Episodic CAMN: Contextual Attention-based Memory Networks With Iterative Feedback For Scene Labeling

Abrar H. Abdulnabi[1,2] Bing Shuai[1] Stefan Winkler[2] Gang Wang[3]

[1] School of EEE, Nanyang Technological University (NTU), Singapore
[2] Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
[3] Alibaba Group

Abstract

Scene labeling can be seen as a sequence-sequence prediction task (pixels-labels), and it is quite important to leverage relevant context to enhance the performance of pixel classification. In this paper, we introduce an episodic attention-based memory network to achieve the goal. We present a unified framework that mainly consists of a Convolutional Neural Network (CNN), specifically, Fully Convolutional Network (FCN) and an attention-based memory module with feedback connections to perform context selection and refinement. The full model produces context-aware representation for each target patch by aggregating the activated context and its original local representation produced by the convolution layers. We evaluate our model on PASCAL Context, SIFT Flow and PASCAL VOC 2011 datasets and achieve competitive results to other state-of-the-art methods in scene labeling.

1. Introduction

This paper deals with the problem of scene labeling (or semantic segmentation), which typically aims to relate a unique semantic class label to each pixel in an image. It is a challenging task in computer vision, as different scenes may be full of complex and occluded objects, and the images can be captured under various lighting conditions and viewpoints. Also, objects in an image may appear at any location and scale. In order to classify each pixel in a typical scene labeling pipeline, a feature vector is first extracted from a patch of adequate size containing that pixel. The patch contains the surrounding context to locally discriminate the pixel during labeling. But these local feature representations can be ambiguous, e.g. a sofa patch can be visually indistinguishable from a bed patch. As a natural solution, many papers utilize context to distinguish locally ambiguous patches [27,50,53]. However in their works, the long-range context is not effectively leveraged. Besides, the surrounding contextual information is incorporated without selection. We believe that different contextual information is not always equally important or useful to one local region. For example, when recognizing a local region belonging to computer monitor, it is more helpful to leverage contextual information from desk rather than from other regions such as window. Hence, special attention should be paid to certain context conditioned on the target region.

In this paper we propose a unified framework that includes FCN and episodic attention-based memory network layers to address the problem. In details, it consists of three major components: (1) a basic FCN to generate local convolutional patch-level feature representations; (2) a Contextual Attention-based Network (CAN) that adaptively selects the relevant contextual patches for each referenced patch; (3) an 'episodic' recurrent memory module that accumulates and records the selected context over multiple episodes, where its recurrent feedback connections allow the CAN module to refine the activated/selected context over multiple iterations. CAN alongside the recurrent memory network (CAMN) appears to be an effective contextual modeling method that can easily activate useful long-range context. However, CAMN may falsely activate irrelevant context and wrongly inhibit the relevant context in one shot; therefore we propose to refine the selected context in an iterative framework with the recurrent feedback mechanism. We instantiate the episodic memory as the hidden representation of a recurrent neural network [19], which is able to efficiently memorize local information over a short time (several iterations). Then, we add the feedback connections from the episodic memory to CAN, so that the CAN module can refine the context selection in the next iteration. The episodic memory converges over only few early iterations. We describe our method as 'episodic' because it accumulates and refines the selected context over a series of consecutive iterations. Either all of the generated context of these episodes or only the last episode can be utilized.

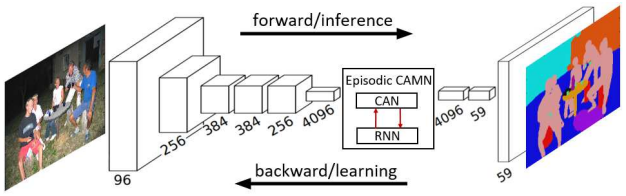

Figure 1. An illustrative overview of our unified framework. The model consists of FCN convolution layers, our episodic contextual attention-based memory network, and the upsampling alongside the FC (classification) layers. Given an input image, our model generates a pixelwise label map.

An overview of the framework is shown in Figure 1. The full network is end-to-end trainable. Even without post-processing, we achieve state-of-the-art labeling results on several challenging scene labeling benchmarks. The main contributions of this paper are summarized as follows:

- We propose the contextual attention-based memory network model that is able to adaptively select relevant context for each target patch. Importantly, it is a novel contextual modeling method that can effectively engage long-range context.
- We introduce an episodic memory module to accumulate contextual selections and aid the CAMN to refine the previous context selections through recurrent feedback connections over multiple iterations ('*episodes*').
- We achieve very competitive results on three public scene labeling benchmarks.

2. Related Work

2.1. Contextual Modeling and Scene Labeling

Contextual cues from other patches are usually important for local patch reasoning and prediction. Many recent successful semantic segmentation systems are developed based on CNNs, specifically, FCNs. Usually, the receptive fields of the neurons in the convolution layer correspond to local regions of the input image. They implicitly engage contextual information through cascading multiple layers. However, it would be ineffective to enlarge the receptive fields of the neurons to explicitly and directly model long-range context as it may degrade the local discriminative features [66]. FCNs try to overcome this limitation by applying skip connections from earlier layers and ultimately aggregating the intermediate features for classification [40]. Other earlier works try to learn hierarchical and multi-scale CNN features to capture the context across a multi-scale image pyramid [21, 22, 47]. Another line of work revisited convolution operations and applied 'atros' and dilated convolutions to perform contextual modeling [11, 12, 68]. Recently, RNN-like layers, e.g. Long-short Term Memory (LSTM) and Gated Recurrent Units (GRUs) [14]) are commonly inserted after convolution layers to explicitly capture and encode long-range context into local representations [5, 51, 59, 66]. Sharma *et al*. [49, 50] exploit the recursive neural network architecture to propagate context. However, RNN-like models are limited in modeling very long-range sequences. In contrast, our attention-based model can effectively encode the long-range contextual information between patches.

