

Contextualizing Object Detection and Classification

Zheng Song^{1*}, Qiang Chen^{1*}, Zhongyang Huang², Yang Hua², Shuicheng Yan¹

¹ Department of Electrical and Computer Engineering, National University of Singapore, Singapore

² Panasonic Singapore Laboratories, Singapore

{zheng.s, chenqiang, eleyans}@nus.edu.sg, {zhongyang.huang, yang.hua}@sg.panasonic.com

Abstract

In this paper, we investigate how to iteratively and mutually boost object classification and detection by taking the outputs from one task as the context of the other one. First, instead of intuitive feature and context concatenation or postprocessing with context, the so-called Contextualized Support Vector Machine (Context-SVM) is proposed, where the context takes the responsibility of dynamically adjusting the classification hyperplane, and thus the context-adaptive classifier is achieved. Then, an iterative training procedure is presented. In each step, Context-SVM, associated with the output context from one task (object classification or detection), is instantiated to boost the performance for the other task, whose augmented outputs are then further used to improve the former task by Context-SVM. The proposed solution is evaluated on the object classification and detection tasks of PASCAL Visual Object Challenge (VOC) 2007 and 2010, and achieves the state-of-the-art performance.

1. Introduction

Object detection and classification are two key tasks for image understanding, and have attracted much attention in the past decades. The object classification task aims to predict the existence of objects within images, whereas the object detection targets localizing the objects. Several image databases tailored for these two tasks have been constructed, such as Caltech-101 [16]/256 [17] and PASCAL Visual Object Challenge (VOC) [9] and many efforts [10][15] have been devoted for these two tasks.

Beyond various image descriptors and modeling methods, the usage of context has become more and more popular for enhancing the algorithmic performance. Many recent studies demonstrated considerable improvement for object detection and classification by using external information, which is independently retrieved and complementary with traditional image descriptors. Specifically, the external context includes user-provided tags [5][14], surrounding texts

* indicates equal contributions

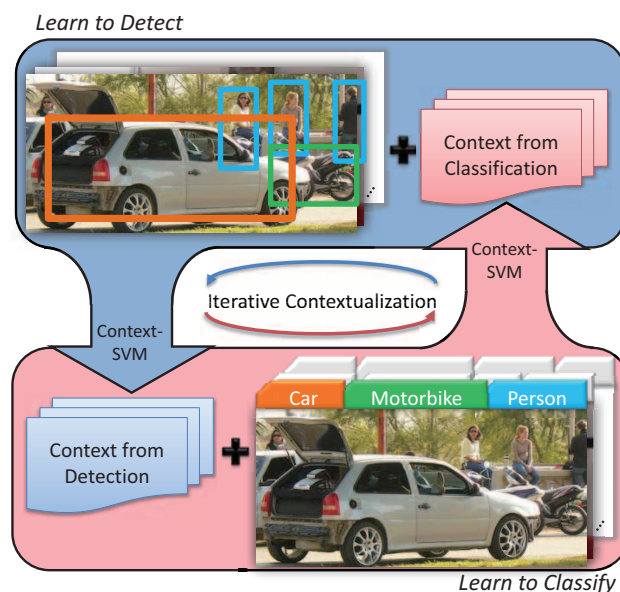


Figure 1: Illustration of the iterative contextualizing procedure. The object detection and classification tasks utilize context from each other and mutually boost performance iteratively. For better viewing, please see original color PDF file.

from Internet [3][1], geo-tags and time stamps [8], etc.

The context may also be the information lying within individual images. Intuitively, the spatial locations of objects and background scene from global view can be used as inside-image context [12][13]. Further, when we consider object detection and classification tasks together, these two tasks can provide natural comprehensive context for each other without any external assistance, and thus can be mutually contextualized for performance boosting [11].

In this paper, we develop a novel mutual contextualization scheme for object detection and classification based on the so-called *Contextualized Support Vector Machine* (Context-SVM) method. Extensive experiments show that Context-SVM can efficiently learn the context models under various conditions and effectively utilize context information for performance boosting. We implement and evaluate the proposed scheme on object detection and classifi-

cation tasks of the VOC 2007 and VOC 2010 datasets [9], and the results are superior over the state-of-the-art on most object categories.

First, we present a *contextualized learning* scheme via Context-SVM with the following characteristics:

- *Adaptive contextualization*: As many studies have shown [24][23], context should be activated to be supportive mostly for those *ambiguous samples* and thus the context effectiveness should be conditional on the ambiguity of sample classification. The Context-SVM is superior over traditional learning schemes by complying this principle in its mathematical formulation.
- *Configurable model complexity*: The contextualization process should be efficient for both detection and classification tasks, and thus the solution should not involve many parameters. In this work, the Context-SVM with tractable control on the complexity of the context model is well formulated, and thus the generalization capability is guaranteed.

Then we propose an iterative contextualization procedure based on the Context-SVM, such that the performance of object classification and detection can be iteratively and mutually boosted as shown in Figure 1.

