objects from images based on the supervision of a set of image-level labels. To the best of our knowledge, the existing WSOD methods have not involved captions, a type of weaker supervisory information than exact image-level labels, in object detection.

Fined-Grained Image Classification(FGIC) [71][72]: FGIC usually involves classifying the sub-classes of objects belonging to the same class. In each class, objects of different subclasses are both semantically and visually similar to each other.

3.1.1. LEXICO-SEMANTIC ANALYSIS

In the widely-used public image datasets, there is typically a semantic gap between the human-written captions and the categorical annotations of the objects in the images. For example,the annotation of the object in red box is "person" while the caption uses the word "man" in Fig 3.1. A variety of lexico-semantic methods have been proposed to bridge this semantic gap. These methods can be divided into two categories: knowledge-based methods and corpus-based methods [73][16]. Knowledge-based methods rely on external semantic resources (thesauri or lexical knowledge bases) to identify similarities between two words. For example, WordNet [74] and HowNet [75] [76] are used to measure semantic distance between a pair of words. Although these semantic metrics are interpretable and effective, they have as drawbacks that they lack context information and that the similarity can only be computed when both words are present in the lexicon.

Due to the knowledge-based methods limitations, corpus-based methods are then proposed to utilize context information around the center words. Current corpus-based methods train vector representations (called 'embeddings') based on contexts of words in a large text collection. The word similarity study mostly uses a statistical description of the context [77][11]. The most used static word embedding model is Word2Vec [26][25], a highly efficient model proposed by Google. The model can simplify the processing the text context into a K-dimensional vector space, so we can use the spatial similarity to represent similarity in text semantics. Li et al. [78] provide a transferred vector approach, that utilizes a transferred vector for the representation of a word to reveal the word semantics better, not just relying on its own embedding. In our work, we use these two types of model, i.e., WordNet [74] and Word2Vec [26][25], to build a semantic map between the pre-annotated coarse labels and the fine-grained label proposals from captions.

3.1.2. Weakly Supervised Multiple Instance Learning

Most weakly supervised methods for object detection formulate the task as a multiple instance learning (MIL) problem. In this problem, MIL addresses the data objects represented by a bag of instances and associated with a label (a set of labels) for each bag. If the image is labeled as containing an object, at least one of the label proposals will be responsible for providing the prediction of that object. The papers by Oquab et al. [79] and Zhou et al. [80] propose a Global Average (Max) Pooling layer to learn class activation maps. Bilen at al. [61] propose Weakly Supervised Deep Detection Networks (WSDDN) containing classification and detection data streams, where the detection stream weighs the results of the classification predictions. Kantorov et al. [81] improve WSDDN by considering context. Tang et al. [69][82] jointly train multiple refining models together with WSDDN, and show the final model benefits from the online iterative refinement. Diba et al. [57] and Wei et al. [83] apply a segmentation map and Wan et al. [62] incorporate saliency. Finally, Redmon et al. [58] introduce a min-entropy loss to reduce the randomness of the detection results.

Our work is similar to all the above since we also represent the proposals using a MIL weighted representation. However, we go one step further to successfully adopt a more challenging supervision scenario where the captions are utilized as the weak supervision for the learning fine-grained labels in the task of object detection.

3.1.3. Curriculum Learning

Curriculum learning[84] was proposed by Yoshua Bengio in 2009. It formalizes the learning process of humans and animals from easy cases to gradually more complex ones. In recent years, more and more weakly supervised learning methods based on curriculum learning have been proposed and obtain good performance [85][86]. CurriculumNet [41] designs a curriculum learning process by measuring the complexity of data using its distribution density in a feature space for the classification of large-scale weakly-supervised web images without human annotations, where the negative impact of noisy labels is reduced substantially. Wang et al. [87] address the object detection problem by learning an effective object detector using weakly-annotated images with curriculum learning. Hacohen et al. [88] analyze the effect of curriculum learning, which involves the non-uniform sampling of mini-batches, on the training of deep networks. In this chapter, we design a curriculum learning process by defining a new measurement of the degree of difficulty in fine-grained label learning.

3.2. METHODOLOGY

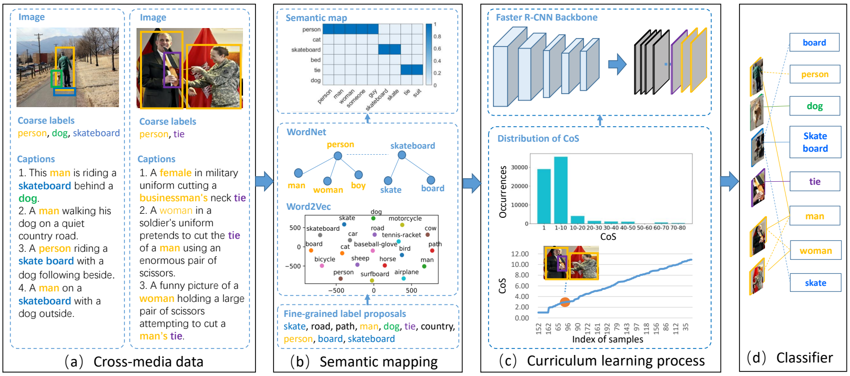
3.2.1. OVERVIEW

In this chapter, we are given a pair consisting of an image and its captions. Formally, we have $\mathscr{D}_{tr} = \{(I_i, R_i, L_i^I, C_i)\}_{i=1}^{M_{tr}}$ and $\mathscr{D}_{te} = \{(I_i, L_i^I, C_i)\}_{i=1}^{M_{te}}$ as the training set and test set, respectively, where I_i and C_i denote the i-th image and caption, respectively, and $L_i^I = \{l_{i1}^I, l_{i2}^I, \cdots, l_{im_i}^I\}$ refers to the annotations of I_i, each considered as a coarse category label for one of the m_i visual object regions $R_i = \{r_{i1}, r_{i2}, \cdots, r_{im_i}\}$ segmented from this image. The caption C_i consists of a set of entities that generally provide more fine-grained category information than L_i^I for the visual object regions R_i and thus we extract them from captions as fine-grained label proposals, denoted by $L_i^C = \{l_{i1}^C, l_{i2}^C, \cdots, l_{in_i}^C\}$. In this manner, we have a coarse label vocabulary V_I and a fine-grained label proposal vocabulary V_C that consist of all coarse labels and fine-grained label proposals, respectively, where $l_{i\cdot}^I \in V_I$ and $l_{i\cdot}^C \in V_C$. Regarding the labels we make two observations: 1) the label proposals L_i^C from captions are generally more fine-grained than the annotations L_i^I preassigned to the image; 2) The correspondence at the granularity of instances (i.e., between a fine-grained label proposal $l_{i\cdot}^C$ and a visual object region $r_{i\cdot}$) is missing. An example can be seen in the second image of Fig. 3.2(a). It is in this image unknown which region corresponds to the fine-grained label "man" or "woman" as extracted from the captions.

