Spatial Memory for Context Reasoning in Object Detection

Xinlei Chen  Abhinav Gupta
School of Computer Science, Carnegie Mellon University
{xinleic,abhinavg}@cs.cmu.edu

Abstract

Modeling instance-level context and object-object relationships is extremely challenging. It requires reasoning about bounding boxes of different classes, locations etc. Above all, instance-level spatial reasoning inherently requires modeling conditional distributions on previous detections. Unfortunately, our current object detection systems do not have any memory to remember what to condition on! The state-of-the-art object detectors still detect all object in parallel followed by non-maximal suppression (NMS). While memory has been used for tasks such as captioning, they mostly use image-level memory cells without capturing the spatial layout. On the other hand, modeling object-object relationships requires spatial reasoning – not only do we need a memory to store the spatial layout, but also a effective reasoning module to extract spatial patterns. This paper presents a conceptually simple yet powerful solution – Spatial Memory Network (SMN), to model the instance-level context efficiently and effectively. Our spatial memory essentially assembles object instances back into a pseudo “image” representation that is easy to be fed into another ConvNet for object-object context reasoning. This leads to a new sequential reasoning architecture where image and memory are processed in parallel to obtain detections which update the memory again. We show our SMN direction is promising as it provides 2.2% improvement over baseline Faster RCNN on the COCO dataset so far.

1. Introduction

Context helps image understanding! Apart from strong psychological evidence [4, 37, 64, 65] that context is vital for humans to recognize objects, many empirical studies in the computer vision community [10, 11, 16, 22, 23, 43, 60, 62, 73, 82, 83] have also suggested that recognition algorithms can be improved by proper modeling of context.

But what is the right model for context? Consider the problem of object detection. There are two common models of context often used in the community. The first type of model incorporates image or scene level context [5, 36, 46, 58, 63, 74, 81]. The second type models object-object relationships at instance-level [15, 27, 32, 58, 67, 94]. Take the the top-left image of Fig. 1 as an example, both the person and the tennis racket can be used to create a contextual prior on where the ball should be.

Of these two models, which one is more effective for modeling context? A quick glimpse on the current state-of-the-art approaches, the idea of single region classification [40, 49, 51, 53, 69] with deep ConvNets [33, 76] is still dominating object detection. On the surface, these approaches hardly use any contextual reasoning; but we be-
lieve the large receptive fields of the neurons in fact do incorporate image-level context (See Fig. 1 for evidences). On the other hand, there has been little or no success in modeling object-object relationships or instance-level context in recent years.

Why so? Arguably, modeling the instance-level context is more challenging. Instance-level reasoning for object detection would have to tackle bounding boxes pairs or groups in different classes, locations, scales, aspect ratios, etc. Moreover, for modeling image-level context, the grid structure of pixels allows the number of contextual inputs to be reduced efficiently (e.g. to a local neighborhood [11, 43, 73] or a smaller scale [76, 87]), whereas such reductions for arbitrary instances appear to be not so trivial. Above all, instance-level spatial reasoning inherently requires modeling conditional distributions on previous detections, but our current object detection systems do not have any memory to remember what to condition on! Even in the case of multi-class object detection, the joint layout [15] is estimated by detecting all objects in parallel followed by non-maximal suppression (NMS) [20]. What we need is an object detection system with memory built inside it!

Memory has been successfully used in the recognition community recently for tasks such as captioning [13, 17, 59, 84, 85, 92] or visual question answering [2, 3, 24, 44, 55, 57, 72, 90, 91, 93, 97]. However, these works mostly focus on modeling an image-level memory, without capturing the spatial layout of the understanding so far. On the other hand, modeling object-object relationships requires spatial reasoning – not only do we need a memory to store the spatial layout, but also a suitable reasoning module to extract spatial patterns. This paper presents a conceptually simple yet powerful solution – Spatial Memory Network (SMN), to model the instance-level context efficiently and effectively. Our key insight is that the best spatial reasoning module is a ConvNet itself! In fact, we argue that ConvNets are actually the most generic\(^1\) and effective framework for extracting spatial and contextual information so far! Inspired by this observation, our spatial memory essentially assembles object instances back into a pseudo “image” representation that is easy to be fed into another ConvNet to perform object-object context reasoning.

However, if ConvNets are already so excellent at modeling context, why would we even bother something else? Isn’t the image itself the ultimate source of information and therefore the best form of “spatial memory”? Given an image, shouldn’t an ultra-deep network already take care of the full reasoning inside its architecture? In spite of these valid concerns, we argue that a spatial memory still presents as an important next step for object detection and other related tasks, for the following reasons:

- First, we note that current region-based object detection methods are still treating object detection as a perception problem, not a reasoning problem: the region classifier still produces multiple detection results around an object instance during inference, and relies on manually designed NMS [69] with a pre-defined threshold for de-duplication. This process can be sub-optimal. We show that with a spatial memory that memorizes the already detected objects, it is possible to learn the functionality of NMS automatically.

- Second, replacing NMS is merely a first demonstration for context-based reasoning for object detection. Since the spatial memory is supposed to store both semantic and location information, a legitimate next step would be full context reasoning: i.e., infer the “what” and “where” of other instances based on the current layout of detected objects in the scene. We show evidence for such benefits on COCO [52].

- Third, our spatial memory essentially presents as a general framework to encode instance-level visual knowledge [48], which requires the model to properly handle the spatial (e.g. overlaps) and semantic (e.g. poses) interactions between groups of objects. Our approach follows the spirit of end-to-end learning, optimizing the representation for an end-task – object detection. Both the representation and the idea can be applied to other tasks that require holistic image understanding [3, 42, 98].

\(^1\)Many context models can be built or formulated as ConvNets [89, 96].

