Importance of object selection in Relational Reasoning tasks

Kshitij Dwivedi
kshitij_dwivedi@mymail.sutd.edu.sg

Gemma Roig
gemma_roig@sutd.edu.sg

Singapore University of Technology and Design

Abstract

Relation reasoning is a task that showcases basic intelligence and reasoning ability in humans. Deep Neural Networks (DNNs) achieve superhuman performance on classification tasks, yet fail to outperform humans on relational reasoning. Relation networks (RN) were introduced to explicitly model relations using pairs of objects, and outperform humans on some relational tasks. In this work, we note that current RNs applied to visual question answering tasks lack reasoning over the objects present in images. Such RN models consider every location on a low dimensional representation of the image as possible object candidate, which inhibits their ability to reason about concrete objects pairs. We tackle this issue by incorporating an attention model and show that even with fewer parameters, our model outperforms standard RN by a large absolute margin. When using the same number of parameters, our proposed model with object selection achieves an accuracy that is 23.89% higher in absolute terms than the baseline RN model without object selection.

1 Introduction

The reasoning ability of humans to understand and identify relations between different objects and their attributes is central to their intelligent behavior. Deep Neural Networks (DNN) have been shown higher accuracy performance than humans in some tasks, such as image classification [6]. DNN to perform on par with humans in relational tasks, require specific relation modules [8], outperforming baseline neural network models without a relational reasoning module. The relation networks (RN) have been successfully applied to text-based question answering [8], complex reasoning about a dynamic physical system [8], temporal reasoning [9], and have shown superhuman performance on visual question answering task [8],

Most RN models consider every location on the spatial feature map as a possible candidate object. The objects are generally located sparsely in the scene, and the rest of the locations are background. Since RN are designed specifically to operate on object pairs, including background locations as the object candidates introduces unnecessary comparisons of object vs. background and background vs. background. This also adds noise to the relevant object vs. object information. We overcome this by adding an object selection module which enables object selection within Relation Networks, thus reducing the total number of comparisons for relational reasoning.

Models such as attend infer repeat(AIR) [3], DRAW [4] are possible candidates for attention modules in RN. Yet, those are too expensive as they require at least n steps to detect n objects, and thus, require multiple steps to generate candidate objects. Other recent works [1, 2] use object detectors pretrained in a supervised manner on a different image dataset as the object selection modules. One drawback of such an approach is that additional supervision is provided for the object detection model which might not be easily available for all object categories. We argue that to fairly compare

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

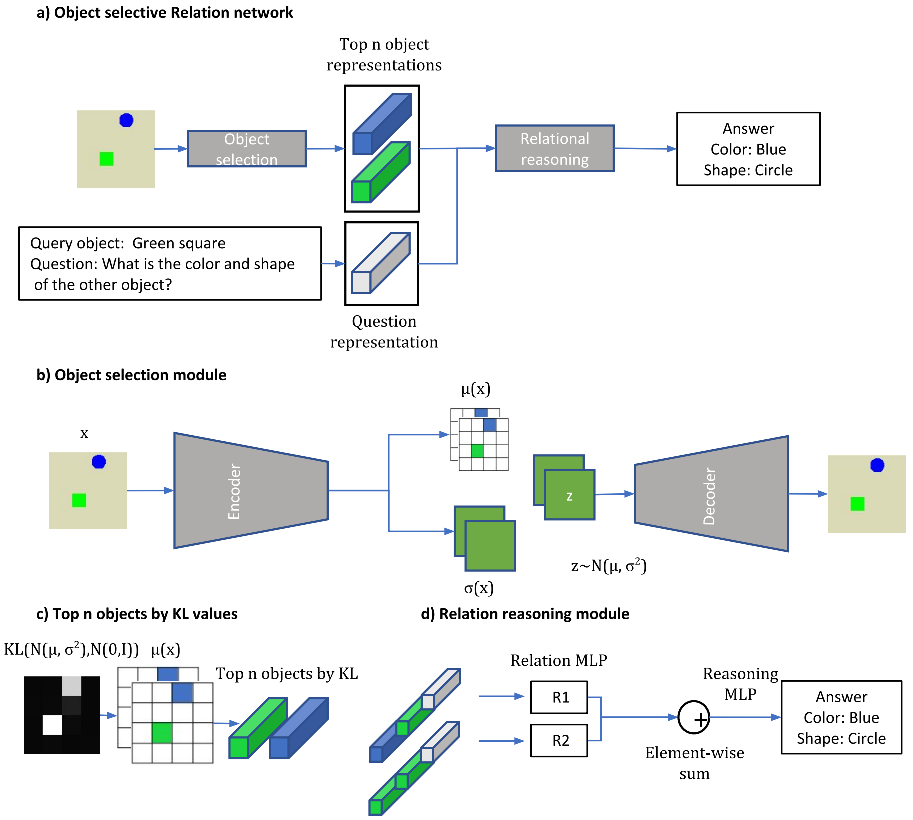

Figure 1: a) *Object selective relation network:* it consist on an object selection module and a relational reasoning module. b) *Object selection module:* VAE for the object selection module. c) *Top n objects by KL values:* Object representation from the output of the encoder at locations with high KL values. d) *Relation reasoning module:* It generates all combination of object pairs and concatenates with the question representation to use it as an input to an MLP. The outputs of all combinations are summed and fed to another MLP to predict the final answer.

whether object selection is crucial for relational reasoning, the object selection module should have the following properties: i) it should not introduce any additional parameters, ii) it should not use additional supervision other than provided for the relational task.

