OMNIA Faster R-CNN: Detection in the wild through dataset merging and soft distillation

Alexandre Rame * [1], Emilien Garreau † [1], Hedi Ben-Younes ‡ [1,2], and Charles Ollion § [1]

[1]Heuritech [2]Sorbonne University, CNRS, LIP6

Abstract

Object detectors tend to perform poorly in new or open domains, and require exhaustive yet costly annotations from fully labeled datasets. We aim at benefiting from several datasets with different categories but without additional labelling, not only to increase the number of categories detected, but also to take advantage from transfer learning and to enhance domain independence.

Our dataset merging procedure starts with training several initial Faster R-CNN on the different datasets while considering the complementary datasets' images for domain adaptation. Similarly to self-training methods, the predictions of these initial detectors mitigate the missing annotations on the complementary datasets. The final ***OMNIA Faster R-CNN*** *is trained with all categories on the union of the datasets enriched by predictions. The joint training handles unsafe targets with a new classification loss called* ***SoftSig*** *in a softly supervised way.*

Experimental results show that in the case of fashion detection for images in the wild, merging Modanet with COCO increases the final performance from 45.5% to 57.4% in mAP. Applying our soft distillation to the task of detection with domain shift between GTA and Cityscapes enables to beat the state-of-the-art by 5.3 points. Our methodology could unlock object detection for real-world applications without immense datasets.

1. Introduction

Convolutional Neural Networks (CNNs) [34, 26] has become the default method for any computer vision task, and is widely used for problems such as image classification, semantic segmentation or visual relationship detection. One of the key computer vision task is certainly object detection. This task aims at localizing specific objects in an image. Best-performing detectors are Fully Supervised Detectors (FSDs): instance annotations are needed for each object in each image, composed of a category and its location.

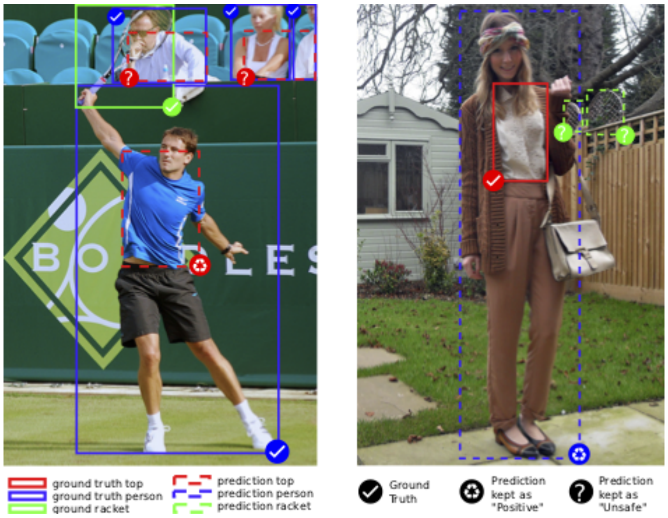

Figure 1. Dataset merging between *COCO* (left) and *Modanet* (right). We fill missing annotations with predictions.

As it has been remarked in [65, 54, 44, 52], FSDs are sensitive to noisy and missing annotations. In particular, the performance of these models are deteriorated when categories are not labeled in some images. Thus, the cost of adding one new category in a dataset is very high: a manual enrichment throughout the whole dataset is required to find all the occurrences of this category. Moreover, the instance annotation process for object detection is expensive and time-consuming. For these reasons, most detection datasets are constrained to a small set of object categories, and are limited in size compared to datasets for other tasks.

The problem of training CNNs for classification on small datasets is usually tackled by transfer learning methods

*alexandre.rame@heuritech.com
†emilien.garreau@heuritech.com
‡hedi.benyounes@gmail.com
§ollion@heuritech.com