2. Related Work

As we already mentioned most related work for context and memory in Sec. 1, in this section we mainly review ideas that use sequential prediction for object detection. A large portion of the literature [28, 45, 56] focuses on sequential approaches for region proposals (i.e., foreground/background classification). The motivation is to relieve the burden for region classifiers by replacing an exhaustive sliding-window search [20] with a smarter and faster search process. In the era of ConvNet-based detectors, such methods usually struggle to keep a delicate balance between efficiency and accuracy, since a convolution based 0/1 classifier (e.g. region proposal network [69]) already achieves an impressive performance when maintaining a reasonable speed. Sequential search has also been used for localizing small landmarks [77], but the per-class model assumes the existence of such objects in an image and lacks the ability to use other categories as context.

Another commonly used trick especially beneficial for reducing localization error is iterative bounding box refinement [25, 26, 69, 95], which leverages local image context to predict a better bounding box iteratively. This line of research is complementary to our SMN, since its goal is to
locate the original instance itself better, whereas our focus is on how to better detect other objects given the current detections.

An interesting recent direction focuses on using deep reinforcement learning (DRL) to optimize the sequence selection problem in detection [6, 9, 50, 61]. However, due to the lack of full supervision signal in a problem with high-dimensional action space\(^2\), DRL has so far only been used for bounding box refinements or knowledge-assisted detection, where the action space is greatly reduced. Nevertheless, SMN can naturally serve as an encoder of the state in a DRL system to directly optimize average precision [34].

Note that the idea of using higher-dimensional memory in vision is not entirely new. It has resemblance to spatial attention, which has been explored in many high-level tasks [47, 68, 91, 92]. To bypass NMS, LSTM [35] cells arranged in 2D order [78] and intersection-over-union (IoU) maps [39] have been used for single-class object detection. We also notice a recent trend in using 2D memory as a map for planning and navigation [31, 66]. Our work extends such efforts into generic, multi-class object detection, performing joint reasoning on both space and semantics.

3. Background: Faster RCNN

Our spatial memory network is agnostic to the choice of base object detection model. In this paper we build SMN on top of Faster R-CNN [69] (FRCNN) as a demonstration, which is a state-of-the-art detector that predicts and classifies Regions of Interest (RoIs). Here we first give a brief review of the approach.

3.1. Base Network

We use VGG16 [76] as the base network for feature extraction. It has 13 convolutional (conv), 5 max-pooling (pool), and 2 fully connected (fc) layers before feeding into the final classifier, and was pre-trained on the ILSVRC challenge [71]. Given an image \(I\) of height \(h\) and width \(w\), feature maps from the last conv layer (conv5_3) are first extracted by FRCNN. The conv5_3 feature size \((h', w')\) is roughly \(\gamma = 1/16\) of the original image in each spatial dimension. On top of it, FRCNN proceeds by allocating two sub-networks for region proposal and region classification.

3.2. Region Proposal

The region proposal network essentially trains a class-agnostic objectness [1] classifier, proposing regions that are likely to have a foreground object in a sliding window manner [20]. It consists of 3 conv layers, one maps from conv5_3 to a suitable representation for RoI proposals, and two \(1 \times 1\) siblings on top of this representation for foreground/background classification and bounding box regression. Note that at each location, anchor boxes [69] of multiple scales \((s)\) and aspect ratios \((r)\) are used to cover a dense sampling of possible windows. Therefore the total number of proposed boxes is \(K \approx h' \times w' \times s \times r^3\). During training and testing, \(k \ll K\) regions are selected by this network as candidates for the second-stage region classification.

3.3. Region Classification

Since the base network is originally an image classifier, region classification network inherits most usable parts of VGG16, with two caveats. First, because RoI proposals can be arbitrary rectangular bounding boxes, RoI pooling [26, 40] is used in place of pool on conv5_3 to match the the square-sized \((7 \times 7)\) input requirement for fc6. Second, the 1,000-way fc layer for ILSVRC classification is replaced by two fc layers for \(C\)-way classification and bounding box regression respectively. Each of the \(C\) classes gets a separate bounding box regressor.

3.4. De-duplication

We want to point out the often-neglected fact that a standard post-processing step is used in almost all detectors [20, 49, 53, 69] to disambiguate detections – NMS. For FRCNN, NMS takes place in both stages. First, for region proposals, it prunes out the overlapping RoIs that are likely corresponding to the same object (“one-for-all-class”) to train the region classifier. Second, for the final detection results, NMS is applied in an isolated, per-class manner (“one-for-each-class”). In this paper, we still use NMS for RoI sampling during training [12], and mainly focus on building a model to replace the per-class NMS, with the hope that the model can encode the rich interplay across multiple classes when suppressing redundant detections.

4. Spatial Memory Network

To better motivate the use of spatial memory network, we resort to a mathematical formulation of the task at hand. For object detection, the goal is to jointly infer and detect all the object instances \(O = \{O_1, O_2, O_3, \cdots, O_N\}\) given an image \(I\), where \(N\) is the maximum number of object instances for any image\(^4\). Then the objective function of training a model (e.g. FRCNN) \(M\) to maximize the log-likelihood:

\[
\arg \max_M \mathcal{L} = \log \mathbb{P}(O_{1:N} | M, I) = \sum_{n=1}^N \log \mathbb{P}(O_n | O_{0:n-1}, M, I), \tag{1}
\]

where \(O_{0:n-1}\) is short for \([O_1, O_2, O_3, \cdots, O_n-1]\) and \(O_{0:0}\) is an empty set. Note that this decomposition of the joint

---

\(^2\)Jointly reason about all bounding boxes and all classes.

\(^4\)O_o denotes both the class and location of the object instance. When there is not enough foreground objects, the sequence can be padded with the background class.
With the above formulation, the inference procedure for object detection is as follows: Once an object instance is detected, each detected object triggers a local detector (Sec. 4.1) to output a context representation, which is then concatenated to the previous context output in the memory. This process goes on until a fixed number of iterations have reached (See Fig. 2 for an overview).

We now describe each module, beginning with a description of the memory itself.

