Spatial Semantic Regularisation for Large Scale Object Detection

Damian Mrowca[1,2], Marcus Rohrbach[1,3], Judy Hoffman[1], Ronghang Hu[1], Kate Saenko[4], Trevor Darrell[1]

[1]UC Berkeley, USA; [2]TU Muenchen, Germany; [3]ICSI, Berkeley, USA; [4]UMass Lowell, USA

Abstract

Large scale object detection with thousands of classes introduces the problem of many contradicting false positive detections, which have to be suppressed. Class-independent non-maximum suppression has traditionally been used for this step, but it does not scale well as the number of classes grows. Traditional non-maximum suppression does not consider label- and instance-level relationships nor does it allow an exploitation of the spatial layout of detection proposals. We propose a new multi-class spatial semantic regularisation method based on affinity propagation clustering [8, 22], which simultaneously optimises across all categories and all proposed locations in the image, to improve both the localisation and categorisation of selected detection proposals. Constraints are shared across the labels through the semantic WordNet hierarchy. Our approach proves to be especially useful in large scale settings with thousands of classes, where spatial and semantic interactions are very frequent and only weakly supervised detectors can be built due to a lack of bounding box annotations. Detection experiments are conducted on the ImageNet and COCO dataset, and in settings with thousands of detected categories. Our method provides a significant precision improvement by reducing false positives, while simultaneously improving the recall.

1. Introduction

Human assistance technologies or question answering require a precise and detailed object recognition of a visual scene. Recently, large scale detection approaches have been proposed which aim to distinguish hundreds or thousands of object categories [1, 3, 11, 12, 23]. While impressive progress has been shown, they suffer from competing object category candidate detections as can be seen in Figure 1 (a). Commonly, non-maximum suppression (NMS) is used to select the bounding boxes with the highest detection score for each category. This method is not globally optimal as only locally overlapping boxes are suppressed by the highest scoring box. Further, in the multi-class case, it does not take semantic relations between objects into account, e.g.

(a) Competing box proposals from different categories

(b) Non-Maximum Suppression (NMS)
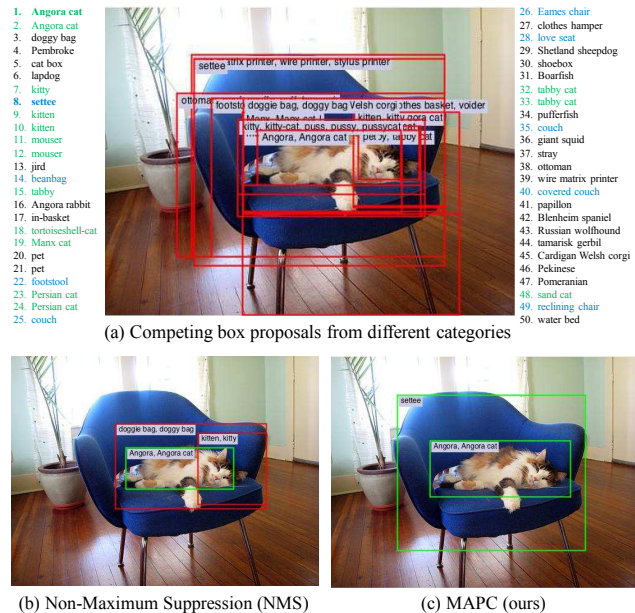
(c) MAPC (ours)

Figure 1. Raw and spatially regularised detection of 7,404 classes using the LSDA extension [12] of the R-CNN method [9]. (a) Top 50 scoring candidate detections and associated categories are listed: all proposals which support the depicted *cat* are green, for *chair* blue. Black entries do not describe any object in this picture. (b) NMS clusters boxes only according to their overlap not according to their class leading to multiple detections of different fine-grained classes for the same object. (c) Our approach (MAPC) exploits category relationships, clustering overlapping boxes with similar classes together which results in less false positives on the same object and enables to detect classes which are suppressed by NMS because of their overlap.

the *couch*, *floorstool* and *beanbag* proposals should support the *settee* candidate detection box in Figure 1, such that it is not suppressed by *doggy bag* as in Figure 1(b).

With thousands of different object categories, semantic relationships become a valuable source of information. Using semantics, consistency can be ensured across different detections. Hence, this work examines the benefit of a semantic hierarchy to object detection of thousands of object categories. We show that in such a large scale setting semantic constraints significantly improve detection.

The key contribution of this work is a large scale spatial semantic regulariser for the correct selection of candidate object detection proposals. Under the framework of Affinity Propagation Clustering (APC) [8], our developed method is characterised by two new ideas.

