Fusing Scene Context to Improve Object Recognition

Leandro P. da Silva[1], Roger Granada[1], Juarez Monteiro[1], Duncan D. Ruiz[2]

Programa de Pós-Graduação em Ciência da Computação
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, 90619-900, Porto Alegre, RS, Brazil
[1] Email: {leandro.silva.007, roger.granada, juarez.santos}@acad.pucrs.br
[2] Email: duncan.ruiz@pucrs.br

Abstract. Computer vision is a branch of science that seeks to give computers the capability of seeing the world around them. Among its tasks, object recognition aims to classify objects and to identify where each object is in a given image. As objects tend to occur in particular environments, their contextual association can be useful for improving the object recognition task. To address the contextual awareness in object recognition tasks, our approach aims to use the context of the scenes in order to achieve higher quality in object recognition, by fusing context information with object detection features. Hence, we propose a novel architecture composed of two convolutional neural networks based on two well-known pre-trained nets: Places365-CNN and Faster R-CNN. Our two-streams architecture uses the concatenation of object features with scene context features in a late fusion approach. We performed experiments using public datasets (PASCAL VOC 2007, MS COCO and a subset of SUN09) analyzing the performance of our architecture with different threshold scores. Results show that our approach is able to raise in-context object scores, and reduces out-of-context objects scores.

Categories and Subject Descriptors: I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: convolutional neural networks, neural networks, object recognition

1. INTRODUCTION

Human brains are designed from birth to understand the visual world with ease, receiving information about objects, contexts and their associations. Observing a scene context, a human can easily recognize objects that belong to that context, while requiring a longer time to identify out-of-context objects. As described by [Biederman et al. 1982], the contextual information, and the relative size between objects and location, are important cues used by humans to detect objects. In fact, even when using low resolution images humans can distinguish whether an object on a table is a plate or an airplane, since it is unlikely to have an airplane on a table.

As pointed out by [Oliva and Torralba 2007], in the real world, objects co-occur with other objects and in particular environments, providing a rich source of contextual associations to be exploited. These relations between an object and its surroundings are classified by [Biederman et al. 1982] into five different classes: *interposition*, *support*, *probability*, *position* and *familiar size*. *Interposition* and *support* refer to physical space, *probability*, *position* and *familiar size* are defined as semantic relations since they require access to the referential meaning of the object being considered. The semantic relations include information about detailed interactions between objects in the scene and they are often used as *contextual features* [Galleguillos and Belongie 2010]. In this paper, our approach is closely related to the *probability* class [Biederman et al. 1982], which considers that objects tend to be found in some scenes but not others, although the other two classes are usually seen in training data:

Copyright©2018 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

position, *i.e.*, given that an object is probable in a scene, it is often found in some positions and not others, and *familiar size*, *i.e.*, objects have a limited set of size relations with other objects.

Although many research projects have been proposed to improve computer vision algorithms, they are still far from a human's ability to recognize objects despite their pose, illumination and occlusions. In this paper, we address the problem of object recognition by using the contextual features (features of the scene context) as indicative of the presence of the object. We argue the context will help the object recognizer to improve the classification of objects that are context dependent. Unlike in previous studies [Liu et al. 2016], our approach deals with raw images without the need of occlusion of the object in order to detect the context. We propose an approach that relies on a deep neural architecture that comprises two well known pre-trained Convolutional Neural Networks (CNNs): Places365-CNN [Zhou et al. 2017] and Faster R-CNN [Ren et al. 2015]. The two networks run in parallel as a two-stream configuration [Simonyan and Zisserman 2014a] and are fused to improve results in the object recognition task. We performed experiments using PASCAL VOC 2007 [Everingham et al. 2010], MS COCO [Lin et al. 2014] and a subset of the SUN09 [Jin et al. 2010] datasets, and examined the influence in the final classification when adding scene context to the objects. We concluded from these results that it is possible to improve the detection of objects by properly considering the context.

The rest of this paper is structured as follows. In Section 2, we introduce object and scene recognition and how it is usually performed nowadays. We describe the architecture we use to recognize objects using scene contexts in Section 3. Section 4 describes the datasets we used in this study as well as our experimental settings. In Section 5, we report the corresponding results achieved in our experiments, as well as analyze and discuss our findings on representative images of the dataset. Finally, the paper ends with our conclusions and future work possibilities in Section 6.

2. OBJECT AND SCENE RECOGNITION

Object recognition is widely used in computer vision from simple tasks such as inspection, registration and manipulation to complex tasks as autonomous robots operating in unstructured, real world environments. Traditional pipelines consist of extracting local feature descriptors, such as SIFT [Lowe 1999] or SURF [Bay et al. 2008], or a set of descriptors in a Bag-of-Visual-Words (BOV) [Perronnin 2008] which are then fed into a classifier such as Support Vector Machine (SVM). Recently, approaches based on deep learning [LeCun et al. 2015] where features are learned from data have advanced state-of-the-art results. Deep learning algorithms have surpassed by a significant margin the results achieved by hand-crafted features in object recognition tasks [Krizhevsky et al. 2012; Simonyan and Zisserman 2014b]. Many architectures that use these deep learning algorithms have been developed over the past few years, such as Faster R-CNN.

Faster R-CNN is a deep learning architecture designed to detect objects and has two modules: a Region Proposal Network (RPN) and a Fast R-CNN detector [Girshick 2015]. The RPN module contains a deep fully connected convolutional network that receives an image as input and outputs a set of region proposals (*i.e.*, object bounds), each with an objectness score for object detection. The region proposals are discovered by sliding over the last shared convolution feature map to determine whether the region is an object or not.

Scene recognition is a fundamental problem in computer vision and recently has been receiving increased attention [Guo et al. 2017; Wang et al. 2017]. As posed by [Wang et al. 2017], a scene provides rich semantic information about its overall structure, resulting in a meaningful context. On the other hand, the concept of scene is more subjective and complicated when compared to the concept of object and a consensus does not exist on how to assign a category to a scene. For example, a category *baseball field* may be labeled as *baseball stadium.* Thus, although a scene can be classified by the specific objects that are arranged in the layout, in this paper we use the concept of scene as the background environment where objects appear. In order to use the background environment, a CNN can be trained using only images that contain scenes. Hence, when testing with images containing objects, the attention will focus only on the environment.

3. ARCHITECTURE DESIGN

Early studies have shown the importance of context for giving cues about the recognition of certain objects. As described by [Galleguillos and Belongie 2010], context can be exploited in computer vision in three main forms: (1) the semantic context is described by the likelihood of an object present in the scene and not in other scenes, *e.g.*, a computer and a keyboard are more likely to appear in the same image than a computer and a car; (2) the likelihood of finding an object in a certain position in relation to other objects in the scene, *e.g.*, a keyboard is expected to be below the monitor; and (3) the size of the object in the context in relation to the other objects in the scene, *e.g.*, a keyboard is expected to be smaller than a person in the scene. In this study, we mainly exploit the semantic context as the likelihood of an object be in a scene and not in other scenes.

