# CS 676A Project

Group -G26

L.S. Vishnu Sai Rao (12376) and Saurabh Kataria (12637)

# Chosen Paper

**Topic**

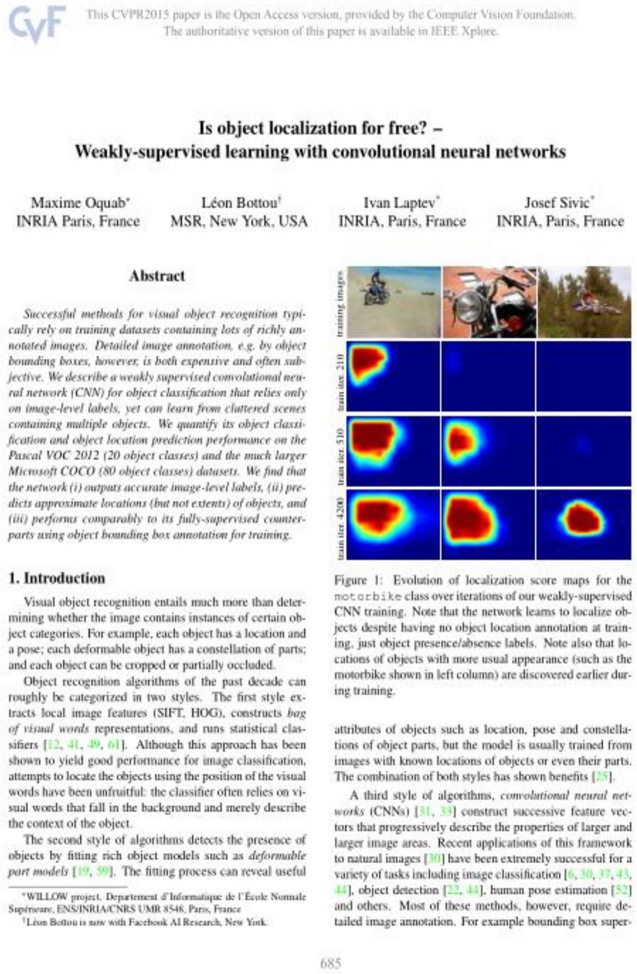
- Weakly Supervised Object Detection and localisation

**Paper title and authors**

- Is object localization for free?-weakly-supervised learning with convolutional neural networks
- Oquab, Maxime, Léon Bottou, Ivan Laptev, and Josef Sivic

**Published in**

- Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015





**Is object localization for free? –
Weakly-supervised learning with convolutional neural networks**

Maxime Oquab* INRIA Paris, France
Léon Bottou† MSR, New York, USA
Ivan Laptev* INRIA, Paris, France
Josef Sivic* INRIA, Paris, France

**Abstract**

*Successful methods for visual object recognition typically rely on training datasets containing lots of richly annotated images. Detailed image annotation, e.g. by object bounding boxes, however, is both expensive and often subjective. We describe a weakly supervised convolutional neural network (CNN) for object classification that relies only on image-level labels, yet can learn from cluttered scenes containing multiple objects. We quantify its object classification and object location prediction performance on the Pascal VOC 2012 (20 object classes) and the much larger Microsoft COCO (80 object classes) datasets. We find that the network (i) outputs accurate image-level labels, (ii) predicts approximate locations (but not extents) of objects, and (iii) performs comparably to its fully-supervised counterparts using object bounding box annotation for training.*

**1. Introduction**

Visual object recognition entails much more than determining whether the image contains instances of certain object categories. For example, each object has a location and a pose; each deformable object has a constellation of parts; and each object can be cropped or partially occluded.

Object recognition algorithms of the past decade can roughly be categorized in two styles. The first style extracts local image features (SIFT, HOG), constructs *bag of visual words* representations, and runs statistical classifiers [12, 41, 49, 61]. Although this approach has been shown to yield good performance for image classification, attempts to locate the objects using the position of the visual words have been unfruitful: the classifier often relies on visual words that fall in the background and merely describe the context of the object.