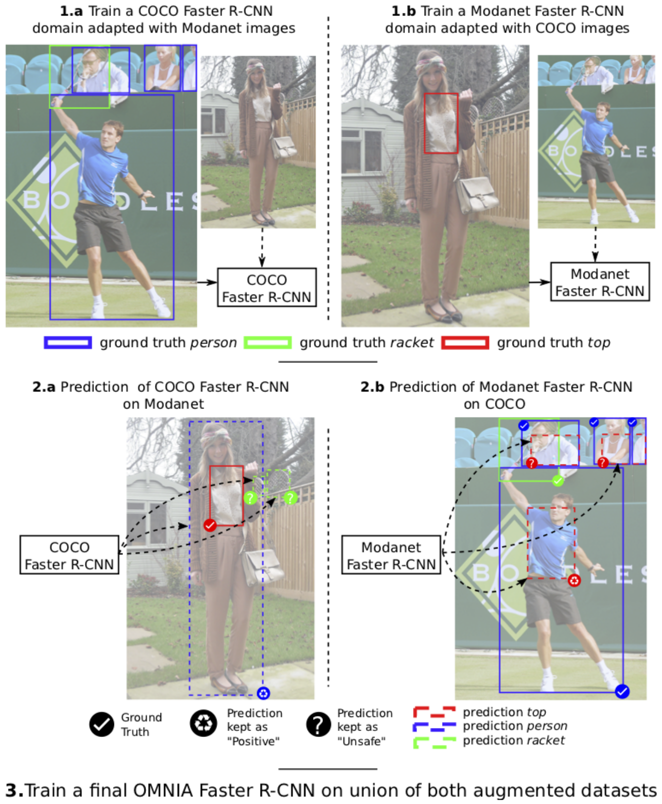
1.a Train a COCO Faster R-CNN domain adapted with Modanet images
1.b Train a Modanet Faster R-CNN domain adapted with COCO images
COCO Faster R-CNN
Modanet Faster R-CNN
ground truth person
ground truth racket
ground truth top
2.a Prediction of COCO Faster R-CNN on Modanet
2.b Prediction of Modanet Faster R-CNN on COCO
COCO Faster R-CNN
Modanet Faster R-CNN
Ground Truth
Prediction kept as "Positive"
Prediction kept as "Unsafe"
prediction top
prediction person
prediction racket
3.Train a final OMNIA Faster R-CNN on union of both augmented datasets

Figure 2. Dataset merging procedure, leveraging Faster R-CNN, domain adaptation and self-training. Best seen in color.

3.2. Merging Procedure

Self-training We aim at training a detector simultaneously for all categories $\mathcal{C}_a \cup \mathcal{C}_b$ on all images $\mathcal{I}_a \cup \mathcal{I}_b$. $\mathcal{G}_a$ provides instances annotations for images from $\mathcal{I}_a$, but only for categories from $\mathcal{C}_a$. Ideally we would like to have access to the ground truths of categories from $\mathcal{C}_b \setminus \mathcal{C}_a$ for images in $\mathcal{I}_a$. The predictions of Detect_b on I_a will replace a new expensive labelling step. For the next training on categories from $\mathcal{C}_a \cup \mathcal{C}_b$, the new targets for I_a are $\mathcal{G}_a \cup \text{Detect}_b(\mathcal{I}_a)$.

Prediction Selection We want to alleviate the fact that some predictions in $\text{Detect}_b(\mathcal{I}_a)$ are erroneous. A prediction with a high detection score is more trustworthy than another one with a lower score: the classification predicted score can be used as proxy for annotation quality [48]. We only take into account predictions with a score higher than $threshold\_low$. Another threshold $threshold\_high$ is defined: all predictions with a higher score will be considered as ground truths.

- $score \leq threshold\_low$, the prediction is discarded
- if $score > threshold\_high$, the instance is a safe prediction and will be considered as a ground truth
- $threshold\_low < score \leq threshold\_high$, the instance is an unsafe prediction

Furthermore, if a prediction has a very high IoU overlap with a human labeled ground truth, we assume that the initial detector made a mistake and we simply discard the prediction. The same procedure is applied to images from I_b.

3.3. OMNIA Faster R-CNN

We introduce our **OMNIA Faster R-CNN**, trained on union on both augmented datasets and that handles unsafe predictions: some of them are correct and should not be considered as background.

3.3.1 RPN

To train the RPN, a binary class label for foreground/background classification is usually assigned to each anchor. In our custom RPN, we add a new *undefined* label. There are now 3 possibilities:

- a *positive* label is given if the anchor matches with **any** ground truth or safe prediction, i.e it has an IoU overlap higher than a certain threshold
- an *undefined* label is given if the anchor matches with an unsafe prediction
- a *negative* label is given if the anchor has a max IoU overlap lower than a certain threshold with **all** ground truths and safe predictions

Positive and *negative* anchors are sampled given a fixed probability. In conclusion, the anchors that match with unsafe predictions will contribute neither to the classification loss nor to the regression loss of the RPN.