It is important to note that not all objects have a strong relationship with their context, *e.g.*, a person may appear in many contexts such as in a house or on the street, while a fire hydrant tends to appear always on a sidewalk. In this research study, the likelihood of an object being detected in a scene will tend to increase when the object has a strong relationship with the context, and decrease when the object is out of its context. In order to do that, we propose an approach that fuses two pre-trained deep learning architectures to separately extract information from the objects in the scene, and from the context. Hence, the likelihood of detecting an object may vary depending on the context in which it is inserted.

Figure 1 illustrates our approach, which receives raw images as input and pre-processes them to fit into the input of each deep learning architecture (Faster R-CNN and Places365-CNN). While Faster R-CNN extracts object features from the raw input image, Places365-CNN extracts features from the scene from a pre-processed version of the original image, since the scene recognition network requires a fixed size for the input image. Our late fusion approach is similar to previous studies [Simonyan and Zisserman 2014a; Li et al. 2017], where features from both streams are merged in the last fully connected layer. Unlike previous studies that used two-stream architecture to identify actions occurring in the scene [Simonyan and Zisserman 2014a], we intend to use two-stream architecture to detect features from objects and scene contexts.

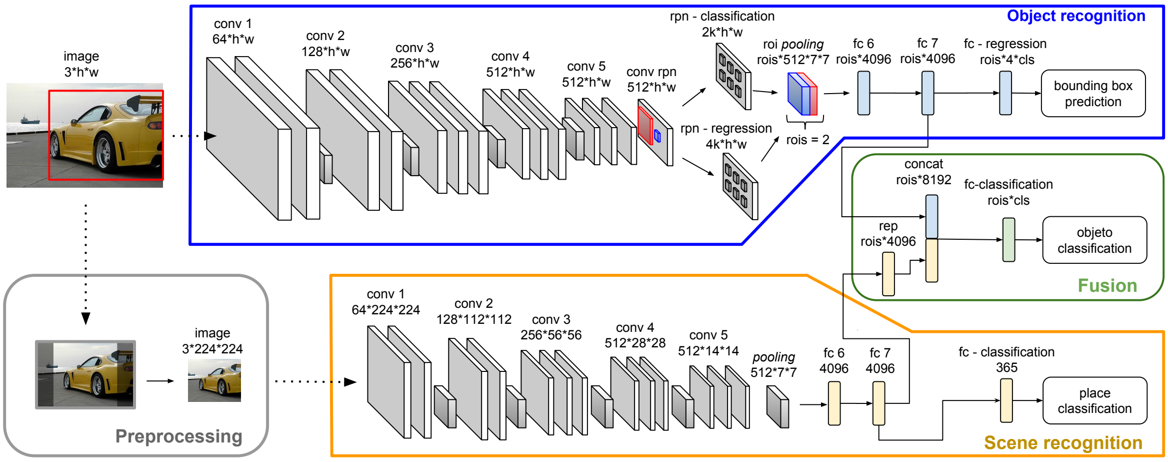


Fig. 1. Architecture from our approach, which consists of the concatenation of Faster R-CNN and Places365-CNN.

3.1 Pre-Processing Images

The pre-processing step intends to adapt raw images from the input to fit into each deep learning architecture (Faster R-CNN and Places365-CNN). While Faster R-CNN receives high resolution images with different sizes as input, Places365-CNN receives images with low resolution and fixed input sizes. In order to meet Places365-CNN requirements, we first transform the input image into a square by cropping the largest dimension and maintaining the lowest dimension, *e.g.*, an input image of 460×640 is cropped to 460×460. The cropped image is then resized to 224×224, which corresponds to the input image of Places365-CNN.

3.2 Object Recognition

Transfer learning aims to transfer knowledge between related *source* and *target* domains [Pan and Yang 2010]. In practice, few people train an entire CNN from scratch since it usually needs a lot of training data. For example, learning more than 60 million parameters of AlexNet [Krizhevsky et al. 2012] from just a few thousand training images is problematic. Instead, it is common to use a network that was pre-trained on a large-scale fully-labelled dataset (*e.g.*, ImageNet [Deng et al. 2009], PASCAL VOC 2007 [Everingham et al. 2010], Places365 [Zhou et al. 2017], *etc.*) as a feature extractor or as an initialization in a fine-tuning strategy. The primary goal of transfer learning is that the internal layers of a CNN can act as a feature extractor of the image representation, which can be pre-trained on one large dataset and re-used on other smaller datasets.

Our object recognition stream is composed by a network that extracts regions of interest (RoI) from each input image. Although this stream allows us to use the most common object recognition networks (*e.g.*, YOLO [Redmon et al. 2016], YOLO9000 [Redmon and Farhadi 2017], SSD [Liu et al. 2016], *etc.*), we use the Faster-RCNN [Ren et al. 2015] network, since it has pre-trained models and good results. As reported by [Ren et al. 2015], Faster-RCNN is available in two flavours: using ZFNet [Zeiler and Fergus 2014] or VGG-16 [Simonyan and Zisserman 2014b] as Convolutional Neural Network. In this work, we use the Faster-RCNN containing the VGG-16 CNN for the object recognition stream of our architecture. The expected outputs are: (a) a set of regions of interest (we set to a maximum of 300), each containing a list of 4096 features of each RoI (*fc7*); and (b) the prediction of bounding boxes (position in x, in y, height and width) in the original image.

3.3 Scene Recognition

The scene recognition stream intends to extract features and classify the scene (background) of the input image. In the output of the stream is expected a list of features corresponding to the scene and its label. As in the stream of object recognition, the stream of scene recognition allows us to use pre-trained models from different datasets (*e.g.*, SUM [Xiao et al. 2010], Places205 [Zhou et al. 2014], MIT Indoor 67 [Quattoni and Torralba 2009], *etc.*). In this work, we also apply transfer learning, using a network that was pre-trained in the Places365 dataset. [Zhou et al. 2017] make available pre-trained models using three popular CNNs: AlexNet , GoogLeNet [Szegedy et al. 2015], and VGG-16. We employed the VGG-16 version of the pre-trained network (hereafter called Places365-CNN) since we intended to keep the output of the network with the same dimension as the output of the object recognition stream, favoring the concatenation of both streams. Thus, the scene recognition stream outputs 4096 features from the last fully connected layer (*fc7*) corresponding to the activations generated by the identification of the image context, and a label is assigned.

3.4 Fusion Network

Our fusion method seeks to merge the information from the scene recognition stream with the information from the object recognition stream. The fusion is inspired by the approach proposed in [Li et al.

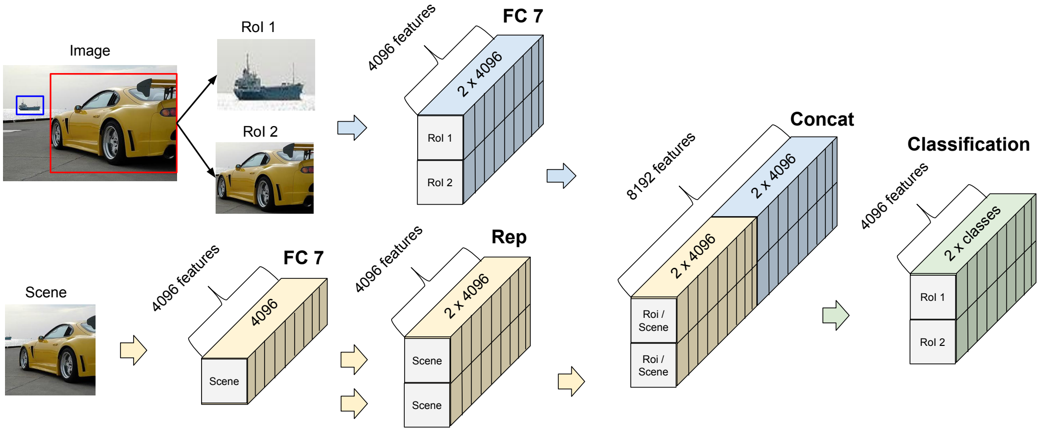


Fig. 2. Concatenation of object recognition and scene recognition streams.