2. Related Work

Harzallah et al. [11] introduced the pioneering work for object detection and classification contextualization through probability combination in postprocessing. In this work, we instead develop the learning scheme which seamlessly integrates the context information for collaborative learning.

Traditionally, the context is considered as special features. Most of the existing strategies [14][8][11] utilize the context via feature concatenation, model fusion or confidence combination, and take the context as another independent component. However, context may have instable distribution, and its reliability and noise level are not controllable. Therefore adaptive integration of context is required to avoid the inappropriate usage of context information. In this work, we follow this line to design the learning scheme for utilizing context information.

Also, some methods have been proposed to model the context in a comprehensive manner, e.g. [25], but they are served for a more specific purpose and not easily generalized to our requirement.

3. Contextualized SVM

In this work, the *context* is generally defined as certain extra supportive information for one task, which is retrieved independently from the *subject* task ¹. In the section, we

¹We refer the main/principal task concerned as the *subject* task.

first introduce the probabilistic motivation of the contextualized SVM (Context-SVM) and then derive its linear formulation based on the probabilistic motivation. Finally, we extend the linear Context-SVM to the kernel version for more general usage.

3.1. Probabilistic Motivation

Let $x_i^f \in \mathbb{R}^n$ denote the features of a sample for the subject task, $x_i^c \in \mathbb{R}^m$ denote the features of the corresponding context, and y_i denotes the ground-truth class label. Then the entire training data can be expressed as

$$\{X_i = \{x_i^f, x_i^c\}, y_i; i = 1, 2, \dots, N\}. \quad (1)$$

Generally, the objective of a discriminative learning model can be defined as to maximize:

$$\prod_{i=1}^N P(y = y_i | X_i),$$

namely the Maximum a Posteriori (MAP).

There are two components within X_i , and often the independent assumption of the subject features x_i^f and the context x_i^c is made and then the probability of label y for a given sample X_i can be approximated as:

$$p(y|X_i) \approx p(y|x_i^f)p(y|x_i^c). \quad (2)$$

The inference based on (2) is right for the traditional solution of confidence combination [11][8] or multiple feature/model fusion [14].

The independence assumption, however, is often invalid for real data, and hence we propose to infer the label probability by (3) which explicitly models the conditional usage of context with respect to the given subject features:

$$p(y|X_i) = p(y|x_i^f, x_i^c) \propto p(y|x_i^f) \cdot p(y, x_i^c|x_i^f). \quad (3)$$

More specifically, we aim to infer the label probability via two components simultaneously. The first one is based on the subject features, i.e. $p(y|x_i^f)$, and the second one is based on the context features, which contribute to the inference while only ambiguous decision from the first component is expected, i.e. $p(y, x_i^c|x_i^f)$.

The second component is critical for a contextualized learning model. For object detection, the context of scene information from object classification is nearly the same for all detected windows within one image and might not be necessary for many windows. Instead, only the most ambiguous detections need the assistance from context.

For object classification, the context from object detection generally shows low reliability due to the possible false alarms and the selective usage of context can effectively avoid the disturbance caused by the false context to those already high-confident object patterns.

3.2. Context-SVM: Formulation and Solution

3.2.1 General Formulation

For ease of formulation, we only concern the binary classification problem for object detection or classification task, i.e. $y_i \in \{+1, -1\}$ and the N_c -class problem can be decomposed into N_c binary classification problems through one-vs-all strategy. SVM [4] provides a general supervised learning framework by maximum margin optimization, and in this work, we extend SVM by introducing a novel parameterized model to describe the dependence between the context features and the subject features.

The general SVM learns a classifier over the subject feature space and obtains a fixed hyperplane:

$$w_0^T \cdot x^f + b = 0. \quad (4)$$

As the corresponding context features x_i^c can provide extra supportive information for the classification of x_i^f , we propose to utilize x_i^c to adapt w_0 for sample X_i . Then a sample-specific w_i can be obtained to substitute w_0 , which essentially optimizes the margin of sample i and can consequently improve the discriminative power of the classifier. More specifically, we introduce a transformation matrix $P \in \mathbb{R}^{n \times m}$ to utilize x_i^c for the subject classification, and then

$$w_0 \longrightarrow w_i = Px_i^c + w_0. \quad (5)$$

The number of parameters brought by P is very large, which may easily make the derived model overfitting, and thus we introduce a complexity constraint over P . That is, the matrix P is constrained as a low-rank matrix, expressed as the sum of R rank-1 matrices in (6) in which $u_r \in \mathbb{R}^n$ and $q_r \in \mathbb{R}^m$,

$$P = \sum_{r=1}^R u_r \cdot q_r^T, \quad (6)$$

and then the complexity of the context model could be well controlled with $R \times (m + n)$ parameters, where R is the rank of P . As latter introduced, the P in constrained form will better interpret how the proposed contextualized learning model adaptively utilizes the context for inference.

