Residual Attention Network:A new baseline model for visual question answering

*

1st Salma Louanas
Laboratory of Informatics and its Applications of M'sila
Department of Computer Science, University of M'sila
M'sila, Algeria
salma.louanas@univ-msila.dz

2nd Hichem Debbi
Laboratory of Informatics and its Applications of M'sila
Department of Computer Science, University of M'sila
M'sila, Algeria
hichem.debbi@univ-msila.dz

***Abstract*—Answering questions over images is a challenging task, it requires reasoning over both images and text. In this paper, we introduce Residual Attention Network(RAN), a new visual question answering model, and compare it with baseline models such as stacked attention model and CNN-LSTM model. We find that our model performs better than these baseline models. In addition to our model, we also evaluate several holistic models and compare them with neural module networks frameworks, and the results show that neural modules networks perform better in questions reasoning. All the experiments have been done on the CLEVER dataset, which is a recent VQA dataset for evaluating multiple-step reasoning VQA models.**

***Index Terms*—Visual question answering, baseline, Holistic, Neural module network**

I. Introduction

Visual Question Answering (VQA) is a new multi-disciplinary task that was first proposed in [1]. VQA has captured the attention of both communities, computer vision as well as natural language processing.

A VQA system requires an image and a related question as an input, and determines the correct answer in natural language as an output (See Fig. 1).

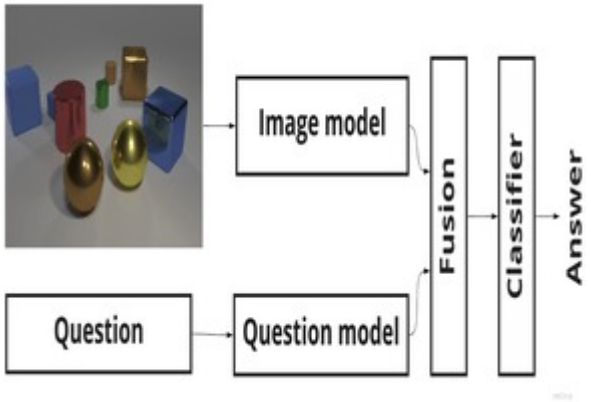

Fig. 1. Visual question answering system.

II. Introduction

VQA is a challenging task because it requires extracting and understanding semantic information in the visual and textual channels.

The success of deep learning methods in both computer vision and NLP motivates the researchers to use the concept of jointly embedding both image and text into features space, which was first used in image captioning [2]–[4].

The basic idea of VQA models is to extract features vector from image using Convolution neural networks (CNNs), and encode the related question as features vector using Long Short Term Memory (LSTM), and then, it fed the both representations into a common space to infer the answer.

To achieve high accuracy and interpretability in the VQA task, the model needs more than just processing image and text. The model is also required to reason over the two representations.

Reasoning in VQA needs to understand subtle relationships among multiple objects, as well as to focus on the specific regions that are relevant to the answer.

Deep learning methods are highly accurate in term of performance, because they are optimized to learn easily on dataset bias [1], [5], but not in term of reasoning [6].

In this paper we introduce a new baseline model Residual attention network (RAN)(See Fig.2) for the VQA task. The proposed model is adapted from [7] that was mainly used for image classification, and gives promised results in image classification. To better show the efficiency and limitations of attention-based methods, we also perform heavy and large experiments on several holistic and neural modules frameworks on the CLEVR dataset [6]. We targeted main baselines models such as SAN [8] and XNM [9].

The structure of this paper is as follows. In section 2 we cover the most VQA related works. Section 3 presents CLEVR dataset. In section 4 our method RAN. Next we show experiments of our model and the other compared methods in section 5. Finally we conclude this study in section 6.

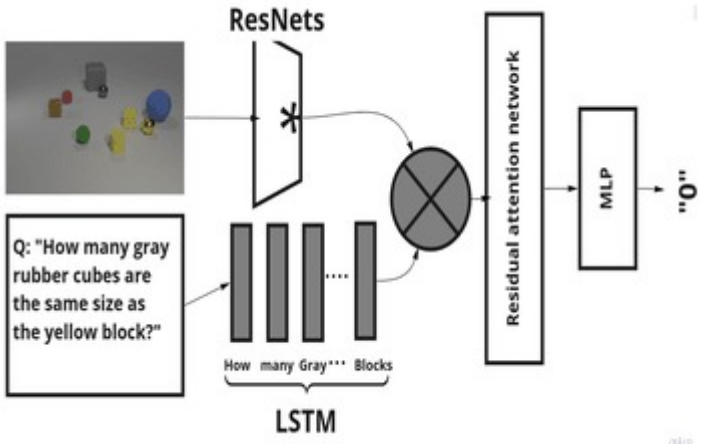

Fig. 2. Residual attention network .