2017], which concatenates two fully connected layers prior to classification. The input of the network receives a set of RoIs, each containing 4096 features from the object stream, and a list containing 4096 features of the object context from the scene recognition stream. As the input from both streams have different shapes, *i.e.*, (number of RoIs, 4096) from object recognition stream and (1, 4096) from scene recognition stream, we have to multiply the number of features from scene recognition stream by the number of RoIs from the object recognition stream. For example, Figure 2 illustrates the identification of 2 RoIs by the object recognition stream and the replication layer (layer "rep") duplicating the list of features from the scene recognition stream to perform the correct concatenation (layer "concat") prior to object classification. As explained by [Li et al. 2017], the information of the scene does not contribute towards predicting the coordinates of bounding boxes. Thus, we use the prediction of the coordinates of the bounding boxes directly from the output of the object recognition stream.

4. EXPERIMENTS

4.1 Datasets

PASCAL VOC 2007[1] [Everingham et al. 2010] is a dataset that consists of images representing realistic scenes, where each image has at least one object. This dataset was published through the PASCAL Visual Object Classes Challenge 2007 and contains 20 different object classes, having a total of 9,963 images with 24,640 annotated objects, being grouped as shown in Table I. The dataset has 2,501 images in training set, 2,510 images in validation set, and 4,952 images in testing set. It is important to mention that as the main objective of PASCAL VOC is object recognition, most images contain little information about context, with objects filling almost the whole image.

Category	**Classes**
Person	person
Animal	bird, cat, cow, dog, horse, sheep
Vehicles	aeroplane, bicycle, boat, bus, car, motorbike, train
Indoor	bottle, chair, dining table, potted plant, sofa, tv/monitor

Table I. Category and classes of PASCAL VOC 2007 dataset.

[1] `http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html`

Category	**Classes**
Animal	bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe
Appliance	microwave, oven, toaster, sink, refrigerator
Electronics	tv monitor, laptop, mouse, remote, keyboard, cell phone
Food	banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake
Furniture	chair, sofa, potted plant, bed, dining table, toilet
Indoor objects	book, clock, vase, scissors, teddy bear, hair drier, toothbrush
Kitchenware	bottle, wine glass, cup, fork, knife, spoon, bowl
Outdoor Obj.	traffic light, fire hydrant, stop sign, parking meter, bench
Person & Acc.	person, backpack, umbrella, handbag, tie, suitcase
Sport	frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket
Vehicle	bicycle, car, motorbike, aeroplane, bus, train, truck, boat

Table II. Category and classes of MS COCO dataset.

MS COCO[2] [Lin et al. 2014] is a dataset that contains images of complex everyday scenes with common objects in their natural context. The dataset addresses core research problems in scene understanding, such as detecting non-iconic views, *i.e.*, objects in background, partially occluded and amid clutter. Objects' spatial locations are annotated using bounding boxes and pixel-level segmentation. The dataset contains 80 classes to define objects grouped as shown in Table II, with a total of 123,287 images, which are divided into 82,783 for training models, and 40,504 for validation and testing models. Compared to PASCAL VOC 2007 [Everingham et al. 2010], the MS COCO dataset contains considerably more objects per scene, and smaller labeled objects.

SUN09[3] [Jin et al. 2010] is a dataset for general object recognition. In this study, we used a small subset of SUN09 containing images previously selected as *out of context*. Although the subset is very small, it can give us an idea about the performance of our approach. The selected subset contains 42 images with objects in the context, and 42 images with objects that are out of their context, generating a subset of the SUN09 dataset containing 84 images.

4.2 Network Settings

In order to perform our experiments we developed our architecture containing two VGG-16 using the Caffe[4] framework. Our architecture allowed us to load two pre-trained models running in parallel and a fully connected layer that performs the fusion of both networks. In the scene recognition stream, we use the Places365-CNN pre-trained model, which is freely available in MIT CSAIL Computer Vision Website[5]. Places365-CNN is a pre-trained model using the 1,803,460 images of the Places365-Standard dataset. It contains a list of the categories of environments encountered in the world, such as bedroom, train station platform, conference center, veterinarians' office *etc.* For model fitting, we froze all layers of Places365-CNN, using the network as a feature extractor of the input images.

In object recognition stream, we performed two experiments: (a) loading a pre-trained model in PASCAL VOC 2007 [Everingham et al. 2010] dataset and (b) training from scratch a model using the MS COCO [Lin et al. 2014] dataset. The pre-trained version of the VGG-16 network contains weights learned using the PASCAL VOC 2007 dataset and was downloaded from the Faster R-CNN Website[6]. In this pre-trained network, we froze all layers and used the network as a feature extractor from the input images.

[2] http://mscoco.org/dataset/

[3] http://people.csail.mit.edu/myungjin/HContext.html

[4] http://caffe.berkeleyvision.org/

[5] https://github.com/CSAILVision/places365

[6] https://github.com/rbgirshick/py-faster-rcnn

The network trained from scratch in the MS COCO dataset contains the same parameters as described by [Ren et al. 2015] and consists of initializing weights from a zero-mean Gaussian distribution with standard deviation 0.01. All images have their pixels subtracted by the mean pixel values per channel of all training images. We used a learning rate of 10^{-3}, dropping it to 10^{-4} after 331.132 iterations (4 epoch), and a momentum of 0.9 with a weight decay of 5×10^{-4}. All convolutions used rectified linear activation units (ReLU). To minimize the chances of overfitting, we applied dropout on the fully-connected layers with a probability of 50%. We fixed the number of RoI generated by Faster R-CNN up to 300, since this value achieved the best results by [Ren et al. 2015]. Each iteration forwarded a single image and the network training stopped after 496.698 iterations (6 epoch).

4.3 Fusion Settings

As the object recognition stream generates a maximum of 300 RoIs for each image in the output, we had to replicate the number of features generated by the scene stream. The *fc-classification* of the network was trained for 496.698 iterations (6 epoch) when using the MS COCO dataset, and for 70.154 iterations (6 epoch) when using the pre-trained version of the PASCAL VOC dataset. The fully connected layer was trained using a learning rate of 10^{-3}, reducing it to 10^{-4} after 75% of the training in both datasets, and a dropout of 50% for both networks.

5. EVALUATION AND RESULTS

In order to evaluate our approach, we compared the results using the testing set of the MS COCO [Lin et al. 2014] and of the PASCAL VOC 2007 [Everingham et al. 2010] datasets. We verified the mean Average Precision (mAP) using Intersection over Union (IoU) scores. As the perfect match of a predicted bounding box with its ground truth is very unlikely, IoU is an evaluation metric that takes into account the overlap between predicted bounding box and the ground truth, as well as their area of union. Thus, we considered not only the perfect match between the predicted bounding box and the ground truth as correct, but also a threshold of the overlap between the two, as measured by IoU. We varied the threshold of IoU from 0 to 100%, increasing by 10% each step for each tested network, as illustrated in Figure 3.