### 4.1. Memory

Different from previous works that either mixes memory with computation [14, 18, 35] or mimics the one-dimensional memory in the Turing machine/von Neumann architecture [86], we would like to build a two-dimensional memory for images. This is intuitive because images are intrinsically 2D mappings of the 3D visual world. But more importantly, we aim to leverage the power of ConvNets for context reasoning, which “forces” us to provide an image-like 2D input.

How big the memory should be spatially? For object detection, FRCNN that operates entirely on conv5_3 features can already retrieve even tiny objects (e.g., the ones in Fig. 1), suggesting that a resolution 1/16 of the full image strikes a reasonable balance between speed and accuracy. At each location, the memory cell is a \(D=256\) dimensional vector that stores the visual information discovered so far. Ideally, the initial values within the memory should capture the photographic bias of a natural image, i.e., prior about where a certain object tend to occur (e.g., sun is more likely to occur in the upper part). But the prior cannot be dependent on the input image size. To this end, we simply initialize the memory with a fixed spatial size \((20 \times 20 \times 256)\) cells, and resize it according to the incoming conv5_3 size using bilinear interpolation. In this way, the memory is fully utilized to learn the prior, regardless of different image sizes.

### 4.2. Indexing

The most difficult problem that previous works [29, 79] face when building an differentiable external memory is the design of memory indexing. The core problem is which memory cell to write to for what inputs. Luckily for our problem, strong correspondence between memory and 2D images solves this problem. Specifically, the target regions to look up in 2D memory are already provided. Furthermore, RoI pooling [26, 40] is precisely the operations needed to read off from the spatial memory\(^6\). The only remaining task is to create a write function that updates the

---

\(^6\)Although RoI pooling only computes partial gradients, backpropagation w.r.t. bounding box coordinates are not entirely necessary [69] and previously found unstable [40].
memory given a detection. This can be divided into two parts, “what” (Sec. 4.3), and “how” (Sec. 4.4).

4.3. Input Features

It may appear trivial, but the decision of what features to insert into the memory requires careful deliberation. First, since conv5\_3 feature preserves spatial information, we need to incorporate it. Specifically, we use RoI pooling (without taking Max) to obtain the feature map at the location, and resize it to 14×14. However, merely having conv5\_3 is not sufficient to capture the higher-level semantic information, especially pertaining which object class is detected. The detection score is particularly useful for disambiguation when two objects occur in the same region, e.g., a person riding a horse. Therefore, we also include fc8 SoftMax score as an input, which is appended at each conv5\_3 locations and followed by two 1x1 conv layers to fuse the information (see Fig. 3). We choose the full score over a one-hot class vector, because it is more robust to false detections.

4.4. Writing

Given the region location and the input features \( x_n \), we update the corresponding memory cells with a convolutional Gated Recurrent Unit [14] (GRU), which uses 3x3 conv filters in place of fc layers as weights. The GRU has a reset gate, and an update gate, shared at each location and activated with Sigmoid function \( \sigma(\cdot) \). Hyperbolic tangent \( \tanh(\cdot) \) is used to constrain the memory values between \( -1 \) and \( 1 \). For alignment, the region from the original memory \( S_{n-1} \) is also cropped with the same RoI pooling operation to 14×14. After GRU, the new memory cells are placed back to \( S_n \) with a reverse RoI operation.

4.5. Context Model

Now that the detected objects are encoded in the memory, all we have to do for context reasoning is stacking another ConNet on the top. In the current setup, we use a simple 5-layer all-convolutional network to extract the spatial patterns. Each conv filter has a spatial size of 3x3, and channel size of 256. Padding is added to keep the final layer m-conv5 same size of conv5\_3. To ease back-propagation, we add residual connections [33] every two layers.

4.6. Output

As for the module that outputs the reasoning results, we treat m-conv5 exactly the same way as conv5\_3 in FRCNN: 3 conv layers for region proposal, and 2 fc layers with RoI pooling for region classification. The fc layers have 2048 neurons each.

We design another residual architecture to combine the memory scores with the FRCNN scores (see Fig. 4): in the first iteration when the memory is empty, we only use FRCNN for detection; from the second iteration on, we add the memory predictions on top of the FRCNN ones, so that the memory essentially provides the additional context to close the gap. This design allows a handy visualization of the prediction difference with/without context. But more importantly, such an architecture is critical to let us converge the full network. Details for this are covered in Sec. 5.1.

4.7. Selecting Next Region

Since spatial memory turns object detection into a sequential prediction problem, an important decision to make is which region to take-in next [7]. Intuitively, some objects are more useful serving as context for others (e.g. person) [21, 30, 32, 94], and some object instances are easier to detect and less prone to consequent errors. However, in this paper we simply follow a greedy strategy – the most confident foreground object box is selected to update the memory, leaving more advanced models that directly optimize the sequence [80] as future work.
Figure 5: Four design choices for learning the functionality of de-duplication. \( M \) is FRCNN features, and \( S_{n-1} \) represents memory features. Each design is shown by two gray panels showing the information flow of Iteration 0 (left) and Iteration \( n>0 \) (right). We find it hard to even converge the network when the gradient is back-propagated to FRCNN in all iterations (a) & (b). Stop the gradient in later iterations (c) can successfully converge the network, and our final design (d) separates perception from reasoning and makes it easy to visualize the effect of context. All design choices are abstract and apply to both region proposal and classification. Please see Sec. 5.1 for more details.

5. Training the Spatial Memory

Like a standard network with recurrent connections, our SMN is trained by back-propagation through time (BPTT) [88], which unrolls the network multiple times before executing a weight-update. However, apart from the well-known gradient propagation issue, imposing the conditional structure on object detection incurs new challenges for training. Interestingly, the most difficult one we face in our experiment, is the “straightforward” task of de-duplication.