III. Related works

The existing methods on VQA task can be divided into two categories: holistic approaches [8] and neural module networks (NMN) [10]. The main difference between the two categories is that NMNs approaches decompose the question into a set of sub-tasks and use specific modules to handle them, while holistic approaches handle all the question homogeneously.

A. Holistic approaches

A model that can answer a question over an image with several concepts (objects, relations) needs first to locate and identify concepts referred in the question, then excludes irrelevant objects, and finally identifies the most important regions to infer the answer. The steps referred above led to build models conduct the sequential interaction. SAN [8] uses stacked attention to extract the visual information, FiLM [11] introduces multiple conditional batch normalization layers to combine the image and question representations, and MAC [12]performs multi-steps reading-writing operations to extract visual information and update its memory. All these models use implicit reasoning to interact between image and question and do not dismantle the question.

B. Neural module networks

Existing methods for visual reasoning attempt to directly map inputs to outputs using black-box architectures without explicitly modeling the underlying reasoning processes. In this context [10] proposed Neural module network (NMNs), which explicitly dismantle the reasoning procedure into several sub-tasks, and have specialized modules to handle the sub-tasks easily and more transparently. Existing neural module networks (NMNs) answer natural language questions about images using sets of jointly-trained neural modules [9], [10], [13]–[18].

NMN [10],N2NM [14],PG+EE [17] and TbD-net [16] share the same prediction procedure that is performed by parsing the question and dismantling the reasoning procedure into a set of sub-tasks(Fig 3). XNM [9] defines a model that contains three components: a scene parser (scene graph), a question parser (module program) and the third component (program executor) tries to execute the program over the scene graph.

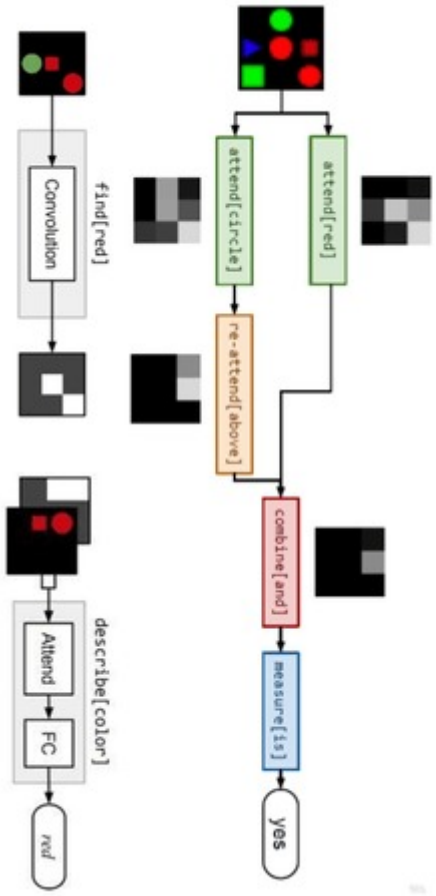

Fig. 3. Neural module networks system [10].

NS-VQA [19] introduces a model that first parses the image into (de-renderer) to obtain a structural image representation, and parse the question into (program generator) to generate a hierarchical program, then it uses a program executor to execute the program generator over the structural scene representation to infer the answer.

TbD-nets [16]pars the input question into a layout policy, which is used to assemble the question into a module layout(neural network), that learns how to perform each sub-tasks.

IV. CLEVR Dataset

CLEVR [6] (Compositional Language and Element Visual Question Reasoning)(Fig.4) . The CLEVR dataset was introduced to analyse the ability of reasoning of visual systems. it is a collection of 100,000 synthetic images of 3D shapes such as spheres and cylinders. The questions included in this dataset are used to test the visual reasoning capabilities (See Fig.5) of a VQA model. There are different categories of questions associated with each image. In the training set, there are 70,000 images with 699,989 question-answer pairs. The

