Modeling Visual Question Answering as Hard Attention over Semantic Image Regions

Griffin Adams, Tejas Nama, Saksham Singhal, Soumya Wadhwa
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{gta,tnama, sakshams, soumyaw}@andrew.cmu.edu

Abstract

Visual Question Answering (*VQA*) has emerged as a powerful testbed for machine comprehension, attention-based neural modeling, and more generally, multimodal machine learning. Existing approaches to VQA treat the problem as a classification problem. Feature vectors for images are derived from Convolutional Neural Networks (CNNs). In parallel, Recurrent Neural Networks (RNNs) encode the word-level question sequence. Similarity between question and image regions allows the network to attend to image and question regions before prediction. Our approach deviates from the status quo on several fronts. We incorporate hard attention over image regions using reinforcement learning, which allows the model to focus on a single 'glimpse' at a time in a recurrent manner. This approach aims to reduce information loss from averaging over image regions, as well as improve efficiency. Secondly, we leverage semantic image segmentation maps, which produce pixel-wise semantic labels, to extract glimpse features in the form of semantic histograms. This localized *soft* classification offers a bridge between single-label classification and CNN-based image features, which lack semantic interpretability. Additionally, we explore VQA as a Zero Shot Learning (ZSL) task. We demonstrate that ZSL can properly predict unseen cases without the need for additional parameters. Questions centered on object detection stand to gain the most from ZSL. While not delivering state of the art performance, our approach is guided by intuitive and grounded principles we hope can motivate future research.

1 Introduction

Our research tackles the problem of free form and open-ended visual question-answering (VQA). The potential to uncover patterns, content, activities, knowledge, and context in images without supervision represents not only a fascinating multimodal challenge, but is a vital component to an "AI-complete" task [13]. Image captioning represents a gateway to open-ended image comprehension. Yet image captioning is limited by two obstacles: lack of a clear learning objective and lack of a clear information-seeking goal. Image captioning can be deemed a success with a coarse understanding of the image. Metrics for evaluating machine generated passages (such as BLEU) require multiple reference solutions and are subject to substantial bias. VQA, however, has a clear learning objective and information seeking goal. Questions target precise information and elicit concise answers. We note that $> 89\%$ of the Virginia Tech's VQA dataset consists of one word answers. While multiple choice VQA presents a more well-defined task, we believe open-ended is more useful in real-world applications. Some of the more socially aware VQA applications relate to support for the visually impaired. A potential application incorporates natural language dialogue into automated surveillance systems, allowing for a human to periodically inquire about the place of interest.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Our main contribution to Visual QA research centers on enhancing the performance of a hard attention model with semantically interpretable image features. Concretely, we extend the glimpse network model described by Minh et al. [16] to the VQA domain by adding a text channel for the question and a semantic segmentation channel to map the image into word space. Incorporating hard attention increases efficiency since the computation of image features is over a patch of fixed size rather than the entire image. The semantic channel produces a semantic class label for each pixel. We construct these glimpse feature vectors by computing a histogram over pixel-wise labels. We call it 'soft' regional classification. Our secondary contribution is the exploration of Zero Shot Learning (ZSL) for the VQA task. We leverage massive text corpora to extend the predictive range of our model and train it to directly predict an answer in word space. We demonstrate that for the *other* question category - which mostly represents object detection - ZSL outperforms the other models. It is able to predict semantically close answers with limited training examples. We hypothesized that richer input representations could provide clearer context clues for the ZSL model. We incorporate a summary of the captions into the Deep CNN-LSTM ZSL model and confirm a noticeable uptick in performance on the *other* type questions.

