Stochastic Relational Network

Kang Min Yoo, Hyunsoo Cho, Hanbit Lee, Jeeseung Han, and Sang-goo Lee
Seoul National University
{kangminyoo, johyunsoo, skcheon, jshan, sglee}@europa.snu.ac.kr

Abstract

Reasoning about relations among a set of objects is one of the key aspects of human intelligence, and Relational Networks (RNs) [24] are one of the classes of architectures that specializes in such relational reasoning. However, RNs are limited in their general applicability due to significant (quadratic) complexity of all-pair comparative operations. In this paper, we propose Stochastic RN (SRN) that learns to prune distractors and pick task-related objects that are crucial for relational reasoning, thereby reducing forward and backward computation costs with minimal sacrifice. We empirically show that our approach is effective in a real-world visual question-answering task, where vanilla RNs might be computationally expensive to run due to the sheer number of candidate objects for each image.

1. Introduction

Relational reasoning exploits relationships among a set of objects or entities to conduct higher order of inference and is considered crucial in achieving better compositionality in machine intelligence. [24] is one of the earlier works to identify the potential of relational reasoning and has proposed Relational Network or Relation Network (RN), a simple but effective neural network architecture that has been designed to explicitly carry out binary relational reasoning.

This specialized architectural design has been shown to be effective notably in visual relational reasoning (VQA) [24, 21]. However, since RNs take pairwise relationships among a group of entities, the network complexity is at least in the order of n^2 where n is the number of objects in question. The quadratic complexity imposes a significant challenge in employing RNs in tasks where the number of objects is substantial, such as dynamic feature maps extracted from objective detection algorithms (e.g. Faster R-CNN [23]) on real-world images.

Despite said difficulties, there have been notable efforts with varying success in applying RNs in a more challenging setting, i.e. real-world visual question answering (VQA) [9, 6]. These efforts circumvent the network scalability issue by strictly restricting the number of objects subject to relational reasoning. For example, in the latest work by [6], only the top 36 feature maps detected with the bottom-up mechanism [1] have been utilized for relational reasoning, trading off objects or object relationships that potentially hold key evidence to answering corresponding visual questions. Meanwhile, the latest version of Pythia [14], the state-of-the-art attention-based model for the VQA task, leverages feature maps extracted from both pre-trained and task-specifically fine-tuned Faster R-CNN models to achieve superior performance in VQA 2.0 [10]. The union of the object detectors produces 296 feature maps for each image, supporting the notion that a large candidate pool of detected objects is crucial for extracting richer and deeper visual semantics. However, considering that each feature map is also a 2048-dimensional vector, the sheer number of pairwise relations and their features remains a challenge for RNs to be applied directly.

In this preliminary work, we explore the feasibility of employing stochastic neural network architecture to perform relational reasoning discriminatively. Ideally, the proposed model must not only reduce forward pass computation cost but also reduce computation cost during backpropagation, without losing its expressive power. Based on the intuition that not all objects detected by a general object detector are useful for conducting visual reasoning, we propose a variant of RN called Stochastic Relational Network (SRN), in which the model stochastically learns to select most relevant objects to be passed onto more costly relational reasoning operation. The contributions of the paper are two folds:

1. We introduce Stochastic RN (SRN) as a means to address the scalability issue of relational reasoning, especially in the context of real world visual relational reasoning.

2. We conduct preliminary experiments to support our motivation, and we empirically show that suppressing distractors in relation subjects not only alleviates the inherit scalability issue of RNs but also further improves performances in the VQA task.

2. Relational Network

Deep neural networks have made remarkable progress at recognizing objects[23, 19], but teaching models to reason with relations between objects is ongoing research. Relational Network (RN) is the leading approach in this regard, achieving super-human performances in some tasks [24].