We aim to learn and infer the fine-grained label $l_{i\cdot} \in V_I \cup V_C$ for each visual object region based on the supervision from the training data D_{tr}. As illustrated in Fig. 3.2, our framework includes two main processes: semantic mapping and curriculum learning-based fine-grained label learning. In the semantic mapping, we extract the entities from captions as the fine-grained label proposals $l_{i\cdot}^C \in V_C$ and measure the semantic similarity between the extracted label proposals and the coarse labels $l_{i\cdot}^I \in V_I$ based on the combination of the knowledge base WordNet and data-driven embedding techniques. To learn the fine-grained label for each object, we propose a curriculum learning-based method to train the model by adding data in an ascending order of example complexity.

3.2.2. SEMANTIC MAPPING

The purpose of the semantic map is to build the relationship between the coarse label $l_{i\cdot}^I$ and the fine-grained label proposals $l_{i\cdot}^C$ by measuring their similarity over the training set. We extract all nouns from captions with the CoreNLP toolkit [89]



1. This man is riding a skateboard behind a dog.
2. A man walking his dog on a quiet country road.
3. A person riding a skate board with a dog following beside.
4. A man on a skateboard with a dog outside.

1. A female in military uniform cutting a businessman's neck tie
2. A woman in a soldier's uniform pretends to cut the tie of a man using an enormous pair of scissors.
3. A funny picture of a woman holding a large pair of scissors attempting to cut a man's tie.

(a) Cross-media data (b) Semantic mapping (c) Curriculum learning process (d) Classifier

Figure 3.2: The framework of the proposed LICN approach. (a) The input data in the form of image-captions pairs, where the image, coarse labels and captions are provided in training sets. (b) Semantic mapping between the coarse labels and fine-grained label proposals based on embedding techniques and a knowledge base. (c) Curriculum learning process for the Faster R-CNN backbone, where the complexity of samples is defined to measure the degree of difficulty in learning the fine-grained labels. (d) The classifier for predicting fine-grained labels.

as the candidates for the fine-grained label proposals. In order to get a semantic map we pass three steps.

Semantic Mapping Based on Knowledge Base

We employ WordNet as the knowledge base to measure the semantic similarity between annotations and fine-grained label proposals. WordNet can represent relations between word senses with an ontology. For an annotation $l^I_{i\cdot}$, we can obtain the synset $W_{kb}(l^I_{i\cdot})$ from WordNet in the form of:

$$W_{kb}(l^I_{i\cdot}) = \{H_{per}(l^I_{i\cdot}), H_{pon}(l^I_{i\cdot}), S_{non}(l^I_{i\cdot})\}, \tag{3.1}$$

where $H_{per}(\cdot)$, $H_{pon}(\cdot)$ and $S_{non}(\cdot)$ refer to the hypernym, hyponym and synonym, respectively, for a given word in the WordNet.

Semantic Mapping Based on Embedding

We use Word2Vec as the embedding technique to measure the similarity of labels in $V_I \cup V_C$. We fine-tune the pre-trained Word2Vec model [26] on all captions in the data. In this chapter, we extract all words in captions to build a vocabulary. By our analysis, all coarse labels preassigned to images appear in this vocabulary, so that we can obtain the feature vector of each coarse label in embedding space. As the fine-grained labels are extracted from the captions, we can obtain the feature vectors of fine-grained labels as well. For a coarse label $l^I_{i\cdot}$ and a fine-grained label proposal $l^C_{i\cdot}$, we achieve their d_e-dimensional embedding vectors $\mathbf{l}^I_{i\cdot}$ and $\mathbf{l}^C_{i\cdot}$, respectively, with the Word2Vec model. The similarity between two vectors in the embedding space is measured by the cosine similarity $S(\cdot,\cdot)$.

Building the Semantic Map

As analyzed above, we build a semantic mapping between the annotations $l^I_{i\cdot}$ and the fine-grained label proposals $l^C_{i\cdot}$ with the following matrix:

$$W(l^I_{i\cdot}, l^C_{i\cdot}) = \begin{cases} 1,\ if\ l^C_{i\cdot} \in W_{kb}(l^I_{i\cdot})\ and\ S(\mathbf{l}^I_{i\cdot}, \mathbf{l}^C_{i\cdot}) > \varepsilon \\ 0,\ otherwise, \end{cases} \tag{3.2}$$

where ε is a threshold in $[0,1]$. With Eq.3.2, we can find one or multiple fine-grained label proposals that are semantically similar with the given annotation. Since a visual object region strictly corresponds to an annotation in the dataset, we can achieve a weak correspondence between visual object regions and fine-grained label proposals.

3.2.3. FINE-GRAINED LABEL LEARNING BASED ON CURRICULUM LEARNING

Curriculum learning is an effective learning framework that imposes structure on the training set relying on a notion of "easy" and "hard" examples [84]. In the following subsection, we will find that the examples are of different difficulties to learn and infer the fine-grained labels. Therefore, we perform the fine-grained object label learning based on the curriculum learning framework.

Backbone for Object Detection

Based on the semantic mapping introduced in Subsection 3.2.2, we have achieved the correspondence between each visual object region $r_{i\cdot}$ in the i-th image and its fine-grained label proposals (a subset of L^C_i). Without ambiguity, we redenote them by r_k and $\widetilde{L}^C_k$ by removing the subscript i (the index of images), where k is the index of a visual object region in the dataset, $r_k \in R_i$ and $\widetilde{L}^C_k \subset L^C_i$. Thus, our objective is to localize the visual object and learn its fine-grained label with the weak supervision of a set of fine-grained label proposals $\widetilde{L}^C_k$ to the visual object region r_k.

We use the Faster R-CNN model [30], denoted by $F_{det}(I_i)$, as the backbone of our work. The Faster R-CNN consists of three modules: a convolutional neural network for generating the feature map of an image, an RPN (region proposal network) for generating a set of rectangular object proposals performed on the feature map, and a classifier for learning the category label of each region. The output of the backbone can be described as follows:

$$(\mathbf{P}_i, R_i) = F_{det}(I_i), \tag{3.3}$$

where $R_i = \{r_{ij}\}^{m_i}_{j=1}$ denotes the set of m_i visual object regions extracted from the image I_i, where the location of each region is described by four coordinates of the bounding box, and $\mathbf{P}_i = [\mathbf{p}_{i,1}, \mathbf{p}_{i,2}, \cdots, \mathbf{p}_{i,m_i}]$ denotes the probabilities that all object regions in R_i are predicted to categories. Without ambiguity, we rewrite

$\mathbf{p}_{i,j}$ as $\mathbf{p}_k = [p_{k,1}, p_{k,2}, \cdots, p_{k,C_C}]^T$ by removing the index of images, where $p_{k,c}$ denotes the probability that a visual object region r_k is categorized into the c-th class and C_C denotes the cardinality of V_C (the same as the cardinality of $V_I \cup V_C$ as all annotations appear in the fine-grained label proposals). In our work, we define the space of categories with the fine-grained label proposals, i.e., V_C.