2 Dataset and Input Modalities

We train and test our models with Virginia Tech's Visual Question Answering (VQA) dataset v1.0 (http://www.visualqa.org/index.html) [13]. The dataset is composed of 204,721 real world images from the MS COCO dataset [22], decomposed into 123,287 for training/validation and the remaining 81,434 for testing. For each image, three open-ended questions were collected by Amazon Mechanical Turk workers. Over six million ground truth answers are provided for the image-question pairs, in addition to almost two million plausible answers. Each MS COCO image has a corresponding caption, which we use as auxiliary input to the zero-shot learning module. We considered other datasets [19][20][23][24], but bypassed them in favor of VQA because of VQA's size, liberal QA collection method, and use of MS COCO images.

The main challenge of the open-ended VQA task is that it does not pose a singular challenge. Open-ended questions come in different forms and require diverse, flexible solutions. Some questions require granular object detection, while others require external knowledge, and yet others involve commonsense. The VQA task poses the classic challenge of modality fusion. One must combine a representation of a question with a representation of an image to formulate an answer. Yet there is more visual data than text. Accommodating this imbalance requires generating rich textual features. On the flip-side, images usually contains redundant and/or irrelevant features. To avoid unnecessary computation, an effective VQA system must isolate relevant regions through an attentive process.

3 Related Work

3.1 Zero-Shot Learning

Most of the techniques for VQA center on classification over answer candidates. As highlighted by Antol et al. [13], a typical VQA pipeline employs a recurrent neural network to generate a question summary, and a CNN to generate image features. The status quo model passes the fused image and text features to a multilayer perceptron. The output is a softmax distribution over a fixed set of candidate answers. Zero shot learning (ZSL) represents a method for expanding the set of candidate answers without expanding the model. We can leverage massive text corpora, train on a single dataset, and test on many [13][20][21] without any network modification. We can also evaluate on softer metrics: cosine proximity, and compare distances of prediction and ground truth. ZSL projects the image into word space, and predicts the class based on proximity in the space [7][8]. Palatucci et al. employ a knowledge base over the descriptive class properties to perform classification [9]. Bakhshandeh et al. construct a generative semantic graph model over the questions to create a candidate answer list [10]. The authors model a deep binary classifier over tuples of question, image, and answer. Although an innovative approach, the model only achieved an accuracy of 55.9%. Teney et al.'s model relies on test time exemplar retrieval [11], which uses potential answers as query terms for image search. The authors pass the top four images returned by Google through a CNN, and average the global features for improved joint visual semantic embedding. Additionally, they incorporate explicit object detection [12] and use the GloVe representation [32] of the object class to maintain coherence with the language modality.

3.2 Attention Models

Spatial attention mechanisms have been proven to support a wide variety of image processing tasks, particularly VQA [3][4][5]. These models "attend to" or "focus on" the appropriate region of the image to answer the question [28][29]. Lu et al. [30] construct a co-attention model to jointly train both the question and the image in a hierarchical fashion at three different levels: word, phrase and question. The authors propose two co-attention models: parallel and alternating. The parallel co-attention model generates both images and questions parallely, while the alternating attention model alternates between image and question generation. Recently, Nam et al. [31] proposed a novel co-attention model, a Dual Attention Network (DAN), which iteratively refines the attention using the memory of previous attention outputs.

Until recently, hard (i.e. binary) attention modeling was largely ignored given the difficulty in training non-differentiable models. Yet the past few years have witnessed an emergence of reinforcement learning based approaches for hard attention modeling. Caicedo and Lazebnik [15] propose a dynamic attention-action model using a deep Q-learning algorithm to estimate the action-value function for object localization. The agent is trained to focus attention on candidate regions for identifying the correct locations of relevant regions. Reinforcement learning provides a more computationally efficient way to train deep networks for tasks such as image classification and even for dynamic environments, since it enables the modeling of non-differentiable hard attention. Mnih et al. [16] propose an alternative to computationally expensive CNNs for processing high dimensional image data: an RNN trained using the REINFORCE algorithm [25] with variance reduction. The input to the RNN is a representation of a glimpse of the image (a fixed size high resolution region and low resolution for the rest of the image). The model learns to detect important glimpses even in the presence of noise.