We represent the core RN module as a function that takes a query vector $\mathbf{q}$ and an array of objects $\mathbf{V}$ and returns an output vector $\mathbf{o}$ that encodes the relational information about the objects queried by $\mathbf{q}$:

$$\mathbf{o} = \mathrm{RN}\left(\mathbf{q}, \mathbf{V}\right)$$

The first step is to fuse each object vector $\mathbf{v}_i$ with the query vector to produce a relation subject vector $\mathbf{s}_i$ through a fusion function f_s: $\mathbf{s}_i = f_s\left(\mathbf{q}, \mathbf{v}_i\right) \forall 0 \leq i \leq k$, where k is the number of objects. Next, each permutation of subject vector pairs goes through pairwise relation layer f_r, producing relation vectors that embed relational knowledge about the paired objects. A pooling operator $\bigsqcup$ (e.g. sum-pooling) pools k^2 relation vectors to produce a single vector that goes through the final output layer f_o:

$$\mathrm{RN}\left(\mathbf{q}, \mathbf{V}\right) = f_o\left(\bigsqcup_{i,j} f_r\left(s_i, s_j\right)\right) \tag{1}$$

3. Stochastic Relational Network

A discriminative network (p_θ), parametrized by θ, predicts whether an object is worthy of being subject to relational reasoning in Equation 1. $\mathbf{z} \in \{0, 1\}^k$ is the discrete latent variable that indicates whether the object is selected. Objects suppressed by the discriminative network will be dropped out from the pool of object vectors, forming a new object matrix $\mathbf{V}' \in \mathbb{R}^{k' \times d_o}$ where $k' = \sum_i z_i \leq k$:

$$\mathbf{V}' = \left[\begin{array}{c|c} \mathbf{v}_i & \\ \dots & z_i = 1 \\ \mathbf{v}_k & \end{array}\right] \tag{2}$$

The goal of the discriminative network is to select the most probable set of objects that maximizes the following expectation.

$$\mathbb{E}_{p_\theta(\mathbf{z})}\left[g\left(\mathbf{z}\right)\right] \tag{3}$$

Where g is a function that evaluates the value of a particular combination of object indicators. For classification problems, g is set to be the log-likelihood of target labels, i.e. $\log p\left(y|\mathbf{x}, \mathbf{z}\right)$, which is a natural by-product of categorical cross-entropy loss during training. For subsequent sections, we assume g to be the log-likelihood parameterized by ϕ, the parameters in the main model excluding those of the discriminative network. ϕ might be omitted for clarity.

Gradients of Equation 3 respect to θ cannot be estimated directly, as g is not continuous: object selection operation described in Equation 2 is not differentiable. We must turn to stochastic gradient estimators to allow the network to be trainable using gradient descent methods.

3.1. Score-function Gradient Estimator

The score-function gradient estimator [29], also known as the likelihood ratio (LR) estimator, is an unbiased stochastic gradient estimator that uses the log derivative trick (i.e. $\triangledown_\theta p_\theta\left(z\right) = p_\theta\left(z\right) \triangledown_\theta \log p_\theta\left(z\right)$) to derive the following equation:

$$\triangledown_\theta \mathbb{E}_{p_\theta(z)}\left[f\left(z\right)\right] = \mathbb{E}_{p_\theta(z)}\left[f\left(z\right) \triangledown_\theta \log p_\theta\left(z\right)\right]$$

By using the identity, the partial derivative of Equation 3 can be estimated using Monte Carlo simulations:

$$\mathbb{E}_{p_\theta(\mathbf{z})}\left[g\left(\mathbf{z}\right) \triangledown_\theta \log p_\theta\left(\mathbf{z}\right)\right] \approx \frac{1}{M} \sum_i^M g\left(\tilde{\mathbf{z}}\right) \triangledown_\theta \log p_\theta\left(\tilde{\mathbf{z}}\right) \tag{4}$$

Where $\tilde{\mathbf{z}}$ is sampled from multidimensional Bernoulli distribution with parameters $p_\theta\left(z_1\right), \dots, p_\theta\left(z_k\right)$. In practice, for each mini-batch, the model is first trained respect to ϕ using the classification error (i.e. $\triangledown_\phi g_\phi\left(\mathbf{z}\right)$), then we perform backpropagation respect to θ using the gradient estimated in Equation 4 with ϕ being fixed.