By substituting P into (5), we obtain (7), and the so-called margin for sample X_i could be derived as in (8):

$$w_i = w_0 + \sum_{r=1}^R (q_r^T x_i^c) \cdot u_r, \quad (7)$$

$$\gamma_i = y_i (w_0^T x_i^f + \sum_{r=1}^R (q_r^T x_i^c) \cdot (u_r^T x_i^f) + b). \quad (8)$$

These two equations well show the more insightful meaning of the contextualized SVM formulation:

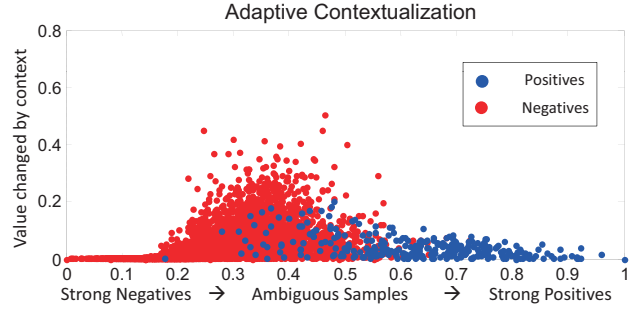


Figure 2: Illustration of the relationship between original sample confidence and confidence variation amount from context. The blue and red dots represent positive and negative samples respectively. The x-axis denotes the sample confidence in subject feature space and y-axis denotes the absolute amount of confidence changed by the contextualization procedure. The confidences are converted into probabilistic values within 0 and 1 indicating strongest negative and positive decisions respectively. For better viewing, please see original color PDF file.

- The adaptive hyperplane w_i is the combination of the subject hyperplane w_0 and R rectifications via $\{q_r, u_r\}$'s with the corresponding contributions determined by the context feature x_i^c . Intuitively, we can treat $u_r^T x_i^f$ as a switch to determine whether the context should be activated while the value $q_r^T x_i^c$ determines how to rectify w_0 .
- The refined margin expression corresponds exactly to our probabilistic motivation. The $\{u_r\}$ and $\{q_r\}$ collaboratively model the component $p(y, x_i^c | x_i^f)$ in (3). The decomposition of P helps us better understand that $\{u_r\}$ serve to judge the discrimination ambiguity of x_i^f , and $\{q_r\}$ are utilized to integrate the context feature x_i^c for the classification of the samples with different ambiguities.

3.2.2 Instantiate $\{u_r\}$

As aforementioned, we design $\{u_r\}$ to highlight samples which are classified ambiguously with their subject features $\{x_i^f\}$. Practically, we instantiate $\{u_r\}$ as a set of hyperplanes parallel to a learned hyperplane w_0 in subject feature space by traditional SVM:

$$u_r = \alpha_r w_0 + \beta_r, \quad r = 1, 2, \dots, R. \quad (9)$$

Intuitively, for $\alpha_r > 0$, if we set α_r and β_r properly such that all $\{u_r^T x_i^f\}$ are within $[0, 1]$, those samples classified as negative by w_0 with high confidences shall be suppressed, namely their corresponding values of $\{u_r^T x_i^f\}$ shall be small. At the same time, for $\alpha_r < 0$, if we set α_r and β_r properly such that all $\{u_r^T x_i^f\}$ are within $[0, 1]$, those samples classified as positive by w_0 with high confidences shall be suppressed, namely their corresponding values of $\{u_r^T x_i^f\}$ shall be small. Therefore we can sample multiple

combinations of α_r and β_r , and both strong negative and positive samples shall be suppressed by $\{u_r\}$ such that the samples with ambiguous decisions by w_0 are highlighted.

Our empirical experiments show that using larger R may derive better ambiguity modeling but may also lead to overfitting, and it is a good trade-off by setting $R = 2$, i.e. using two auxiliary hyperplanes u_1 and u_2 and set $\alpha_1 > 0$ and $\alpha_2 < 0$. Then the combination of u_1 and u_2 can provide a rough yet efficient judgment for the decision ambiguity of a sample and force the context model to concentrate on the samples with large ambiguities.

We illustrate one exemplar contextualization result by Context-SVM on object classification task of the ‘‘aeroplane’’ category in Figure 2. This figure shows the adaptive contextualization with respect to the sample ambiguity: the samples with higher ambiguities (i.e. samples lying in the middle of the figure) are changed largely by the contextualization procedure while the well-classified samples (i.e. samples lying on the two sides of the figure) are nearly not affected.

3.2.3 Optimization for Context-SVM

Based on the instantiated $\{u_r\}$, we can formulate the Context-SVM as a max-margin optimization problem with the margin described as the average of the rectified individual margins related to $\|w_i\|$'s, namely,

$$\min_{w_0, \{q_r\}} \frac{1}{2N} \sum_{i=1}^N \|w_i\|_2^2 + C \sum_{i=1}^N \xi_i, \quad (10)$$

$$s.t. \quad y_i(w_i^T x_i^f + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \quad \forall i,$$

where C is a tunable parameter for balancing two items and ξ_i 's are relaxation parameters.