5.1. Learning De-duplication

Simply put, the functionality of de-duplication is: how can the network learn that a detected instance should no longer be detected again? More specifically, we need to design the output module (Sec. 4.6) to fuse the memory (\( S \)) and FRCNN (\( M \)) beliefs and predict intelligently: when the memory is empty, the FRCNN score should be used; but when the memory has the instance stored, the network needs to ignore, or negate the cue from FRCNN.

Since multi-layer networks are universal function approximators [38], our first attempt is to fuse the information by directly feeding into a multi-layer network (Fig. 5 (a)). However, joint-training fails to even converge FRCNN. Suspicious that the longer, weaker supervision might be the cause, we also added skip connections [5] to guide the FRCNN training directly (Fig. 5 (b)). Yet it still does not help much. Tracking the learning process, we find where the actual problem lies – because the network needs to de-duplicate, it keeps receiving contradicting signals: the normal one that guides perception, and the adversarial one that prevents more perception. And because \( S \) also starts off from scratch, the signal it can provide is also weak and unreliable. As a result, part of both error signals are back-propagated to \( M' \), causing trouble for learning further.

Realizing where the issue is, a direct solution is to just stop the adversarial signal from flowing back and canceling the normal one. Therefore, we stopped the gradient to FRCNN from second iteration on (Fig. 5 (c)), and the network can successfully converge.

To make it easy for training and showing the confidence changes for consequent detections given the context, we further reduced the architecture to exclude all memory related weights in the first iteration (Fig. 5 (d)). This way, the change in predictions with/without memory can be read-off directly\(^7\), and training can be done separately for \( M \) and \( S \).

5.2. RoI Sampling

To avoid getting overwhelmed by negative boxes, FRCNN enforces a target sampling ratio for foreground/background boxes. The introduction of a spatial memory that learns to de-duplicate, brings in another special type – regions whose label is flipped from previous iterations. To keep these regions from being buried in negative examples too, we changed the sampling distribution to include flipped regions.

It is important to point out that RoI sampling greatly enhances the robustness of our sequential detection system. Because only \( k<<K \) regions are sampled from all regions, the overall most confident RoI is not guaranteed to be picked when updating the memory. This opens up chances for other highly confident boxes to be inserted into the sequence as well [80] and reduces over-fitting.

5.3. Multi-Tasking

We also practiced the idea of multi-task learning for SMN. The major motivation is to force the memory to memorize more: the basic SMN is only asked fulfill the mission of predicting the missing objects, which does not necessarily translate to a good memorization of previously detected objects. E.g., it may remember that one region has an object in general, but does not store more categorical information beyond that. To better converge the memory, we also added a reconstruction loss [13, 70], i.e., letting the network in addition predict the object classes it has stored in the memory. Specifically, we add an identical set of branches on top of

\[^7\]Since there are two sets of scores (from \( M \) and fused \( f.c \)) added together for prediction in Fig. 5 (b), we find the conflicting signals are also propagated to the biases of these predictions: resulting in one going up and the other down while essentially canceling each other.

\[^8\]Otherwise we have to run the inference again with \( S_0 \).
the m-conv5 features as FRCNN, for both region proposal and region classification in each iteration. These weights are used to predict only the previously detected objects.

5.4. Stage-wise Training

Thanks to the design of our memory augmented prediction, so far we have trained the full model in two separate stages, where FRCNN M, the perception model can be optimized independently at first; then the reasoning model with spatial memory S is learned on top of fixed M. This helps us isolate the influence of the base model and focus directly on the study of SMN.

For efficiency, we also follow a curriculum learning [8] strategy: bootstrap a SMN of more iterations (e.g. \( N=10 \)) with a pre-trained SMN of fewer iterations (e.g. \( N=5 \)). As \( N \) gets larger, the task becomes harder. Curriculum learning does not require re-learning de-duplication (which we learn with \( N \) from 2 to 4), and allows the network to focus more on object-object relationships instead.

5.5. Hyper-parameters

Given a pre-trained FRCNN or SMN (in the case of curriculum learning), we train a fixed number of 30k steps. The initial learning rate is set to \( 1\cdot10^{-3} \) and reduced to \( 1\cdot10^{-4} \) after 20k steps. Since we do not use automatic normalization tricks [41, 54], different variances are manually set when initializing weights from scratch, in order to let different inputs contribute comparably (e.g. when concatenating fc7 and m-fc7). Other hyper-parameters are kept the same to the ones used in FRCNN.

6. Experimental Results

We highlight the performance of our spatial memory network on COCO [52]. However, for ablative analysis and understanding the behaviour of our system, we use both PASCAL VOC 2007 [19] and COCO [52]. For VOC we use the trainval split for training, and test for evaluation. For COCO we use trainval35k [5] and minival. For evaluation, toolkits provided by the respective dataset are used. The main metrics (mAP, AP and AR) are based on detection average precision/recall.

**Implementation Details:** We use TensorFlow to implement our model, which is built on top of the open-sourced FRCNN implementation \(^9\) serving as a baseline. For COCO, this implementation has an AP of 29.1% compared to the original one 24.2% [69].

Original FRCNN uses NMS for region sampling as well. However, NMS hurts our performance more since we do sequential prediction and one miss along the chain can negatively impact all the follow-up detections. To overcome this disadvantage, we would ideally like to examine all \( K \) regions in a sliding window fashion. However, due to the GPU memory limit, the top 5k regions are used instead. We analyze this choice in ablative analysis (Sec. 6.2). Due to the same limitation, our current implementation of SMN can only unroll \( N=10 \) times in a single GPU. At each timestep in SMN, we do a soft max-prediction for the top box selected, so that a single box can be assigned to multiple classes. We will also justify and analyze this choice in Sec. 6.2.