The second style of algorithms detects the presence of objects by fitting rich object models such as *deformable part models* [19, 59]. The fitting process can reveal useful attributes of objects such as location, pose and constellations of object parts, but the model is usually trained from images with known locations of objects or even their parts. The combination of both styles has shown benefits [25].

A third style of algorithms, *convolutional neural networks* (CNNs) [31, 33] construct successive feature vectors that progressively describe the properties of larger and larger image areas. Recent applications of this framework to natural images [30] have been extremely successful for a variety of tasks including image classification [6, 30, 37, 43, 44], object detection [22, 44], human pose estimation [52] and others. Most of these methods, however, require detailed image annotation. For example bounding box super-



Figure 1: Evolution of localization score maps for the `motorbike` class over iterations of our weakly-supervised CNN training. Note that the network learns to localize objects despite having no object location annotation at training, just object presence/absence labels. Note also that locations of objects with more usual appearance (such as the motorbike shown in left column) are discovered earlier during training.

*WILLOW project, Departement d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548, Paris, France
†Léon Bottou is now with Facebook AI Research, New York.

# What is Weakly supervised object detection?

- Training data-set contains images labelled only with lists of objects they contain and not their locations
- Weakly supervised object detection is important because
  - Annotating locations in an image is an expensive process.
  - ‘label-only’ annotations are often readily available in large amounts, e.g. in the form of text tags or full sentences even geographical meta-data
- The paper employs a CNN architecture to achieve this task.

# Contributions of the paper

Accurate image level labels

comparable to its fully-supervised counterparts

Approximate locations of objects

Is Object localisation free ?