Conditional Random Fields (CRFs) and Markov Random Fields (MRFs) are also used to model the label-level context [11, 12, 32, 38, 69]. They characterize co-occurrence relationships between labels. Alternatively, we aim to encode contextual information into local feature representation.

2.2. Attention-based Models

Attention-based models have been successfully applied on a broad range of Natural Language Processing (NLP) related tasks, which include machine translation [18] [43], speech recognition [13], image caption generation [65], reading comprehension [28], sentence summarization [48], part-of-speech (POS) tagging [34], question answering [30, 64], etc. Some works benefited from attention models for image classification tasks [7, 44, 62] and object recognition [3]. The attention mechanism was first introduced in neural machine translation [18] to automatically align the words in the original sentence with the words in the target sentence using encoder-decoder RNN model. The aligned (attention-based) model is embedded in the sequence to sequence learning framework [55]. Concurrently, in the image caption generation task, Xu *et al*. [65] use the attention-based model to roughly localize image regions of interest, which are deemed as relevant in producing the next word. Another interesting work [10] use attention to select from different multi-scale features. Hermann *et al*. [28] and Rush *et al*. [48] employ the attention-based model to discover the keywords/sentences that are informative to comprehend the paragraph or sentence. Chorowski *et al*. [13] also leverage the attention-based model to filter noisy frames in a short window-based audio to interpret the desired phoneme. Neural Memory Networks [24, 54, 57, 60] became popular with the aid of attention mechanism where separate memory structures are involved to stably store information.

To the best of our knowledge, ours is the first work to approach scene labeling utilizing the differentiable soft attention mechanism alongside recurrent memory networks. Our model can dynamically and interactively select and refine the relevant and informative contextual patches for each referenced patch, therefore its local representation is contextualized.

2.3. Recurrent Feedback

Recurrent feedback has been explored in the tasks of scene labeling and human pose estimation. Auto-context [70] is the pioneering work that utilizes the output of classifiers as feedback to the next classification model. Meanwhile, Pinheiro *et al*. [47] add feedback connections between the output and the input of the convolutional neural networks (CNNs). Namely, the output of the CNN in the previous iteration is fed back to the input of the same CNN in next iteration. This recurrent CNN has been successfully applied to scene labeling. Besides, Carreira *et al*. [8] also add the feedback connection in the CNNs to enable the network to learn from past errors, so that it can predict the location of human joints more accurately in the next iteration. The idea of our episodic memory feedback is similar to these works. Engaging the recurrent feedback connections can be viewed as cascading multiple attention layers that can form a deep attention model. However, instead of simply adding the feedback from the output of the previous iteration, we introduce an episodic memory to record and accumulate all past information leveraging RNN layers, and then add its feedback to the attention-based module such that the selected context is refined over multiple iterations.

3. Framework

Given a scene image **S**, the task is to map from the pixel-space to label-space. We divide each image into patches and train our episodic CAMN model to extract contextually-aware convolutional features to better represent each local patch. Suppose patches in a scene image **S** are represented in terms of convolutional features as $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where N is the number of patches, our goal is to predict the semantic label of the referenced patch $\mathbf{x}_i$. Specifically, given a referenced raw local patch from **S**, the data flow in our model is as follows (cf. Figure 1):

Input feature map: the convolutional layers in the FCN convert input patches into local feature representations $\mathbf{x_i}$.

Output feature map: the episodic CAMN performs contextual modeling and compute the output which is the contextually-aware features for the input referenced patch. It adaptively and progressively selects the relevant contextual patches for the referenced patch. Two major networks are employed within: (1) the CAN model contains a feed-forward network to calculate the compatibility/similarity scores to perform soft attention between the input $\mathbf{x_i}$ and the remaining patches in the image $\mathbf{x}$; (2) the feedback mechanism is comprised of a recurrent layer that records and aggregates/accumulates the previous generated context (activated context) and links it back to the feed-forward layers so that the CAN can further refine the previous selections over multiple iterations. This also allows the model to generate deep contextual representations. The recurrent feedback connections with the CAN module allow it to adapt (refine) the activated context over multiple iterations. The episodic CAMN module is run for multiple iterations per training epoch until the selected context converges.