First, we present an approach which unifies within and across class selection of candidate object detections. Our new multi-class affinity propagation clustering (MAPC) allows for global reasoning over the whole image simultaneously, rather than reasoning locally over image parts or single classes separately, to determine the correct setup of an image. Unlike NMS or [20], which perform the selection separately for each class, our algorithm uses the relationships of highly related fine-grained categories in a large scale detection setting. Based on WordNet relationships, our algorithm knows that *golden retrievers*, *dalmatians* and *dachshunds* are all different dog breeds and should support each other, rather than suppress, if the corresponding boxes cover almost identical regions of interest in the image.

Second, we propose a large scale detection evaluation including over a thousand categories, which requires discriminating among competing classes, in contrast to standard detection challenges, which focus on a per category mean Average Precision (mAP) evaluation. We demonstrate that our algorithm improves performance in two challenging scenarios. First, for a large number of objects per image, we show results on COCO. Second, for a large number of categories, we evaluate on a subset of ImageNet, which is labeled with bounding boxes of 1,825 categories, a large scale detection scenario, which has not been previously evaluated.

2. Related Work

Our work is most related to spatial regularisation over detection proposals. In most detection methods, detection proposals (raw bounding box outputs with scores from detectors) need to be regularised over space to remove double detections on the same object, prune false positives, and improve localisation. Although greedy non-maximum suppression (NMS) is the most often used spatial regularisation approach, other approaches, such as merging nearby detection boxes, are sometimes shown to be more robust [25]. In [28], overlapping detections are averaged and a threshold is set based on overlapping box numbers. In [25], a greedy merge strategy is proposed to group detection proposals together and reward bounding box coherence. Spatial and cooccurrence priors are introduced in [2, 27] to prune detection results. In [7], labels of detection proposals are obtained via approximate inference over several types of spatial relationships instead of greedy NMS. Recently, Affinity Propagation Clustering (APC) [8], an unsupervised clustering method based on message passing, has been used to cluster proposed bounding boxes of the same class based on their overlap [22]. In [22], background and repellence

terms are introduced to APC to allow the suppression of false positives and to avoid selecting object proposals lying too close to each other. Our work builds on [22], but is different in that: (1) our algorithm clusters object proposals of the same and different classes simultaneously, whereas [22] is applied only within each class, (2) we introduce new constraints to ensure that one label per detection proposal is selected, and (3) we design our similarity measure such that semantically close objects get clustered together.

Another line of related work is exploiting semantic category hierarchies in visual recognition and detection [4, 6, 11, 14, 15, 17, 20, 24, 29]. Real world object categories often form a hierarchical structure, which can provide useful information for large scale detection. Such hierarchical relationships can be obtained from predefined semantic structures such as WordNet, or learned by data-driven methods. In [4], a conditional random field based hierarchy-exclusion Graph is proposed to represent subsumption and exclusion relationships between classes. In [11, 14], the ImageNet hierarchy, which is based on WordNet, is used to transfer bounding box annotations and segmentations to semantically close categories. In [6], an accuracy-specificity trade-off based on the ImageNet hierarchy is optimised through the DARTS algorithm. PST [20] uses the WordNet hierarchy to transfer knowledge from known to novel categories and propagates information between instances of the novel categories. In [24] a visual hierarchy is discovered based on the Chinese Restaurant Prior and used to share detector parameters between classes. [15] learn a semantic hierarchy based on visual and semantical constraints. Our work is complementary to previous methods in this area, as we integrate a semantic hierarchy into Multi-class Affinity Propagation Clustering (MAPC) for spatial regularisation, while hierarchies have been only used to train classifiers or share features in previous methods.

Our work is also related to large scale detection. In [3], large scale detectors on over 100,000 classes are trained based on hashing. In [1], NEIL, a semi-supervised learning system, is proposed to train detectors from Internet images. One major obstacle for large scale detection is the lack of bounding box annotations, which has been recently partially resolved by weakly supervised methods such as knowledge transfer [11], Multiple Instance Learning [26, 18], domain adaptation [12] or combined approaches [13]. Among these methods, LSDA [12] is a framework for classifier-to-detector adaptation, and was shown to effectively train large scale detectors based on image-level labels. Thus, in this paper we use LSDA to train a baseline detector on 7,404 leaf classes of the ImageNet hierarchy. However, we note that our spatial regularisation method does not depend on how detectors are trained, and can be applied to arbitrary sets of detectors.