This formulation can be further compiled with respect to $\{q_r\}$ and w_0 as:

$$\min_v \frac{1}{2N} \sum_{i=1}^N v^T U_i^T U_i v + C \sum_{i=1}^N \xi_i, \quad (11)$$

$$s.t. \quad y_i[(U_i v)^T x_i^f + b] - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \quad \forall i,$$

where the matrices $U_i = [I_n, u_1 x_i^{cT}, u_2 x_i^{cT}, \dots, u_R x_i^{cT}]$, $v = [w_0; q_1; q_2; \dots; q_R]$ and I_n is an $n \times n$ identity matrix.

Note that in this optimization problem, there are only $(R \times m + n)$ parameters to optimize, and generally R is small. Therefore the overfitting issue can be well alleviated. It is easy to prove² that (11) can be converted to a standard SVM problem and its solution can be derived with standard SVM solvers.

3.3. Kernel Extension

For many visual understanding problems, image descriptors are further encoded as similarity measurements or kernel matrices, and there is no explicit vector representation

²Details are omitted here due to the space limitation.

for each image. Therefore, it is necessary to generalize the Context-SVM formulation to the case with only kernel matrices available. We consider the problem in a feature space \mathcal{F} induced by certain nonlinear mapping function $\phi : \mathbb{R}^n \rightarrow \mathcal{F}$. For a properly chosen ϕ , an inner product $\langle \cdot, \cdot \rangle$ can be defined on \mathcal{F} which induces a Reproducing Kernel Hilbert Space (RKHS). More specifically, $\langle \phi(x_i^f), \phi(x_j^f) \rangle = \mathcal{K}(x_i^f, x_j^f)$ where $\mathcal{K}(\cdot, \cdot)$ is a positive semi-definite kernel function.

The context-adaptive hyperplane for each sample can be defined as:

$$(w_0 + \sum_{r=1}^R u_r \cdot (q_r^T x_i^c))^T \cdot \phi(x_i^f) + b = 0, \quad (12)$$

which is similar to (7).

By Representer Theorem [22], u_r and w_0 can be expressed as linear combinations of $\{\phi(x_i^f)\}$. Thus, there exist sets of coefficients such that $u_r = \sum_{i=1}^N \beta_{ri} \phi(x_i^f)$ and $w_0 = \sum_{i=1}^N \alpha_i \phi(x_i^f)$. Let $\beta_r = [\beta_{r1}, \dots, \beta_{rN}]^T$, $\alpha = [\alpha_1, \dots, \alpha_N]^T$ and $\Phi(X^f) = [\phi(x_1^f), \dots, \phi(x_N^f)]$. The context-aware hyperplane can then be expressed as:

$$\left(\sum_{r=1}^R \Phi(X^f) \beta_r q_r^T x_i^c + \Phi(X^f) \alpha \right)^T \cdot \phi(x_i^f) + b = 0, \quad (13)$$

namely, $(\sum_{r=1}^R \beta_r q_r^T x_i^c + \alpha)^T \cdot K(:, i) + b = 0$, where K is the kernel matrix with $K_{ij} = \langle \phi(x_i^f), \phi(x_j^f) \rangle$ and $K(:, i)$ is the i -th column vector of the matrix K .

Then the overall formulation for kernel Context-SVM is:

$$\min_z \frac{1}{2N} \sum_{i=1}^N z^T B_i^T K B_i z + C \sum_{i=1}^N \xi_i, \quad (14)$$

$$s.t. \quad y_i[(B_i z)^T K(:, i) + b] - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \quad \forall i,$$

where $B_i = [I_N, \beta_1 x_i^{cT}, \beta_2 x_i^{cT}, \dots, \beta_R x_i^{cT}]$, $z = [\alpha; q_1; q_2; \dots; q_R]$, and I_N is an $N \times N$ identity matrix. The main differences between the kernel version and the linear version include: 1) the original subject feature vector x_i^f is replaced by the column vector of the kernel matrix K , and 2) l_2 regularizer in the objective contains a kernel matrix. Thus, the same optimization approach can be used for solving the kernel extension of Context-SVM.

4. Application: Contextualizing Object Detection and Classification

In this section, we apply the Context-SVM to contextualize two prevalent tasks of image understanding, namely object detection and classification.

4.1. Initializations

The initial object detection and classification models $M_{det}(0)$ and $M_{cls}(0)$ for the first iteration are learned based

Algorithm 1 Contextualizing Classification and Detection

Input:

$M_{det}(0)$: Initial object detection model,
 $M_{cls}(0)$: Initial object classification model,
 $\{I_i\}$: Training images,
 R : Rank of the matrix P .

For $t = 1, 2, \dots, T_{max}$

1. Extract detection features and context for each image,

$$\begin{aligned}x_i^f(t) &\leftarrow \text{extract}(I_i), \forall i, \\x_i^c(t) &\leftarrow \text{eval}(M_{cls}(t-1), I_i), \forall i.\end{aligned}\quad (15)$$

2. Instantiate $\{u_r\}$ with $\{\{x_i^f(t)\}, R\}$ and $M_{det}(t-1)$.
3. Learn $M_{det}(t)$ via Context-SVM on $\{x_i^f(t), x_i^c(t)\}$.
4. Similarly, learn $M_{cls}(t)$ via Context-SVM by using the outputs from $M_{det}(t)$ as context.