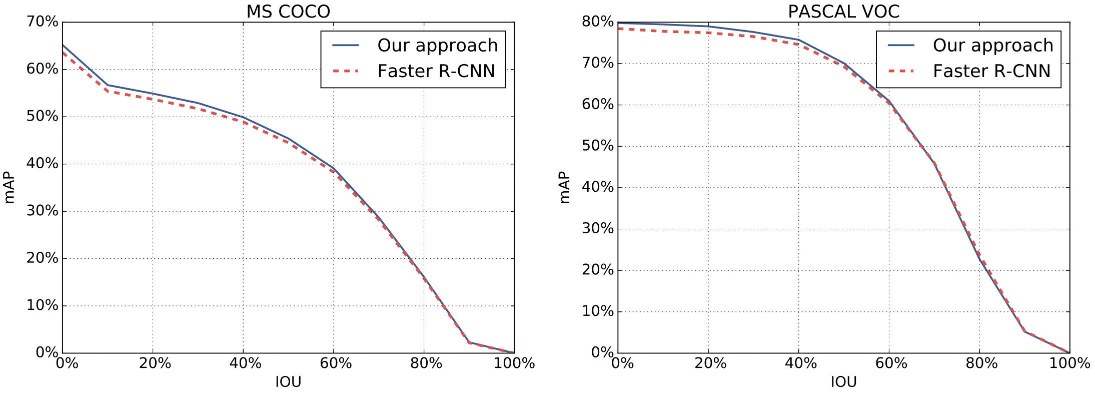


Fig. 3. Mean Average Precision for PASCAL VOC 2007 and MS COCO datasets with different values of Intersection over Union

As illustrated in Figure 3, we can compare the addition of scene features to object features with the Faster R-CNN approach that does not use the scene. When adding scene recognition to our approach we achieved better results than running Faster R-CNN alone. Our approach achieved better results for low thresholds, equating to Faster R-CNN closer to 100% of IoU. In fact, in both datasets

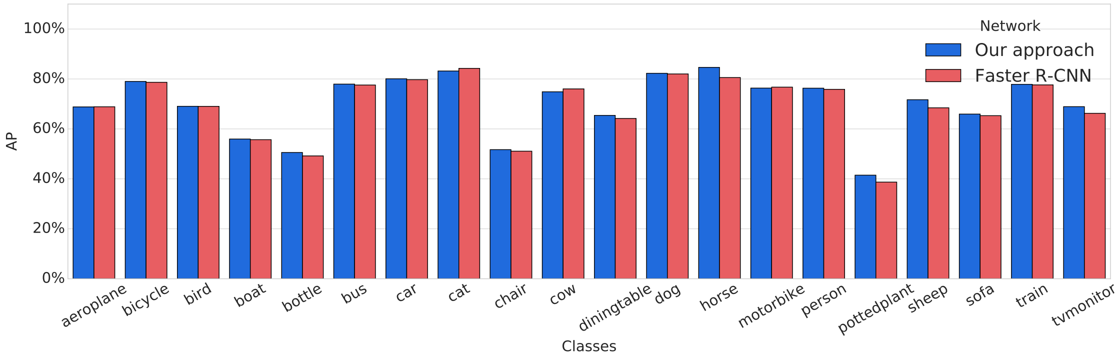


Fig. 4. Average precision per class using an intersection over union (IoU) fixed in 50% in the PASCAL VOC 2007 dataset.

our approach obtained better results up to IoU=60%. Although our approach got better results for smaller IoU values, this threshold was set to 50% as usual. Hence, we performed an analysis in tests using both datasets with the threshold set to 50% of IoU.

5.1 PASCAL VOC 2007 Results

Setting the IoU to 50%, we can compare the average precision achieved by each approach for each class using images from the PASCAL VOC 2007 dataset, as illustrated in Figure 4. We can see that our approach achieved higher values of precision for 16 out of 20 classes, having lower results only for classes *cow* (−1.16%pts), *cat* (−1.06%pts), *motorbike* (−0.38%pts) and *aeroplane* (−0.05%pts). The classes with the highest differences were *horse* (4.06%pts), *sheep* (3.23%pts), and *potted-plant* (2.78%pts) in favor of our approach. We observed that the context does not increase the score of all classes because some objects fill whole images, making it difficult to recognize the scene.

Since the dataset can be grouped by categories, we evaluated the results achieved by our approach and the Faster R-CNN using the categories of both streams. Thus, we can see in which places our approach is improving the object recognition, and which objects are being improved by the addition of the scene. In order to do so, we selected only the objects containing the highest IoU in relation to the ground truth bounding boxes for both approaches. As show in Table III, we evaluated: the percentage of correct assigns that both approaches achieved (Both Correct); the number of incorrect assigns of both approaches (Both Incorrect); the divergences in favor of Faster R-CNN, *i.e.*, when Faster R-CNN assigned the correct label and our approach missed it (Pro divergence Faster R-CNN); and the divergences in favor of our approach *i.e.*, when our approach assigned the correct label and Faster R-CNN missed it (Pro divergence Our approach).

Observing Table III, we can see that our approach achieved better results than the Faster R-CNN when compared the Pro divergences, *i.e.*, our approach can detect the object correctly and the Faster R-CNN cannot detect it. In fact, the Faster R-CNN obtained better results only for objects that belong to the *indoor* category (*i.e.*, objects that tend to occur in indoor environments such as tv/monitor, sofa, dinning table *etc.*), with the *outdoor, man-made* and *outdoor, natural* categories to the scene. Hence, our approach was able to learn that objects which are associated with indoor category should appear in indoor environments, such as dinning rooms and bedrooms, and not in outdoor environments, such as decks and sidewalks.

5.2 MS COCO Results

Having the IoU set to 50%, we compared the average precision (AP) achieved by each approach for each class using images from the MS COCO dataset, as illustrated in Figure 5. Due to page

Category		Pro divergence		Both	
Objects	Scenes	Our approach	Faster R-CNN	Correct	Incorrect
animal	*indoor*	**14.44%**	11.02%	48.19%	26.34%
	outdoor, man-made	**12.51%**	10.46%	51.60%	25.42%
	outdoor, natural	**14.10%**	10.75%	45.06%	30.09%
indoor	*indoor*	**12.30%**	10.00%	54.44%	23.60%
	outdoor, man-made	8.96%	**10.70%**	61.50%	18.85%
	outdoor, natural	9.75%	**14.98%**	51.99%	23.28%
person	*indoor*	**17.32%**	4.58%	31.41%	49.69%
	outdoor, man-made	**13.28%**	4.95%	41.17%	40.60%
	outdoor, natural	**14.46%**	6.24%	50.98%	28.32%
vehicles	*indoor*	**5.34%**	4.90%	81.47%	8.29%
	outdoor, man-made	**5.79%**	4.98%	79.06%	10.17%
	outdoor, natural	**6.67%**	4.97%	77.44%	10.92%

Table III. Number of correct, incorrect and divergence between the Faster R-CNN and our model, grouped by category of objects to the PASCAL VOC 2007 and category of places to the Places365-Standard dataset

constraints, we split the 80 object classes into 4 plots with 20 classes each. In Figure 5, we can see that our approach achieved higher values of average precision for all but the *train* class, where Faster R-CNN achieved an AP 0, 1%pts higher than our approach. The best results were achieved by the classes *remote* (5.7%pts), *hair drier* (2.5%pts) and *bed* (2.5%pts). By analyzing the images, we observe that the objects *remote* and *hair drier* are small objects occupying a minor part of the image, with a larger space for scene context detection. The larger the space of the scene context, the larger its influence on the final classification. On the other hand, objects that fill in the majority of the image, leaving a small scene context to detect, tend to have the same average precision in both approaches. When comparing the results of our approach between the two datasets (PASCAL VOC 2007 and MS COCO), our approach achieved better results in the latter, since its images contained more scene context than the images of the former.