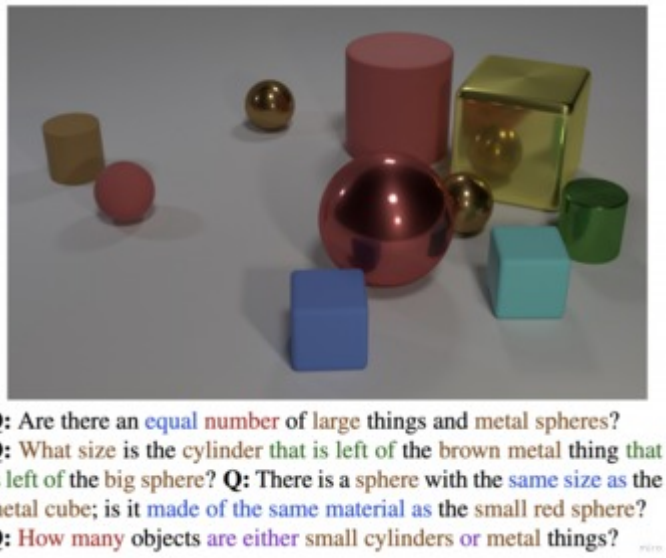

Fig. 4. Compositional Language and Element Visual Question Reasoning dataset [6].

validation and test sets contain 15,000 images with 149,991 and 14,988 question answer pairs respectively.

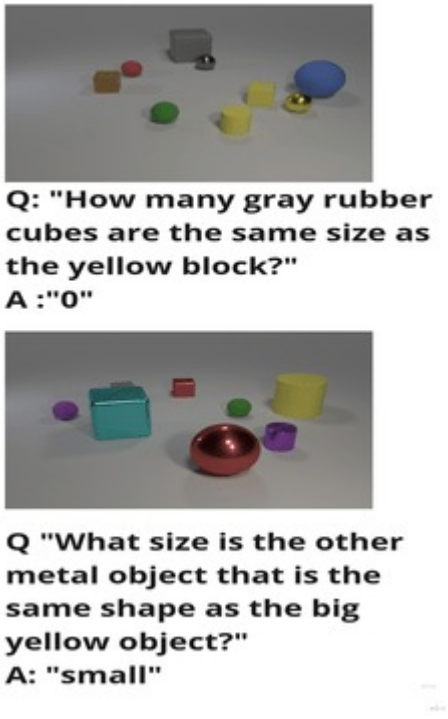

Fig. 5. Sample of CLEVR reasoning questions.

V. Residual attention networks (RANs)

The overall architecture of RANs is shown in (Fig 2) (section 1) : image representation, question representation and the residual attention representation.

A. Image representation:

Residual neural networks or ResNets [20] are proposed in order to solve the vanishing gradient problem. The ResNets architecture introduced the residual blocks concepts, this architecture based on skip connection techniques(Fig 6), that if any layer disadvantages the performance then it will be skipped by regularization. Where x is the input and the residual mapping to be learned is represented by the function $F(x)$.

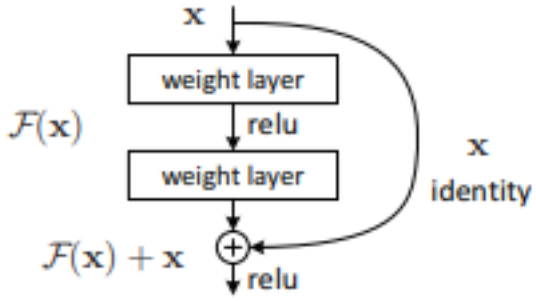

Fig. 6. Residual network block [20]

In our model we use ResNet in three levels: first in image feature extraction, and then in image projection, and also in the fusion phase we use RANs that it based on ResNets.

B. Question representation:

In question representation we use the same representation used in SAN [8] based on LSTM encoder.

C. Residual attention representation

RANs is a fusion attention model that fuses image feature matrix and question feature vector by element-wise multiplication and fed them into attention layer.

The RANs is an attention layer that consists of staked residual blocks and a softmax function. The RANs return the relevant image regions, that will be fed into a multi-layer perceptron (MLP) that predicts the answer distribution (Fig 7), and then use softmax to infer the final answer.

VI. Experiments

The results reported in our paper were obtained by performing the experiments on a single GPU RTX2070 Super with 8GB of GPU memory and 32GB of RAM.

All experiments in this work are based on CLEVR validation set(the annotations of CLEVR test set are not public [6]). Baseline models play important role in testing datasets complexity. We first compared the base line models in the same conditions (batch size=16,iteration=100,000) and on the same environment (pytorch). SANs and RANs are considered as holistic approaches, because they use the multi-steps reasoning concepts. For the neural modules networks we evaluated XNMs [9] and Clevr-iep [17].

