Research Proposal - Deep Convolutional Neural Network with Deep Reinforcement Learning For Visual Question Answering

Nishank Singhal
B.E(honour) Computer Science
Bits Pilani Dubai Campus

Computer vision(CV) and Natural Language Processing(NLP) both play an important role in transforming industries and help in reshaping the economy [1]. To understand the society in a better way, social sciences and humanities experts today can enhance their capabilities by extracting information from large scale image and text corpora [2]. However, a fundamental problem that still exists in Artificial Intelligence (AI) is teaching it to reason and understand the world's enormous knowledge contained in the videos, images, audios, and text languages [3]. While it is always believed that the learning models should be general and flexible, in practice, most of the progress has been achieved is by using classical method of supervised learning, which requires a large amount of training data (annotated) and a lot of human intervention. Having said that, recent advances in machine learning, such as the use of deep learning networks, have made it capable of solving various problems, which have enabled the training - AI using enormous amounts of data. Advances such as these have opened an excellent opportunity for expanding the ability of non-experts to make use of sophisticated models to not only accelerate the social and scientific research cycle, but also deliver camouflaged knowledge that. This will be useful in future research in answering real-world policy questions [4]. More importantly, there is now an opportunity to move beyond classic supervised learning methods. According to me, upcoming developments should involve deployment of deep networks to focus more on creative tasks and common-sense reasoning. These methods can unravel the structure of textual and visual world from the data itself, and learn to synthesize realistic high-dimensional outputs directly. The ultimate goal would be to build machines that can recreate the visual world and help everyone tell visual stories with precision.

The core research question that I have addressed in my journey of projects so far is, "How can we design self-supervised deep learning methods to operate over rich language and knowledge representations?" by working in advancing the state-of-the-art methods and success of deep learning tasks, Scene Classification, Diabetic Retinopathy detection, Sentiment Analysis, and Action Recognition.

This idea created the need for a 'semantic' understanding of images and led my research towards an evolution of a potential Visual Turing Test called Visual Question Answer (VQA) [5]. Essentially the task involves a machine to answer any question about an image such as: 'what this boy is doing' answer: 'playing football'. However, since the questions are free formed and open-ended, the

space of answers is also not limited. It is a task that is currently pushing the boundaries of Artificial Intelligence by attempting to train a model on hundreds and thousands of images and questions that can attain a perfect score. VQA can find its applications in any area which require humans to elicit information from virtual and textual/audio data [6],[7]. Many approaches to object recognition task have been proposed as well as implemented over the past few decades. Yet, there still lacks a general and comprehensive solution to the modern recognition challenges, such as those in security surveillance domain, where the number of CCTV cameras is growing exponentially, and in digital devices that require efficient detection techniques.

I propose to research in three directions within deep learning involving novel tasks that can be performed on VQA [8] using Deep Convolutional Neural Networks (DCNNs)[9] for image classification and Deep Reinforcement Learning (DRL)[10] model for understanding the query asked by the user. I aim to extend the DCNNs to video understanding followed by DRL network to natural language understanding by modelling spatial and temporal information, as well as arranging the amount of data required to train it.

The three directions of my research would be:

a) Preparing a deep neural network which can learn and infer from bi-model visual datasets, like CCTV footages.

b) Deploying natural language tasks with the help of deep networks for which the order of the words in a sentence matters.

c) Training a deep network for VQA to learn generic features using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

The amount of visual data along with queries to learn generic features is essential in investigating how these questions can be used as an aid for developing an improved computer vision and natural language processing model. This would help in examining how much information can be contained in a visual question and also demonstrate how this information can be effectively used. The results of this research can have potential applications for situations such as detecting untoward/violent activity in CCTV footage, developing an automated quality assurance/quantity check for industries and providing customized solutions to the institutions/organizations.

References

[1] Gauthier, I., & Tarr, M. J. (2016). Visual Object Recognition: Do We (Finally) Know More Now Than We Did?. Annual Review of Vision Science, 2, 377-396

[2] Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. JAIR, pages 409–442.

[3] Tobias, L., Ducournau, A., Rosseau, F., Fablet, R., & Mercier, G. (2016). Convolutional Neural Networks for Object Recognition on Mobile Devices: a Case Study. International Conference on Pattern Recognition (pp. 2-7). Cancun: ResearchGate.
[4] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.
[5] A. K. Gupta, "Survey of Visual Question Answering: Datasets and Techniques," Computing Research Repository (CoRR), 2017.
[6] D. Teney, Q. Wu and A. v. d. Hengel, "Visual Question Answering: A Tutorial," IEEE Signal Processing Magazine, pp. 63-65, Nov 2017.
[7] K. Kafle and C. Kanan, "Visual Question Answering: Datasets, Algorithms, and Future Challenges," Computer Vision and Image Understanding (CVIU), vol. 163, pp. 3-20, 2017.
[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016.
[9] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In ICCV, 2017.
[10] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalanditis, Y., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: arXiv 1602.07332. (2016)