**Initial Results:** Table 1 shows the initial results of our approach as described. As it can be seen for \( N=5 \) detections per image our SMN give an AP of 24.5% and for \( N=10 \) if gives an AP of 28.1%. When the baseline is allowed the same number of detections \((N=5, 10)\), the AP is 23.8% and 27.1%. Therefore, while we do outperform baseline for fixed number of detections per image, due to limited roll-out capability we are still \( \sim 1\% \) below the baseline [12].

### Table 1: Baseline and initial analysis on COCO 2014 mini-val when constraining the number of detections \( N=5/10 \).

| \( N \) | Method | AP | AR-10 | AR-S | AR-M | AR-L |
|-------|--------|----|-------|------|------|------|
| -     | FRCNN [69] | 24.2 | 33.7 | 11.7 | 39.5 | 54.1 |
| -     | Baseline [12] | 29.1 | 38.7 | 17.7 | 44.9 | 56.9 |
| \( N=5 \) | Baseline | 23.8 | 27.8 | 7.0 | 28.7 | 48.4 |
|       | SMN | 24.5 | 28.9 | 7.3 | 29.7 | 50.6 |
| \( N=10 \) | Baseline | 27.1 | 33.5 | 10.8 | 36.7 | 53.8 |
|       | SMN | 28.1 | 35.0 | 11.5 | 38.1 | 56.4 |

**GPU memory limit, the top 5k regions are used instead. We analyze this choice in ablative analysis (Sec. 6.2). Due to the same limitation, our current implementation of SMN can only unroll \( N=10 \) times in a single GPU. At each timestep in SMN, we do a soft max-prediction for the top box selected, so that a single box can be assigned to multiple classes. We will also justify and analyze this choice in Sec. 6.2.**

**6.1. SMN for Hard Examples**

In this section, we want to go beyond \( N=10 \) detections and see if the overall detection performance can be improved with SMN. Intuitively, for highly confident detections, ConvNet-based FRCNN is already doing a decent job and not much can be learned from an additional memory. It is the “tails” that need help from the context! This means two things: 1) with a limited resource budget, SMN should be used in later iterations to provide conditional information; and 2) at the beginning of the sequence, a standard FRCNN can work as a proxy. Given these insights, we experimented with the following strategy: For the first \( N_1 \) iterations, we use a standard FRCNN to detect easier objects and feed the memory with a sequence ordered by FRCNN confidence (after per-class NMS). Memory gets updated as objects come in, but does not output features to augment prediction. Only for the later \( N_2 \) iterations it acts normally as a context provider to detect harder examples. For COCO, we set \( N_1=50 \) and bootstrap from a \( N_2=10 \) SMN model.

Although SMN is trained with the goal of context reasoning and learns new functionality (e.g. de-duplication) that the original FRCNN does not have, it does have introduced more parameters for memory-augmented prediction. Therefore, we also add a MLP baseline, where a 5-layer ConvNet (Sec. 4.5) is directly stacked on top of

---

\(^9\)https://github.com/endernewton/tf-faster-rcnn
Table 2: Final comparison between SMN and baselines. We additionally include MLP baseline where the number of parameters are kept the same as SMN for context aggregation and output. Top 5k regions are used to select proposal instead of NMS.

| Method         | AP   | AP-.5 | AP-.75 | AP-S | AP-M | AP-L | AR-S | AR-M | AR-L |
|----------------|------|-------|--------|------|------|------|------|------|------|
| Baseline [12]  | 29.4 | 50.0  | 30.9   | 12.2 | 33.7 | 43.8 | 18.5 | 45.5 | 58.9 |
| MLP            | 30.1 | 50.8  | 31.7   | 12.5 | 34.2 | 44.5 | 19.2 | 47.0 | 59.8 |
| SMN            | 31.6 | 52.2  | 33.2   | 14.4 | 35.7 | 45.8 | 20.5 | 48.8 | 55.8 |

Table 3: Ablative analysis on VOC 2007 test and COCO 2014 minival. All approaches constrained by detections N=5/10. mAP is used to evaluate VOC, AP and AR numbers are from COCO.

| N   | Method                      | mAP | AP      | AR-10 | AR-S  | AR-M  | AR-L  |
|-----|-----------------------------|-----|---------|-------|-------|-------|-------|
| 5   | Baseline (FR-CNN)           | 65.8| 23.6    | 27.6  | 7.0   | 29.1  | 47.4  |
|     | SMN Base                    | 63.6| 23.3    | 27.2  | 6.7   | 28.0  | 46.1  |
|     | + Sample Flipped            | 64.4| 23.5    | 27.2  | 6.9   | 28.4  | 46.4  |
|     | SMN Full                    | 64.6| 23.8    | 27.7  | 6.9   | 28.5  | 47.4  |
| 10  | Baseline                   | 70.3| 26.9    | 33.2  | 10.9  | 36.6  | 52.7  |
|     | SMN Full                   | 67.5| 26.6    | 32.6  | 10.3  | 35.6  | 52.1  |
|     | + Tune from N=5            | 67.8| 27.1    | 32.7  | 10.3  | 35.9  | 52.3  |

Table 4: Investigating the recall issue. S stands for SoftMax based testing, and H for HardMax. N is short for Non-aggressive NMS, where top 5k RoIs are directly selected without NMS.

| N   | Method | N | Max | mAP | AP  | AR-10 | AR-S  | AR-M  | AR-L  |
|-----|--------|---|-----|-----|-----|-------|-------|-------|-------|
| 5   | Baseline | × | S   | 65.8| 23.6| 27.6  | 7.0   | 29.1  | 47.4  |
|     | SMN Full | × | S   | 66.4| 24.1| 28.8  | 7.5   | 29.7  | 50.0  |
| 10  | Baseline | × | H   | 65.4| 23.5| 27.2  | 6.7   | 28.4  | 46.9  |
|     | SMN Full | × | H   | 64.6| 23.8| 27.7  | 6.9   | 28.5  | 47.4  |
|     | Baseline | √ | S   | 66.0| 23.8| 27.8  | 7.0   | 28.7  | 48.4  |
|     | SMN Full | √ | S   | 66.6| 24.5| 28.9  | 7.3   | 29.7  | 50.6  |

Figure 6: Examples of context has helped improve scores by reasoning. Left: the score of sheep is increased due to presence of other sheep in background. Right: the score of horse is decreased due to the detection of cake and table. conv5_3 for context aggregation, and the same output modules (Sec. 4.6) are used to make predictions.