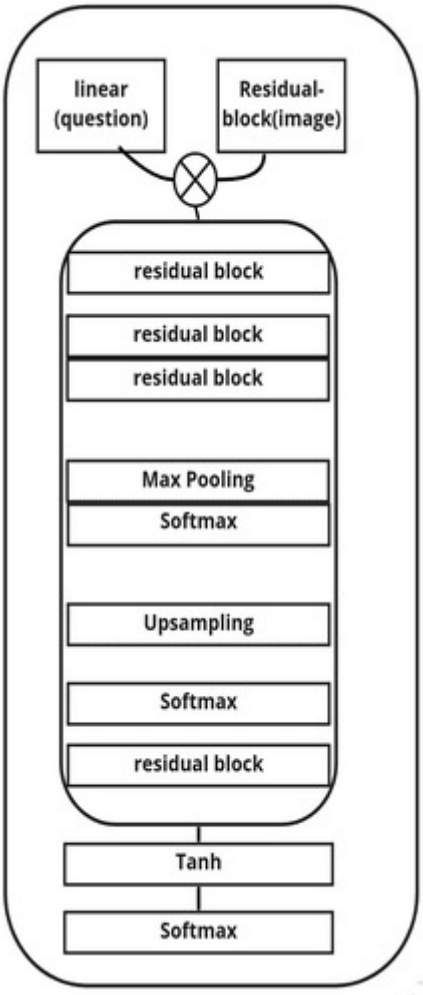

Fig. 7. Residual attention module block.

A. Baseline models

CNN-LSTM and SANs baseline were first tested on the CLEVR dataset by [6]. We reproduce these tow baselines in the same conditions as in [17].

CNNLSTM: uses CNN to encode the image and LSTM's last hidden state to encode the question, and concatenate them. Then the last representation is fed to MLP classifier to infer the answer.

SANs [8]: encodes image and question as the same in CNN-LSTM model,then fused using two blocks of soft spatial attention; a linear transform to infer the answer.

RANs: we used Residual block to encode image, and LSTM to encode question, and fuse the two representations using residual attention networks; and MLP classifier to infer the answer.

Table 1 shows comparison of our model against baseline models. RANs show better performance.

Table 2 presents NMNs methods comparison, we run XNMs in the same conditions referred in [18]; High resolution image features(512,28,28) and 256 batch size. On the contrary, in TbD-net we first use high resolution features and 128 batch size the same as in [16] the training stopped early in the first epoch, for that we use image features(1024,14,14) and batch size=16. The performances on NMNs shows that TbD-nets perform better then XNM.

TABLE I
BASELINE MODELS COMPARAISON

Methods	Accuracy%
LSTMCNN	49.34
SAN	51.78
SAN+MLP	52.74
RAN	**61.89**

TABLE II
NEURAL MODULE NETWORKS MODELS COMPARAISON

Methods	Accuracy%
TbD-Net	**98.55**
XNMs	94.29

Table 3 shows an overall comparison between holistic and neural module approaches.The results shows the out-performance of neural module networks over the holostic ones.

Neural modules perform well because they use specialized modules for each sub-task, for example, for count or compare task they introduce modules that perform well on these specific tasks.

TABLE III
OVERALL COMPARISON

Methods	Accuracy%
TbD-Net	**98.55**
LSTMCNN	49.34
SAN	51.78
SAN+MLP	52.74
RAN	61.89
XNMs	94.29

VII. Conclusion

We have presented Residual attention networks, which is a VQA model composed of residual blocks. We also reproduce several holistic and neural modules approaches and make a comparative study. The obtained results on the baseline models showed the complexity of the CLEVR dataset. The results also show the performance of neural modules approaches on this dataset over the holistic approaches.

As future work, we plan to consider other challenging datasets such as VQA 2.0. VQA 2.0 is another VQA dataset that is challenging from other aspects, thus, the comparison between our technique and the holistic approaches in general from a side, and the neural modules from another would be fair enough.

References

[1] S. Antol, A. Agrawal, 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*, 2015, pp. 2425–2433.

[2] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[3] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems*, vol. 27, 2014.

[4] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[6] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[7] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[8] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[9] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8376–8384.

[10] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 39–48.

[11] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[12] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," *arXiv preprint arXiv:1803.03067*, 2018.

[13] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," *arXiv preprint arXiv:1601.01705*, 2016.

[14] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 804–813.

[15] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable net: an efficient subgraph-based framework for scene graph generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 335–351.

[16] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4942–4950.

[17] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2989–2998.

[18] R. Hu, J. Andreas, T. Darrell, and K. Saenko, "Explainable neural computation via stack neural module networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 53–69.

[19] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding," *Advances in neural information processing systems*, vol. 31, 2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.