3.3 Semantic Image Segmentation

Extending the reinforcement learning RNN-based approach from Mnih et al. [16] requires construction of richer visual features and introduction of spatial textual features. Recent advances in performance on object detection (bounding box) [34, 43, 44] and semantic segmentation (pixel-wise classification) [34, 35, 36, 37, 38, 38, 40, 41, 42, 44] offer efficient means to explore the enhancement of spatial image features with textual/semantic overlays. Girshick et al. [34] create a model called R-CNN which exploits the availability of large scale datasets with image-level annotations. They fine tune a CNN, pre-trained on the ILSVRC2012 dataset, on the VOC object detection dataset. They independently generate candidate region proposal and extract feature vectors using the pre-trained CNN. The region's feature vector is used to train class-independent linear SVM classifiers against the ground truth region labels provided by PASCAL VOC. Their work has motivated several enhancements [43][44]. In 2015, Long, Shelhamer, and Darrell performed semantic image segmentation by transforming an ILSVRC CNN classifier into a Fully Convolutional Network (FCN) [35, 36]. The authors remove the fully connected layers and append an extra convolutional layer, which proxies as a linear classifier over candidate class labels. The authors upsample the final pixel-wise scores to recover the original image dimensions. Then, they take the per pixel softmax over each candidate class, compare to the ground truth class label, and back propagate through the full network. They train and validate on the PASCAL VOC 2011 segmentation challenge. Other authors pick up where Long, Shelhamer, and Darrell left off and focus on enhancing output resolution via a combination of post-processing CRF inference [41][42], dilated convolutions to widen receptive fields [40][41], incorporating global context [38], and deconvolutional layers [39]. These new approaches share the same express goal of the skip architecture employed Long, Shelhamer, and Darrell: to combine wide/coarse features with narrow/fine features.

4 Problem Statement & Evaluation

"Given an image and a natural language question about the image, the task is to provide an accurate natural language answer" [13]. Each Question is represented as $Q = \{q_1, q_2..q_T\}$, where q_t is the one-hot representation the t^{th} word. The image features are represented as $V = \{v_1, v_2..v_N\}$ where v_n denotes the spatial feature (such as the pixel value or VGG feature) at the n^{th} location. The task is to predict the right answer A given Q and V.

$$A^* = \max P(A|Q, V)$$

We evaluate each proposed model on the slacked accuracy given as:

$$slacked\_accuracy = \min\left(\frac{\#humans\; who\; provided\; that\; answer}{3}, 1\right)$$

5 Proposed Methods

5.1 Zero Shot Learning

Small errors in the semantic word space can be more useful than classification errors - given that class label indices hold no intrinsic value. For questions relating to object detection/retrieval, zero shot learning produces a central value for which candidate labels are ranked by proximity. This process produces highly clustered answers at the top. The model is very similar to the multimodal baseline of CNN and Deep LSTM described in [13] where instead of a softmax distribution, we learn a feature representation. The model aims to transfer the semantic knowledge learned from the text domain to the multimodal domain of VQA. Training minimizes a pairwise hinge loss, defined as:

$$Loss(y^*, \hat{y}) = \sum_{y!=y^*} max(0, margin - sim(y^*, \hat{y}) + sim(y, \hat{y}))$$

Here, y^* is the ground truth embedding, $margin$ is a hyperparameter which controls how far predictions should be from false labels, y is the false label embedding and $\hat{y}$ is the predicted embedding. sim is the used similarity function (cosine similarity in our case).