In this work, we introduce a variational autoencoder (VAE) [5], inspired by a multi-entity variational autoencoder [7], as the object selection module for relational networks. The spatial KL divergence map of VAE correctly localizes the objects present in the images, as illustrated in Figure 1. Using a VAE allows analyzing the performance improvement without the ambiguity that might be due to an increase in model complexity or to pretraining on other image datasets. We replace the CNN part of the relational networks with a VAE encoder while keeping the same number of parameters.

Results on a dataset similar to SortofCLEVR [8] for Visual Question Answering (VQA) tasks show that our object-selective model outperforms baseline relation networks by a large margin.

2 Object Selective Relation Networks

The proposed model consists of a convolutional neural network (CNN) encoder that transforms input image to features, an object selection module to select possible object candidates, and a relational module to perform relational reasoning tasks (Figure 1a). The input to the model is an image and a question represented by a hard-coded binary string. The object selection module extracts object representations from the image which are combined with question representation and fed as input to the relational reasoning module. Finally, the relational reasoning module predicts the answer based on object relations. In the following, we detail each of the building blocks of our model.

Object selection module We use a VAE model, as depicted in Figure 1b, to obtain an object level latent representation of the input image. The VAE model transforms the input data, denoted as x, to approximate the posterior distribution, denoted as q_ϕ, using an encoder. The encoder outputs the means and variances of the conditional Gaussian distribution. Using the reparametrization trick, a sample is generated from the posterior distribution, and a generative model is used to reconstruct the input image from the sample. The model is trained by optimizing the variational lower bound:

$$L(x) = D_{KL}\left(q_\phi(z|x)||p_\theta(z)\right) - \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right], \tag{1}$$

where θ and ϕ are the parameters of the generative model and the encoder respectively, D_{KL} denotes the Kullback-Leibler(KL) divergence and $p_\theta(z)$ is the standard Gaussian prior($N(0, I)$).

We use the encoder of the VAE as the representation of the image and each location on the feature map is considered as a possible object candidate. It has been argued in previous work [7], that the locations with objects have high KL-divergence values as compared to non-object locations. Following this findings, we select top N locations by KL-divergence values as our object candidates. The representations from the last layer of the convolutional encoder at the aforementioned locations with high KL values are used as object representations. These representations can be directly used as the input to the relational module which we describe in the sequel.

Relation reasoning module Relation reasoning module(Figure 1d) consists of two modules. The first module, called relation module, takes a pair of object representations as the input and transforms it into a relation. In the case of question answering tasks the relation module also receives the question representation as the input. This is done to condition the processing of the module on the question. The second module, named reasoning module, performs the relational reasoning task by taking the summation of all possible relations as the input. Both these modules are implemented with Multi-Layer Perceptrons (MLPs).

3 Experimental Set-up

3.1 Dataset and Benchmark

We create a dataset similar to SortofCLEVR [8] that consists of images ($128x128$) with 6 objects of size $32x32$ per image. The objects were selected from one of the twelve possible color-shape combinations (2 shapes (circle and square) and 6 colors (Red, Blue, Green, Yellow, Magenta, and Cyan)). The relational tasks for the dataset were to predict the color of the farthest and nearest possible objects from a query object. Since the positions of the object were selected randomly there could be multiple possible answers to the farthest and nearest tasks. We also included the default non-relational questions already present in the RN publicly available implementation[1] to compare the performance in the non-relational tasks. The dataset consists of 8,000 training and 2,000 test samples.

Each question answering task for an image in the dataset has multiple correct answers provided as groundtruth. Let n denote the number of correct answers for an image, which varies for each image. The accuracy performance is calculated by taking the top n values from the last layer of the network, each of them mapping to a predicted answer. Then, we count the number of predicted answers that match with the n correct answers from the groundtruth for that particular question, and average across the testing set images.

3.2 Models

We compared different variants of object-selective models with a baseline and the RN model, all described below.