# Example location predictions for images from the Microsoft COCO dataset

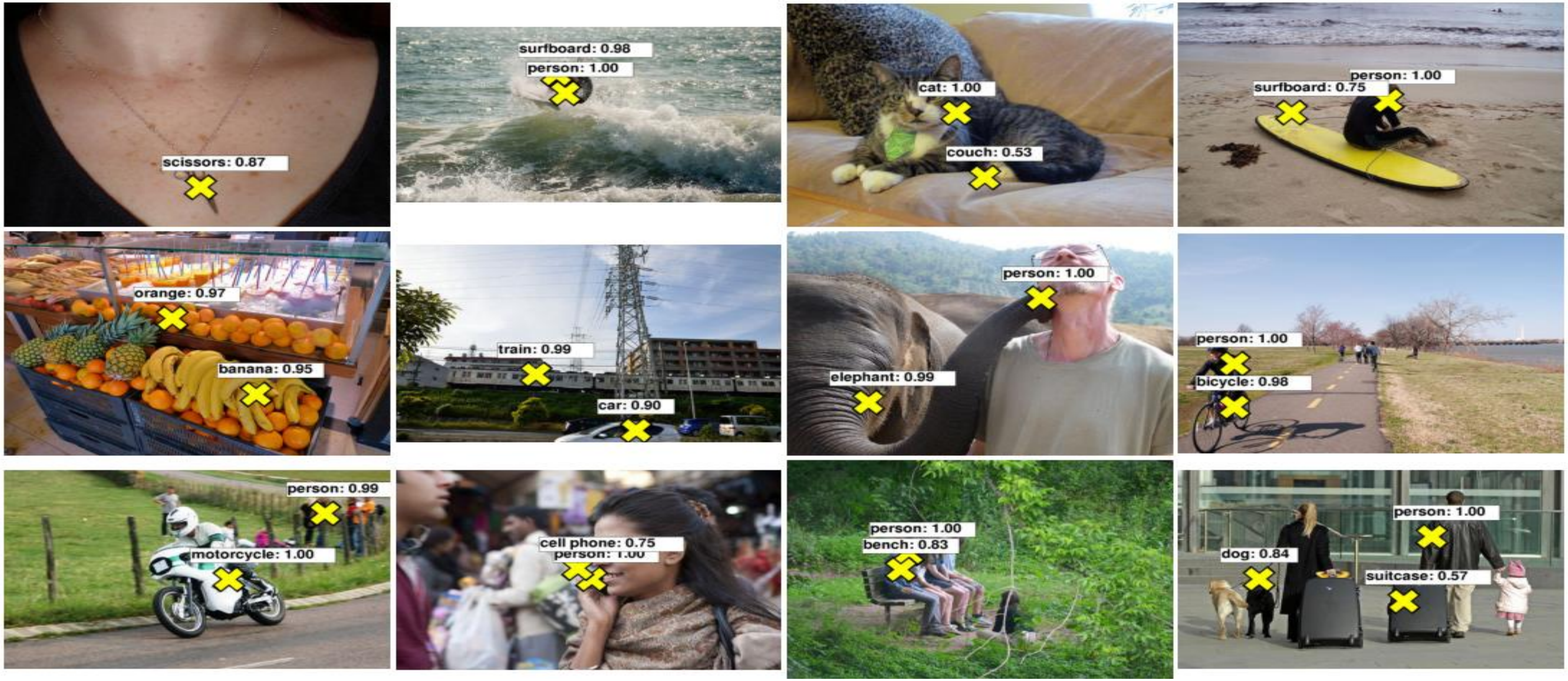



Image credit: Oquab, Maxime, et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." CVPR 2015