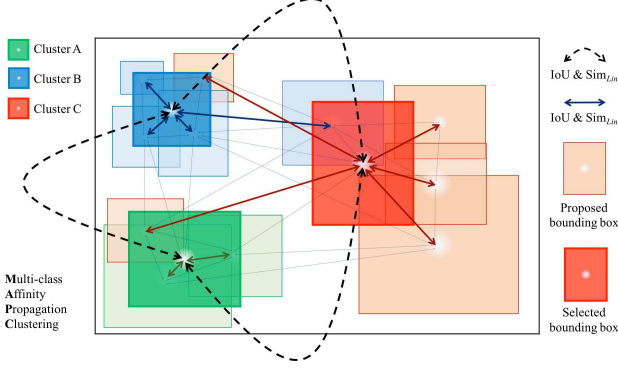To our knowledge, this is the first time that hierarchical

Figure 2. MAPC message passing. Messages are passed between all candidate detections until a subset of detections gets selected as exemplars. IoU stands for Intersection over Union, and sim_{Lin} is the Lin measure. For simplicity not all messages are depicted.

semantic relationships are used together with spatial information to determine the correct scene configuration from contradicting candidate object detections. Furthermore, it is even more challenging to apply this algorithm on a large scale setting, as it requires inference over thousands of fine-grained and diverse categories. Our detection system is unique in its amount of categories, both in terms of the degree of fine-grained detail, for instance incorporating different dog breeds, and the variety of categories, including various animals, plants, fruits, and man-made structures.

3. Spatial semantic regulariser

In this section we describe our spatial semantic regulariser. Our method is based on Affinity Propagation Clustering (APC), which has shown to outperform other clustering techniques such as k-means clustering [8]. [22] successfully adapted APC to the task of selecting candidate object detections of the same class. This method is denoted as Single-class APC (SAPC) in the following.

Our main contributions are to extend the previous work on APC [8, 10, 22] to multi-class detection and a large scale setting with thousands of fine-grained classes. Therefore, we incorporate a new constraint ensuring that each bounding box exemplar gets assigned only one label. Similar to [22], we use an intercluster repellence term and a background category to remove false positives. Additionally, in order to leverage the visual similarity of semantically related fine-grained classes, we introduce hierarchical label relations into APC to cluster semantically similar objects. The resulting Multi-class APC (MAPC) algorithm is presented in Figure 2 after introducing standard APC.

3.1. Standard affinity propagation clustering

APC is a message passing based clustering method. It uses data similarities to identify exemplars such that the

sum of similarities between cluster exemplars and cluster members is maximised. Let $s(i,j)$ denote the similarity between data points i and $j \in \{1,...,N\}$ with N being the number of data points. $s(i,j) \leq 0$ indicates how well j would serve as an exemplar for i [8]. The self-similarity $s(i,i)$ indicates how likely a certain point will be chosen as an exemplar. Using the binary formulation of [10], we encode the exemplar assignment with a set of N^2 binary variables c_{ij}: $c_{ij} = 1$ if i is represented by j and $c_{ij} = 0$ otherwise. A valid clustering must hold two constraints: (i) each point is represented by exactly one exemplar and (ii) when j represents any other point i, then j must be an exemplar representing itself. In the following objective function, I represents constraint (i) and E represents constraint (ii):

$$E_{APC}(\{c_{ij}\}) = \sum_{i,j} S_{ij}(c_{ij}) + \sum_i I_i(c_{i1},...,c_{iN})$$
$$+ \sum_j E_j(c_{1j},...,c_{Nj}) \quad (1)$$

$$S_{ij}(c_{ij}) = \begin{cases} s(i,j) & \text{if } c_{ij} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$I_i(c_{i1},...,c_{iN}) = \begin{cases} -\infty & \text{if } \sum_j c_{ij} \neq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$E_j(c_{1j},...,c_{Nj}) = \begin{cases} -\infty & \text{if } c_{jj} = 0 \text{ and } \exists i \neq j \\ & \text{s.t. } c_{ij} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Max-sum message passing is applied to maximise equation (1) [8, 10] consisting of two messages: The responsibility ρ_{ij} (sent from i to j) describes how suited j would be as an exemplar for i. The availability α_{ij} (sent from j to i) reflects the accumulated evidence for point i to choose point j as its exemplar:

$$\alpha_{ij} = \begin{cases} \sum_{k \neq j} \max(\rho_{kj}, 0) & \text{for } i = j \\ \min(0, \rho_{jj} + \sum_{k \notin \{k,j\}} \max(\rho_{kj}, 0)) & \text{for } i \neq j \end{cases} \quad (5)$$

$$\rho_{ij} = s(i,j) - \max_{q \neq j}(s(i,j) + \alpha_{iq}) \quad (6)$$

3.2. Affinity propagation clustering for multi-class object detection

We introduce our novel Multi-class Affinity Propagation Clustering (MAPC) algorithm, which extends SAPC [22] from single-class to multi-class detection. In multi-class detection most object detectors propose multiple category labels with a certain confidence score for each bounding box. However, the label with the highest confidence is not always the correct one. Hence, not only the correct location but also

the correct class for each box has to be inferred. Therefore, we redefine each data point i or j as a combined box-class detection, e.g. *box$_1$-dog*, *box$_1$-cat*, or *box$_2$-cat*. This allows us to define a similarity measure $s(i, j)$ between detections which includes both the spatial relation between bounding boxes and the relation between their labels (7):

$$s(i, j) = \lambda \, \mathrm{IoU}(i, j) + (1 - \lambda) \, \mathrm{sim}_{Lin}(i, j) \qquad (7)$$