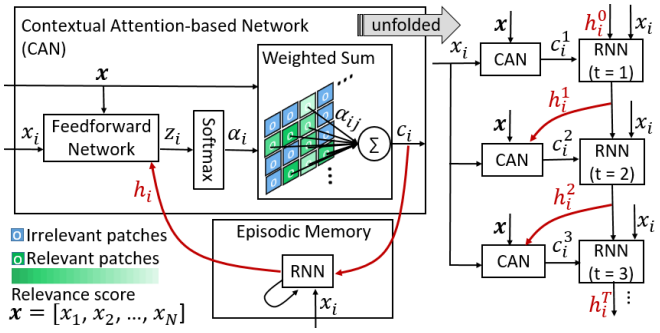

Figure 2. An illustration of our Episodic CAMN including its contextual attention-based network, the episodic recurrent memory and the feedback. The model consists of two major components: (1) the CAN adaptively selects the relevant context to contextualize the reference patch representation ($\mathbf{x}_i$); (2) the episodic memory summarizes the activated context in the past. The feedback from the episodic memory enables the CAN to iteratively refine the selected context over multiple iterations.

Response: To calculate the response (label) of the input patch, the last classification layers (FCs) alongside the upsampling layers and Softmax are engaged to decode the final contextually-aware dense representations.

3.1. FCN for Local Patch Representation

Fully convolutional networks (FCN) originally adapt classification networks like the VGG net [52] into fully convolutional networks and transfer their learned representations to the segmentation task. FCN architecture combines/fuses semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. In their work, they design the basic 32s model that has no intermediate skip connections from earlier layers, and the 16s version has skip connections from $pool4$ layer and finally the 8s version that has skip connections from both $pool4$ and $pool3$ layers. All FCN models have upsampling layers which upsample the downsampled label prediction maps (due to the application of the the pooling layers) to the original image resolution. The proposed episodic CAMN model can be inserted into any location between the last pooling layer and the final FC layer. In our work, we placed it between FC6 and FC7. The input patch representations to episodic CAMN are generated by the early convolution layers and the output of episodic CAMN is the contextually-enriched representations.

3.2. Contextual Attention-based Memory Model

Context is of great significance in local prediction. In this paper, we model context from the perspective of feature representation. Our attention-based module is introduced to learn to activate certain contextual features for each local patch. Concretely, it aggregates selected features from surrounding patches with the referenced patch to produce its contextualized representation, thus encodes useful contextual information for local classification. However, for a referenced patch, not all the surrounding patches are equally useful. For example, in terms of the contextual support to a pillow patch in a bedroom image, bed patches are more useful than the wall patches. Hence, the attention mechanism is developed to adaptively select the relevant patches and assign them proper weights. The episodic CAMN model has feedback connections to operate over multiple iterations, thus refining the selected context. Our model can naturally engage long-range contextual dependencies, which makes it very suitable for scene labeling.

3.2.1 Contextual Attention-Based Network

In order to attend to patches which produce useful contextual information to $\mathbf{x}_i$, our CAN uses a feed-forward network (Attention model) to evaluate the high-level relevance between referenced patch representation $\mathbf{x}_i$ and the other surrounding patches $\mathbf{x}_j$. Mathematically, it is expressed:

$$z_{ij} = w_b^T \tanh(W_a\mathbf{x_i} + V_a\mathbf{x_j} + b), \tag{1}$$

where W_a, V_a are embedding matrices and w_b is a vector to capture feature similarities; they are jointly learned. z_{ij} reflects the degree of relevance or compatibility (activation/inhibition) between patch representations $\mathbf{x}_i$ and $\mathbf{x}_j$.

There are alternative methods to generate the relevance scores, e.g. concatenation: $W_a[\mathbf{x_i}; \mathbf{x_j}]$, or cosine similarity: $[\mathbf{x_i}^T\mathbf{x_j}]$, or absolute distance: $|\mathbf{x_i} - \mathbf{x_j}|$. However, we observed that all these functions perform almost the same, while the feed-forward layers perform the best.

CAMN attends and generates the context vector $\mathbf{c}_i$ by soft attention based on $\mathbf{c}_i = \sum_j \alpha_{ij}\mathbf{x}_j$, where $\alpha_{ij} = softmax(z_{ij})$, s.t. $\sum_j \alpha_{ij} = 1$. α_{ij} is a non-negative scalar that modulates how much information from the patch representation $\mathbf{x}_j$ is exposed to $\mathbf{c}_i$. It is obtained by applying the typical softmax function to generate the relevance distribution over all locations. Then, the weighted summarization of the activated patch representations gives rise to the context vector $\mathbf{c}_i$, which is calculated through the convex summation of $\mathbf{x}$ weighted by α_{ij}. Interestingly, as the context vector is able to include features from distant image patches, it can capture the long-range contextual information easily.

We also experiment with the ReLU function instead of Softmax. The ReLU operates directly on the unnormalized probabilities produced by the feed-forward layers. We observe that in some cases the ReLU performs as well as Softmax. However, Softmax produces a more visualization-friendly distribution of weights over the patches.

The final 'context-aware' representations of the referenced image patches are generated by aggregating the original features $\mathbf{x}_i$ with their specific context vectors $\mathbf{c}_i$:

$$\mathbf{h}_i = \text{ReLU}(U_{hx}\mathbf{x}_i + U_{hc}\mathbf{c}_i + z_h), \tag{2}$$

where U_{hx} and U_{hc} are the transformation matrices that map the input and the context vector respectively to new hidden representation space, z_h is a bias vector, and $\mathbf{h}_i$ is the hidden representation forwarded to the classification layers. All of the parameters in Equation 1 and 2 are jointly learned. In this scenario, the CAN is trained to maximize the labeling performance on the training images by inserting useful context into the local representations.