Grouping the MS COCO classes and Places365-Standard classes into categories for both streams, we evaluated the results achieved by our approach and the Faster R-CNN. As performed in the PASCAL VOC 2007 dataset, we selected only the objects containing the highest IoU in relation to the ground truth bounding boxes for both approaches. Table IV shows the percentage of correct assigns that both approaches achieved (Both Correct), the percentage of incorrect assigns of both approaches (Both Incorrect), the divergence in favor of Faster R-CNN (Pro divergence Faster R-CNN) and the divergence in favor of our approach (Pro divergence Our approach).

Observing Table IV, we can see that in most categories both approaches achieved good results, *i.e.*, assigned the correct label to the object. When comparing the pro divergence between approaches, our approach obtained better results for most categories. As occurred in the PASCAL VOC 2007 dataset, our approach wrongly classified objects that were out of context, *e.g.*, *kitchenware* in *outdoor* scenes, *indoor* objects in *outdoor* scenes, and *furniture* in *outdoor* scenes. As our approach expects that objects should always appear in their context, it will tend to reduce the probability of identification of an object when it is out of its context. For example, our approach learns that *kitchenware* category is related to *indoor* scenes, probably because in most of the images kitchenware appears in a kitchen. When *kitchenware* appears out of its contexts, such as outdoor scenes or nature, our approach reduces its score in classification.

An unexpected result is presented by the category *furniture* in *indoor* scenes, since furniture does tend to occur in indoor places such as house, bedroom, office *etc.*. This is the only case that an object in its context did not improve the results when using our approach. By analyzing the images

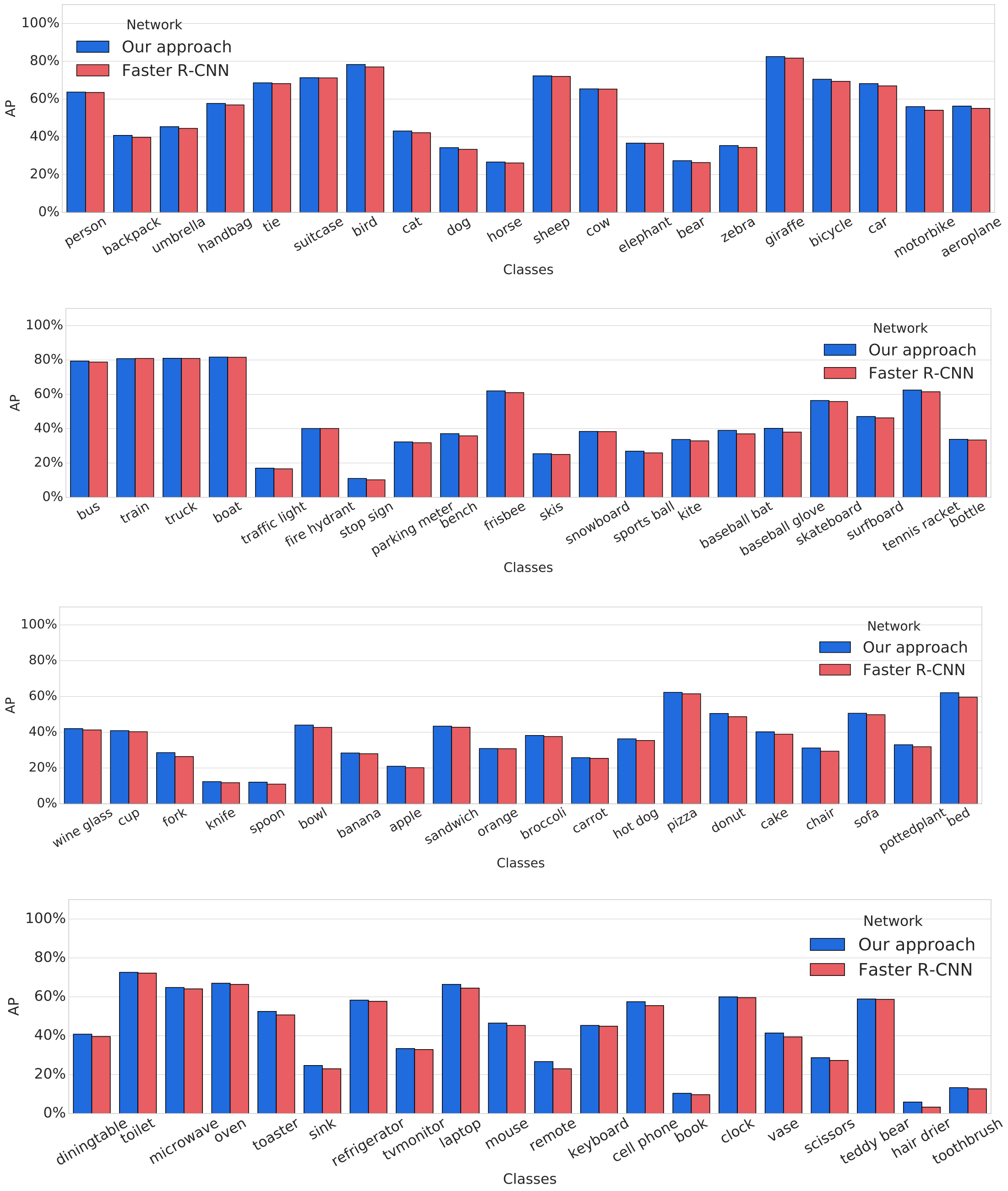


Fig. 5. Average precision (AP) per class using an intersection over union (IoU) fixed in 50% in the MS COCO dataset.

of the dataset that are described in the categories *furniture* and *indoor* environment, we figured out that some images also contain objects from outdoor environments (*e.g.*, potted plant), while others contain images that are close-up (*e.g.*, potted plant and dinning table), which may induce bias in the network. By studying the statistics of the dataset, we see that Places365 classifies many images as *stadium/baseball*, *stadium/soccer*, *athletic field/outdoor* and *restaurant patio*. Observing such images, we see that many of them are outdoor environments, but are classified in the dataset as *indoor*.