The Complexity of Samples

Different samples have different difficulty in the learning of fine-grained labels. For example, if there is only an object region annotated by “person” in an image and only an fine-grained label proposal “man” in the caption is related to the annotation according to the semantic mapping in Eq.3.2, it is easy to infer the fined-grained label for the object region. In contrast, if there are multiple fined-grained label proposals corresponding to the annotation according to the semantic mapping, it is much difficult to discriminate which one is the true fine-grained label of the object region in the image. We introduce a term called *the complexity of samples* (CoS) to describe the difficulty in the task. We define the CoS of a sample $D_i \in \mathscr{D}$ as follows:

$$H_{CoS}(D_i) = -\sum_{l^I_{i\cdot}} \sum_{l^C_{i\cdot}} \Pr(l^C_{i\cdot}|l^I_{i\cdot}) \log(\Pr(l^C_{i\cdot}|l^I_{i\cdot})), \tag{3.4}$$

where $\Pr(l^C_{i\cdot}|l^I_{i\cdot})$ is the conditional probability of the fine-grained label proposal $l^C_{i\cdot}$ given the annotation $l_{i\cdot}$ and can be achieved by:

$$\Pr(l^C_{i\cdot}|l^I_{i\cdot}) = \frac{W(l^I_{i\cdot}, l^C_{i\cdot})}{\sum_{l^C_{i\cdot} \sim l^I_{i\cdot}} W(l^I_{i\cdot}, l^C_{i\cdot})}, \tag{3.5}$$

where $l^C_{i\cdot} \sim l^I_{i\cdot}$ denotes all fine-grained label proposals $l^C_{i\cdot}$ related to the annotation $l^I_{i\cdot}$ according to Eq.3.2. As shown in Eq.3.4, CoS is defined based on the Shannon's Entropy that is mainly used to measure the uncertainty of a discrete random variable. In this chapter, we consider $l^I_{i\cdot}$ as the random variable and $l^C_{i\cdot}$ as its values with non-zero probability. If more fine-grained label proposals are related to the annotation, the correspondence between them is more uncertain and the label proposal is thus more intractable. Moreover, if there are multiple visual objects detected in an image, the CoS tends to be a larger value accordingly based on Eq.3.4.

Curriculum Learning Process

Based on the semantic mapping, we have obtained the fine-grained label proposals $\widetilde{L}^C_k$ for each visual object region r_k. Here we transform $\widetilde{L}^C_k$ to a binary vector $\mathbf{y}_k = [y_{k,1}, y_{k,2}, \cdots, y_{k,C_C}]^T \in \{0,1\}^{C_C}$. $y_{k,c} = 1$ ($y_{k,c} = 0$) means the c-th fine-grained label proposal of V_C is present (absent) in $\widetilde{L}^C_k$.

In the curriculum learning process, the training data are fed to the Faster R-CNN in the order of easy samples (with low CoS) to hard samples (with high CoS). The loss for the learning of fine-grained labels is defined as follows:

$$L_{ws}^{k} = \sum_{c=1}^{C_C} y_{k,c} \cdot \log p_{k,c} + (1 - y_{k,c}) \cdot (1 - \log p_{k,c}), \tag{3.6}$$

where L_{ws}^{k} refers to the weakly supervised loss. Different from the original Faster R-CNN, the ground truth of label vector, i.e., $\mathbf{y}_k$, may consist of multiple ones corresponding to multiple fine-grained label proposals, rather than being a one-hot vector.

3.3. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate the effectiveness of the proposed model LICN by answering the following two questions. Q1: What is the quality of the learnt fine-grained label proposals semantic map reasonable for this weakly supervised object detection model? Q2: How effective the proposed LICN approach is in terms of the fine-grained label learning based on weakly supervised paradigm learning?

3.3.1. EXPERIMENTAL SETUP

For the experimental setup, we first describe the dataset and then the implementation details.

Datasets

The experiments are conducted on the MS COCO 2017 dataset, Visual Genome, the Pascal VOC 2007 test dataset, and our constructed datasets based on the three datasets. Table 3.1 shows an overview of these datasets.

- The *MS COCO 2017* dataset contains 118,287 training images and 5,000 validation images. It provides 5 human-annotated captions per image and a

Table 3.1: An overview of the datasets.

datasets	# of images	# of categories	# of objects
Visual Genome	107,228	80,138	3,909,697
MS COCO	118,287	80	860,001
FG-COCO	118,287	169	860,001
sCOCO training	76,631	69	200,962
FG-sCOCO training	76,631	150	200,962
FG-sCOCO test	13,175	150	29,169
FG-sCOCO val.	2,000	150	14,090
Visual Genome test	54,212	150	496,809

total of 80 category labels for the object regions segmented from all the images. The category labels play the role of the annotations L_i^I and the captions are used for the building of the semantic map and the extraction of fine-grained label proposals L_i^C for image I_i.

- *Visual Genome* contains 107,228 images, 3,909,697 objects from 80,138 categories, and other information such as the relationships between objects. The categories in Visual Genome are much more fine-grained than those in MS COCO, and thus we use this dataset for testing the performance of fine-grained label inference and the category labels as the ground truth.

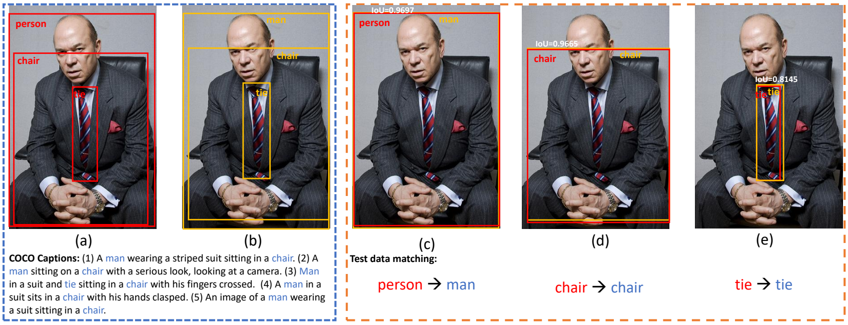
- The *Pascal VOC 2007 test* dataset has 4,952 images and 20 categories of objects. It is utilized as as the test dataset.

Based on the above datasets, We construct the following datasets for training and testing our approach from different aspects:

- *FG-COCO*: We replace the coarse category labels of the objects in each image in MS COCO by the fine-grained label proposals appearing in the corresponding caption based on the semantic map and thus obtain FG-COCO. A total of 169 category labels (including the original coarse labels from MS COCO and new fine-grained category labels) are generated for the objects in the dataset.