EndFor**Output** $M_{det}(T_{max}), M_{cls}(T_{max})$.

on the state-of-the-art algorithms. We follow the part-based model proposed by Felzenswalb et al. [10] for the initial detection model training. The Histogram of Gradient (HOG) [6] and Local Binary Pattern (LBP) [20] features are used for object description and the number of part models for each object category is set to be 6.

For object classification task, the traditional Bag-of-Words (BoW) model [18] is employed. We first extract the low-level features including SIFT and its color variants [21], LBP and HOG by dense sampling strategy in three scales. Each image is represented by BoW model with spatial pyramid matching [15]. The kernel function is based on χ^2 distance for each type of feature, and then all kernels are combined to an average kernel for kernelized Context-SVM.

4.2. Iterative Mutual Contextualization

The detailed algorithm for contextualizing object detection and classification by iterative Context-SVM is listed in Algorithm 1. More specifically, the context features for detection and classification refer to the probabilities of object existence in each image. And each object category is represented in one probabilistic value. Thus the context feature values are within $[0, 1]$ and the dimension of context feature vector is the number of object categories. The context from the object classification task is obtained by converting classification scores on each image to probabilities via sigmoid scaling. And the context features from the object detection task are obtained by converting the detected highest score for each object category to the probability in the same manner as for object classification. If there is no object detected for certain category, the corresponding entry in context feature vector is set as 0.

At the t -th step, the context features of one task (e.g. detection) are obtained by evaluating the $(t-1)$ -th model of the other task (e.g. classification) on the training data $\{I_i\}$.

We use cross validation method to obtain context from object classification in (15) as kernel model is easy to overfit on its training data. 10-fold of training data are used and we evaluate each fold via the model trained on all other folds. Then we instantiate $\{u_r\}$ based on the extracted subject features and the learnt model from the previous step, and finally proceed to conduct Context-SVM based on $\{u_r\}$, subject features and the corresponding context features for all training images.

For training stage of iterative contextualization, the additional computation cost of optimization for the Context-SVM is trivial comparing to the cost of the subject task, i.e. the feature extraction and kernel vector calculation for object classification and the mining of training samples from sub-windows of each image for object detection.

5. Experiments

5.1. Datasets and Metrics

The PASCAL Visual Object Challenge (VOC) datasets [9] are widely used as testbeds for evaluating algorithms for image understanding tasks and provide a common evaluation platform for both object classification and detection. These datasets are extremely challenging since the objects vary significantly in size, view angle, illumination, appearance and pose. We use PASCAL VOC 2007 and 2010 datasets for experiments in this paper. The twenty object categories of VOC datasets are as illustrated in Table 1.

VOC 2007 and VOC 2010 datasets contain 9,963 and 21,738 images respectively. The two datasets are divided into “train”, “val” and “test” subsets, i.e. 25% for training, 25% for validation and 50% for testing. The annotations for the whole dataset of VOC 2007 and “train”, “val” set of VOC 2010 are provided while the annotations for “test” set of VOC 2010 are still confidential and can only be evaluated on the web server with limited trials. The employed evaluation metric is *Average Precision* (AP) complying with the PASCAL challenge rules.

In the following experiments, we first evaluate the performance boosting capability from iterative mutual contextualization on VOC 2010 “train/val” dataset (i.e. “train” set for training and “val” set for test) since frequent evaluations of the performance are required. Then several traditional methods for contextualizing object detection and classification are compared with our iterative Context-SVM on the VOC 2010 trainval/test dataset. Finally, we evaluate the optimal configuration on PASCAL VOC 2007 and 2010 trainval/test datasets and compare it with the state-of-the-art performance ever reported.

5.2. Iterative Performance Boosting via Mutual Contextualization

To evaluate the effectiveness of our proposed iterative mutual contextualization process, we conduct three experi-

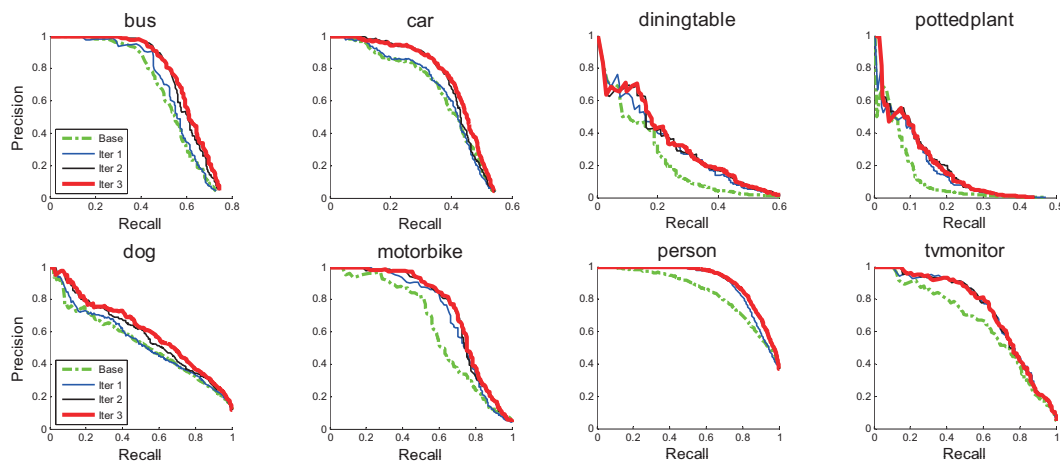


Figure 3: Illustration of performance improvement with comparison Precision-recall curves of object detection (upper row) and classification (lower row). The performance of baseline (without contextualization) and those of Context-SVM at iteration 1-3 are plotted.