To generate the response/class likelihood for each referenced patch, we apply $\mathbf{r}_i = softmax(V_r\mathbf{h}_i + z_r)$, where V_r is the classification matrix, z_r is a bias vector, and $\mathbf{r}_i$ is the class likelihood for $\mathbf{x}_i$.

3.2.2 Episodic Recurrent Memory and Feedback

The CAN produces the context-aware vector that summarizes the relevant contextual information, which is selected based on the attention model. However, if we only forward the attention model once to calculate the attention weights, some irrelevant patches may be mis-activated while some relevant patches may be falsely inhibited. To address this issue, we propose to refine the attention weights iteratively such that the CAN model can attend to undiscovered relevant context and remove irrelevant context. Specifically, we introduce an episodic memory module that is able to memorize the context selections in the past. We employ the feedback from the episodic memory to refine the selected context by CAN. Biologically, it is inspired by discoveries in neuroscience that neuron adaptation actually happens across multiple time-scales through the feedback connections between the dendrites and the axon [39, 42]; feedback of ionic currents are flowing back from the dendrites to the axon to allow refining its adoption of the stimulus [39].

Suppose we generate the corresponding relevance score vector $z_i^t = [z_{i1}^t, z_{i2}^t, \ldots, z_{iN}^t]$ of a reference patch representation $\mathbf{x}_i^t$, and a context vector $\mathbf{c}_i^t$ during the t-th iteration. The episodic memory is intended to accumulate and record the relevant context selected by the CAN over multiple iterations. In this paper, we instantiate the episodic memory as a recurrent neural network [19] (RNN), considering that RNNs are well-known for their memory capability for short sequences. The activities in the episodic memory module can be formulated as:

$$\mathbf{h}_i^t = \text{ReLU}(U_{hx}\mathbf{x}_i + U_{hc}\mathbf{c}_i^t + U_{hh}\mathbf{h}_i^{t-1} + z_h), \tag{3}$$

where $\mathbf{c}_i^t$ is the context vector generated in the t-th iteration, $\mathbf{h}_i^t$ represents the memory state in the t-th iteration. $\mathbf{h}_i^t = \mathbf{x}$ as $t = 0$, U_{hh}, U_{hx} and U_{hc} are hidden-hidden, input-hidden and context-hidden transformation matrices, respectively. In each iteration, the episodic memory module accumulates the context vector produced from the CAN module $\mathbf{c}_i^t$ into the memory hidden representations.

With the availability of the episodic memory, we add the feedback connections between the memory module and the attention module. In this case, the episodic memory feedback allows the CAN to refine the selected context in the subsequent iteration. Concretely, the equation that calculates the relevance score is adapted to the following form:

$$z_{ij}^t = w_b^T \tanh(W_a \mathbf{x}_i + V_a \mathbf{x}_j + U_a \mathbf{h}_i^{t-1} + b), \quad (4)$$

where z_{ij}^t is the relevance score between patch representations $\mathbf{x}_i, \mathbf{x}_j$ and $\mathbf{h}_i^{t-1}$ in the t-th iteration. The memory state ($\mathbf{h}_i^{t-1}$) encodes the context selections over previous iterations. Thus, when they are fed back, the context selection is expected to be refined by allowing $\mathbf{x}_i$ and $\mathbf{x}_j$ to interact with all previous context summarization $\mathbf{h}_i^{t-1}$.

Feeding back the memory vector that contains a summarization of the previous contextual selections allows context-context interactions, therefore new indirect relevant contextual patches can be activated. However, we observe that a few iterations are enough to discover new related context as the selection saturates quickly. We call these forward passes *episodes*. Specifically, after T forward passes by the CAN, context-aware representations generated by the final episode or all the episode states are forwarded to the classification layers. We call our model Episodic CAMN (see Figure 2).

3.3. Model Optimization

Given an image $\mathbf{S}$, we derive the class likelihood for each constituent patch in $\mathbf{S}$. Next, we calculate the cross entropy loss to train the full model (including the parameters in episodic CAMN). The error signal of an image is averaged across all valid (i.e. semantically labeled) constituent patches $\mathbf{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{b=1}^{B} \delta(\mathbf{y}_i = b) \log \mathbf{r}_i(b), \quad (5)$$

where $\delta(\cdot)$ is the indicator function, B is the number of semantic classes, y_i is the ground truth label for the patch representation $\mathbf{x}_i$. $\mathbf{r}_i$ is the class likelihood for the patch representation $\mathbf{x}_i$, which is a B-dimensional vector. We ignore the contribution of unlabeled (invalid) patches in the loss calculation. The whole model is differentiable, thus it is end-to-end trainable by using the back-propagation algorithm. The weights in recurrent layer (episodic memory) are shared, and they are updated via back-propagation through time (iteration) after unfolding the episodic memory network in time (iteration).

4. Experiments

In this section we describe our experimental evaluation and we provide an ablation study of our proposed architecture. We apply our framework to the task of scene labeling, and we show competitive performance to other state-of-the-art works on PASCAL-Context, SIFT Flow and PASCAL VOC 2011. We also perform a diagnostic evaluation of our convolutional episodic CAMN and other hyperparameters/design choices.