Whereas SAPC bases its similarities solely on the IoU between bounding boxes [22], our similarity measure clusters overlapping detections, represented by the $\mathrm{IoU}(i, j)$ term, as well as semantically similar detections, represented by the $\mathrm{sim}_{Lin}(i, j)$ term. An example can be seen in Figure 3. λ is a weighting factor trading off spatial and semantic similarity. The Intersection over Union is defined as $\mathrm{IoU}(i, j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$, where A_i is the area of the image covered by the bounding box of i. It is used to describe the overlap and hence the visual similarity of two detections. The Lin measure $\mathrm{sim}_{Lin}(i, j) = \frac{2 IC(lcs(C_i, C_j))}{IC(C_i) IC(C_j)}$ denotes how semantically similar the labels of two detections are. $lcs(C_i, C_j)$ denotes the lowest common subsumer of the classes C_i of i and C_j of j in the WordNet hierarchy and $IC(C) = \log p(C)$ equals to the information content of a class, where $p(C)$ is the probability of encountering an instance of the class C in a corpus. The relative corpus frequency of C and the probabilities of all child classes that C subsumes are used to estimate the probability $p(C)$ [19, 21].

The self-similarity is defined as $s(i, i) = -\frac{1}{p - \theta_{bg}}$, where p is the detection score generated by the object detector and θ_{bg} is a background threshold used to discard detections scoring lower than θ_{bg} before APC inference.

To further avoid that contradicting detections are chosen as exemplars, we introduce a new constraint: If class C_i is an exemplar for a specific box k (i.e. $c_{ii} = 1$), no other class can be an exemplar for box k:

$$\widetilde{E}_k(c_{11}, ..., c_{NN}) = \begin{cases} -\infty & \text{if } \sum_{j \text{ with box } k} c_{jj} > 1 \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

The remaining algorithm exactly follows [22], which uses a repellence term $R = \sum_{i \neq j} R_{ij}(c_{ii}, c_{jj})$, but with $r(i, j) = -(s(i, j) + 1)$ to avoid selecting semantic-spatially close exemplars, and a background category to allow for false positives to be suppressed, denoted by the $\widetilde{I}_i(c_{i1}, ..., c_{iN})$ term in equation (9). Linearly combining all of the terms presented yields in the following objective function to be maximised:

$$\widetilde{E}_{APC} = w_a \sum_i S_{ii} + w_b \sum_{i \neq j} S_{ij} + w_c \sum_i \widetilde{I}_i$$
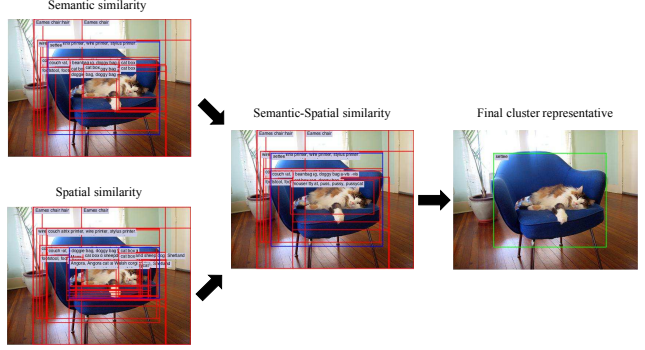$$+ w_d \sum_{i < j} R_{ij} + w_e \sum_j E_j + w_f \sum_k \widetilde{E}_k \qquad (9)$$

Figure 3. Combining spatial and semantic similarity in MAPC. All red boxes form one cluster in which the blue box emerged as their exemplar. With a semantic-spatial similarity, semantically similar and spatially localised detections get clustered which finally results in a well localised true positive detection.

All function arguments in equation (9) were left out for the sake of clarity. To solve this optimisation problem the message passing paradigm of [22] is used. All messages are initialised with zero and iteratively updated until convergence.

4. Experiments

In this section we evaluate the performance of MAPC in a large scale setting. At this time, there is no standardised large scale dataset with both a large amount of object instances within one image as well as a large amount of different object categories. Hence, we evaluate MAPC on two different datasets. We use the Microsoft COCO dataset [16] for the evaluation on a large amount of object instances within one image. To evaluate on a large amount of fine-grained categories, we create a new dataset built of images with bounding box annotations from ImageNet [5]. This dataset covers 1,825 categories, but contains only a few object instances per image due to incomplete annotations.