3.2. Variance Reduction

Various methods to reduce the variance of the score-function gradient estimator [20] have been proposed over the years, but in this work we use exponential moving average of g as the baseline for its simplicity and effectiveness:

$$\mathbb{E}_{p_\theta(\mathbf{z})}\left[\left(g\left(\mathbf{z}\right) - b\right) \triangledown_\theta \log p_\theta\left(\mathbf{z}\right)\right]$$

Baseline b is re-calculated at each mini-batch step by updating it with a new reward: $b' = \lambda_b \cdot b + \left(1 - \lambda_b\right) \cdot g\left(\mathbf{z}\right)$, where λ_b is the baseline decay rate. We also employ entropy regularization to encourage exploration [30].

3.3. Comparison to "Hard" Attention

In contrast to the more ubiquitous ”soft” attention mechanism [3, 28], stochastic attention models attention mechanism as a discrete action of choosing a particular object rather than the normalized weighted sum of the candidates [30]. The probability of choosing an object is modeled by a multinoulli parameterized by the attention distribution. [30] uses variational lower bound to analytically derive the partial derivative equation of the log-likelihood:

$$L = \log p\left(y\right) = \log \sum_{z} p\left(z\right) p\left(y|z\right)$$
$$\geq \sum_{z} p\left(z\right) \log p\left(y|z\right) = L_z$$

$$\begin{aligned}\nabla L_z &= \sum_{z} p\left(z\right) \left(\nabla \log p\left(y|z\right) + \log p\left(y|z\right) \nabla \log p\left(z\right)\right) \\ &\approx \frac{1}{M} \sum_{i}^{M} \left(\nabla \log p\left(y|\tilde{z}\right) + \log p\left(y|\tilde{z}\right) \nabla \log p\left(\tilde{z}\right)\right)\end{aligned} \tag{5}$$

where $\tilde{z}$ is sampled from $p\left(z\right)$. The equation for Monte Carlo approximation (Equation 5) is not new, as the first term corresponds to the classification error and the second term corresponds to the gradient estimator (Equation 4) in our approach. It shows that the same gradient equation can be derived from two perspectives (variational bayes and stochastic gradient estimation).

The key difference between stochastic attention and the selection mechanism in our approach is that the distribution of $\mathbf{z}$ is categorical in stochastic attention (i.e. $\sum_i z_i = 1$), whereas it follows a multidimensional Bernoulli (i.e $\sum_i z_i \geq 0$) in our approach.

4. Related Work

Relational Reasoning and RNs. Earlier works [4] have identified the importance of relational reasoning in machine intelligence. Since then, RNs' specialty in various relational reasoning tasks has inspired the emergence of variations of the architecture. One of the shortcomings of RNs is that they can only capture binary relationships, inherently limited by the model structure. Some works [21, 6] address the issue by proposing stacking RNs one another either recursively or separately, which allows RNs to capture higher order of relationships. On the other hand, there are efforts [25] to generalize relations into predicates, which brings models closer to how humans process knowledge.

Attention Mechanisms. Since the advent of attention mechanism [3] in the machine translation task, many domains have adopted the mechanism for various other tasks, such as image captioning [30], visual question-answering [15], language understanding [28, 18], and speech recognition [8], etc. The "soft" variant of attention, in which object features are summed up by normalized attention weights, is ubiquitous due to the ease of integration into existing models, but the stochastic variant is rarer despite empirical evidence that supports its relative superiority. Some recent works have further explored employing binarized hard attention in the domain of object detection [13] and sequence modeling [26], but none has explored the idea in the context of object selection for relational reasoning.