4.1. Datasets

PASCAL-Context [46] comprises 4998 training images and 5105 testing images. Originally, the images are sampled from PASCAL VOC 2010 dataset and re-labeled at pixel-level for the segmentation task, where there are in total 540 classes. Each image has a resolution of about 375×500 pixels. In our experiments, we only consider the task of labeling the 59 most frequent classes for evaluation.

SIFT Flow [35] consists of 2688 images. We follow the training/testing split protocol (2488/200) provided by [35] in our experiments. The images are captured from 8 typical outdoor scenes with a resolution of 256×256 pixels. The segmentation task in this dataset is to assign each pixel to one of the 33 semantic classes. Statistically, this dataset has an imbalanced class distribution. We found that applying class balancing helps similar to [51, 67].

PASCAL VOC 2011 [20] involves 21 categories, including 20 foreground object classes and one background class. There are 736 images as non-intersecting validation set and 1111 testing images provided by their server. In our experiments, we only train our final model from the training set and did not include any images from the validation set.

Following recent literature, we report three performance evaluation scores: Pixel Accuracy (**PA**) (percentage of all correctly classified pixels), Per-class Accuracy (**CA**) and the Intersection-Over-Union (**IU**).

4.2. Implementation Details

We adopt the stochastic gradient descent (SGD) with momentum. The learning rate starts at 10^{-5} and decays exponentially with the rate of 10% after 10 epochs. The momentum is fixed as 0.9. The internal dimension of the RNN network (episodic memory h_i^t) is set to the same as that of the input feature vector (dimension = 4096). The reported results are based on the model trained for 50 epochs. The feedforward, FC layers and RNN model parameters are initialized either randomly or with zeros. For fair comparison, we select the strongest FCN models (best performing among 8s, 16s and 32s versions) as our baselines in all

Algorithm	PASCAL-Context			SIFT Flow			PASCAL VOC 2011		
	PA(%)	CA(%)	IU(%)	PA(%)	CA(%)	IU(%)	PA(%)	CA(%)	IU(%)
Baseline-FCN	65.90	46.53	35.11	85.19	54.68	41.45	90.30	75.90	62.70
FCN+CRF	69.19	47.36	38.37	85.70	52.44	43.12	90.98	76.30	64.01
FCN+CAMN	71.27	52.91	39.64	85.90	57.10	43.88	91.41	76.99	66.91
FCN+Episodic-CAMN	**72.11**	**54.28**	**41.18**	**86.20**	**58.69**	**45.22**	**92.26**	**78.59**	**68.18**

Table 1. Results on the PASCAL-Context [46], SIFT Flow [35] and PASCAL VOC 2011 (validation set) [20]. All models are trained jointly with FCN models (best performing ones).

datasets. We follow [40] to train the fully convolution network layers. We initialize all our convolutional models with the ImageNet-pretrained model [17]. We didn't engage any additional training data (such as Microsoft Common Objects in Context 'COCO' dataset [33]). We used the publicly available MatConvNet MATLAB implementations [58].

We jointly train FCN and CAMN as one unified model. We introduce the following baselines to prove the effectiveness of our proposal.

Baseline-FCN performs full training of the FCN model [40] for scene labeling. In this baseline, we examine and report the best model among 8s, 16s and 32s models for each dataset.

FCN+CRF performs post-processing on the strongest trained FCN model using the Fully-connected Conditional Random Fields (CRF) implementations [29]. We choose to engage fully-connected CRF because it is one of the strongest contextual modeling methods in scene labeling.

FCN+CAMN performs joint training of our model; FCN with one-episode CAMN. Here, only one iteration of CAMN is engaged.

FCN+Episodic-CAMN is our final overall proposed framework with episodic memory ($T = 3$).

4.3. Evaluation Results

Comparison with FCN: The quantitative labeling results on the PASCAL-Context, SIFT Flow and PASCAL VOC 2011 datasets are summarized in Table 1. Our FCN+CAMN outperforms FCN on the three datasets by 4.53%, 2.43% and 4.21%, respectively (IU). The results show that the discovered context is indeed informative towards understanding the semantic classes of the referenced patches and strengthen the discriminative power of the local representations. For Baseline-FCN, we perform class weighting while evaluating on SIFT Flow dataset. The original results in [40] are lower than ours on this dataset.

Recurrent Feedback: The introduction of the feedback from the episodic memory module is to steer the CAMN to iteratively refine the selected contextual memories and produce more powerful deep contextual vectors. The performance of our overall model FCN+Episodic-CAMN is better than that of FCN+CAMN. In Table 1, the IU accuracy is increased around 1.54%, 1.34% and 1.27% over one-episode CAMN in all the datasets respectively. The most significant performance jump originates from the CAN alone, which indicates that the attention model is able to select very useful contextual patches just in one-shot. Moreover, as we introduce the episodic memory feedback to iteratively refine the context selections, the recognition performance for some classes can be further improved. As shown in Figure 6, the context vector usually saturates after around 2-3 iterations; the performance starts to degrade slightly after many iterations ($T > 5$). This may be due to the effect of over-contextualizing the local features as after several contextual aggregations, some local discriminative features may be overwhelmed by the global contextual features.