# Procedure: Earlier architecture

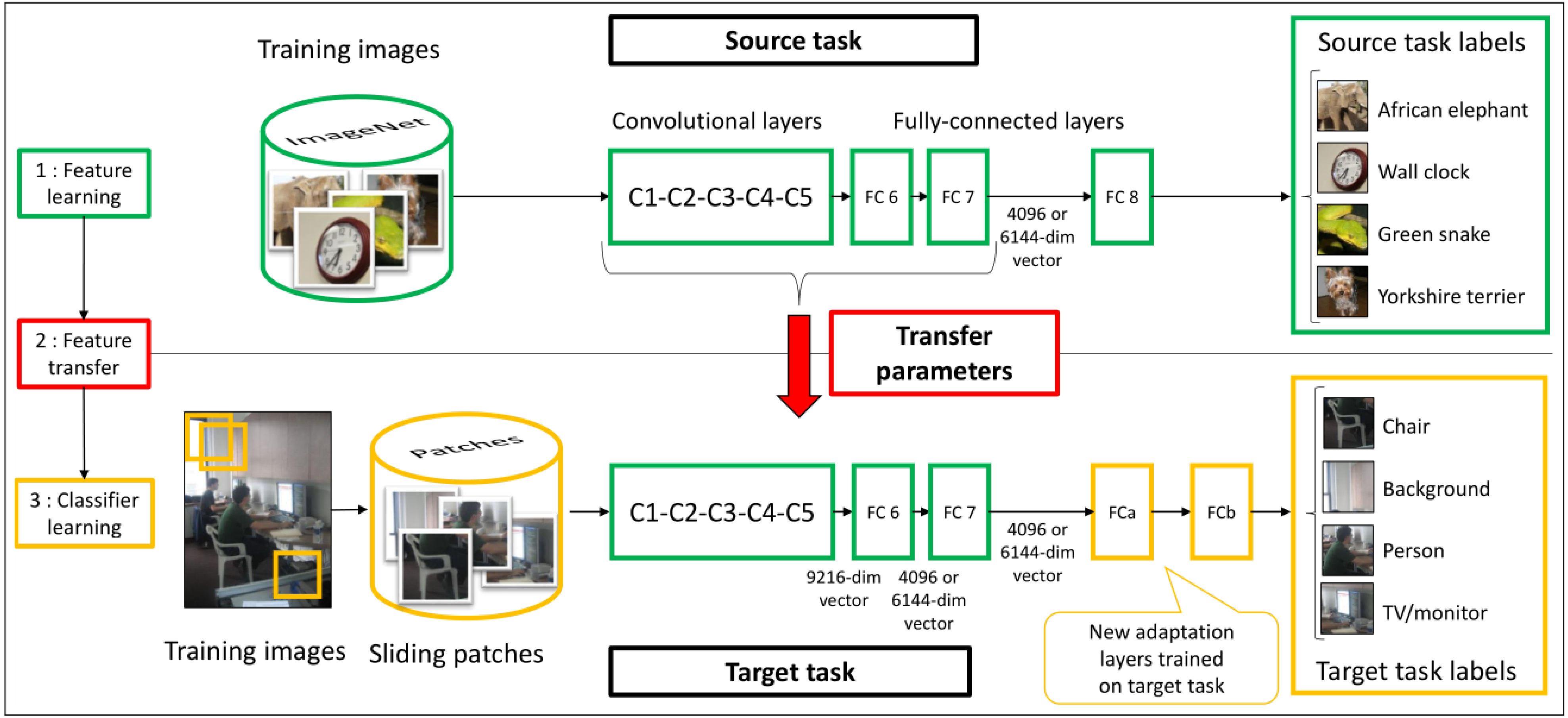
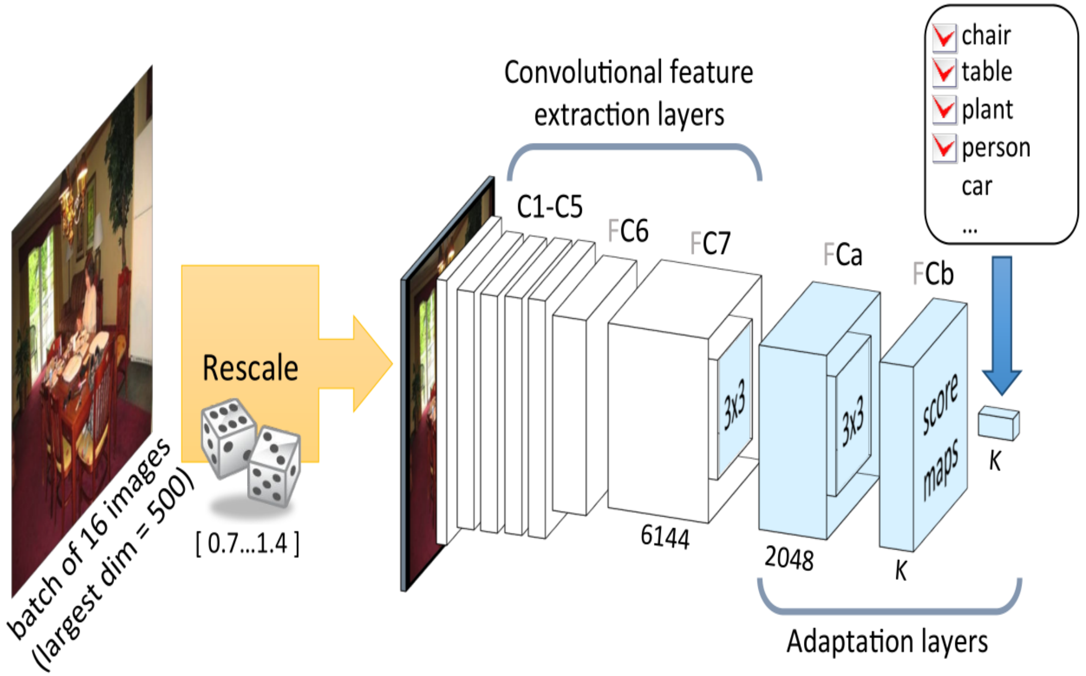



Image credit: Oquab, Maxime, et al. "Learning and transferring mid-level image representations using convolutional neural networks." CVPR 2014

# Description of previous architecture

- The procedure is divided into two steps: pre-training and training.
- In the pre-training step, the earlier convolutional layers are trained using the ImageNet database, which consists of tightly cropped images of single objects. This step enables the architecture to recognise individual objects.
- In the training step, two fully connected adaptation layers are added at the end of architecture, which adapts the new architecture to recognise individual objects in a cluttered image with multiple objects in it. A sliding window method with fixed patch size is used to look at different sections of the image.