However, we believe that in a setting with both, thousands of fine-grained categories as well as dozens of object instances per image, our method would perform best. Hence, we also present qualitative results in the supplemental material, where we show the performance of our MAPC algorithm on all 7,404 LSDA categories [12].

We mainly use precision and recall as well as the F1-score, which is the harmonic mean of precision and recall, to evaluate MAPC on these datasets. The mAP metric, which is usually used to evaluate the performance on detection tasks, is not an appropriate performance measure for our multi-class detection setup. mAP is a metric for retrieval tasks. Traditionally, single-class detection has been seen as a retrieval task: all window detections that contain an object of the given class are to be retrieved. As most object detectors were designed as window-scoring methods it was obvious to rank all window detections according to

their scores. With the clustering view, there is no absolute score which could be used for a global ranking and mAP can not be used correctly. The multi-class setting makes it even less suited. mAP favors multiple detections for each class and overall punishes across class selection of object proposals. In contrast, our method actually tries to provide a better way of selecting detections across classes. Hence, we can not use mAP to evaluate this task. For a true understanding of a depicted scene we have to focus especially on a high precision and F1-score for selecting object proposals across classes, while trying to maintain the recall. It is obvious that a high recall could also be achieved by selecting many object proposals without doing across class suppression. As can be seen in Figure 1(a) within class suppression alone—which would be desirable for the mAP measure—still leaves the question unanswered which objects are actually depicted in an image. For a more detail investigation of wrong detections, we examine whether a false positive occurred due to a wrong localisation or classification. *Wrong label* is the amount of all false positives with wrong labels of all false positives. *Wrong overlap* is the amount of all false positives with a wrong location of all false positives.

To setup MAPC and determine all of its parameters, we use grid search on a training set obtained from ImageNet [5] as follows: First, we search for all ImageNet categories with available bounding box annotations. Next, we determine which of these categories overlap with the 7,404 categories of the LSDA detector [12]. This results in 1,825 categories with annotated images. Next, we discard all images used during the LSDA training and in the ImageNet test set described in section 4.2. We obtain our final training set by randomly selecting two annotated images per category from the remaining images. After performing grid search on this training set the MAPC parameters are set such that recall and precision are maximised.

In all our experiments common non-maximum suppression (NMS) is used as the baseline. More specifically, detections of the same category overlapping more than a defined IoU threshold are suppressed in a first step. Then, all the remaining detections are suppressed across all classes with a different IoU threshold. Both NMS thresholds were determined using grid search as previously described. The best configuration resulted in a higher IoU threshold for within class suppression than for across class suppression. The intuition for this is that detections of the same class instance are typically located at similar positions in the image. Thus, in order to suppress within classes a higher threshold is necessary. This baseline will be denoted as Within Class and Across Class NMS (WC+AC-NMS). MAPC is also compared against SAPC [22]. However, SAPC was designed for single-class detection. As we evaluate in a multi-class detection scenario, we simply accumulate the per class output of SAPC across all classes for a first SAPC version.

However, accumulating all detections without suppressing across classes is more suitable for an object retrieval task than for multi-class object detection. Thus, in a second version, we use across class NMS (AC-NMS) on top of the accumulated SAPC output to select object detections also across classes. This makes SAPC [22] better comparable to our method. The IoU threshold for this across class NMS was also determined using grid search.

4.1. Multiple instance detection on COCO

The Microsoft COCO dataset [16] consists of images that depict objects in their real world context rather than centered objects. Because of this, the detection on COCO is much more challenging than on the mostly centered ImageNet pictures. Hence, this dataset is chosen to evaluate the performance of our semantic spatial regulariser in a contextual setup with numerous object instances per image.

4.1.1 Experimental setup

COCO consists of 80 different categories with on average 7 object instances per image. In a first experiment, we use the latest LSDA setup with 7,404 fine-grained categories [12]. 15 COCO categories neither overlap with the leaf node categories of LSDA nor with either of their parents in the Word-Net hierarchy[1]. For those of the remaining 65 categories which overlap with a parent category, we use all of their children as an input to our method and the baselines. For example we detect *beagle* and *dachshunds* instead of their parent category *dog*. This results in 1,133 fine-grained child categories, which all methods have to infer on. We simply relabel the children output after inference to their parent categories to compare it with the COCO ground truth. We neither train LSDA nor adapt the MAPC paramters to COCO.

In a second experiment, we fine-tune our detection network on the COCO training set using all 80 COCO categories as input to our method and the baselines. Both experiments are evaluated on the COCO validation set.