5.2 Recurrent Attention Model

We closely follow Mnih et al. [16], who propose a recurrent attention model for processing the visual modality as in Figure 1. Applying CNNs to images is computationally expensive; the proposed RNN is capable of extracting information from an image by adaptively focusing on a sequence of locations. Like CNNs, the proposed model has some translation invariance built-in, but the amount of computation is independent of image size. While the model is non-differentiable, it can be trained using reinforcement learning methods to learn task-specific policies. The problem is modeled as a Partially Observable Markov Decision Process (POMDP). At each step, the agent can extract information from the image via a bandwidth limited sensor, which generates a high resolution representation for the region of interest and successively lower resolutions for surrounding regions. The glimpse sensor produces a feature vector. At each time step, the agent processes the sensor data, integrates information over time, and chooses how to act (which answer class to choose) and how to deploy its location sensor at the next time step. The model can also be augmented with an action that decides when the glimpse generation will stop instead of restricting it to a fixed number of steps. Since the model is non-differentiable, we train using REINFORCE with variance reduction.

We incorporate a text channel (GloVe embeddings and a pre-trained LSTM) in the glimpse network so that the hard image attention is conditioned on the question. We introduce the question representation at the glimpse selection stage so that the next location is chosen based on a joint representation of the previous glimpse location and question. We add a pixel-wise semantic label channel to the glimpse setup. We fine-tune an image-level CNN classifier to the task of semantic segmentation by training on the Pascal VOC 2012 segmentation task. This closely follows the work of Long, Shelhamer, and Darrell [35, 36]. We pre-processed each training and test image in the VQA dataset by generating a same sized semantic map. Local image features are extracted at train and test time by computing a semantic label histogram over the glimpse of the highest resolution.

5.3 Counting Module

Most VQA neural models struggle with counting-based questions since counting is really a computer vision task. Our error analysis supports this notion: even our best deep models revert to predicting the dominant number label (2). To address these limitations, we leverage the semantic segmentation maps generated for the main model to develop a counting mechanism. To determine how many numbers

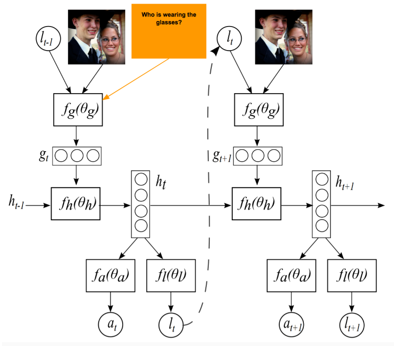


Figure 1: The glimpse network of Mnih et al [16]

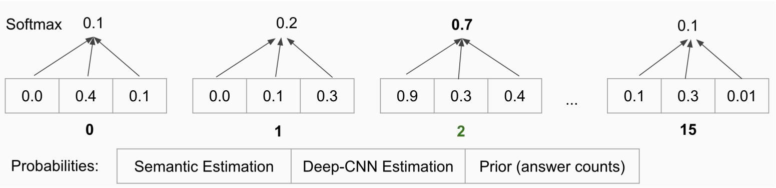


Figure 2: Our proposed count network with tied weights for each answer class. The three sources of evidence are derived from semantic estimation, deep LSTM + CNN estimation, and a prior.

to consider for prediction, we compute a cumulative density function (CDF) over all the number labels in the training dataset and apply a cutoff at 96%. This amounts to considering the inclusive range [0, 15]. Given the difficulty of the task, we take a multiple-evidence based approach, where we extract probability estimates over the number classes, [0, 15], from several sources and learn an appropriate weighting for each source's estimate. The first feature is a prior generated from the training set. The second source of evidence is the Deep LSTM + CNN model's output. We pass each counting question image through the model, extract the softmax probabilities for just the considered labels [0, 15], and renormalize. We did not expect this feature to offer much in the way of additional information, considering that the deep model overweights the dominant class anyway. To incorporate explicit computer vision, we take semantic segmentation maps and derive a probability distribution over [0, 15] for each counting question. We first manually extract the counting entity **x** according to the template: "how many **x**...?". This approach reliably extracts the correct entity. We then compute an affinity vector for the entity **x** with respect to the 150 semantic labels provided by the MIT Scene Parsing Benchmark (SceneParse150) [51]. Ideally, we could take the MIT label with the greatest affinity but this approach produces too many mismatches and zero counts. So we expand the search to consider the closest 5 classes. To produce the count distribution, we dilate the segmentation maps at 3 resolutions to denoise the label map. Then we compute the number of connected components and use it as a proxy for that label's count. Each candidate label provides a vote for its own count scaled by affinity to the counting entity. From these features, we construct a linear classifier whose weights are tied across each output number class. The model learns which evidence source is most reliable: semantic segmentation, deep LSTM + CNN, and prior. Figure 2 visualizes our approach.