Category		Pro divergence		Both	
Objects	Scenes	Our approach	Faster R-CNN	Correct	Incorrect
Animal	*indoor*	**6.30%**	3.12%	55.74%	34.84%
	outdoor, man-made	**5.79%**	2.79%	62.68%	28.49%
	outdoor, natural	**4.96%**	3.01%	69.68%	22.35%
Appliance	*indoor*	**4.42%**	1.94%	63.06%	30.58%
	outdoor, man-made	**4.69%**	3.91%	41.41%	50.00%
	outdoor, natural	0.00%	0.00%	26.67%	73.33%
Electronics	*indoor*	**3.92%**	2.69%	58.43%	34.95%
	outdoor, man-made	**4.44%**	2.44%	36.29%	56.83%
	outdoor, natural	4.29%	4.29%	31.90%	59.52%
Food	*indoor*	**4.88%**	2.73%	58.78%	33.60%
	outdoor, man-made	**4.13%**	1.65%	64.37%	29.84%
	outdoor, natural	**3.86%**	1.75%	69.16%	25.24%
Furniture	*indoor*	3.95%	**4.02%**	61.55%	30.49%
	outdoor, man-made	3.01%	**4.52%**	50.29%	42.19%
	outdoor, natural	3.14%	**4.93%**	52.80%	39.13%
Indoor	*indoor*	**4.05%**	1.88%	63.85%	30.22%
	outdoor, man-made	**1.92%**	1.55%	65.05%	31.48%
	outdoor, natural	2.54%	2.54%	54.96%	39.95%
Kitchenware	*indoor*	**4.03%**	3.43%	50.36%	42.18%
	outdoor, man-made	**4.22%**	3.00%	39.59%	53.19%
	outdoor, natural	1.85%	**4.25%**	34.94%	58.96%
Outdoor	*indoor*	**5.10%**	1.80%	45.56%	47.54%
	outdoor, man-made	**2.98%**	1.63%	51.78%	43.61%
	outdoor, natural	**5.30%**	2.59%	55.83%	36.29%
Person & Accessory	*indoor*	**2.10%**	1.39%	79.81%	16.69%
	outdoor, man-made	**1.63%**	1.07%	81.13%	16.17%
	outdoor, natural	**1.69%**	1.43%	82.23%	14.65%
Sport	*indoor*	**3.24%**	2.46%	49.13%	45.17%
	outdoor, man-made	**3.62%**	1.54%	56.13%	38.72%
	outdoor, natural	**3.57%**	1.59%	63.90%	30.94%
Vehicle	*indoor*	**4.44%**	3.67%	63.14%	28.75%
	outdoor, man-made	**4.59%**	3.07%	64.90%	27.44%
	outdoor, natural	**4.35%**	3.81%	58.43%	33.42%

Table IV. Number of correct, incorrect and divergence between the Faster R-CNN and our model, grouped by category of objects to the MS COCO and category of places to the Places365-Standard dataset

5.3 Balanced Dataset

As the experiments may be biased towards images occurring in context, we created a subset of the SUN09[7] [Jin et al. 2010] dataset. To the best of our knowledge, this is the only dataset that contains images with objects out of their context. We tested the addition of context and their impact in the results on this dataset. In order to do so, we compared the results achieved by the network containing only the Faster R-CNN and the network with the addition of the context. Table V shows the results achieved for each network in terms of Precision ($\mathcal{P}$), Recall ($\mathcal{R}$), F-measure ($\mathcal{F}$) and mean Average Precision (mAP), where *In context* contains 42 images where the objects are in their context, and *Out of context* contains 42 images with objects that are out of context.

As Table V shows, our approach improved confidence score results for objects in their context, while out-of-context objects have lower confidence scores assigned. As our approach will decrease the confidence score for objects that are out of context, it will tend to decrease the number of identified objects in the image, keeping only the ones with high probabilities. Hence, the precision score will tend to increase while the recall will tend to decrease.

[7] http://people.csail.mit.edu/myungjin/HContext.html

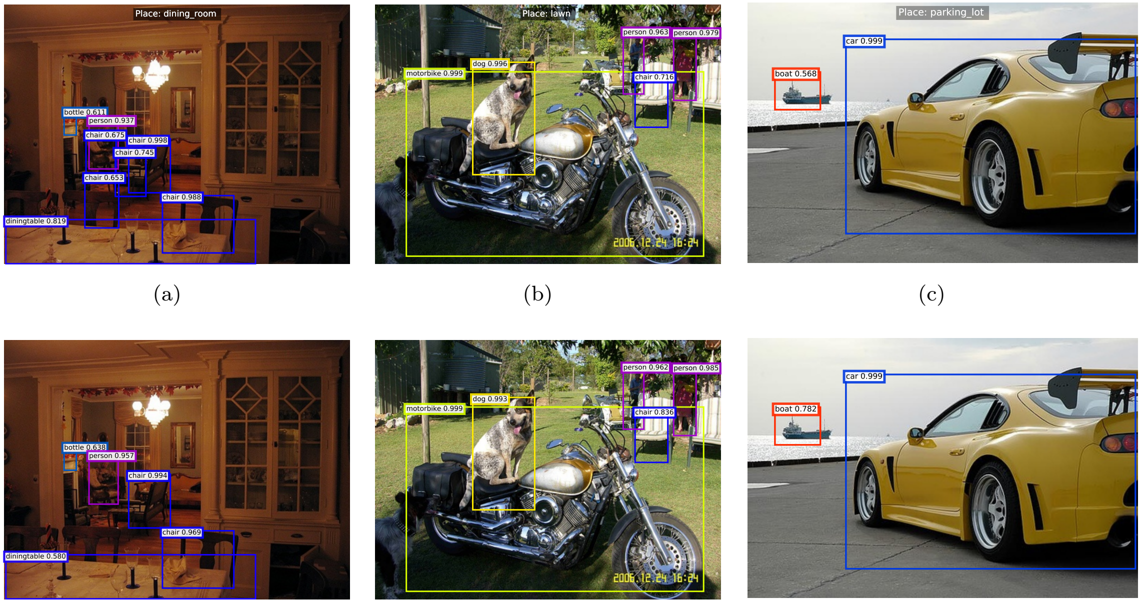
Network	**In context**				**Out of context**			
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	mAP	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	mAP
Faster R-CNN	18.27%	17.56%	17.46%	33.13%	**48.98%**	72.34%	**55.33%**	**39.86%**
Our approach	**49.50%**	**72.59%**	**55.62%**	**36.16%**	47.35%	**72.54%**	54.60%	38.29%

Table V. Results for Faster R-CNN and our approach using *in context* and *out of context* images with $IoU = 50\%$.

5.4 Image Analysis

Although in some cases it seems worse to use the context to identify objects, our objective in this paper is to increase the confidence in objects that are context related and decrease the confidence when objects are not related to their context. For instance, suppose a classifier without using the context predicts an arm chair in the image with a confidence of 50%. When using the scene context, our goal was to increase this confidence if the arm chair is located in a living room or decrease this confidence if the arm chair was located in a scene where it should not appear frequently, such as a forest. In order to perform such an analysis, we selected some images from the PASCAL VOC 2007 and the MS COCO datasets and analyzed the probability scores of each object in the scene when classifying them with and without context.