4.1.2 Experimental results

Table 1 shows the detection results of our first experiment without finetuning our detector on COCO on the COCO validation set. As can be seen MAPC outperforms WC+AC-NMS by 3.16% in terms of precision when maintaining the recall. This performance gain can be explained by less wrongly labeled (65.13%) and wrongly localised (74.31%) detections. The F1 score for the chosen setup is 13.46% for WC+AC-NMS versus 15.09% for MAPC. In Figure 5(c) & (d) we vary the IoU evaluation threshold above which a detection is counted as a true positive. As can be seen MAPC

---

[1]*traffic light, fire hydrant, stop sign, snowboard, person, kite, fork, sandwich, hot dog, pizza, donut, cake, potted plant, book, teddy bear*
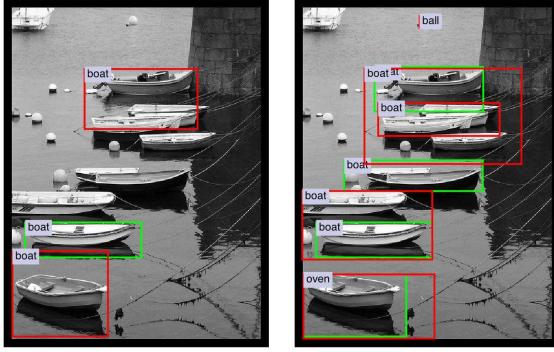
MAPC (optimal precision) MAPC (optimal F1)

Figure 4. Different optimisation criteria. When optimised for F1 score instead of precision, MAPC selects more detections, resulting in more true and false positives.

is always better than WC+AC-NMS. In general, almost all operating points of MAPC lie above WC+AC-NMS as can be seen in the precision-recall curve depicted in Figure 5(a). These results clearly show that MAPC is superior to WC+AC-NMS in scenarios with a lot of object instances per image. Also when compared to SAPC [22] our MAPC method shows an improvement over all numbers, except for the recall of 20.72% since no across class suppression is applied. Hence, many detections are selected resulting in a cluttered outcome, which manifests in the low precision value of 5.25% and decreases the F1-score to 8.38%. As [22] was designed for within class suppression and does not suppress across classes, these results are not surprising. However, when across class NMS (AC-NMS) is applied on the accumulated outcome of [22] the precision increases to 14.66% at the cost of a recall decrease. Overall the F1-score increases to 13.12%. However, MAPC performs best on the COCO validation set amongst all tested methods.

The greater precision of MAPC can be especially seen when we look at example images. The pictures in Figure 6 show the output of WC+AC-NMS and MAPC after optimising both algorithms for the highest precision with comparable recall. The detector not fine-tuned on COCO was used. Green boxes are true positive detections. Red boxes are false positive detections. WC+AC-NMS reaches its precision limit after suppressing all overlapping boxes, while MAPC can also suppress non-overlapping boxes. At the same time, MAPC still enables the selection of overlapping object proposals as can be clearly seen in the example pictures. Allowing a greater overlap for WC+AC-NMS would increase true positives at the cost of lower precision and a cluttered detection output. In general, MAPC outputs less false positives and better localised true positives.

If required, MAPC can also be optimised towards a higher recall. Figure 4 examplarily compares a F1 score optimised MAPC to a precision optimised MAPC. Clearly

Method	Pre-cision	Re-call	Wrong Label	Wrong Overlap	F1 Score
WC+AC-NMS	13.44	13.47	79.39	88.97	13.46
SAPC [22]	5.25	20.72	74.79	72.73	8.38
SAPC + AC-NMS	14.66	11.86	81.36	87.15	13.12
MAPC (ours)	16.60	13.84	65.13	74.31	15.09

Table 1. Detection results on COCO without finetuning, in %.

Method	Pre-cision	Re-call	Wrong Label	Wrong Overlap	F1 Score
WC+AC-NMS	23.50	24.80	62.99	94.97	24.10
SAPC [22]	15.66	32.61	69.01	72.43	21.17
SAPC + AC-NMS	30.01	21.97	74.95	92.90	25.39
MAPC (ours)	37.64	24.23	55.47	71.79	29.50

Table 2. Detection results on COCO, fine-tuned on COCO, in %.

Method	Pre-cision	Re-call	Wrong Label	Wrong Overlap	F1 Score
WC+AC-NMS	8.34	11.29	91.90	85.53	9.59
SAPC [22]	3.46	22.57	93.69	68.05	6.00
SAPC + AC-NMS	9.76	10.34	91.02	81.54	10.04
MAPC (ours)	10.94	16.22	86.41	68.57	13.07

Table 3. Detection results on ImageNet without finetuning, in %.

more *boats* get detected when we optimise towards F1, but also more false positives are selected. All in all, MAPC can be optimised towards a high recall and a high precision, while WC+AC-NMS reaches its precision limit when trying to suppress non overlapping boxes. Thus, MAPC can be preciser in selecting the correct bounding box proposals.