6 Experimental Setup & Results

6.1 Zero Shot Learning

We use pre-trained 300 dimensional GloVe features for the text domain which had been trained on 6 billion tokens from Wikipedia and GigaWord-5 corpus. For the image representation, we use the output from VGGNet, which was pre-trained on ImageNet dataset. We append a real-valued output layer to the VGGNet model. The output of this network is the predicted word embedding of the answer. While training, we restrict the false labels to the top 1000 answer candidates. We fix the hinge loss margin to 0.9. The model was trained using stochastic gradient descent on p2.2xlarge

Table 1: Zero Shot Learning Validation Results

Exact Match Accuracy	Top - 3 Accuracy	Top - 5 Accuracy
16.33	29.34	37.16

Table 2: Zero Shot Learning - Comparison

Model	Yes / No	Count	Other	Overall
Deep LSTM + CNN	78.96	34.67	34.53	53.63
ZSL (top 5)	54.12	7.34	42.30	37.16
ZSL (with captions) (top 5)	52.17	7.34	44.63	37.02

instance on AWS which had a GPU support from Nvidia Tesla K80. The model ran for 60 epochs. At test time, for each question and image pair, we list the top-5 answers with closest to the generated output vector in word-embedding space. The results obtained are shown in Table 1 and compared with the baseline approaches in Table 2.

ZSL improves significantly in the *other* category of questions with co-attention between the image and the question. As expected, ZSL does not perform very well on “yes/no” and “counting” categories, since there is no coherent semantic relation between the question, image, and answer in these cases. We trained a separate model which incorporated image captions. The additional context clues from the captions boosted performance on the ‘other questions’ by over 2 percent. We model captions using a bidirectional LSTM and concatenate the summarized caption to the output layer input.

6.2 Recurrent Attention Model

The design choices made for these experiments were as follows:

Retina and location encodings: The retina encoding extracts square patches centered at location l, with each successive patch having twice the width of the previous. The patches are then resized and concatenated. Glimpse locations are encoded as real-valued (x,y) coordinates with (0, 0) being the center of the image. [16]

Glimpse network: The glimpse network has one fully connected layer with ReLU nonlinearity. It takes pixel values of the image and the location as input and outputs a suitable representation.

Text network: We experiment with GloVe word embeddings as well as pre-trained LSTM embeddings for the question representation, and both gave comparable results. The text network had one fully connected layer with ReLU nonlinearity, producing a text representation for input to the core RNN.

Location network: The policy for the locations is defined by a two-component Gaussian with a fixed variance. The location network outputs the mean of the location policy given the hidden state values of the core network.

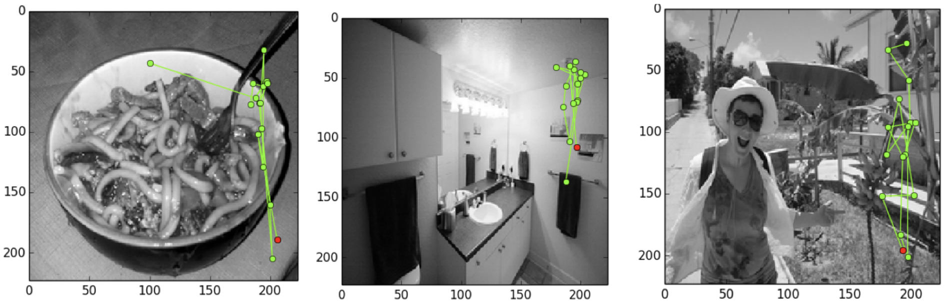
Core network: The core network is an RNN with 16 glimpses (timesteps) and other features. Its output is the next location and action (answer label, in our case).