Relational Network (RN) We use the publicly available RN implementation[1] for the experiments. The question is represented as a hard-coded binary string. The CNN part consists of 4 convolutional layers with 24 channels per layer and kernel size 5x5. For the first two convolutional layers the stride was set to 3 and the next two convolutional layers the stride was 2. The object representation was obtained by concatenation of relative x,y coordinates with the features extracted from the last convolutional layer. The relational module was implemented as a 4 layer MLP with 256 neurons in each layer and reasoning module was implemented as 3 layer MLP with 256 neurons in each layer.

Object Selective Relation Networks(ORN) and variants The VAE encoder architecture for ORN was kept same as the CNN part of the baseline RN model. The output of the last convolutional layer was split into two feature maps of dimension 12 each to represent the parameters of the posterior. The

[1] https://github.com/gitlimlab/Relation-Network-Tensorflow

Model	Relational Questions	Non Relational Questions
ORNx7y7	**79.62**	**100.0**
ORNxy0	67.68	**100.0**
ORNx7y7bl	55.73	**100.0**
RNx7y7bl	54.36	76.35
RN	75.85	93.98

Table 1: Accuracy on Farthest Nearest task.

decoder of VAE transforms the sample generated from the posterior distribution to the reconstructed image of the same dimension as the original input image. It consists of 5 upsampling layers each followed by a convolutional layer and batch normalization layer. The relational and reasoning modules are same as the RN model.

Since we split the last convolution layer into 2 to represent the mean and variance of the VAE posterior, the dimension of the object representation of our model is not consistent with the baseline RN model. The object representation, in this case, is 14-dimensional(12 from CNN and 2 x,y coordinates) while in the default RN model the object representation is 26-dimensional(24 from CNN and 2 x,y coordinates).To keep the dimension of object representation consistent with the RN model we create two variants of ORN with different object representations. In the first, we concatenate the object features and the x,y coordinates and pad with zeros to match the dimensions. We call this model ORNxy0. In the second representation we concatenate x,y coordinates 7 times and we call this model ORNx7y7.

ORN baseline We compare to a baseline where the object representation is extracted from the pretrained VAE encoder similar to ORN but instead of selecting top-N objects all the locations on feature maps are considered as objects. We represent the object coordinates by concatenating x,y coordinates 7 times. We call this model (ORNx7y7bl).

RN baseline We compare with another RN baseline(RNx7y7) where instead of 24-dimensional object representation concatenated with x,y coordinates the object representation of RN is 12 dimensional and padded with x,y coordinates similar to ORNx7y7.

3.3 Training details

First, a VAE was trained on the same dataset for $200,000$ minibatch iterations with batch size equal to 16 using ADAM optimization with learning rate = $2.5e-4$. The same VAE encoder was used for all the object selection variants, and the weights of the VAE encoder were fixed during the training on VQA task. The optimization parameters in VQA training phase were kept same as VAE training phase for all the variants and the baseline.

The RN model required no pretraining, and all the modules were optimized for the VQA task using the same optimization parameters as above. The last output layer for the farthest nearest task is a sigmoid to incorporate multilabel setting and the binary cross-entropy loss is used for optimization.

4 Results and Discussion

We compare the accuracy performance of object selection variants with all baselines and RN described in section 3.2, of predicting the correct answer. Results are in Table 1. The ORNx7y7 model shows the highest accuracy performance (100% on relational questions and 80% on non- relational questions). ORNx7y7 model, which have a 12-dimensional object representation, outperforms by a large margin(3.77%) the baseline RN model which does not have an object selection module and has 24-dimensional object representation. The comparison between ORNx7y7 vs. ORNx7y7bl demonstrates the importance of the object selection module (ORNx7y7 accuracy is 23.89% higher than ORNx7y7bl), since the baseline ORNx7y7bl uses all locations as possible objects, while ORNx7y7 uses the object selection module, and both use the same padded object coordinates representation.

In Table 1, we also observe that ORNx7y7bl vs. RNx7y7bl, both take all possible locations for object candidates, obtain similar accuracy. The above comparison suggests that training task-specific object representations is not required for the relation network. Moreover, we found that training the encoder reduces performance on non-relational questions.

The above comparison results of ORNx7y7 vs. ORNx7y7bl and ORNx7y7bl vs. RNx7y7bl suggest that for visual question answering using relation networks, an object detection model is more important than learning a high dimensional image representation for solving the task.

Acknowledgement

This work was funded by the MOE SUTD SRG grant (SRG ISTD 2017 131). Kshitij Dwivedi was also funded by SUTD President's Graduate Fellowship.

References

[1] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. *arXiv preprint arXiv:1803.11189*, 2018.

[2] Mikyas T Desta, Larry Chen, and Tomasz Kornuta. Object-based reasoning in vqa. *arXiv preprint arXiv:1801.09718*, 2018.

[3] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016.

[4] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Charlie Nash, SM Ali Eslami, Chris Burgess, Irina Higgins, Daniel Zoran, Theophane Weber, and Peter Battaglia. The multi-entity variational autoencoder. *NIPS Workshops*, 2017.

[8] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[9] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017.