3.3.2 SoftSig Box Classifier

The Box classification loss needs to handle the ROIs that match with unsafe predictions. First, they will not contribute to the regression loss. In the classification loss, considering them as background regions would be too conservative in terms of exploration and may be equivalent to having lot of missing annotations. On contrary, discarding them totally from the loss would not exploit all the available information. In particular, even though we are not fully confident that the category c is true for this ROI, it is very likely that all other classes are false. Indeed, the network should only predict either background or c, but not any other category. We propose a simple and efficient custom classification function that takes advantage of all the available information during the training. We named this loss **SoftSig** because it combines **Soft**max and **Sig**moid activation fonctions while handling **Soft Sig**ns. It is a mixed loss between a masked categorical and binary cross-entropy.

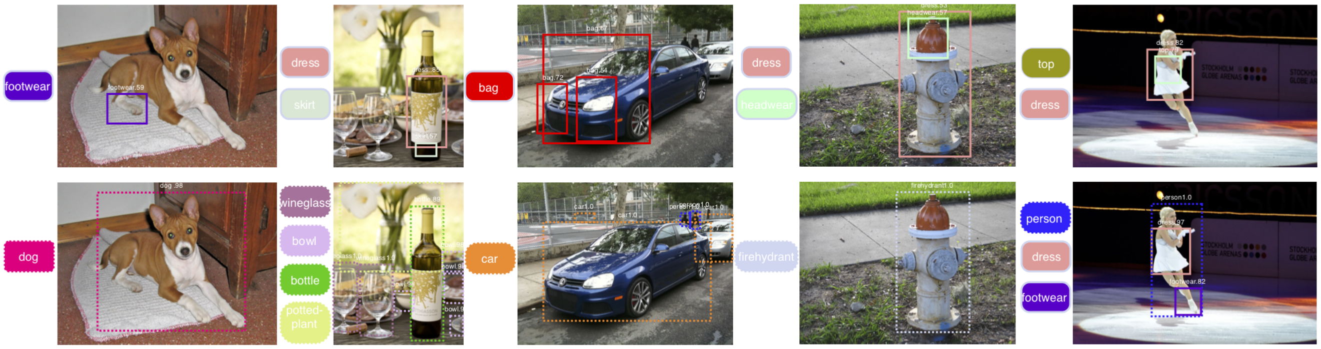

Figure 3. Predictions comparison between *Modanet* Faster R-CNN (top) and **OMNIA Faster R-CNN** (bottom) on *OpenImages*. Our procedure reduces the number of hard false positives for fashion objects. Dotted boxes represent *COCO* categories whereas the solid lines are for *Modanet* garments. Best seen in color.

	Contributions							Results				
	Dataset Creation Procedure					Handling Unsafe Predictions		Mean		Per Category		
Method	DA	Merging	Self training	Source	Target	Masking	Binary	mean mAP/MoLRP	mean w/o bag&tie mAP/MoLRP	dress AP/oLRP	footwear AP/oLRP	pants AP/oLRP
Domain Adaptive [12]	✓			S_{source} U_{target}	U_{source} U_{target}			45.5/77.8	48.9/75.6	67.5/66.8	19.8/91.9	42.6/78.1
Hard Distil. [48]	✓		✓	S_{source} U_{target}	P_{source} U_{target}			51.5/75.7	48.7/76.4	68.4/67.2	20.6/92.0	44.3/77.8
Naive Merging	✓	✓		S_{source} U_{target}	U_{source} S_{target}			37.0/82.8	31.9/84.3	55.7/75.2	7.7/93.6	28.1/83.7
OMNIA Hard Distil.	✓	✓	✓	S_{source} P_{target}	P_{source} S_{target}			56.8/71.8	54.5/72.1	72.4/62.8	29.3/86.3	46.8/74.6
OMNIA Discard Unsafe ROI	✓	✓	✓	S_{source} P_{target}	P_{source} S_{target}	✓		57.0/71.5	54.9/71.8	72.8/62.7	29.8/85.9	46.9/75.1
OMNIA SoftSig	✓	✓	✓	S_{source} P_{target}	P_{source} S_{target}	✓	✓	**57.4/71.0**	**55.2/71.6**	**74.6/62.0**	**30.4/85.6**	**48.3/74.3**

Table 1. Complete ablation study. *Source* dataset only has supervised (S) annotations for *source* categories (S_{source}). *Target* dataset is initially unsupervised for *source* categories (U_{source}) while being supervised for the *target* categories (S_{target}). Predictions (P) are used for replacing missing annotations. Only pictures with image-level labels are considered.

the currently biggest detection dataset with objects in context, *COCO* [39], with 80 objects (such as *glasses*, *firehydrant*) labeled on 117,281 training images. The scheduling from Tensorflow [28] is applied: the learning rate is reduced by 10 after 900K iterations and another 10 after 1.2M iterations. As *COCO* is twice as big as *Modanet*, the fashion images will be sampled 2 times more.