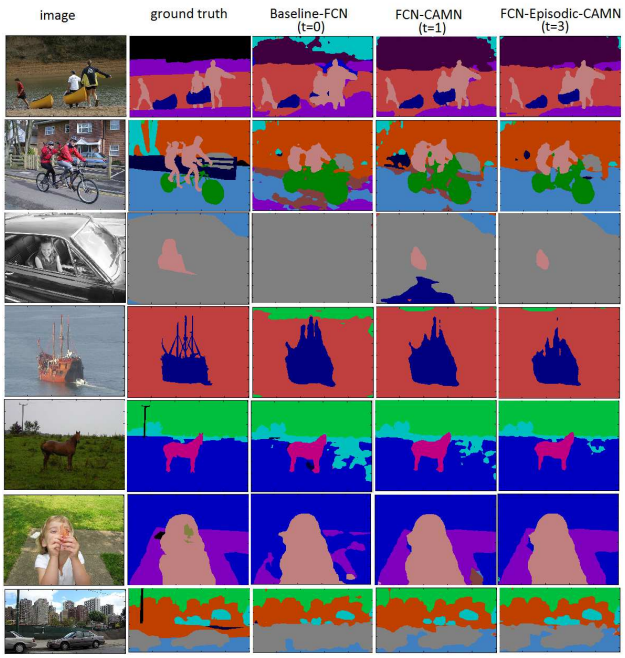

Figure 3. Examples of qualitative labeling results (best viewed in color). Each row shows the original test image, its ground truth label map, and three label prediction maps of our convolutional episodic attention-based contextual memory network with different iterations. The test images are from the PASCAL-Context [46]. The quality of the label prediction maps is remarkably improved as CAN is engaged (iteration 1), and is gradually refined after additional iterations.

Figure 3 show the qualitative labeling results and their

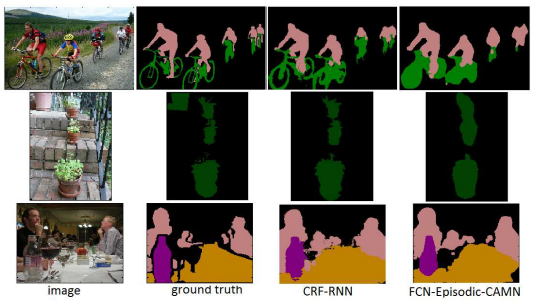

Figure 4. Comparison with CRF-RNN method [69]. The samples are from PASCAL VOC 2011 dataset [20].

gradual improvement over iterations. The quality of label prediction is significantly improved over FCN as CAMN is engaged for the first time, and it is further improved as the context refinement evolves. We also visualize in Figure 5 the weighted selections (Heatmaps) for several patches belonging to building, sky and computer respectively at episode 1 and 3. In the first row, the heat increases in some areas (e.g. building, road) and decreases in other areas (e.g. sky). The second row visualizes heatmaps for a sky patch belonging to the same scene as in the first row. This shows that attention model is indeed location/patch sensitive. The third row presents heatmaps for computer monitor patch, showing how some patches from table are activated as related context. Our method models the associations between features at patch level, i.e. not necessarily the whole object is activated as related context. In other words, different parts of an object/stuff can be activated for different target object/stuff parts. We also report (train /test) run time on the validation set of PASCAL VOC 2011 (averaged over 20 trials for an input image of 384×384 on an NVIDIA Tesla K40m with a companion CPU Intel Xeon E5-2643 v3 at 3.40GHz); FCN+Episodic-CAMN: ($\sim 775ms$ /$\sim 270ms$) and Baseline-FCN: ($\sim 546ms$ /$\sim 157ms$).

Comparison with State-of-the-art: We compare the performance of our full model with other state-of-the-art methods for contextual modeling. The quantitative result comparison on PASCAL-Context, SIFT Flow and PASCAL VOC 2011 datasets are listed in Tables 2, 3 and 4 respectively. There are other methods that use different settings like adopting Residual Networks [26] (e.g. [12, 61]) or engaging extra training data to train their models. In our case, we only compare with VGG-based networks with similar settings. Table 4 shows evaluation results on PASCAL VOC 2011 on the test set (provided by their server). Our model achieve significant improvement over FCN-8s.

Comparison with Fully Connected CRF and other CRF-based Models: We run the fully-connected CRF model [29] that is used in other state-of-the-art works [11, 12] as a post-processing step based on FCN prediction maps. As shown in Table 1, our method outperforms

Algorithm	PA(%)	CA(%)	IU(%)
FCN-8s [40]	65.9	46.5	35.1
FCN-8s [41]	67.5	52.3	39.1
DeepLab [12]	-	-	37.6
DeepLab-CRF [12]	-	-	39.6
HO-CRF [2]	-	-	41.3
CNN-CRF [32]	71.5	53.9	**43.3**
CRF-RNN [69]	-	-	39.3
ParseNet [37]	67.5	52.3	39.1
ConvPP-8 [63]	-	-	41.0
PixelNet [4]	-	51.5	41.4
O2P [9]	-	-	18.1
CFM [16]	-	-	34.4
BoxSup [15]	-	-	40.5
FCN+Episodic-CAMN	**72.1**	**54.3**	41.2

Table 2. Performance comparison on PASCAL-Context [46].

it consistently. This performance gap illustrates that the proposed method is more effective than CRF on leveraging relevant context to perform local classification. Plus, our model is very simple in terms of optimization compared with graphical-based models. It is purely constructed from primary feed-forward and recurrent layers so it is fast and easy to optimize.