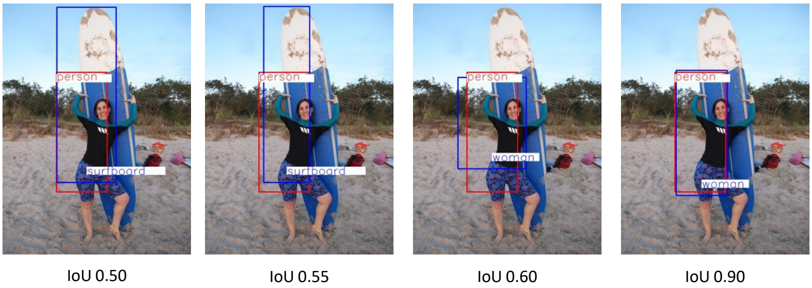
- *FG-sCOCO test dataset*: There are a set of images appearing both in MS COCO and in Visual Genome. For an image in the set, if the Intersection over Union (IoU) between a bounding box from MS COCO and a bounding box from Visual Genome is larger than 0.90, we keep the image as an image example, and the bounding box from MS COCO and category labels from Visual Genome (must appear in the corresponding caption from MS COCO as well) as the ground truth of the location and fine-grained label for an object, respectively. We randomly choose 2000 images from the set for validation (called FG-sCOCO val. as shown in Table 3.1), and the rest is for test. As a result, the FG-sCOCO test dataset consists of 13,175 images and 29,169 objects with 150 category labels (including the original coarse labels from MS COCO and new fine-grained category labels). In the experiments, we adopt the FG-sCOCO test dataset to evaluate the performance of fine-grained label learning and inference.

- *FG-sCOCO training dataset*: It is a subset of FG-COCO, which excludes all the images appearing in the FG-sCOCO test and FG-sCOCO val. dataset. This dataset consists of 76,631 images and 200,962 objects with 150 category labels (including the original coarse labels from MS COCO and new fine-grained category label proposals from the semantic map). To make the learning robust, we keep only the categories consisting of more than 200 examples of object regions in the dataset.



COCO Captions: (1) A man wearing a striped suit sitting in a chair. (2) A man sitting on a chair with a serious look, looking at a camera. (3) Man in a suit and tie sitting in a chair with his fingers crossed. (4) A man in a suit sits in a chair with his hands clasped. (5) An image of a man wearing a suit sitting in a chair.

Figure 3.3: Test data example: (a) shows an example from MS COCO with object bounding boxes and the associated category labels (red color); (b) shows the same image in the Visual Genome dataset with object bounding boxes and the associated category labels (blue color); (c), (d) and (e) show the matching between the object regions from MS COCO and Visual Genome with an IoU value larger than 0.90. We see that "person" matches to "man", "chair" to "chair" and "tie" to "tie".



IoU 0.50 IoU 0.55 IoU 0.60 IoU 0.90

Figure 3.4: IoU example: The red color box and blue color box come from MS COCO and Visual Genome, respectively, and the IoU value of two different boxes of the same object should be high.

- *sCOCO training dataset*: As a subset of MS COCO, it consists of all the images in FG-sCOCO training dataset, and its bounding boxes and category labels are from MS COCO. As a result, the dataset consists of 76,631 images and 69 category labels for 200,962 objects.
- *Visual Genome test dataset*: Different from FG-sCOCO test dataset, Visual Genome test dataset is the subset of Visual Genome that excludes all the images appearing in MS COCO. In this dataset, we only keep those objects whose category labels appear in the FG-sCOCO training dataset. As a result, the dataset consists of 54,212 images and 496,809 objects with 150 category labels.

The following is an analysis of the building of the FG-sCOCO test dataset. We assume that the IoU value is high for a paired bounding boxes of the same object

Table 3.2: The characteristics of the interaction of MS COCO and Visual Genome with different IoU threshold values.

IoU	# of images	# of objects	# of categories in MS COCO	# of categories in Visual Genome
0.50	30,983	96,529	79	2,004
0.55	29,337	85,468	79	1,680
0.60	27,621	75,772	79	1,407
0.65	26,118	67,248	79	1,143
0.70	24,890	59,503	78	940
0.75	23,591	51,222	77	787
0.80	21,958	41,848	76	654
0.85	19,603	31,303	76	537
0.90	15,529	19,702	74	413
0.95	7,306	7,957	72	281

in the same image from the overlapping part between Visual Genome and MS COCO. As shown in Fig.3.3, the paired bounding boxes with high IoU value has the same semantics, but may have different object labels. As Visual Genome has 80K category labels which contain all fine-grained categories, we use these object labels as ground truth label to evaluate the semantic map (Q1). We illustrate the role of the threshold on the IoU value in Fig. 3.4. Table 3.2 shows the effect of different IoU threshold values on the data. For example, for the IoU of 0.90, there are 19,702 paired objects with an IoU larger than 0.90 from 15,529 images, and these objects belong to 74 categories in MS COCO and 413 categories in Visual Genome. Considering the count of test data and the count of the object categories, we will evaluate our model on the FG-sCOCO test dataset with IoU in $[0.90, 1]$. For the object detection (Q2), we found that size of images are a little different between MS COCO and Visual Genome for the image with same id. We resize the size of Visual Genome images to make them equal to the size of same image in MS COCO.

Implementation Details

We train the proposed models on two different datasets: FG-COCO and FG-sCOCO, and thus generate the following four configurations:

- LICN-E2C$_{FG-COCO}$: learned on the FG-COCO dataset by feeding training examples from easy to complex;
- LICN-C2E$_{FG-COCO}$: learned on the FG-COCO dataset by feeding training examples from complex to easy;
- LICN-E2C$_{FG-sCOCO}$: learned on the FG-sCOCO training dataset by feeding training examples from easy to complex;

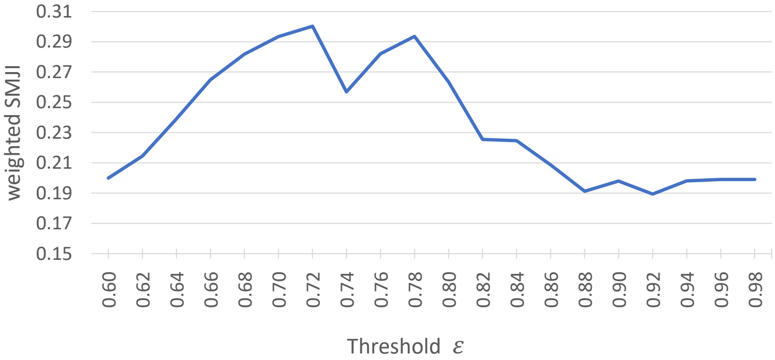


Figure 3.5: The effect of Word2Vec similarity parameter ε in Eq.2 on the performance (weighted SMJI) of semantic mapping for the FG-sCOCO validation set.