We evaluate our results on *OpenImages* V4 validation dataset [33]. Because of the large scale, the 125,436 images are only annotated for a category if this category has been detected by a image-level CNN classifier: that's why the ground truth annotations are not exhaustive. In the first experiment, we only consider the 13,431 pictures image-level labeled for at least one fashion garment. For example, only 1,412 pictures were verified at image-level for the category *dresses* (1164 contain a dress, 248 do not), and contribute to the computation of the *dress* AP.

Baseline & Results Our experiments are summarized in Table 1. Our OMNIA detector is better on all categories. It was expected for *bag* and *tie* as OMNIA benefits from additional annotations from *COCO* (even though the labelling rules do not perfectly match) and we have a 6.3 points gain (+12.8%) in average on all other categories.

Our first contribution is our **dataset creation procedure**. The Domain Adaptive [12] trained on *Modanet* and adapted to *COCO* does not generalize well to *OpenImages* (see Figure 3). The Hard Distillation is trained similarly to [48] and use *COCO* images as unlabelled data for bootstrapping. Finally, the Naive Merging's training datasets is the naive concatenation of initial datasets. Its low score confirms our initial intuition: the missing annotations of all fashion items in *COCO* are detrimental. Our merging procedure enables to beat all three baselines by a large margin.

Our second contribution is the **handling of unsafe predictions**. The first component is the **masking**: in OMNIA Hard Distillation, unsafe predictions are considered as

Method	mean mAP/MoLRP	mean without bag and tie mAP/MoLRP	bag AP/oLRP	belt AP/oLRP	dress AP/oLRP	footwear AP/oLRP	headwear AP/oLRP	outer AP/oLRP	pants AP/oLRP	tie AP/oLRP	shorts AP/oLRP	skirt AP/oLRP	sunglasses AP/oLRP	top AP/oLRP
Domain Adaptive (0)	25.6/86.3	27.6/85.3	27.5/84.6	4.9/98.2	51.5/75.1	17.0/92.7	41.1/77.9	20.2/87.7	28.2/83.8	2.9/98.3	38.3/77.9	39.8/77.4	27.7/88.3	7.7/94.0
Hard Distil. [48]	31.3/84.7	27.4/86.4	42.4/79.9	6.2/98.2	50.8/76.4	17.6/92.8	39.6/84.0	19.0/88.0	28.6/83.9	59.5/73.2	35.9/78.7	39.3/79.7	28.5/88.2	8.2/93.8
Naive Merging	24.4/88.0	19.0/90.3	42.2/79.5	6.0/96.4	44.5/80.2	6.8/97.0	29.9/84.3	22.6/88.5	10.3/93.7	60.9/72.8	15.1/91.4	33.4/83.0	13.1/94.0	7.9/94.8
OMNIA Hard Distil.	36.9/81.1	33.7/82.1	**43.5/79.3**	6.9/97.1	60.5/**70.1**	23.2/90.3	**48.0/73.3**	32.0/81.7	31.3/82.0	**62.8/72.7**	**43.9/75.7**	49.5/71.8	31.5/86.6	9.9/92.4
OMNIA Discard Unsafe ROI	36.1/81.5	32.9/82.7	43.3/78.6	**10.2**/97.8	58.2/71.6	20.7/91.0	46.4/74.2	32.0/81.6	31.1/82.0	61.3/73.1	41.0/76.0	49.7/72.4	31.6/86.4	7.6/93.9
OMNIA SoftSig	**37.2/80.8**	**34.4/81.7**	43.2/79.1	**10.2/95.7**	**61.0**/70.3	**24.0/89.5**	45.8/73.8	**33.2/81.1**	**31.4/81.8**	60.1/73.1	42.6/76.2	**52.2/71.3**	**32.4/85.6**	**10.6/91.9**

Table 2. Fashion detection results on *OpenImages*. In addition to the pictures with image-level labels, we randomly sampled 10000 images from *OpenImages* and assumed that they did not contain any garment. This assumption is often false but enables to unbias the selection procedure of the validation images.