The results for different experiments are shown in Table 3. We almost achieve baseline accuracy, and our model is computationally efficient since it operates on fixed-size glimpses. We also experiment with the addition of VGG features for the entire image at every recurrent step. However, this didn’t lead to much improvement. This is probably because the glimpse network already sufficiently captures the required image features. We also add a semantic segmentation channel comprised of the normalized distribution of the values of semantic labels as auxiliary input to the core network, after being passed through a dense layer. Yet this produces very similar results, and is very computationally expensive (we could get results for 1 epoch only). We also experiment with discrete locations.

We perform qualitative analysis on the Recurrent Attention Model (with LSTM) by hand-picking examples from the validation dataset. We look at the trajectory followed by the recurrent model to further understand its working. We find that the answer predicted by the model is very sensitive towards the 16 glimpses it chooses during testing. For example, in the leftmost image in Figure 3, the chosen patches contain a fork, and thus the models predicts “fork”. For the center image in Figure 3, it can be seen that the model uses the textual features from the question to understand the context in

Table 3: Recurrent Attention Model Validation Results

Model	Yes / No	Count	Other	Overall	Exact Match
RAM + VGG	75.13	32.95	31.18	50.60	43.12
RAM + LSTM	76.46	34.89	29.50	50.70	43.31



What kind of cuisine is this?	chinese	fork

What room is this?	bathroom	bathroom

How does the woman feel?	happy	yes

Figure 3: The trajectory of the chosen glimpses is shown for each figure. The first column is the question and the second and third columns are the true and predicted answers.

order to predict the answer. The answer “bathroom” is predicted using both the features from the glimpse and also from the question. The low performance of the model might be because of the lack of navigation capability. In the rightmost image in Figure 3, the model does not even localize on the person on the left and hence predicts the wrong answer.

6.3 Counting Module

We train on all questions starting with “how many” whose answers lie in the inclusive range 0-15 for 5 epochs and compute a binary cross entropy loss. We test the model on the same class of questions (0-15) in the validation set. Importantly, since the counting model only predicts in the range 0-15, these numbers are inflated by several percentage points. We test on 11,852 counting questions, which represents $> 96\%$ of the total “how many” questions with numeric answers. ($< 1\%$ of the counting questions have qualitative ground truth answers, which we do not consider for validation purposes). We test on a model with no hidden layers, and one with a single hidden layer with 5 neurons. The SEG model shows the output of directly using the count generated from the highest confidence class as the predicted count. When the top-5 label counts showed an object count of 0, we defer to the dominant class (2) for inference. The results in Table 4 show no demonstrable improvement over the baseline. Digging deeper, the confusion matrix for the first model reveals that the model weights the prior as the only reliable indicator. The segmentation predictions (SEG Prediction), when considered alone, do not always produce the dominant class but are too imprecise to improve results. We show the confusion matrix for the segmentation predictions in Figure 4. The model fails to predict large numbers in part due to inability of the entity to be counted to be consistently represented in the top k semantic labels. We considered additional labels but this resulted in boosting the count scores for incorrect objects. Coverage of the count entities by the segmentation map label categories remains a major roadblock to any FCN-based counting module. Yet even when the 5 nearest semantic labels were present, counting precision was badly affected by the the agglomeration problem, which relates to the inability of vanilla segmentation maps to delineate overlapping objects.