The results can be found in Table 2. As can be seen, on our final system, we are 2.2% better than the baseline FRCNN. This demonstrates our ability to find hard examples. It is worth noting that here hard does not necessarily translate to small. In fact, our reasoning system also helps big objects, potentially due to its ability to perform de-duplication more intelligently and benefit larger objects that are more likely to overlap.

Qualitative Results: We show a couple of examples of how context using spatial memory can help improve the performance and detections. In the first case, the score of sheep gets boosted due to other sheep. The score of horse decreases due to the detection of cake and table. Please check the supplementary material for more examples.

6.2. Ablative Analysis

We now perform ablative analysis to explain all our choices for the final implementation. For ablative analysis, we use both VOC and COCO datasets. The numbers are summarized in Table 3. For the comparisons shown here, we switch back to the standard NMS-based region sampling and select top k=300 RoIs as in original FRCNN. Also, when we do the roll-out, at each step we choose one detection and perform HardMax (rather than SoftMax): make the hard decision about what class does the selected box belong to – a natural idea for sequential prediction.

For N=5, we compared three models. First, SMN Base, where we simply train the network as is done in FRCNN. Next, regions with flipped labels (Sec. 5.2) are added to replace some of the negative example – for training region proposal the ratio for positive/flipped/negative is 2:1:1, and for region classification it is 1:1:2. Third, SMN Full, where we keep the previous sampling strategy and in addition include the reconstruction loss (Sec. 5.3). Overall, both strategies help performance but with a seemingly different strength: sampling flipped regions helps more on small objects, and multi-task learning helps more on bigger ones.

However, our best performance in Table 3 is still behind the baseline and judging from the COCO AR we believe the biggest issue lies in recall. Therefore, we take the best SMN Full model and conduct two other investigations specifically targeting recall. Here we only list the final results, please see supplementary material for more discussions.

SoftMax vs. HardMax: First, we address a subtle question: if we take top N detections with the memory and compare them directly with top N detections of Faster R-CNN: are these results comparable? It turns out to be not! As mentioned in Sec. 3.4, because NMS is applied in a per-class manner, the actual number of box candidates it can put in the final detection is $k \times C$. To make it more clear, for a confusing region where e.g. the belief for laptop is 40% and keyboard is 35%, NMS can keep both candidates in the top N detections, whereas for SMN it can only keep the max-
imum one\textsuperscript{10}. Therefore, to be fair, we try: a) HardMax for baseline; and b) SoftMax for SMN.

**Non-aggressive NMS:** Finally, we also evaluate our choice of non-aggressive NMS during RoI sampling. Both baseline and SMN perform better with 5k proposals; however our boost on AP is more significant due to sequential prediction issues.

### 7. Conclusion

This paper is our first step towards instance-level reasoning in object detection with ConvNets. We introduce a simple yet powerful framework of spatial memory network, to model the instance-level context efficiently and effectively. Our spatial memory essentially assembles object instances back into a pseudo “image” representation. This memory can simply be fed into another ConvNet to extract context information and perform object-relationship reasoning. We show our SMN direction is promising as it provides 2.2% improvement over baseline Faster RCNN on the COCO dataset so far. We believe our framework is generic and should promote research focusing on knowledge-based reasoning on images.

### Acknowledgements

XC would like to thank Jiwei Li for helpful discussions on reinforcement learning, and everyone who took time to provide useful feedback. This research is supported in part by ONR MURI N000141612007, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) and hardware donations from NVIDIA. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of ODNI, IARPA or the US government. The US Government is authorized to reproduce and distribute the reprints for governmental purposes notwithstanding any copyright annotation thereon.

### References

\textsuperscript{10}It’s also a result of our current input feature design, where we only used fc8 and conv5_3 features to update the memory without a top-down notion \cite{75} of which class is picked, so there’s no more need for SMN to return.

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 34(11):2189–2202, 2012.

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv:1601.01705*, 2016.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[4] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[5] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016.

[6] M. Bellver, X. Giró-i Nieto, F. Marqués, and J. Torres. Hierarchical object detection with deep reinforcement learning. *arXiv:1611.03718*, 2016.

[7] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[9] J. C. Caicedo and S. Lazebnik. Active object localization with deep reinforcement learning. In *ICCV*, 2015.

[10] P. Carbonetto, N. De Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[12] X. Chen and A. Gupta. An implementation of faster rcnn with study for region sampling. *arXiv:1702.02138*, 2017.

[13] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*, 2014.

[15] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95(1):1–12, 2011.

[16] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[17] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv:1411.4389*, 2014.

[18] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[19] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[21] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and A. Zisserman. People watching: Human actions as a cue for single view geometry. *IJCV*, 110(3):259–274, 2014.

[22] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *CVIU*, 114(6):712–722, 2010.