# Procedure: Modifications




- Treated the fully connected layers as **convolutions** which helps to deal with nearly arbitrary-sized images as input.
- Explicitly searched for the highest scoring object position in the image by adding a single global **max-pooling layer** at the output.
- Used a sum of K binary logistic regression based **cost function** that can explicitly model multiple objects present in the image

Image credit: Oquab, Maxime, et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." CVPR 2015

# Descriptions of the modifications

- To achieve the same effect of sliding window, the fully connected layers are treated as convolutional layers.
- The global max-pooling layer added at the end of the architecture converts the n x m x K result to 1 x 1 x K. We can use the n x m x K information to make prediction on the location of detected objects. And, the 1 x 1 x K information is used to predict the presence of objects.
- The new cost function enables the architecture to detect multiple objects from one scene collectively, instead of searching for single objects individually.

# Procedure: Multi-scaling

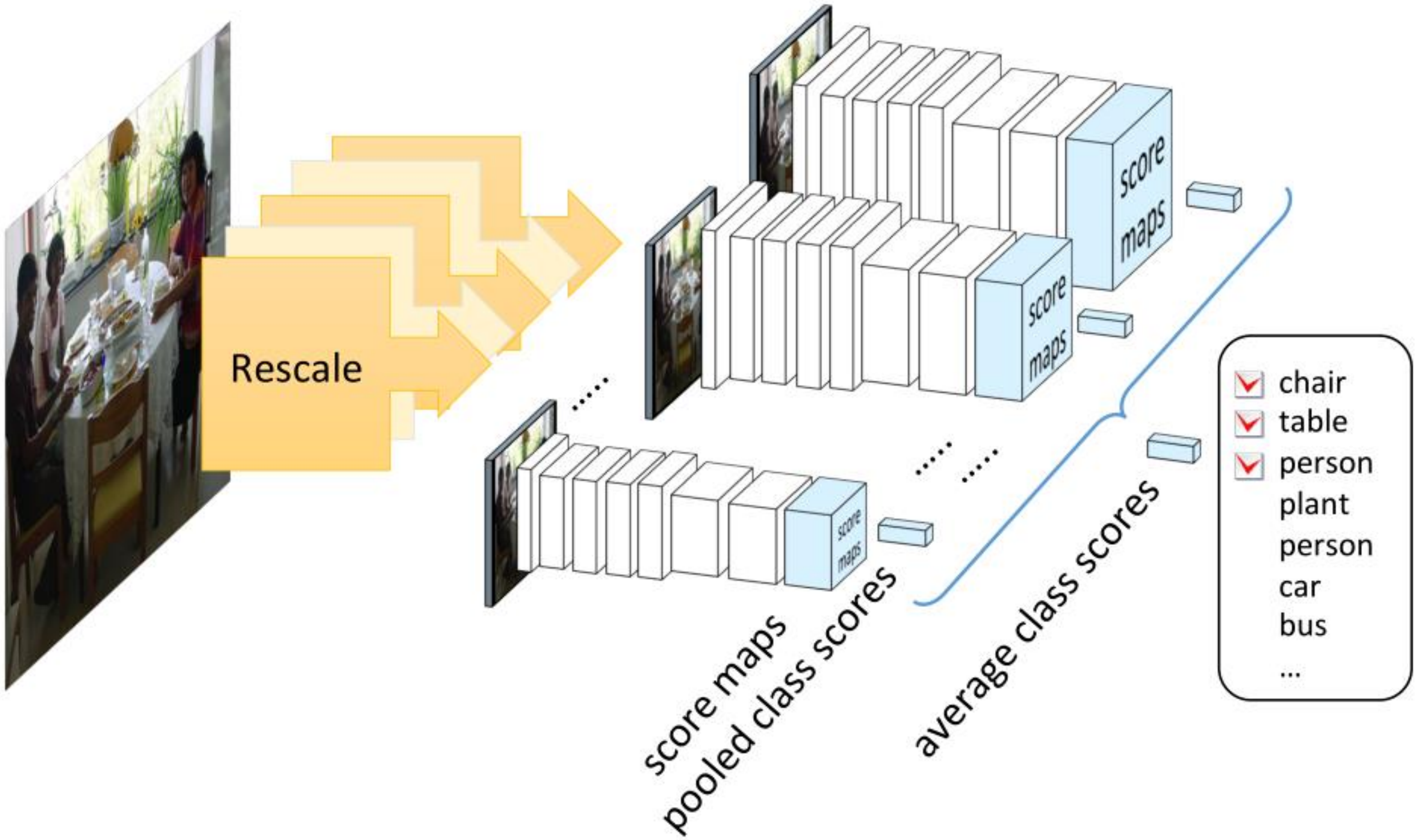



Image credit: Oquab, Maxime, et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." CVPR 2015