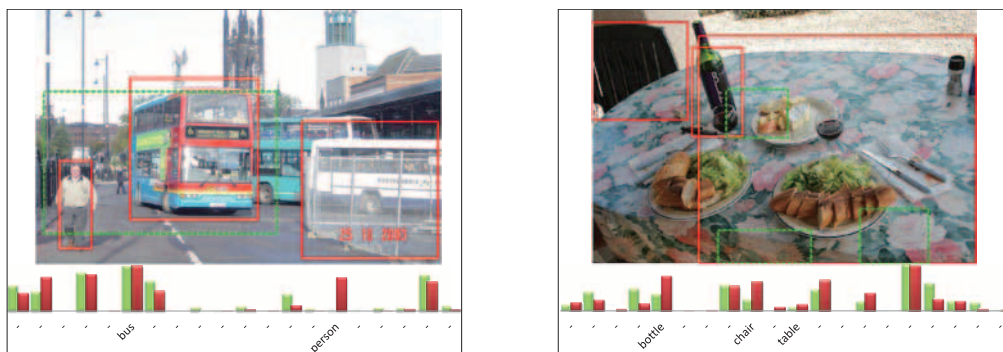


Figure 4: Representative examples of the baseline (without contextualization) and Context-SVM at iteration 3. The detections are shown via the detected bounding boxes on images (with proper threshold): the green boxes with dashed lines denote the false alarms from baseline, which are further removed by contextualization and red boxes denote the true detections of both methods. The classification results are compared by the confidences for each object category before (green) and after (red) contextualization. For better viewing, please see original color PDF file.

ments on VOC 2010 “train/val” dataset. Firstly, we demonstrate the performance improvement measured by mean AP for all the 20 classes in Figure 5. In this experiment, the mutual contextualization is conducted for 3 iterations, and obvious performance improvement is observed for the first and second iteration. As the improvement from the third iteration becomes trivial, we set the maximum iteration number, namely T_{max} to 3 for all the experiments in this work.

In the second experiment, we show exactly how the mutual contextualization process benefits each class by Precision-Recall curves of several representative classes in Figure 3, and also we show the representative object detection and classification results in Figure 4 for the third experiment. As can be observed from Figure 3, great performance improvement can be achieved for the first two iterations and in the 3rd iteration, certain amount of improvement can still be achieved for several classes such as “bus” and “dog”. From Figure 4, it may be observed that the

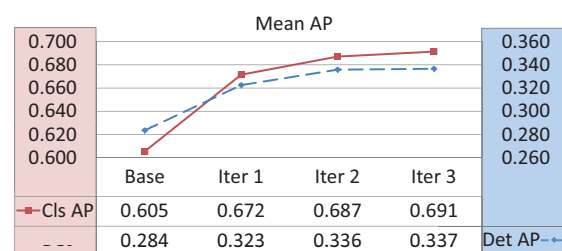


Figure 5: Mean AP values of 20 classes on VOC 2010 train/val dataset along iterative contextualization.

Context-SVM shows good stability in refining the classes even without accurate context such as “pottedplant”. The example detection results show that the improvement of object detection is mainly achieved by effective removal of the ambiguous negatives while the object classification benefits from detection context by calling back those missing objects, e.g. the “person” and “chair” missed in the baseline results as shown in Figure 4.

Table 1: Contextualization method comparison on the PASCAL VOC 2010 (trainval/test) dataset. “Det” and “Cls” respectively denote object detection and classification tasks.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
Det_Fuse [11]	50.5	49.8	16.0	10.4	30.4	54.3	43.3	38.3	15.9	30.0	24.1	23.1	47.8	54.2	42.1	11.8	33.5	27.5	47.3	38.8	34.5
Det_Our Method	53.1	52.7	18.1	13.5	30.7	53.9	43.5	40.3	17.7	31.9	28.0	29.5	52.9	56.6	44.2	12.6	36.2	28.7	50.5	40.7	36.8
Cls_MKL [2]	91.4	76.6	66.7	72.3	53.1	83.7	77.1	75.3	62.9	59.8	57.1	63.6	76.5	81.8	91.2	44.1	64.1	48.4	84.0	75.5	70.3
Cls_Fuse [11]	90.7	74.0	67.2	73.9	53.8	81.7	74.1	73.6	60.9	59.8	60.5	62.3	75.1	80.2	90.4	45.8	61.7	56.0	85.9	76.0	70.2
Cls_Our Method	92.2	77.7	69.2	75.7	53.5	84.7	80.9	76.1	62.8	65.5	63.1	65.6	79.6	83.4	91.2	47.5	71.9	55.2	86.3	76.7	73.0