Figure 6 illustrates the object recognition performed by our approach and by Faster R-CNN using the PASCAL VOC 2007 dataset. Images at the top of the Figure (6(a), 6(b) and 6(c)) contain labels generated by our approach that uses the context of the scene, and images at the bottom of the Figure (6(d), 6(e) and 6(f)) contain labels generated by Faster R-CNN that do not use the context to predict objects.



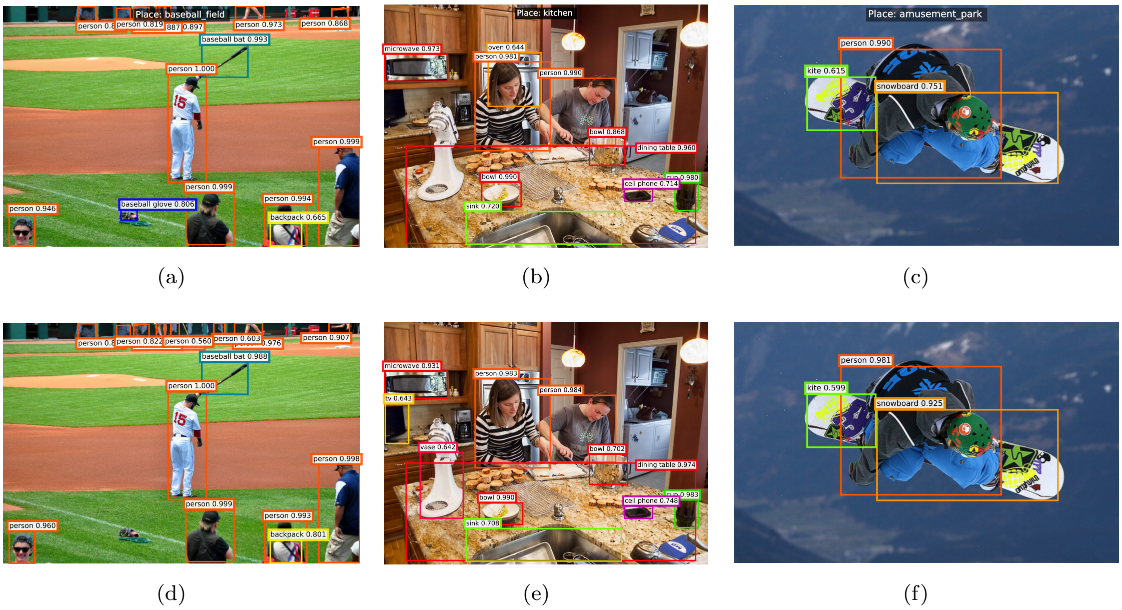
(a) (b) (c)

(d) (e) (f)

Fig. 6. Example of object recognition using PASCAL VOC 2007 dataset with our architecture which uses scene context (a), (b) and (c), and with Faster R-CNN without scene context (d), (e) and (f).

As we can see when comparing labels generated for the first image (Figures 6(a) and 6(d)), our approach increased the probability of the object being a dining table from 58% to ≈82%, *i.e.*, ≈24%pts higher than the score predicted by Faster R-CNN. Our approach also identified more chairs in the scene (identified as *dining room*) that were not predicted by Faster R-CNN. This image illustrates the power of using the context to improve the classification of objects in an image. Analyzing the second images (Figures 6(b) and 6(e)), our approach decreased the score of the chair since the chair is associated with the category of indoor objects. Although the correct object is a chair, our approach learned that it does not have a strong relation with outside places, such as the one classified as *lawn*. Hence, using our approach, the chair score is reduced in 12%pts in relation to the classification realized by Faster R-CNN. The third images (Figures 6(c) and 6(f)) contains an ambiguous context, since the scene could be classified as *harbor* or *parking lot*. Identifying the scene as *parking lot*, the probability of *boat* decreased from ≈78% to ≈57%, since the network learned that it is hard for a boat to appear in a parking lot. Likewise, if the scene were classified as *harbor*, then probably the score of the boat would increase and the score of the car would decrease, since it is unlikely for a car to be in a harbor.

We also performed the analysis of images comparing the results achieved by our approach and Faster R-CNN using the MS COCO dataset. Figure 7 illustrates three images classified by our approach (7(a), 7(b) and 7(c)) and by Faster R-CNN (7(d), 7(e) and 7(f)). In the first image (Figures 7(a) and 7(d)), our approach detected the *baseball glove* on the scene classified as *baseball field* with a probability of ≈81%. Indeed, our approach increased the probability of the *baseball glove* by 56.31%pts. Since the probability found by the Faster R-CNN is under 50%, the object's bounding box does not appear.



(a) (b) (c)

(d) (e) (f)

Fig. 7. Example of object recognition using the MS COCO dataset with our architecture, which uses scene context (a), (b) and (c), and with the Faster R-CNN without scene context (d), (e) and (f).

In the second images (Figures 7(b) and 7(e)), our approach did not identify the *television* on the left side of the image. In fact, it reduced the score of *television* by 40.6%pts (from 64.3% to 23.7%) in relation to the Faster R-CNN. As the score in our approach is below 50%, the bounding box does not appear in the image. Finally, the third images (Figures 7(c) and 7(f)) show the classification of a man on a snowboard. As the scene classification identifies the location as *amusement park*, it reduced the classification score of *snowboard* by 17.4%pts (from 92.5% to 75.1%). On the other hand, it increased the score of the wrongly classified *kite* from 59.9% to 61.5%.

5.5 Comparing with the State of the Art

As our approach uses freely available datasets for testing, it can be compared with other approaches that use the same datasets. Table VI shows a comparison of our results with the ones achieved by recent approaches using the MS COCO and the PASCAL VOC 2007 datasets. As we can see, our approach is in the middle, obtaining lower results when compared with *CNN + Fully Connected CRF* [Chu and Cai 2017], AC-CNN [Li et al. 2017] and ION [Bell et al. 2016]. On the other hand, our approach outperforms Faster R-CNN [Ren et al. 2015]. Although our approach does not achieve the top results, it is able to raise in-context object scores and reduces out-of-context object scores.

The Inside-Outside Net (ION) network [Bell et al. 2016] uses object proposal detectors and dynamic pooling to evaluate the different RoI candidates in an image. A spatial recurrent neural network computes the features from the contextual information outside the region of interest. Object and contextual information are concatenated and pass through several fully connected layers for classification. Experiments using the PASCAL VOC 2007 achieve 75.6% of mAP and those using the MS COCO achieve 55.7% of mAP.

[Chu and Cai 2017] propose the extraction of context information using a CNN and a fully connected conditional random field (CRF) on object proposals. Using a Faster R-CNN, they added a fully connected conditional random field (CRF) after the region proposals generation. The CRF uses the region proposals with context information to perform the inference. The experiments using the PASCAL VOC 2007 achieve 73.5% of mAP.

Attention to Context CNN (AC-CNN) [Li et al. 2017] incorporates global and local contextual information into a CNN using a global contextualized (AGC) subnetwork and a multi-scale local contextualized (MLC) subnetwork. AGC is responsible for highlighting useful global contextual locations through multiple stacked long short-term memory (LSTM) [Hochreiter and Schmidhuber 1997] layers, while MLC capture both inside and outside contextual information, capturing surrounding local context. Both global and local context are fused by fully connected layers prior to classification. Experiments were performed using the PASCAL VOC 2007 achieving 72.4% of mAP and the PASCAL VOC 2012 achieving 70.6% of mAP.