As reported in Tables 2 and 3, our method outperforms other state-of-the-arts that use CRF models such as [32] and [69] in different evaluation metrics. But in Table 4, CRF-RNN [69] outperforms ours on PASCAL VOC 2011 by 2.0% while our method outperforms CRF-RNN on PASCAL Context as shown in Table 2 by 1.9%. CRF-based methods are effective for producing more refined object boundaries which contributes a lot to boost the IU score, especially for the object-centric PASCAL VOC (foreground objects are annotated only). Meanwhile, our method can explicitly model the contextual dependencies between image patches from different categories, hence it can better capture the contextual interactions when images are densely annotated with many different categories (PASCAL Context and SIFT Flow). Our method and CRFs can be used in different applications. In Figure 4, we show some qualitative examples from PASCAL VOC 2011 dataset where CRF-RNN model performs better (third column).

In terms of speed, the inference time of our Episodic-CAMN alone (as a single block) separately is $\sim 0.077s$ on the previously mentioned CPU (no GPU is engaged). Meanwhile the inference time of the dense CRF alone separately is $\sim 3.010s$ on the same CPU. Aside from simplicity, our model is much faster and more efficient ($\sim 39\times$ speedup) compared with CRF-based method, yet it performs competitively (although with high feature dimension 4096), which makes our method suitable for applications that require high speed processing.

Comparison with RNN-like Models: As shown in Ta-

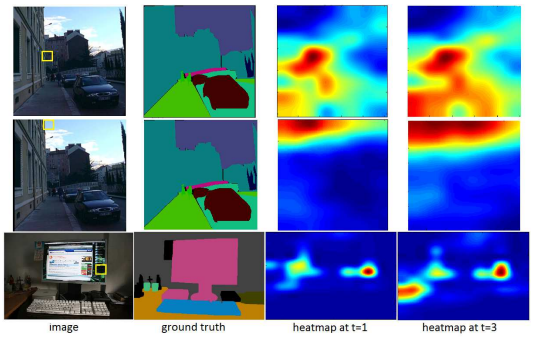

Figure 5. Some sample images from SIFT Flow [35] and PASCAL VOC 2011 [20] alongside the visualization of the weighted selections for specific local patches. From left to right: RGB image, label map, episode 1 heatmap and the final episode 3 heatmap.

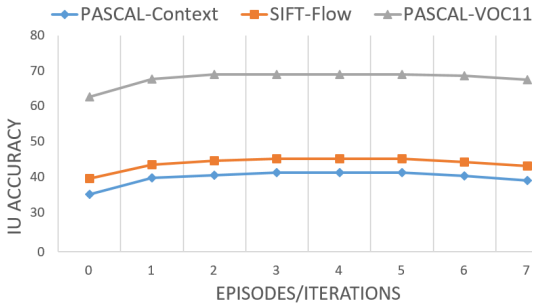

Figure 6. IU performance changes in terms of iterations.

ble 3, our model outperforms many state-of-the-art methods that use RNN on top of convolutional features, e.g. [51] and [5]. Additionally, while it is difficult to visualize the hidden states of an RNN to clearly see whether the network can capture the underlying associations and dependencies among the input sequence, attention-based memory networks can easily reveal this secret. It can simply and explicitly measure the relevance among the patches which eases the visualization and thus help us understand the underlaying hidden relationships and interactions within a scene image, as shown in Figure 5.

5. Conclusions

In this paper, we deal with the problem of scene labeling. In order to effectively leverage the relevant contextual patches to enhance the local classification accuracy, we propose an episodic attention-based contextual memory network. This model presents a unified framework that mainly consists of FCN convolutional layers, attention-based model, and an RNN to perform context selection and refinement. The full model produces context-aware representation for each target patch by aggregating the activated context and its original local representation.

Our experiments show that the proposed method significantly boosts the labeling performance of the original FCN by incorporating the selected context. More importantly, our method achieves competitive performance on the public PASCAL-Context, SIFT Flow and PASCAL VOC 2011 scene labeling benchmarks when compared with other state-of-the-art work. We believe the performance of our model can be further improved by enhancing some functionalities such as relevance learning. With more sophisticated operations, it can potentially capture subtle relationships and interactions between the referenced local patch and its context.

Algorithm	PA(%)	CA(%)	IU(%)
Liu *et al.* [35]	74.8	-	-
Liu *et al.* [36]	76.7	-	-
Tighe *et al.* [56]	75.6	41.1	-
Farabet *et al.* [22]	72.3	50.8	-
Farabet *et al.* [22]	78.5	29.6	-
FCN-16s [40]	85.2	51.7	39.5
FCN-8s [41]	85.9	53.9	41.2
CNN-LSTM [5]	70.1	22.6	-
CNN-CRF [32]	**88.1**	53.4	44.9
DAG-RNN [51]	81.2	45.5	-
Pinheiro *et al.* [47]	77.7	29.8	-
RCNN [31]	83.5	35.8	-
RCNN [31]	79.3	57.1	-
ClassRare [67]	79.8	48.7	-
Sharma *et al.* [50]	79.6	33.6	-
Sharma *et al.* [50]	75.5	48.0	-
ParseNet [37]	86.8	52.0	40.4
JointCalib [6]	-	55.6	-
Tighe-MRF [56]	78.6	39.2	-
FCN+Episodic-CAMN	86.2	**58.7**	**45.2**

Table 3. Performance comparison on SIFT Flow [35].