[23] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[24] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015.
[67] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In ICCV, 2007. 1
[68] M. Ren and R. S. Zemel. End-to-end instance segmentation and counting with recurrent attention. arXiv:1605.09410, 2016. 3
[69] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv:1506.01497, 2015. 1, 2, 3, 4, 7, 13
[70] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In ECCV, 2016. 6
[71] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. IJCV, 115(3):211–252, 2015. 3
[72] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In CVPR, 2016. 2
[73] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In ECCV, 2006. 1, 2
[74] A. Shrivastava and A. Gupta. Contextual priming and feedback for faster r-cnn. In ECCV, 2016. 1
[75] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta. Beyond Skip Connections: Top-Down Modulation for Object Detection. arXiv:1612.06851, 2016. 9
[76] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014. 1, 2, 3
[77] S. Singh, D. Hoiem, and D. Forsyth. Learning a sequential search for landmarks. In CVPR, 2015. 2
[78] R. Stewart, M. Andriluka, and A. Y. Ng. End-to-end people detection in crowded scenes. In CVPR, 2016. 3
[79] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In NIPS, 2015. 4
[80] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In NIPS, 1999. 5, 6
[81] A. Torralba. Contextual priming for object detection. IJCV, 53(2):169–191, 2003. 1
[82] A. Torralba, K. P. Murphy, W. T. Freeman, M. A. Rubin, et al. Context-based vision system for place and object recognition. In ICCV, 2003. 1
[83] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. TPAMI, 32(10):1744–1757, 2010. 1
[84] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In ICCV, 2015. 2
[85] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. arXiv:1411.4555, 2014. 2
[86] J. Von Neumann. First draft of a report on the edvac. IEEE Annals of the History of Computing, 15(4):27–75, 1993. 4
[87] P. Weinaeupfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In ICCV, 2013. 2
[88] R. J. Williams and D. Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity. Back-propagation: Theory, architectures and applications, pages 433–486, 1995. 6
[89] S. Xie, X. Huang, and Z. Tu. Top-down learning for structured labeling with convolutional pseudoprior. In ECCV, 2016. 2
[90] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. arXiv, 1603, 2016. 2
[91] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In ECCV, 2016. 2, 3
[92] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv:1502.03044, 2015. 2, 3
[93] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, 2016. 2
[94] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In CVPR, 2010. 1, 5
[95] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon. Attentionet: Aggregating weak directions for accurate object detection. In ICCV, 2015. 2
[96] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In ICCV, 2015. 2
[97] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In CVPR, 2016. 2
[98] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. arXiv:1507.05670, 2015. 2
A1. More Qualitative Results

We show more qualitative results in Figure 7 to present how the current version of SMN reasons with context for object detection. On the top left, the confidence of skis is increased due to the detection of person and their relative location. On the top right, the confidence of tennis racket is increased despite the motion blur owing to the person and her pose. Similarly, the backpack on the middle left gains confidence due to the person carrying it. On the middle right, we show the example of an occluded sheep detection. SMN is able to go beyond the overlapping reasoning of NMS, which will prevent the sheep in the back from being detected. On the bottom row, we show two failure examples, where the potatoes are mistaken as pizza in the container (left), and the suppression of baby (person) given the person holding it.

A2. Category-wise Ablative Analysis on VOC

Due to space limit, we excluded category-wise numbers in the main paper. However, we believe it is interesting to check the category-wise numbers and get a more insightful idea for the ablative analysis part (Sec. 6.2, main paper).

In Table 3, we listed our ablative analysis on different training strategies, but the best mAP we can reach on VOC is still behind the baseline: for N=5 it is 64.6% compared to 65.8%; for N=10 it is 67.8% compared to 70.3%. Judging from the COCO AR metrics, we speculate the issue lies in recall. This motivates us to take the best SMN Full model and conduct an investigation specifically targeting recall.

Turns out, the biggest issue lies in the SoftMax vs. HardMax strategy. For SMN, we initially deployed a HardMax one: given a bounding box, we proceed with the most confident class (and take the bounding box after regression corresponding to that particular class). This ignores all the rest classes that are potentially competitive. E.g., on COCO snowboard is usually confused with skis when buried in the snow; hot-dog is usually confused with pizza when held in a person’s hands. Note that this does not cause a problem for NMS, because the de-duplication is done in a per-class manner. Therefore, for a confusing bounding box where e.g. the belief for laptop is 50% and keyboard is 35%, NMS can keep both candidates in the top N detections, whereas for SMN keyboard is suppressed because laptop has a higher confidence. While in theory this is not an issue because SMN can learn to revisit the same region, we find it extremely unlikely to happen in practice, mainly due to the heavy burden on SMN to learn de-duplication automatically and may also attribute to our current feature design. To investigate whether it is indeed the case, we added two ablative experiments: a) using HardMax strategy for baseline with NMS, meaning an initial proposal can only be selected once – by the most likely class; and b) using Soft-

Figure 7: Four successful reasoning examples and two failure cases. Please see Sec.A1 for a detailed explanation.

Max for SMN where the final N detections can come from all classes of the N bounding boxes returned by sequential prediction.

As shown in Table 4, we find it worked in both ways: the recall indeed boosts for SMN when a SoftMax strategy is used; and removing the confusing categories for the same bounding box hurts the recall for NMS. For COCO, the improvement on AR metrics directly reflects this finding, however it is less obvious for VOC. Here we additionally include evidence from category-wise results in Table 5 to corroborate the observation. For example, categories like cow and sheep get consistent improvements in SMN since they are more likely to confuse, whereas distinctive categories like person almost remain the same. Overall SoftMax outperforms HardMax in most cases.

Normally, k=300 RoIs are selected by NMS (i.e. region proposal) before feeding into region classification. However, SMN as a sequential prediction method is more vulnerable to such an aggressive region selection scheme, because one miss along the chain can negatively impact all the follow-up detections. Therefore, in addition to the two strategies, we also include the analysis on the impact of the number of regions sampled. Specifically, we include a non-aggressive NMS scheme, where the top 5k proposals are directly selected without NMS.
Table 5: VOC 2007 test object detection average precision. S stands for SoftMax based testing, and H for HardMax. N is short for Non-aggressive NMS, where top 5k RoIs are directly selected without NMS.