[Ouyang et al. 2015] develop the DeepID-Net that captures deformations of the object, *i.e.*, the rounded shape of an object, and its context information. Object deformations are identified by a deformation constrained pooling (def-pooling) layer that are merged with features from an existing deep model (ZFNet). Context is extracted by a network pretrained on 1000-class for image classification. The 1000-class contextual features are concatenated with the 200-class object detection scores to form a 1200 dimensional feature vector. The feature vector is then learned using SVM classifier and used for the inference of each object class. The experiments are realized using the ILSVRC2014 and the PASCAL VOC 2007 datasets achieving 50% and 64.1% of mAP, respectively.

Finally, [Liu et al. 2016] propose a two-stream network to address a fine-grained level classification problem using two-stream contextualized CNN. Their approach uses a network, named Content Net, to capture object features and a network, named Context Net, to capture background features. While the Content Net is fed with images extracted from bounding boxes of objects, the Context Net is fed with the whole image having the bounding box region filled with pixels from the equivalent position of the mean image calculated across the training set. The two-stream networks are merged in a fusion layer that automatically learns weights from content and context features and outputs the final classification. Although their model and experiments are the closest to ours, we cannot compare our results since their experiments were performed on different datasets.

Although our approach does not achieve the highest scores for both datasets, our approach is easy to set since we use two pre-trained networks. The independence of both streams also favors the replacement of any stream with new networks, *i.e.*, the Faster R-CNN in object stream could be replaced by the YOLO or YOLO9000 network. In the scene stream, the actual VGG-16 trained in

Approaches	MS COCO	PASCAL VOC 2007
ION [Bell et al. 2016]	55.7%	75.6%
CNN + Fully Connected CRF [Chu and Cai 2017]	-	73.5%
AC-CNN [Li et al. 2017]	-	72.4%
Our approach	45.4%	70.1%
Faster R-CNN [Ren et al. 2015]	41.5%	69.9%
DeepID-Net [Ouyang et al. 2015]	-	64.1%

Table VI. Comparison of mean Average Precision (mAP) of our approach and related work using IoU fixed at 50% in the MS COCO and the PASCAL VOC 2007 datasets.

Places365 could be replaced by another network such as ResNet trained in the same dataset, or by the same network trained in another dataset, such as Scenes15 [Lazebnik et al. 2016].

6. CONCLUSIONS AND FUTURE WORK

In this paper, we developed a novel architecture for object recognition based on a two-stream convolutional neural network architecture. The pipeline of the architecture includes a CNN focusing on recognizing objects, and another CNN on recognizing the scene context. Finally, we concatenated the object features with the context features in a late fusion scheme to predict the class of each object. We performed experiments using public datasets: MS COCO [Lin et al. 2014], PASCAL VOC 2007 [Everingham et al. 2010] and a subset of the SUN09 [Jin et al. 2010]. The results show that although our object recognition approach does not improve the overall performance when compared to the state-of-the-art recognizers, it performs very well when the object is situated in its proper context.

Preliminary analysis of our architecture demonstrates that the likelihood of an object tends to increase when the scene context is related to the object, and to decrease when the object is out of its context. Attention should be given when setting the IoU threshold to identify objects in the image. For example, we could define that in our system any object with a probability above 65% is considered a correct object. This probability may vary according to the application, *e.g.*, a car detecting obstacles should not have the same threshold as a person counting a crowd. Thus, we can improve the probability of an object occurring in an expected context by using the scene context, and thus more objects may be above the threshold as a correct object.

As future work, we aim to create subsets of the MS COCO and the PASCAL VOC 2007 datasets, focusing on discarding images with objects that are not dependent on the context. These new datasets may indicate whether our approach improves the accuracy of objects that are context-dependent.

ACKNOWLEDGMENT

The authors would like to thank CAPES/FAPERGS for partially funding this research. We gratefully acknowledge the NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. Speeded-up robust features (surf). *Computer Vision and Image Understanding* 110 (3): 346–359, 2008.

Bell, S., Lawrence Zitnick, C., Bala, K., and Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR'16*. pp. 2874–2883, 2016.

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology* 14 (2): 143–177, 1982.

Chu, W. and Cai, D. Deep feature based contextual model for object detection. *Neurocomputing* 275 (31): 1035–1042, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR'09*. pp. 248–255, 2009.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88 (2): 303–338, 2010.

Galleguillos, C. and Belongie, S. Context based object categorization: A critical survey. *Computer Vision and Image Understanding* 114 (6), 2010.

Girshick, R. Fast r-cnn. In *ICCV'15*. pp. 1440–1448, 2015.

Guo, S., Huang, W., Wang, L., and Qiao, Y. Locally supervised deep hybrid model for scene recognition. *IEEE Transactions on Image Processing* 26 (2): 808–820, 2017.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation* 9 (8): 1735–1780, 1997.

Jin, C. M., Lim, J. J., Torralba, A., and Willsky, A. S. Exploiting hierarchical context on a large database of object categories. In *CVPR'10*. pp. 129–136, 2010.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS'12*. pp. 1097–1105, 2012.

Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR'16*. pp. 2169–2178, 2016.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature* 521 (7553): 436–444, 2015.

Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J., and Yan, S. Attentive contexts for object detection. *IEEE Transactions on Multimedia* 19 (5): 944–954, 2017.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV'14*. pp. 740–755, 2014.

Liu, J., Gao, C., Meng, D., and Zuo, W. Two-stream contextualized cnn for fine-grained image classification. In *AAAI'16*. pp. 4232–4233, 2016.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *ECCV'16*. pp. 21–37, 2016.

Lowe, D. G. Object recognition from local scale-invariant features. In *ICCV'99*. pp. 1150–1157, 1999.

Oliva, A. and Torralba, A. The role of context in object recognition. *Trends in Cognitive Sciences* 11 (12): 520–527, 2007.

Ouyang, W., Wang, X., Zeng, X., Qiu, S., et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In *CVPR'15*. pp. 2403–2412, 2015.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 15, 2010.

Perronnin, F. Universal and adapted vocabularies for generic visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (7): 1243–125, 2008.

Quattoni, A. and Torralba, A. Recognizing indoor scenes. In *CVPR'09*. pp. 413–420, 2009.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *CVPR'16*. pp. 779–788, 2016.

Redmon, J. and Farhadi, A. Yolo9000: Better, faster, stronger. In *CVPR'17*. pp. 6517–6525, 2017.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS'15*. pp. 91–99, 2015.

Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *NIPS'14*. pp. 568–576, 2014a.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014b.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *CVPR'15*. pp. 1–9, 2015.

Wang, L., Guo, S., Huang, W., Xiong, Y., and Qiao, Y. Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. *IEEE Transactions on Image Processing* 26 (4): 2055–2068, 2017.

Wang, Z., Wang, L., Wang, Y., Zhang, B., and Qiao, Y. Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *IEEE Transactions on Image Processing* 26 (4): 2028–2041, 2017.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR'10*. pp. 3485–3492, 2010.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *ECCV'14*. pp. 818–833, 2014.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (99): 1–14, 2017.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. Learning deep features for scene recognition using places database. In *NIPS'14*. pp. 487–495, 2014.