7 Conclusion and Future Work

To summarize, we tackle the Visual QA challenge by incorporating hard attention over image regions, allowing the model to focus on one ‘glimpse’ at a time, and leverage semantic image segmentation

Table 4: Counting Module Validation Results

Model	Accuracy (%)
Prior + Deep-CNN + SEG Model (2 FC Layers)	31
Prior + Deep-CNN + SEG Model (1 FC Layer)	31
SEG Prediction	27.5

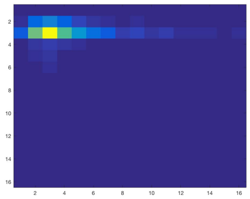


Figure 4: The confusion matrix for the segmentation predictions (SEG Prediction in Table 4). Rows represent predictions and the columns the true labels. The indices represent the respective class labels 0-15. Zeros are represented in dark blue and high counts in bright yellow. 2 is most commonly predicted since we defer to 2 when the top 5 classes produce no connected components.

maps to extract histograms of semantic labels as glimpse features. Although we could only about reach the baseline results, our approach was intuitive and computationally less intensive than the CNN+LSTM baseline. Additionally, we explore VQA as a Zero Shot Learning (ZSL) task (to make the 'other' results better), and implement a counting module (to make the 'counting' results better).

7.1 Semantic Segmentation Standalone

The idea of using semantic segmentation represents an intuitive and interpretable approach to learning localized image features for attention-based VQA models. The results from the glimpse network showed promise but not state of the art performance. Decoupling errors caused by the glimpse network and those caused by segmentation maps is a difficult task. The simplest way to test the efficacy of segmentation maps, ceteris paribus, is to leverage existing state of the art approaches: such as hierarchical co-attention [30], and replace the traditional CNN-feature maps with semantic maps. Summarizing image regions with segmentation maps could be as simple as a histogram over class labels, or as complex as the average final hidden states of row-wise and column-wise RNNs over the region, passing each pixel's semantic label word embedding as the input sequence. The image feature matrix could be seamlessly combined with a question feature matrix to form co-attention matrices. Hierarchical image features could be constructed by applying Gaussian filters with increasing standard deviations to the grid centers. This would be analogous to combining granular and coarse CNN feature maps. Additionally, since glimpse regions are chosen by the action selector online, it is hard to precompute glimpse histograms. Hence, training times are slow. We plan to pre-compute histograms over fixed grids at different scales. Then, for a given glimpse take a weighted average of the precomputed histograms based on mean intersection with the chosen glimpse.

7.2 Counting with Instance-Level Semantic Segmentation

As outlined in Section 6.3, the principal shortcoming of the counting module is the imprecision of the segmentation map predictions. Application of a flood-fill algorithm to a segmentation map as a proxy for a class label's instance count only works if the label maps are smooth and separate instances have clearly delineated label boundaries. Overlapping objects and poor map smoothing, can lead to both over-counts (grainy segmaps) and under-counts (agglomeration of multiple instances). Future work entails generating accurate instance-level segmentation maps with smooth boundaries (post-processing with superpixel-based CRFs). Instance-level segmentation produces separate labels for each object instance, and is more appropriate for counting tightly clustered objects.

Acknowledgments

We would like to thank Professor Louis-Philippe Morency, Tadas Baltrušaitis, and the TAs of Carnegie Mellon class 11-777, Advanced Multimodal Machine Learning, for their expertise, endless stream of fabulous ideas, and undying belief in our work.

References

[1] Kumar, Ankit, et al. "Ask me anything: Dynamic memory networks for natural language processing." CoRR, abs/1506.07285 (2015).

[2] Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic memory networks for visual and textual question answering." arXiv 1603 (2016).

[3] Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ICML. Vol. 14. 2015.

[4] Xu, Huijuan, and Kate Saenko. "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering." European Conference on Computer Vision. Springer International Publishing, 2016.

[5] Yang, Zichao, et al. "Stacked attention networks for image question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[6] Weston, Jason, Sumit Chopra, and Antoine Bordes. "Memory networks." arXiv preprint arXiv:1410.3916 (2014).

[7] Socher, Richard, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. "Zero-shot learning through cross-modal transfer." In Advances in neural information processing systems, pp. 935-943. 2013.