In our second experiment, we fine-tune our object detector on COCO. The results can be seen in table 2. As expected all of our metrics highly improve. Most striking the MAPC precision rises to 37.64%, while the recall remains comparable, which increases the F1 score difference between MAPC and WC+AC-NMS to 5.40%. Also the F1 score of SAPC strongly improves to 21.17%. All methods obviously greatly profit from better detections. Thus, a detector which provides good candidate detections in the first place is crucial for all of the examined methods.

4.2. Fine-grained multi-class detection on ImageNet

In this section we evaluate MAPC on a large scale multi-class detection setup constructed from ImageNet data [5]. Since there is no standardised dataset with thousands of categories, we construct our own dataset to evaluate MAPC on a large amount of fine-grained categories. The final dataset covers 1,825 categories, but only a few object instances per image due to incomplete annotations of ImageNet.

4.2.1 Experimental setup

In order to construct a dataset with numerous fine-grained categories, we search for all ImageNet categories with available detection annotations. As we use the LSDA detector
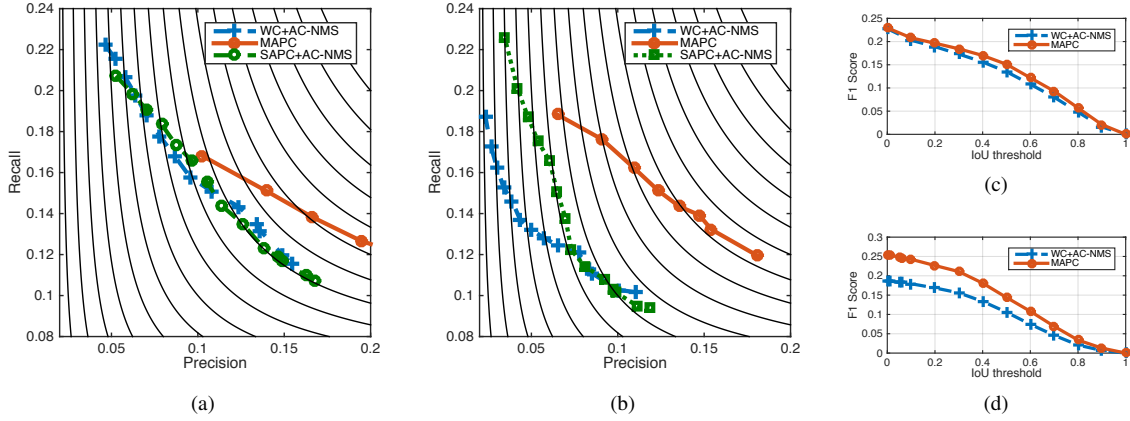
Figure 5. % & (b) Precision-recall curve for WC+AC-NMS, SAPC + AC-NMS and MAPC on COCO (a) and on the set of 1,825 ImageNet categories (b) without finetuning. The curved lines mark points of equal F1 score. The F1 score increases from lower left to upper right. Multiple operating points are obtained by varying the across class IoU threshold in AC-NMS (for WC+AC-NMS and SAPC + AC-NMS) and w_a for MAPC. MAPC consistently outperforms WC+AC-NMS and SAPC + AC-NMS on COCO and ImageNet. SAPC + AC-NMS is superior to WC+AC-NMS in the lower precision range. (c) & (d) F1-score plotted against IoU evaluation threshold for COCO (c) and on the set of 1,825 ImageNet categories (d). MAPC consistently outperforms WC+AC-NMS on both datasets.

[12], we determine which of these categories overlap with its 7,404 categories. This results in 1,825 categories with annotated images. Next, all images used during the training of the LSDA detector are discarded. As most of the remaining images have only one object annotated, we further restrict our test set to images with at least two annotated objects. This way, we ensure that we evaluate on a true detection setup rather than a localisation setup. After this step, we obtain our final fine-grained ImageNet test set.

4.2.2 Experimental results

Table 3 shows the detection results on our large scale ImageNet detection dataset. The same tendencies as on COCO can also be observed here. The precision and recall for MAPC are 2.60% and 4.93% higher than for WC+AC-NMS. The F1-score increases from 9.59% to 13.07%. False positives due to wrong labels drop by 5.49% and localisation errors drop from 85.53% to 68.57%. Again SAPC performs better after applying across class NMS. However, MAPC still performs best, which confirms our results on COCO. What is striking however is that the improvement to WC+AC-NMS and SAPC + AC-NMS is bigger in the fine-grained setting. Also the improvement of MAPC over WC+AC-NMS, when the IoU evaluation threshold is varied, is bigger than on COCO, which can be seen in Figure 5. It seems that the more fine-grained the categories, the more visually similar are semantically similar categories, and thus, the more useful the label relations from the WordNet are. This indicates that our approach is especially useful in a large scale setting when a lot of visually similar fine-grained object categories are competing against each other.