- LICN-C2E$_{FG-sCOCO}$: learned on the FG-sCOCO training dataset by feeding training examples from complex to easy.

We use Faster R-CNN with a backbone of VGG-16 as the basic framework of our work. The VGG-16 backbone is pre-trained on ImageNet and then fine-tuned on our training datasets. In the process of fine-grained label learning, we use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a learning rate of 0.01. We set the maximum epoch to 20 for the convergence of learning process. The minibatch size is set to 1 for the flexible feeding of the examples of different complexity. All the experiments are conducted on a platform of 8 Nvidia Titan V GPUs with Pytorch.

3.3.2. EVALUATION METRICS

Semantic Mapping

We define a weighted semantic map Jaccard index (SMJI) for measuring the closeness between the fine-grained labels mined by semantic mapping and the fine-grained label ground truth provided in the FG-sCOCO validation set. The weighted SMJI is defined as follows:

$$W\_SMJI = \sum_k W_k \cdot \frac{L_k^{SM} \cap L_k^{GT}}{L_k^{SM} \cup L_k^{GT}}, \tag{3.7}$$

where L_k^{SM} and L_k^{GT} denote the sets of fine-grained labels mined by semantic mapping and the fine-grained label ground truth provided in the FG-sCOCO validation set, respectively, corresponding to the k-th coarse category label, and the operators $\cup$ and $\cap$ denote the union and intersection of two sets, respectively. For example, for the coarse category label of “person”, L_k^{SM}= {“guy”, “man”, “person”,

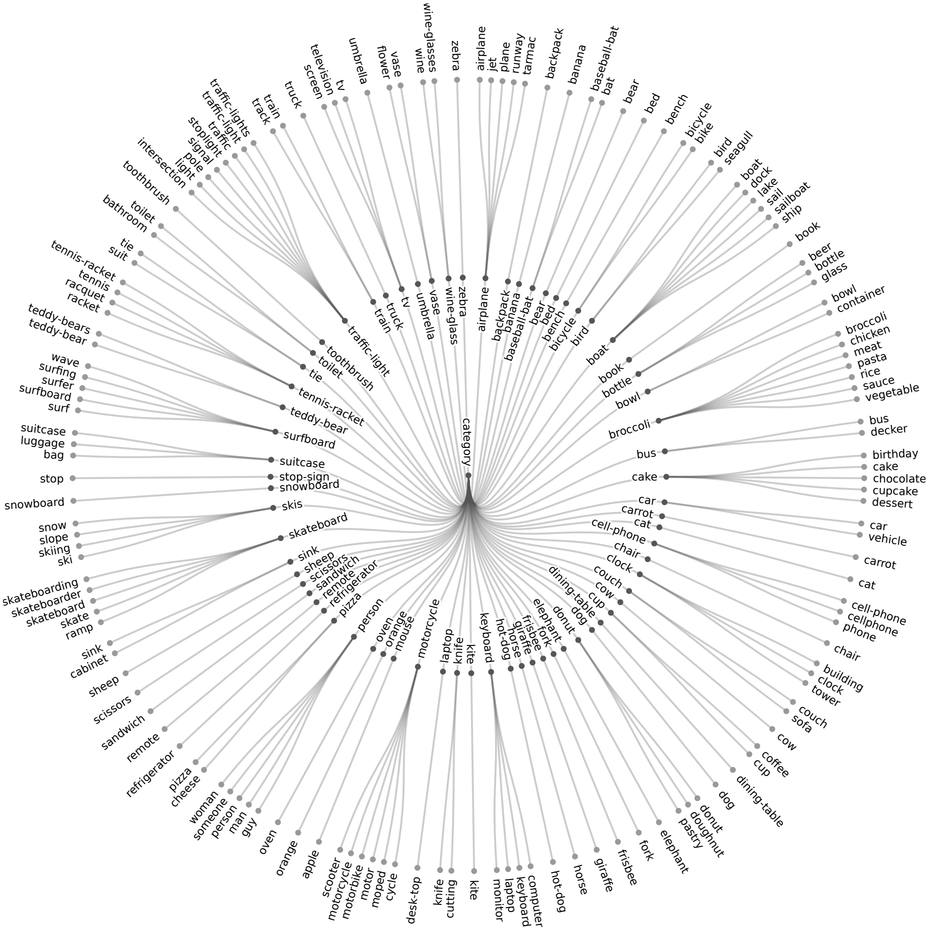


Figure 3.6: The illustration of the semantic map that consists of 69 coarse category labels (points on the inner circle) and 81 fine-grained category labels (points on the outer circle) appearing in the FG-sCOCO validation set.

"woman", "someone"} and L_k^{GT}= {"guy", "man", "person", "skateboarder", "surfer", "woman"}. The weight W_k in Eq. 3.7 is defined as follows:

$$W_k = \frac{\mid L_k^{GT} \mid}{\sum_k \mid L_k^{GT} \mid} \tag{3.8}$$

where $|\cdot|$ denotes the cardinality of a set. Fig. 3.5 reports the weighted SMJI on the FG-sCOCO validation set as the threshold ε changes. From the figure, we observe that the performance of semantic mapping in mining the fine-grained labels is optimal when $\varepsilon = 0.72$. Thus, we choose $\varepsilon = 0.72$ in the following exper-

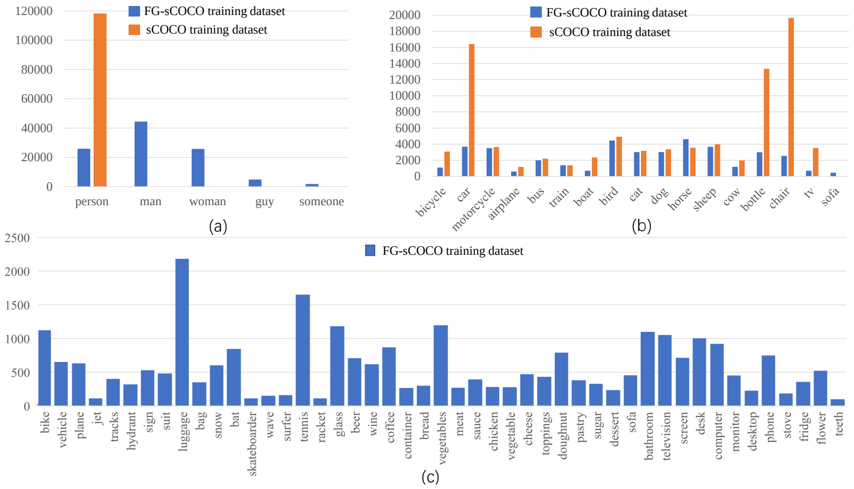


Figure 3.7: The comparison of occurrence frequencies of category labels between before and after semantic mapping, where the orange bars indicate the occurrence frequencies of the coarse labels in sCOCO training dataset and blue bars indicate the occurrence frequencies of the labels (either the original coarse labels or the generated fine-grained label proposals) in the our constructed FG-sCOCO training dataset after semantic mapping. a) Comparison between the coarse label of category "person" and the corresponding fine-grained labels, b) comparison on 17 coarse categories, and c) comparison on the generated fine-grained categories.

iments. Fig. 3.6 illustrates the semantic map that consists of 69 coarse category labels and 81 fine-grained category labels appearing in the FG-sCOCO validation set. From the figure, we observe that most fine-grained label proposals extracting from captions are semantically similar with the coarse labels, while a few noises are introduced by the semantic mapping. For example, the generated "chicken", "meat", "pasta", "rice" and "sauce" are not semantically similar with the coarse label "broccoli". These noises will be reduced with the curriculum learning process.