5.3. Contextualization Methods Comparison

In this subsection, our proposed iterative mutual contextualization method is compared with the method proposed by Harzallah et al. in [11], which combines the confidences from several probabilistic models and is the most representative one among those confidence combination approaches [10] [8]. For object classification, Multiple Kernel Learning (MKL) [2] method used in [14] is also implemented for comparison, which is a general model fusion method and widely used to combine features in kernel form for object classification. An extra linear kernel is constructed for the context features from the object detection task, and then two kernels are combined with MKL. MKL performs very bad for object detection task, and thus we do not report the result of MKL for object detection here. The main reason is that the context is fixed for all candidate windows within an image and the inaccurate context may severely affect the results for quite many candidate windows.

The experiment results are evaluated using the latest VOC 2010 “trainval/test” dataset and shown in Table 1. The comparison results show that the proposed iterative and mutual contextualization method outperforms these two traditional contextualization methods for most object categories.

5.4. Comparison with State-of-the-art Performance

We also compare the proposed contextualization method with the reported state-of-the-art object detection and classification approaches on VOC 2007 and VOC 2010 datasets. The detailed performance comparison is shown in Table 2 and Table 3.

We compare with the best known VOC 2007 performance from several recent papers in Table 2. For object detection, the methods compared include [MIT_2010] by Zhu et al. [27] using latent hierarchical structural learning, [UCI_2009] by Desai et al. [7] using context of object layout, [INRIA_2009] by Harzallah et al. [11] fusing classification scores, and [UoC_2010] by Felzenswalb et al. [10] using part-based model with context of object co-occurrence. For the detection challenge of 2007, our method outperforms 13 classes out of 20 classes and the MAP outperforms the second best [UoC_2010] by 3.6%.

The well-known methods of VOC 2007 object classification task compared are: [INRIA_Genetic] [19] the winner of VOC 2007, [NEC_2010] [26] performing nonlinear

feature transformation on descriptors, [INRIA_2009] fusing detection scores, and [TagModal] [14] using extra tag information of VOC 2007 dataset. Our method significantly outperforms the competing methods for 12 classes out of 20 classes. Note that our MAP achieves leading by 3.8% to the result of [TagModal]. It well validates the effectiveness of the proposed strategy in utilizing detection context for object classification.

For VOC 2010 dataset, we compare with the recently released results from the VOC 2010 challenge [9], which are all obtained through the combinations of multiple methods including mutual combination of detection and classification. Necessary postprocessing is also implemented in these methods. Therefore for a fair comparison, we refine the framework used by Chen et al. in their submission [NUSPSL_KERNELREGFUSING] (NUSPSL) [28] with the following differences: 1) the combination of detection and classification is further refined by the proposed iterative Context-SVM and 2) we exclude the fusion of other learning schemes used in [28] to verify the effectiveness of the Context-SVM.

The comparison results are shown in Table 3, from which we may observe that the classification results from our proposed method outperform in 16 classes out of 20 classes, and 3.3% in mean AP over the second best VOC 2010 submission [NLPR_Context]. Note that the submission [NLPR_Context] combines the best-performed detection results in this challenge for classification. Our proposed method also outperforms the winner submission [NUSPSL] in 12 classes out of 20 classes and achieves the highest mean AP even without the fusion with other learning methods. The object detection results from our proposed method based on Context-SVM also outperform 7 classes out of 20 classes, and our method achieves the highest mean AP together with the winner submission [NLPR_Context], which outperforms 6 classes out of 20 classes in this competition.

6. Conclusions

In this paper, we proposed an iterative contextualization scheme to mutually boost performance of both object detection and classification tasks. We first proposed the so-called Contextualized SVM to seamlessly integrate external context features and subject features for general classification, and then Context-SVM was further utilized to iteratively and mutually boost performance of object detec-

Table 2: Comparison with the state-of-the-art performance of object classification and detection on PASCAL VOC 2007 (trainval/test).

Detection on VOC 2007																					
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
MIT_2010 [27]	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
UCL_2009 [7]	28.8	56.2	3.2	14.2	29.4	38.7	48.7	12.4	16.0	17.7	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.1
INRIA_2009 [11]	35.1	45.6	10.9	12.0	23.2	42.1	50.9	19.0	18.0	31.5	17.2	17.6	49.6	43.1	21.0	18.9	27.3	24.7	29.9	39.7	28.9
UoC_2010 [10]	31.2	61.5	11.9	17.4	27.0	49.1	59.6	23.1	23.0	26.3	24.9	12.9	60.1	51.0	43.2	13.4	18.8	36.2	49.1	43.0	34.1
Our method	38.6	58.7	18.0	18.7	31.8	53.6	56.0	30.6	23.5	31.1	36.6	20.9	62.6	47.9	41.2	18.8	23.5	41.8	53.6	45.3	37.7