Stochastic Networks. Incorporating discrete latent variables is a rising paradigm in deep learning, as there is biology-inspired motivation behind signal binarization [5]. Generally speaking, humans tend to discretize and categorize knowledge and expand new knowledge upon it, hence it is natural for recent deep learning researches to explore the idea. Discrete latent variables have been explored in various tasks [30, 13, 27], and the fundamentals have been visited by [5, 12, 17]

Visual Intelligence. Recently, the breakthrough of deep learning technology based on the convolutional neural network has made remarkable progress in many fields, such as image classification [11] and image segmentation [7]. Breakthroughs in object detection algorithms have contributed significantly to the advance of downstream tasks [23, 19, 1]. VQA is one of the beneficiaries of such advancements.

5. Experiments

To examine the effectiveness of our model in selecting objects meaningful for pair-wise relation reasoning, we design and conduct experiments on the visual question answering dataset.

5.1. Experimental Settings

Visual question answering is a recent task that requires the machine to identify an open-ended answer to a natural language question limited to subjects in a given image. The task encompasses both aspects of computer vision and natural language processing and requires joint comprehension of image and text [2]. Many datasets have been proposed to test and evaluate visual question answering, one of which is VQA 2.0 [10]. The VQA dataset is a popular real-world dataset for studying the visual reasoning problem mainly for its scale and large coverage of question and scene types.

In the VQA dataset, question-answer pairs are annotated by not one but a set of candidate answers aggregated from the responses of ten crowd workers. In order to accurately evaluate model performances, the authors of VQA suggest evaluating models by the number of responders voted for the predicted answer, capping out at 3: Accuracy $= \min\left(N_{\text{votes}}/3, 1\right)$.

5.2. Implementation Details

In practice, we impose minimum and maximum bounds on the number of objects that can be selected by the discriminative network in SRN. The minimum bound ensures that there is at least certain number of objects selected, preventing errors arising from empty samples (i.e. $\sum_i \tilde{z}_i = 0$), while the maximum bound restricts the model from taking too many computational resources. The values are set

	test-dev			
Methods	All	Yes/no	Number	Other
Pythia v0.1 [14]	46.85	**65.26**	33.88	33.97
Vanilla RN	**47.19**	64.72	34.56	**34.98**
SRN (Ours)	47.18	65.03	**34.62**	34.62

Table 1. Pilot results on VQA 2.0 (Accuracy %). The mini-batch size has been limited to 8, the maximum number for vanilla RN to be trainable on our machine configuration without running out of memory. RN-enhanced models outperform the previous state-of-the-art. Our approach (SRN) takes less computation resources with minimal loss in overall performance.

	test-dev			
Methods	All	Yes/no	Number	Other
Pythia (Reported)	*68.05*	-	-	-
Pythia (Reprod.)	67.14	84.55	45.70	57.16
Attentional RN	67.04	84.50	45.28	57.09
SRN (Ours)	**67.42**	**84.61**	**46.35**	**57.53**

Table 2. Main experimental results on VQA 2.0 (Accuracy %). All Pythia version is v0.1. We include the results of our attempt to reproduce the reported Pythia results on github. Our approach (SRN) has been successfully applied to the real-world visual question answering, taking advantage of all candidate object features.

to 10% and 60% of the total number of objects respectively. We found that mean-pooling performs better than sum-pooling in Equation 1. All layers in RNs are two-layer feed-forward networks with 512 as hidden dimensions and a dropout rate of 0.1. We implement baselines and our model on Pythia v0.1 [14], which is a modularized visual reasoning framework based on Python and PyTorch. Our code and model will be made public[1].

We use Adam [16] optimizer with the scheduled learning rate employed in Pythia. Only the top-3000 answer candidates were used. We use BiLSTM of hidden size 512 to encode questions and initialize with pretrained GloVe word embeddings [22]. For attentional RNs, we set the number of objects to be titrated by the attention mechanism to 12. For SRNs, the baseline decay factor λ_b is set to 0.9, the weight for stochastic gradient estiamtor λ_θ to 1.0, and the weight of entropy regularization λ_h to 0.1. All hyperparameters have been determined through grid-based hyperparameter search. Experiments are run on 4x Tesla P100 GPUs.