[8] Frome, Andrea, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov. "Devise: A deep visual-semantic embedding model." In Advances in neural information processing systems, pp. 2121-2129. 2013.

[9] Palatucci, Mark, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. "Zero-shot learning with semantic output codes." In Advances in neural information processing systems, pp. 1410-1418. 2009.

[10] Bakhshandeh, Omid, Trung Bui, Zhe Lin, and Walter Chang. "Proposing Plausible Answers for Open-ended Visual Question Answering." arXiv preprint arXiv:1610.06620 (2016).

[11] Teney, Damien, and Anton van den Hengel. "Zero-Shot Visual Question Answering." arXiv preprint arXiv:1611.05546 (2016).

[12] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[13] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[14] Wu, Qi, et al. "Visual question answering: A survey of methods and datasets." arXiv preprint arXiv:1607.05910 (2016).

[15] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 2488–2496, 2015.

[16] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In Advances in neural information processing systems, pages 2204–2212, 2014.

[17] Peng, Baolin, et al. "Towards neural network-based reasoning." arXiv preprint arXiv:1508.05508 (2015).

[18] Learning to compose neural networks for question answering. Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein. NAACL 2016.

[19] Malinowski, Mateusz, and Mario Fritz. "A multi-world approach to question answering about real-world scenes based on uncertain input." Advances in Neural Information Processing Systems. 2014.

[20] Zhu, Yuke, et al. "Visual7w: Grounded question answering in images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[21] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. (2012)

[22] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European Conference on Computer Vision. Springer International Publishing, 2014.

[23] Ren, Mengye, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering." Advances in Neural Information Processing Systems. 2015.

[24] Yu, Licheng, et al. "Visual madlibs: Fill in the blank description generation and question answering." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[25] Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256, 1992.

[26] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

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

[28] Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[29] Gregor, Karol, et al. "DRAW: A recurrent neural network for image generation." arXiv preprint arXiv:1502.04623 (2015).

[30] Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." Advances In Neural Information Processing Systems. 2016.

[31] Nam, Hyeonseob, Jung-Woo Ha, and Jeonghee Kim. "Dual attention networks for multimodal reasoning and matching." arXiv preprint arXiv:1611.00471 (2016).

[32] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

[33] Karpathy, Andrej. "The unreasonable effectiveness of recurrent neural networks." Andrej Karpathy blog (2015).

[34] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[35] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[36] Shelhamer, Evan, Jonathan Long, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." IEEE transactions on pattern analysis and machine intelligence 39.4 (2017): 640-651.

[37] Hong, Seunghoon, Hyeonwoo Noh, and Bohyung Han. "Decoupled deep neural network for semi-supervised semantic segmentation." Advances in Neural Information Processing Systems. 2015.

[38] Liu, Wei, Andrew Rabinovich, and Alexander C. Berg. "Parsenet: Looking wider to see better." arXiv preprint arXiv:1506.04579 (2015).

[39] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[40] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[41] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." arXiv preprint arXiv:1606.00915 (2016).

[42] Zheng, Shuai, et al. "Conditional random fields as recurrent neural networks." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[43] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[44] He, Kaiming, et al. "Mask R-CNN." arXiv preprint arXiv:1703.06870 (2017).

[45] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[46] https://github.com/jiasenlu/HieCoAttenVQA

[47] Dai, Li et al. “R-FCN: Object Detection via Region-based Fully Convolutional Networks”. https://arxiv.org/pdf/1605.06409.pdf

[48] Goyal, Yash, et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering." arXiv preprint arXiv:1612.00837 (2016)

[49] Thrun, Sebastian. "Lifelong learning algorithms." Learning to learn. Springer US, 1998. 181-209.

[50] Graves, Alex, Greg Wayne, and Ivo Danihelka. "Neural turing machines." arXiv preprint arXiv:1410.5401 (2014).

[51] http://sceneparsing.csail.mit.edu/