Neural Module Networks

Lina Zhu

May 31,2018

1. Introduction

In this paper, a new model structure (NMN) based on neural module network is introduced. This architecture makes it possible to use joint training of neural collection "modules" to answer questions about the natural language of the image, the "module" dynamically according to the language structure of the deep web. Specifically, the visual question answering task has important application value in man-machine interaction, search and accessibility, and is a subject of great research interest in recent years [3]. This work requires a complex understanding of the visual scene and natural language. Recent successful approaches have represented the problem as a bunch of words, or used it to encode a problem as a recursive neural network, training a simple classifier on coding problems and images. Unlike these single methods, another line of work for text QA and image QA USES a semantic parser to break the problem down into logical expressions. These logical expressions are calculated based on pure world logical representations and can be provided directly or extracted from images[1]. In this paper, we propose a technique combining the representation capability of neural network with the flexible component structure provided by the symbolic semantic method. In figure 1, we first focus on the dog, which passes its output to a location descriptor. Depending on the underlying structure, the messages passed between modules may be original image features, considerations, or classification decisions. Each module is mapped from a specific input to an output type. Different types of modules are displayed in different colors. The attention-generating module (such as dog) is shown in green, while the tag module is shown in blue. Importantly, all modules in NMN are independent and composable, which allows the calculation of each problem instance to be different and may not be observed during training. In addition to NMN, our final answer USES a looping network (LSTM) to read the question, which is an additional step that is important for modeling common knowledge and data set bias [2]. We first describe the neural module network, which is a general framework for combining heterogeneous, jointly trained neural modules into deep networks. Next, for the visual QA tasks, we'll show how to construct NMNs based on the output of the semantic parser and use them to successfully complete the set of visual question answering tasks.

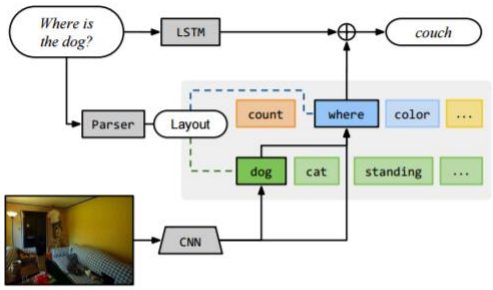

Figure 1. A schematic representation of our proposed modelthe shaded gray area is a neural module network of the kind introduced in this paper.

2. Visualize the neural module network of QA

2.1. From strings to networks

After building the module manifest, we now need to assemble them into the layout specified in the problem. The transformation from natural language problem to instantiated neural network is carried out in two steps. First, we map from natural language problems to layouts,specifies the set of modules used to answer a given question and the connections between them. Next we use these layouts to combine the final prediction network. We use the standard pre-trained tools of existing language resources to obtain a structured representation of the problem. Future work may focus on learning this forecasting process with the rest of the system. These symbols indicate that the structure of the prediction network has been determined, rather than the identification of the modules that make up them. The final allocation of modules is entirely determined by the structure of the parsing. All the leaves become search modules, all the internal nodes become transformation or combination modules based on their dependencies, and the root node becomes a domain based description or measurement module.

Given the mapping from query to network layout described above, we have a network structure, input image, and output tag for each training example. In many cases, these network structures are different, but there are binding parameters. A network with the same high-level structure but with different instantiations of a single module can be processed in the same batch to achieve efficient computation. As mentioned above, the conversion process is part of the task specification - we find relatively simple expressions are the best, for the problem of natural images and synthetic data (by design) need deeper structure. Table 1 provides some summary statistics.

Table 1. : Structure summary statistics for neural module networks used in this paper. types is the set of high-level module types available.

	types	instances	layouts	depth	size
VQA	find, combine	877	51138	3	4
SHAPES	find, transform,	8	164	5	6

3. Conclusion

This paper introduces the neural module network, which provides a general framework for the collection of learning neural modules, which can be dynamically assembled into any deep network. We have shown that this approach achieves the most advanced level of performance in answering visual questions on an existing data set, especially on an object or attribute. In addition, we introduced a new data set for the high composition problem of simple arrangement of shapes and showed that our method is much better than previous work. So far, we have maintained a strict separation between predicting network structure and learning network parameters. It is not hard to imagine that these two problems may be solved jointly, and the network structure remains uncertain throughout the training and decoding process. This can be done by a whole network, through the use of some advanced mechanism to "follow" the relevant part of the calculation, or with the study of semantic parser[1] existing tool to integrate. In a follow-up to this article, we'll show you how to combine learning module behavior and parser.

References

[1] J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics (TACL)*, 2015. 1, 2

[2] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[3] A. Agrawal S. Antol, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1