5.3. Experimental Results

We conduct various experiments on the VQA task to verify our hypothesis and compare different variants of RNs. We confine our interest to single models.

Pilot Experiments. Before carrying out full experiments on VQA 2.0, we investigate how our approach would fare against vanilla RN in a controlled experimental setting. Since vanilla RN cannot be run on our machine configuration (4 x 16GB P100 GPUs) without running out of memory resources, we reduce the mini-batch size to the maximal level where vanilla RN can be run without errors, which was 8. This reduction in mini-batch size has a significant impact on the absolute performance of the models; however, we are mostly interested in the change in relative performances of RN variants ceteris paribus. The results are shown in Table 5.1. The results show that incorporating explicit relational reasoning in the model allows the model to answer visual questions more accurately. The results also show that our approach has similar performance levels with, or even better than in some aspects such as number question types, vanilla RN consuming significantly less computational resources (as much as the square of the average activation rate of $\mathbf{z}$ in the discriminative network).

Full Experiments. Having established the feasibility of SRN in a miniature setting of VQA, we now examine its applicability in the original setting. We try our best to adhere to the hyperparameter settings of the previous state-of-the-art (mini-batch size was 512). For RNs, we set the mini-batch size to 400, as it is the maximal level that can be run without errors in our configuration. The results (Table 5.3) show SRN perform better than the previous state-of-the-art and even the attentional variant of RN, which uses "soft" attention mechanism to aggregate objects into a fixed number of entities. The largest improvement is in number-related question types, which might be attributed to the distractor-suppression effect achieved through discrete object dropout. As number-related question types require more discrete reasoning (counting and identifying numbers, etc.), they have might benefit most from the distractor-suppression effect.

6. Conclusion

In this preliminary work, we proposed a variant of RN where objects are selected based on the relevance to the downstream task before being processed for relational reasoning. Our model alleviates the inherent scalability issue of pair-wise operations in RNs, allowing models to safely reduce computational costs in situations where the number of candidate objects is significant but not all objects are relevant for relational reasoning. Results in VQA show that our approach is not only efficient but could also potentially improve task performances by suppressing distractors in candidate objects. As future work, we hope to explore more applications of our approach, one of which is to explore the possibility of making hierarchical relational networks feasible using the discrete object selection mechanism. We are also interested in analyzing the interpretability of object selectors learned by SRNs.

[1] will be available at https://github.com/kaniblu/pythia-srn

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on CVPR*, pages 6077–6086, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE ICCV*, pages 2425–2433, 2015.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in NIPS*, pages 4502–4510, 2016.

[5] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[6] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE conference on CVPR*, pages 1989–1998, 2019.

[7] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in NIPS*, pages 8699–8710, 2018.

[8] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in NIPS*, pages 577–585, 2015.

[9] Mikyas T Desta, Larry Chen, and Tomasz Kornuta. Object-based reasoning in vqa. In *2018 IEEE WACV*, pages 1814–1823. IEEE, 2018.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on CVPR*, pages 6904–6913, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on CVPR*, pages 770–778, 2016.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[13] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[14] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[15] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[18] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[20] John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

[21] Rasmus Palm, Ulrich Paquet, and Ole Winther. Recurrent relational networks. In *Advances in NIPS*, pages 3368–3378, 2018.

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on EMNLP*, pages 1532–1543, 2014.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in NIPS*, pages 91–99, 2015.

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

[25] Murray Shanahan, Kyriacos Nikiforou, Antonia Creswell, Christos Kaplanis, David Barrett, and Marta Garnelo. An explicitly relational neural network architecture. *arXiv preprint arXiv:1905.10307*, 2019.

[26] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*, 2018.

[27] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in NIPS*, pages 6306–6315, 2017.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in NIPS*, pages 5998–6008, 2017.

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

[30] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.