Classification on VOC 2007																					
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
INRIA_Genetic [19]	77.5	63.6	56.1	71.9	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.6	79.2	53.2	59.4
SuperVec [26]	79.4	72.5	55.6	73.8	34.0	72.4	83.4	63.6	56.6	52.8	63.2	49.5	80.9	71.9	85.1	36.4	46.5	59.8	83.3	58.9	64.0
INRIA_2009 [11]	77.2	69.3	56.2	66.6	45.5	68.1	83.4	53.6	58.3	51.1	62.2	45.2	78.4	69.7	86.1	52.4	54.4	54.3	75.8	62.1	63.5
TagModal [14]	87.9	65.5	76.3	75.6	31.5	71.3	77.5	79.2	46.2	62.7	41.4	74.6	84.6	76.2	84.6	48.0	67.7	44.3	86.1	52.7	66.7
Our Method	82.5	79.6	64.8	73.4	54.2	75.0	87.5	65.6	62.9	56.4	66.0	53.5	85.0	76.8	91.1	53.9	61.0	67.5	83.6	70.6	70.5

Table 3: Comparison with the state-of-the-art performance of object classification and detection on PASCAL VOC 2010 (trainval/test).

Detection on VOC 2010																					
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
NLPR_Context [9]	53.3	55.3	19.2	21.0	30.0	54.4	46.7	41.2	20.0	31.5	20.7	30.3	48.6	55.3	46.5	10.2	34.4	26.5	50.3	40.3	36.8
MITUCLA [9]	54.2	48.5	15.7	19.2	29.2	55.5	43.5	41.7	16.9	28.5	26.7	30.9	48.3	55.0	41.7	9.7	35.8	30.8	47.2	40.8	36.0
NUS_Context [9]	49.1	52.4	17.8	12.0	30.6	53.5	32.8	37.3	17.7	30.6	27.7	29.5	51.9	56.3	44.2	9.6	14.8	27.9	49.5	38.4	34.2
UVA [9]	56.7	39.8	16.8	12.2	13.8	44.9	36.9	47.7	12.1	26.9	26.5	37.2	42.1	51.9	25.7	12.1	37.8	33.0	41.5	41.7	32.9
Our Method	53.1	52.7	18.1	13.5	30.7	53.9	43.5	40.3	17.7	31.9	28.0	29.5	52.9	56.6	44.2	12.6	36.2	28.7	50.5	40.7	36.8

Classification on VOC 2010																					
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
NLPR_Context [9]	90.3	77.0	65.3	75.0	53.7	85.9	80.4	74.6	62.9	66.2	54.1	66.8	76.1	81.7	89.9	41.6	66.3	57.0	85.0	74.3	71.2
NEC_Nonlin [9]	93.3	72.9	69.9	77.2	47.9	85.6	79.7	79.4	61.7	56.6	61.1	71.1	76.7	79.3	86.8	38.1	63.9	55.8	87.5	72.9	70.9
NUSPSL [28]	93.0	79.0	71.6	77.8	54.3	85.2	78.6	78.8	64.5	64.0	62.7	69.6	82.0	84.4	91.6	48.6	64.9	59.6	89.4	76.4	73.8
Our Method	93.1	78.9	73.2	77.1	54.3	85.3	80.7	78.9	64.5	68.4	64.1	70.3	81.3	83.9	91.5	48.9	72.6	58.2	87.8	76.6	74.5

tion and classification tasks. The proposed solution was extensively evaluated on both PASCAL VOC 2007 and VOC 2010 datasets and achieved the state-of-the-art performance for both tasks.

Acknowledgement

This work is supported by NRF/IDM Program, under research Grant NRF2008IDMIDM004-029.

References

- [1] T. Berg, A. Berg, J. Edwards, and D. Forsyth. Who is in the picture. In *NIPS*, 2006.
- [2] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. simpleMKL. In *JMLR*, 2008.
- [3] T. Berg and D. Forsyth. Animals on the web. In *CVPR*, 2006.
- [4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*
- [5] P. Carbonetto, N. Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.
- [6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.
- [8] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009.
- [9] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 Results. <http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html>.
- [10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010.
- [11] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.
- [12] G. Heitz and D. Koller. Learning Spatial Context: Using Stuff to Find Things In *ECCV*.
- [13] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.
- [14] S. Kumar and M. Hebert. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010.
- [15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*.
- [16] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*.
- [17] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report 7694, California Institute of Technology, 2007.
- [18] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*.
- [19] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. In *Visual Recognition Challenge workshop, ICCV*, 2007.
- [20] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 1995.
- [21] K. Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*.
- [22] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. *Computational Learning Theory*, 2001.
- [23] A. Torralba. Contextual priming for object detection. *IJCV*, 2003.
- [24] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 2006.
- [25] B. Yao and F.-F. Li. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.
- [26] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.
- [27] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.
- [28] Q. Chen, Z. Song, S. Liu, X. Chen, X. Yuan, T.S. Chua, S. Yan, Y. Hua, Z. Huang and S. Shen. Boosting classification with exclusive context. The PASCAL VOC Challenge Workshop, 2010.