In Fig. 3.7, we illustrate of the occurrence frequencies of the category labels (including the coarse and fine-grained labels) in the FG-sCOCO training dataset and the sCOCO training dataset, which correspond to the data with and without semantic mapping, respectively. Due to the large difference in the occurrence frequencies of these categories, we report the results separately in three subfigures. From the figure, we find that a large amount of fine-grained label proposals are generated with semantic mapping.

Object Detection

We utilize a widely-used metric, namely average precision (AP), to evaluate the performance of object detection. AP is defined as the average detection precision under different recalls and usually evaluates the performance in a category specific manner. To compare performance over all object categories, the mean AP

Table 3.3: Average precision (AP) (%) results of LICNs trained on FG-sCOCO training dataset. The results are reported on the FG-sCOCO test dataset.

Method	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
LICN-C2E	21.90	37.00	22.80	15.40	16.80	24.00
LICN-E2C	23.60	37.40	25.40	13.10	19.10	25.30

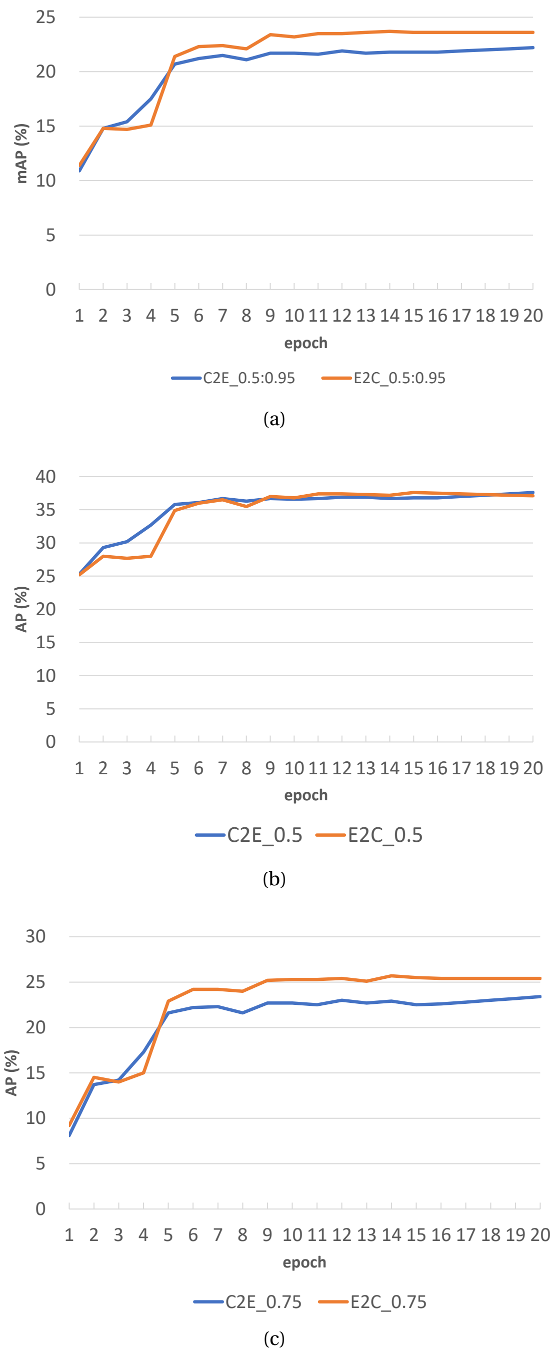
(mAP) averaged over all object categories is usually used as the final metric of performance. To measure the object localization accuracy, the IoU is used to check whether the IoU between the predicted box and the ground truth bounding box is greater than a predefined threshold 0.5. Instead of using a fixed IoU threshold, based on MS COCO AP is averaged over multiple IoU thresholds between 0.5 (coarse localization) and 0.95 (perfect localization).

3.3.3. PERFORMANCE AND ANALYSIS

FG-sCOCO

We first evaluate our method on the FG-sCOCO validation dataset to analyze the importance of curriculum learning, where the proposed LICN models are trained on the FG-sCOCO training dataset.

Fig.3.8 shows the results of the LICN models for the FG-sCOCO validation dataset. We find that the E2C version of LICN improves the performance of fine-grained label learning. As shown in Fig.3.8(a), in terms of the mean AP of 0.5:0.05:0.95 IoU, LICN-E2C performs approximately 0.02 AP improvement better than the LICN-C2E model. However, Fig.3.8(b) for the 0.50 IoU AP, after 7 epochs there is not a large difference between the LICN-E2C and LICN-C2E model. Fig.3.8(c) shows the 0.75 IoU AP, for which LICN-E2C performs approximately 0.03 AP better than the LICN-C2E model. LICN-E2C improves the performance for the predictions of the 0.75 IoU. As IoU means the object location accuracy, IoU close to 1 means that the predicted object location is close to the ground truth. We observe that the improvement is brought by complexity ranked as the IoU increases. This can be explained by the fact that LICN object detection is able to compute the complexity of the images, and thus it is prone to make fine-grained label predictions for the same object. Table 3.3 shows a more detailed experimental result on the FG-sCOCO test dataset. In Table 3.3, " Avg. Precision, Area S M L" means the average precisions for small ($area < 32^2$), medium ($32^2 < area < 96^2$), and large ($area > 96^2$) objects, respectively, where the area is measured as the number of pixels in the segmentation mask. The table shows that in the case of 0.75 IoU, the E2C version improves the performance by 2.6% compared with the C2E version, which demonstrates that it is better to learn the fine-grained labels with the consideration of the complexity of samples defined in the "Methodology" section.