| N  | Method       | N  | Max | mAP | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | persn | plant | sheep | sofa | train | tv  |
|----|--------------|----|-----|-----|-----|-----|-----|-----|--------|-----|-----|-----|------|-----|------|-----|------|------|------|------|------|-----|------|-----|
| 5  | Baseline [12]| ×  | S   | 65.8| 66.8| 71.3| 66.1| 50.0| 42.0   | 74.5| 79.6| 79.3| 42.4 | 76.3| 58.0 | 77.3| 79.4 | 69.1  | 70.1 | 42.5 | 60.0 | 64.8 | 75.0 | 72.2 |
|    | SMN Full     | ×  | S   | 66.4| 66.3| 75.3| 65.4| 53.3| 42.2   | 74.1| 79.5| 81.9| 44.5 | 72.7| 61.9 | 76.5 | 77.0 | 69.7  | 70.1 | 41.8 | 63.9 | 65.5 | 71.2 | 71.6 |
| 5  | Baseline     | H  | H   | 65.8| 67.3| 71.3| 60.1| 50.0| 41.9   | 74.6| 79.6| 79.3| 42.4 | 76.4| 57.9 | 77.2 | 79.4 | 69.3  | 70.1 | 42.5 | 60.0 | 64.8 | 75.1 | 68.6 |
|    | SMN Full     | H  | H   | 64.6| 60.7| 71.4| 65.6| 50.9| 42.2   | 74.2| 79.4| 78.9| 42.3 | 67.9| 59.9 | 79.0 | 72.4 | 69.5  | 70.0 | 41.8 | 59.6 | 65.5 | 75.0 | 68.1 |
| 10 | Baseline     | ×  | S   | 66.0| 67.4| 71.2| 66.8| 51.5| 41.7   | 75.3| 79.8| 79.1| 43.3 | 76.7| 57.6 | 76.6 | 79.6 | 69.8  | 70.3 | 41.2 | 60.3 | 64.7 | 76.0 | 71.5 |
|    | SMN Full     | ×  | S   | 66.6| 65.5| 71.4| 66.1| 54.6| 41.7   | 75.2| 79.6| 82.4| 45.8 | 75.4| 63.1 | 76.6 | 77.6 | 69.2  | 70.5 | 41.0 | 63.7 | 64.8 | 76.0 | 71.1 |
| 10 | Baseline     | H  | H   | 65.8| 67.4| 71.2| 66.8| 51.2| 41.7   | 75.3| 79.8| 79.0| 43.1 | 76.7| 57.6 | 76.8 | 79.6 | 69.8  | 70.3 | 41.3 | 60.3 | 64.8 | 75.8 | 68.2 |
|    | SMN Full     | H  | H   | 65.4| 60.8| 71.3| 66.1| 51.8| 41.9   | 75.0| 79.6| 78.4| 43.3 | 75.3| 62.8 | 76.4 | 78.1 | 69.2  | 70.5 | 40.9 | 59.2 | 64.4 | 75.8 | 67.9 |

Table 6: VOC 2007 test object detection average precision. We use SoftMax and top 5k RoIs during testing for all the methods compared (except [69]).

| Method       | mAP | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | persn | plant | sheep | sofa | train | tv  |
|--------------|-----|-----|-----|-----|-----|--------|-----|-----|-----|------|-----|------|-----|------|------|------|------|------|-----|------|-----|
| FRCNN [69]   | 70.0| 68.7| 79.2| 67.6| 54.1| 52.3   | 75.8| 79.8| 84.3 | 50.1 | 78.3 | 65.1 | 82.2 | 84.8 | 72.9 | 76.0 | 44.9 | 70.9 | 63.3 | 76.1 | 72.6 |
| Baseline [12]| 71.2| 67.6| 78.9| 67.6| 55.2| 56.9   | 78.8| 85.2 | 83.9 | 49.8 | 81.9 | 65.5 | 80.1 | 84.4 | 75.7 | 77.6 | 45.3 | 70.8 | 66.9 | 78.2 | 72.9 |
| MLP          | 70.9| 71.7| 80.0| 70.9| 60.0| 56.6   | 78.2| 85.0 | 85.5 | 47.5 | 72.7 | 64.2 | 76.6 | 83.5 | 75.8 | 77.8 | 45.2 | 72.3 | 68.1 | 76.3 | 70.4 |
| SMN          | 71.1| 67.1| 81.2| 70.3| 55.5| 54.0   | 78.3| 85.1 | 83.7 | 49.4 | 80.9 | 66.1 | 80.1 | 83.5 | 75.7 | 77.7 | 45.1 | 69.7 | 67.1 | 78.4 | 72.6 |

Note that because the baseline feeds more RoIs (k=300 or k=5,000) for final evaluation, it still bears a subtle advantage over SMN when testing with HardMax. For example, if bounding box $B_1$ suppresses cow over sheep, there is still chance that a nearby (e.g. measured by IoU) bounding box $B_2$ where sheep is selected over cow. On the other hand, SMN gets $N \leq k$ chances for picking the candidates. This difference is reflected when we compare $k=300$ vs. $k=5,000$ for baseline: SoftMax based testing has a larger margin over HardMax when $k$ is smaller. Regardless of this, SMN is still able to achieve on-par ($N=10$) or better ($N=5$) results in terms of mAP.

The result can be found in Table 6. As can be seen, our method is on-par with baseline and MLP (~71% mAP). This difference to COCO is reasonable since compared to COCO, there is not much “juice” left for context reasoning, in terms of both quantity (number of images to train SMN on top of FRCNN) and quality (how difficult the detection of objects are in the scene).

### A3. Final Results on VOC

We also excluded the final results on VOC comparing 1) the baseline FRCNN, 2) the MLP where a 5-layer ConvNet is directly stacked on top conv5_3 for context aggregation, and 3) our SMN. We report the results here. For the final evaluation, we use top 5k RoIs for region sampling and the SoftMax strategy for all methods. Due to the memory limitation, the same idea of first using NMS to find easy examples, storing them in the spatial memory and then predicting with SMN is used (Sec. 6.1). For VOC we set $N_{c}=10$ and bootstrap from a $N_{b}=10$ SMN model, so in total $N=20$ bounding boxes are sent for evaluation.