Algorithm	IU(%)
BerkeleyRC [1]	39.1
SDS [25]	52.6
R-CNN [23]	47.9
FCN-8s [40]	62.7
FCN-8s [41]	67.5
Zoomout [45]	64.4
CRF-RNN [69]	**72.4**
FCN+Episodic-CAMN	70.4

Table 4. Average performance comparison on the test set from PASCAL VOC 2011 [20].

Acknowledgments

We gratefully acknowledge the support of the NVIDIA AI Technology Centre for their donation of Tesla K40 and K80 cards used for our research at NTU's ROSE Lab. The work is partially supported by the research grant for the Human Centered Cyber-physical Systems Programme at ADSC from Singapore's Agency for Science, Technology and Research (A*STAR).

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011.

[2] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, 2016.

[3] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015.

[4] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan. PixelNet: Towards a general pixel-level architecture. *arXiv*, 2016.

[5] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with LSTM recurrent neural networks. In *CVPR*, 2015.

[6] H. Caesar, J. Uijlings, and V. Ferrari. Joint calibration for semantic segmentation. In *BMVC*, 2015.

[7] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015.

[8] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016.

[9] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

[10] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.

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

[13] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *NIPS*, 2015.

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

[15] J. Dai, K. He, and J. Sun. BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[16] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[18] B. Dzmitry, C. Kyunghyun, and B. Yoshua. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[19] J. Elman. Finding structure in time. *Cognitive Science*, 1990.

[20] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes challenge 2011 (VOC2011) results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[21] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers. In *ICML*, 2012.

[22] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 2013.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[24] A. Graves, G. Wayne, and I. Danihelka. Neural Turing machines. *arXiv*, 2014.

[25] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[27] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

[28] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *EMNLP*, 2015.

[29] P. Krahenbuhl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.

[30] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016.

[31] M. Liang, X. Hu, and B. Zhang. Convolutional neural networks with intra-layer recurrent connections for scene labeling. In *NIPS*, 2015.

[32] G. Lin, C. Shen, I. Reid, and A. Hengel. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.

[33] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, and L. Z. P. Dollar. Microsoft COCO: Common objects in context. *arXiv*, 2014.

[34] W. Ling, L. Chu-Cheng, Y. Tsvetkov, S. Amir, R. F. Astudillo, C. Dyer, A. W. Black, and I. Trancoso. Not all contexts are created equal: Better word representations with variable attention. In *EMNLP*, 2015.

[35] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[36] C. Liu, J. Yuen, and A. Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011.

[37] W. Liu, A. Rabinovich, and A. Berg. ParseNet: Looking wider to see better. In *ICLR Workshop*, 2016.

[38] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015.

[39] M. London and M. Hausser. Dendritic computation. *Annual Review of Neuroscience*, 2005.

[40] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[41] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2016.

[42] B. Lundstrom, M. Higgs, W. Spain, and A. Fairhall. Fractional differentiation by neocortical pyramidal neurons. *Nature Neuroscience*, 2008.

[43] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[44] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, 2014.

[45] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *CVPR*, 2015.

[46] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[47] P. H. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene parsing. In *ICML*, 2014.

[48] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for sentence summarization. In *EMNLP*, 2015.

[49] A. Sharma, O. Tuzel, and D. Jacobs. Deep hierarchical parsing for semantic segmentation. In *CVPR*, 2015.

[50] A. Sharma, O. Tuzel, and M.-Y. Liu. Recursive context propagation network for semantic scene labeling. In *NIPS*, 2014.

[51] B. Shuai, Z. Zuo, G. Wang, and B. Wang. DAG-Recurrent neural networks for scene labeling. In *CVPR*, 2016.

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[53] G. Singh and J. Kosecka. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, 2013.

[54] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. Weakly supervised memory networks. *arXiv*, 2015.

[55] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[56] J. Tighe and S. Lazebnik. Finding Things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[57] K. Tran, A. Bisazza, and C. Monz. Recurrent memory network for language modeling. *arXiv*, 2016.

[58] A. Vedaldi and K. Lenc. MatConvNet – convolutional neural networks for matlab. In *ACM MM*, 2015.

[59] F. Visin, A. Romero, K. Cho, M. Matteucci, M. Ciccone, K. Kastner, Y. Bengio, and A. Courville. ReSeg: A recurrent neural network for object segmentation. *arXiv*, 2016.

[60] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv*, 2014.

[61] Z. Wu, C. Shen, and A. Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv*, 2016.

[62] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015.

[63] S. Xie, X. Huang, and Z. Tu. Top-down learning for structured labeling with convolutional pseudoprior. In *ECCV*, 2016.

[64] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.

[65] 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. In *ICML*, 2015.

[66] Z. Yan, H. Zhang, Y. Jia, T. Breuel, and Y. Yu. Combining the best of convolutional layers and recurrent layers: A hybrid network for semantic segmentation. *arXiv*, 2016.

[67] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.

[68] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[69] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

[70] T. Zhuowen. Auto-context and its application to high-level vision tasks. In *CVPR*, 2008.