(a)

(b)

(c)

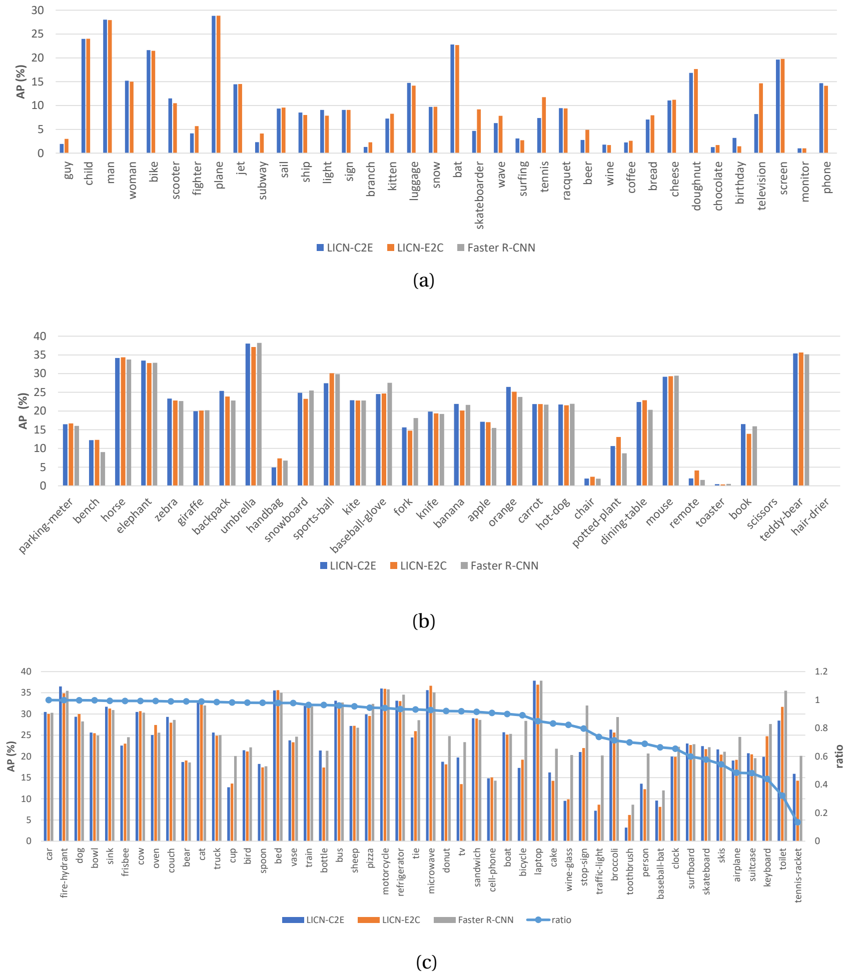
Figure 3.8: Results of LICN-E2C and LICN-C2E on the FG-sCOCO validation set for different training epochs. (a) Mean AP of 0.50:0.95 in steps of 0.05, (b) AP of 0.5, and (c) AP of 0.75.

Table 3.4: Average precision (AP) (%) results for all the 20 categories of the Pascal VOC 2007 test datatet. Faster R-CNN was trained on MS COCO and the sCOCO training dataset consisting of 80 coarse labels and 69 coarse labels, respectively, and LICN was trained on the FG-COCO dataset and FG-sCOCO training dataset with the expanded labels consisting of both the coarse and fine-grained labels.

	Method	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	**mean**
FG-COCO	ratio	0.48	0.89	0.98	0.90	0.96	0.96	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.94	0.69	1.00	0.95	1.00	0.96	0.92	
	Faster R-CNN[30]	**84.0**	**83.1**	**76.5**	**58.9**	**67.7**	**87.4**	77.1	85.6	61.0	**83.9**	**66.3**	78.4	86.3	86.6	**86.2**	50.9	81.7	68.1	86.1	**78.8**	**76.7**
	LICN-C2E	76.8	71.7	74.3	52.7	62.4	87.2	79.7	85.3	60.6	82.5	65.0	79.3	85.5	85.7	70.5	50.2	81.4	68.1	86.5	74.2	74.0
	LICN-E2C	71.6	69.9	75.2	52.9	64.1	87.0	**80.0**	**86.5**	**62.0**	83.6	65.6	**81.0**	**86.4**	**86.7**	68.2	**54.2**	**83.2**	**70.7**	**86.8**	76.9	74.6
FG-sCOCO	ratio	0.26	0.51	0.97	0.67	0.38	0.72	1.00	1.00	1.00	1.00	-	1.00	1.00	0.86	0.11	-	0.93	1.00	0.91	0.59	
	Faster R-CNN[30]	**77.4**	**79.7**	71.5	**58.9**	**52.3**	**85.2**	74.4	**86.3**	38.4	77.5	-	80.4	85.6	81.9	**83.9**	-	81.2	**64.1**	85.2	**64.3**	**73.8**
	LICN-C2E	69.9	76.1	68.6	50.9	41.8	81.4	73.4	85.9	37.3	74.4	-	78.3	84.0	81.2	46.3	-	76.1	63.7	84.1	59.6	68.5
	LICN-E2C	71.7	77.3	**73.8**	48.0	42.7	79.3	**75.4**	86.2	**39.4**	**79.5**	-	**80.5**	**86.2**	**82.7**	47.1	-	**81.4**	63.7	**86.1**	61.7	70.1

VOC 2007

We train our model on FG-COCO and the FG-sCOCO training dataset and test the learned models on the VOC 2007 test dataset to evaluate the object detection performance. Correspondingly, the Faster R-CNN baseline is trained on MS COCO and the sCOCO training dataset. Table 3.4 shows the experimental results for the 20 coarse categories in the VOC 2007 test dataset, where only 18 categories are shown for our model learned on the FG-sCOCO training dataset as the categories of “diningtable” and “pottedplant” do not appear in the training set. The table shows a term called *ratio*, which is defined as the ratio of the number of occurrences for a category in the training set FG-COCO (FG-sCOCO training) to that in the training set MS COCO (sCOCO training) and describes the degree of how many objects in a coarse category of MS COCO (sCOCO training) have not been re-assigned to a corresponding fine-grained category of FG-COCO (FG-sCOCO training) with the semantic mapping. The ratio equal to 1 means that no object in MS COCO (sCOCO training) is re-assigned to a fine-grained category and its coarse label is kept in constructing FG-COCO (FG-sCOCO training). From the table, we observe that for most of the categories with the ratio close to 1, such as “car”, “chair”, “dog” and “train”, the detection result of our proposed LICN-E2C version has better performance than the Faster R-CNN baseline. For these categories, the training examples are almost the same between FG-COCO (FG-sCOCO training) and MS-COCO (sCOCO training). The result demonstrates that our approach improves the label inference performance in the image detection problem. For the categories with the ratio much lower than 1, such as “aero” and “person”, LICN has a lower performance than Faster-RCNN. We note that in this case, there is a large difference between the training sets for LICN and Faster R-CNN: FG-COCO (FG-sCOCO training) has a much larger label space and less training examples for many categories than MS COCO (sCOCO training), which significantly increases the difficulty of label learning and inference and thus results in the the drop of AP of LICN. It is noteworthy that our LICN-E2C achieves improvements of 0.6% and 1.6% compared with LICN-C2E with the training on FG-COCO and the FG-sCOCO training dataset, respectively. The results indicate that it is important to train the model in an ascending order of example complexity in improving the object detection performance.