5. Conclusions and Future Work

We presented MAPC, a large scale multi-class regulariser which globally maximises both the semantic and spatial similarity, and thus, visual similarity of clusters of detection proposals. MAPC reduces false positives significantly in multi-class detection, resulting in an improved classification and localisation. Our results show that the selection of detection proposals can be significantly improved over baseline class-independent non-maximum suppression by formulating a clustering problem across class labels and spatial dimensions, which can be solved by affinity propagation. Overall, we consistently improve precision and recall for different operating points and evaluation thresholds.

As future work, it would be interesting to compare the fine-grained category detection on COCO with detectors trained on all parent categories to see whether training more fine-grained classes to detect the actual parent class helps the detection of objects. MAPC could also be extended to the temporal domain, in order to cluster over consecutive video frames for activity recognition and video description.

WC+AC-NMS	MAPC	WC+AC-NMS	MAPC

Figure 6. Examples, where MAPC outperforms WC+AC-NMS on Microsoft COCO. True positives: green, false positives: red.

Ground Truth	WC+AC-NMS	MAPC

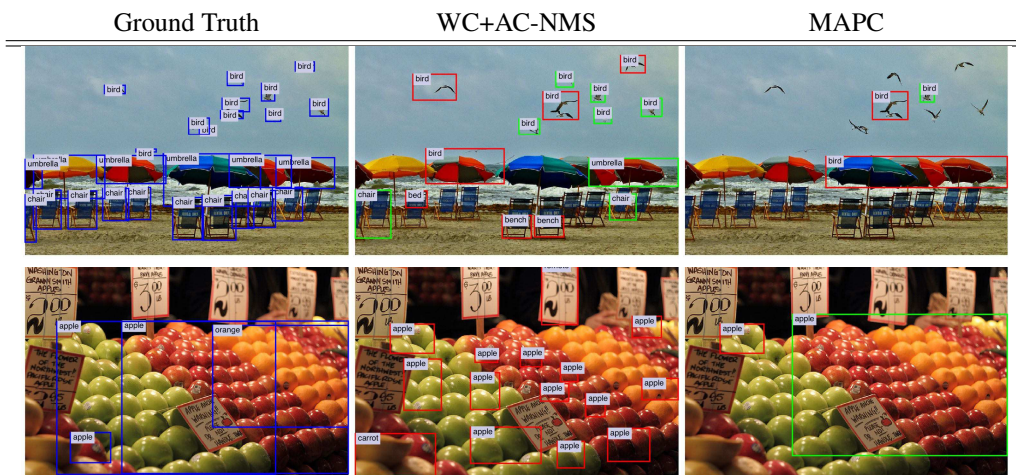
Figure 7. Examples where WC+AC-NMS outperforms MAPC on Microsoft COCO. Ground truth: blue, true positives: green, false positives: red.

References

[1] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[2] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[3] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[4] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 48–64. Springer, 2014.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[6] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative Models for Multi- Class Object Layout. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009.

[8] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587. IEEE, 2014.

[10] I. E. Givoni and B. J. Frey. A binary variable model for affinity propagation. *Neural computation*, 21(6):1589–1600, 2009.

[11] M. Guillaumin and V. Ferrari. Large-scale knowledge transfer for object localization in ImageNet. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[13] J. Hoffman, D. Pathak, T. Darrell, U. C. Berkeley, and K. Saenko. Detector Discovery in the Wild: Joint Multiple Instance and Representation Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation Propagation in ImageNet. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[15] L.-J. Li, C. Wang, Y. Lim, D. M. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[17] M. Marszałek and C. Schmid. Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7. IEEE, 2007.

[18] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Multi-Class Multiple Instance Learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[19] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint*, 1995.

[20] M. Rohrbach, S. Ebert, and B. Schiele. Transfer Learning in a Transductive Setting. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[21] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1641–1648. IEEE, 2011.

[22] R. Rothe, M. Guillaumin, and L. V. Gool. Non-maximum suppression for object detection by passing messages between windows. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2014.

[23] O. Russakovsky, J. Deng, Z. Huang, A. C. Berg, and L. Fei-Fei. Detecting Avocados to Zucchinis: What Have We Done, and Where Are We Going? *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[24] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1488, 2011.

[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[26] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[27] B. A. Torralba, K. P. Murphy, and W. T. Freeman. Using the Forest to See the Trees : Exploiting Context for Visual Object Detection and Localization. *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[28] P. Viola and M. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision (IJCV)*, 2004.

[29] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.