# Role of multi-scaling

- The input image is scaled to various values of $s \in [0.7, 1.4]$. The scaled images are fed to the same network in parallel. The output scores of each network is averaged at the end to give the final score.
- This step enables the architecture to recognise tiny as well as large objects. In other words, this step introduces scale invariance.

# What would we like to contribute?

- Predict number of objects in the image as well.
- Introduce invariance to rotation and slight distortions

# What would we like to contribute?

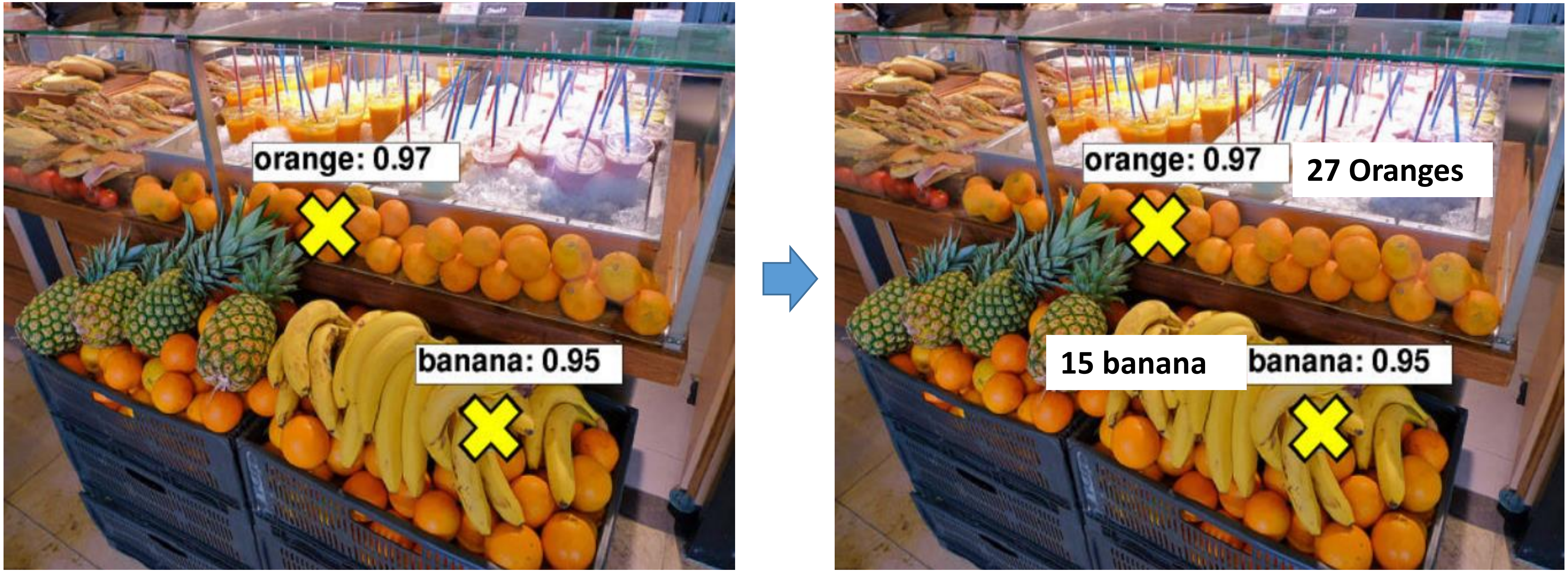



Image credit: Oquab, Maxime, et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." CVPR 2015

# What else is free?

Orientation of the object

Object coverage with the help of a bounding box


Thank You!