(a)

(b)

(c)

Figure 3.9: The comparison of LICNs and Faster R-CNN, where the former is trained on FG-COCO and the latter on MS COCO. As introduced in Subsection 4.1.1, both datasets consist of the same images. The testing results are reported for the Visual Genome test dataset. (a) shows the results for the fine-grained categories whose labels are not appearing in MS COCO. (b) shows results for the coarse categories that have no corresponding fine-grained labels in the semantic map, i.e., $ratio = 1$. (c) shows the results for the coarse categories where different proportions of object samples are re-labeled by new fine-grained labels with semantic mapping, i.e., $ratio \in (0,1)$.

Visual Genome

In this subsection, we evaluate the performance of our approach on the Visual Genome test dataset, where LICNs and Faster R-CNN are trained on FG-COCO and MS COCO, respectively.

Fig. 3.9 reports the comparison results of different methods on the test dataset in three cases: a) Fig. 3.9(a) shows the results for the fine-grained categories that do not appear in MS COCO and do come from the semantic mapping; b) Fig. 3.9(b) is for the coarse categories that have no corresponding fine-grained labels in the semantic map, i.e., the information for these categories in the training set MS COCO is the same as that in FG-COCO, and $ratio = 1$; and c) Fig. 3.9(c) is for the coarse categories, where different proportions of object samples with these category labels in training set MS COCO are re-labeled by new fine-grained labels with semantic mapping in building FG-COCO, i.e., $ratio \in (0,1)$. In Fig. 3.9(a), we see that the proposed LICN-E2C performs better than LICN-C2E for some fine-grained categories, such as “guy", “fighter", “subway", “branch", “skateboarder", “wave", “tennis" , “bear" and “television". The mean AP of the LICN-E2C model over all categories in Fig. 3.9(a) is 10.73, which achieves 0.61 mAP improvement over LICN-C2E (10.12). However, Faster R-CNN baseline training on the coarse categories cannot detect the new fine-grained categories. So its $AP = 0$ for these categories (the gray bars are not visible for that reason). From Fig. 3.9(b), we can see that for those coarse categories that have not been re-annotated with fine-grained category labels, there are no obvious differences between these three models. As shown in Fig. 3.9(c), for each coarse category in which a proportion of object samples have been re-annotated with fine-grained labels from captions by semantic mapping, Faster R-CNN has a better performance because it’s training dataset, i.e., MS COCO, consists of less categories and more examples in each of these categories than the training set of LICN. With the ratio decreases, Faster R-CNN tends to increase the improvement because the number of objects re-assigned from the coarse categories to the fine-grained categories increases continuously. But our LICN model also achieves a performance close to Faster R-CNN for the categories with the ratio close to 1.

Actually, the problem of fine-grained label learning with the weak supervision of captions resolved by our approach is more challenging than the object detection and label inference resolved by the compared method, i.e., Faster R-CNN. The main reason is that the category space coming from captions in our problem (e.g., 150-dim as shown in Table 3.1) is much larger and consists of much more labeling noise than that in the latter problem.

Example Illustrations

Fig. 3.10 shows 5 fine-grained categories, namely “man”, “woman”, “plane”, “bike" and “bat”, predicted in object detection with our approach. For each category, we show 4 representative images with top confidence of category prediction. The illustration shows that our LICN approach can truly predict fine-grain category label with the weak supervision of captions.

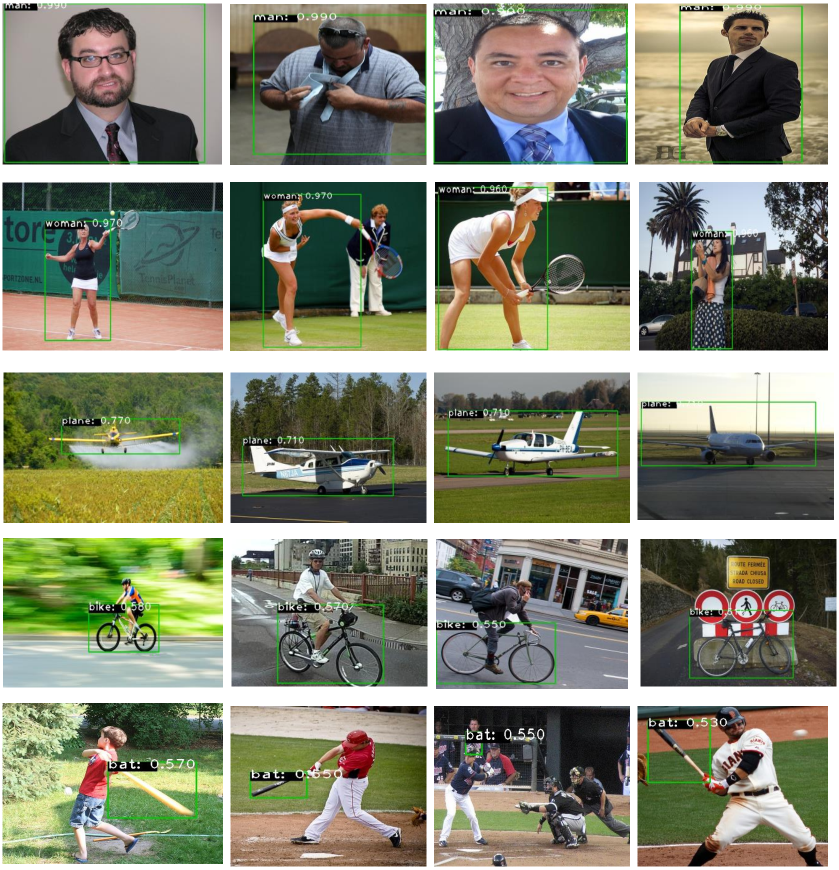


Figure 3.10: Example illustration of 5 fine-grained categories: “man”, “woman”, “bike”, “plane” and “bat”, which correspond to the coarse categories: “person”, “person”, “airplane”, “bicycle” and “baseball bat”, respectively. The values next to bounding boxes indicate the confidences of fine-grained label prediction.

3.4. Conclusion and Future Work

This chapter seeks to answer the question of how to learn the fine-grained object labels in object detection with the help of auxiliary information attached to images. In this chapter, we propose a novel approach called label inference curriculum network (LICN) to the problem of fine-grained object label learning with the weak supervision of captions. First, we construct a semantic map that builds a correspondence between the coarse category labels provided by public datasets and the fine-grained category labels extracted from captions based on the combination of embedding techniques and knowledge bases. Second, we present the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning. To evaluate the performance of fine-grained object label learning in different aspects, we construct multiple datasets based on widely-used public datasets. Experimental results implemented on the public datasets and our constructed datasets demonstrate the effectiveness of our approach and show that it is helpful to structure the training process in the order of easy samples to hard samples in the task under the framework of curriculum learning.