RESEARCH PROJECT

Attentional models and Deep Learning

Erik de Godoy Perillo
Supervisor: Profa. Dra. Esther Luna Colombini

Abstract

Attention is a fundamental mechanism in intelligent beings. It is necessary for filtering the big volumes of stimuli constantly arriving and selecting information that is important for a certain task. Deep Learning is currently broadly applied to Artificial Intelligence. The use of attention concepts in current models has also been increasingly frequent, resulting many times in better results for the task being addressed. In this context, this work proposes the elaboration of attentional models based on Deep Learning for problems in Artificial Ingelligence. We aim at obtaining frameworks more generically applicable in broad problem classes such as computer vision, Natural Language Processing, Differential Computation and others.

State University of Campinas

June 4, 2018

I. Introduction

We are constantly receiving high volumes of multimodal stimuli from both external sources – such as visual, auditive signals – and internal sources – proprioception, memories et cetera. It would be very inefficient to proccess all the information with the same intensity given that a big portion of it is irrelevant for the task being executed at the moment and we have limited cognitive capacity. When we read, our vision does not focus on all words equally, but rather on a small subset of the text at a time. When we're addressing a given subject, it tends to mediate the focus in the memory search process, essentially retrieving memories that are useful for the subject: many other irrelevant memories are not used. It often happens that something conspicuous – such as a bird abruptly appearing in front of us or a sudden sound – quickly draws our focus, ''stealing" it from what was previously being focused on. The ability to filter and select stimuli that is relevant for a given cognitive task, keeping the focus for an extended period of time and directing it to new stimuli when appropriate is fundamental to human beings and other complex forms of life. We name this ability ''attention" [4].

Attention can potentially play an important role in the development of Artificial Intelligence (AI). Areas such as computer vision often involve a big quantity of data and most of time only part of image is relevant to the task at a given moment. In robotics, attention can be substantially useful: robots that navigate in complex and dynamic environments need systems to enable them to handle data from all sensors so that relevant objects and parts of the scene are promoted to further processing and decision making – which needs to be done in real time. Furthermore, paying attention to abrupt changes in the environment that may affect the robot's navigation is important for the robustness, success and safety of the application. Computational models of attention have been elaborated for years. A classic example is VOCUS [5], which was proposed to simulate the visual attention process in humans. Many of its mechanisms are based in concepts from psychology.

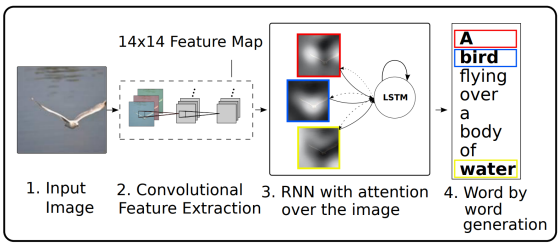

Figure 1: Diagram of natural language image description using attention (from [3]).

In recent years, there have been significant improvements in AI due to the popularization of Deep Learning (DL) [6]. As we will discuss in following sections, the technique consists of artificial neural networks architectured in a hierarquical manner. DL showed to be effective in a variety of tasks in computer vision [9][8], audio processing [11] and Natural Language Processing (NLP) [12] mainly due to its ability to learn what features to be extracted (rather than relying on hand-crafted features). Along with the transposition from classic models to DL approaches, an increasingly high number of works on the field have been using concepts related to attention in combination with DL to achieve better results. One example is image captioning: the task consists of giving a natural language description of a given image. The work in [3] shows that the task benefits from sequentially focusing on different parts of the image in sequence. It is achieved by the use of an attentional component in the model. Other examples include linguistic translation [1], audio recognition [2] and neural computation [7]. More examples will be discussed in-depth in following sections.

A. Objectives

The recent adoption of attention by a variety of Deep Learning models has shown significant improvements in different tasks. However, it is conjectured that many other tasks that still don't use attention would benefit from the concept. It is believed that a variety of tasks related to robotic navigation, for example, can be approached by

using models with attention. Furthermore, we note that attention models currently being used are very specific to each problem in question. Some works propose a higher level of generalization [10], but we believe it is possible to go further than that. Therefore, the specific objectives of this work are:

- To perform an extensive literature review on the use of attention along with modern DL techniques;
- To identify specific problems in different classes (robotics, vision, NLP, differential programming) with improvement potential by the use of attention;
- To study the viability of generalization of attention models to broader problems in different classes;
- To implement the proposed model, evaluating it in an application (preferably related to robotics).

II. Background

A. Deep Learning

TODO:

- brief history and timeline;
- hierarquical features
- types (cnns, rnns, ntms)
- non-convex optimization (SGD etc)

B. Attention

TODO:

- definition
- general concepts
- bottom up
- top down

III. Related Work

TODO: detailed examples on DL + attention. maybe cite our previous work here?

IV. Methodology

TODO:

- description of stages: lit review, search for problems, generalization, application

A. Schedule

TODO: the schedule.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *CoRR* abs/1409.0473 (2014). arXiv: 1409.0473. URL: http://arxiv.org/abs/1409.0473.

[2] William Chan et al. “Listen, Attend and Spell”. In: *CoRR* abs/1508.01211 (2015). arXiv: 1508.01211. URL: http://arxiv.org/abs/1508.01211.

[3] KyungHyun Cho, Aaron C. Courville, and Yoshua Bengio. “Describing Multimedia Content using Attention-based Encoder-Decoder Networks”. In: *CoRR* abs/1507.01053 (2015). arXiv: 1507.01053. URL: http://arxiv.org/abs/1507.01053.

[4] E.L. Colombini, A. da Silva Simoes, and C.H. Costa Ribeiro. “An Attentional Model for Autonomous Mobile Robots”. In: *IEEE Systems* 99 (2016), pp. 1–12.

[5] Simone Frintrop. “VOCUS: a visual attention system for object detection and goal-directed search”. In: *IN LECTURE NOTES IN ARTIFICIAL INTELLIGENCE (LNAI)*. Springer, 2005.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[7] Alex Graves, Greg Wayne, and Ivo Danihelka. “Neural Turing Machines”. In: *CoRR* abs/1410.5401 (2014). arXiv: 1410.5401. URL: http://arxiv.org/abs/1410.5401.

[8] Kaiming He et al. “Mask R-CNN”. In: *CoRR* abs/1703.06870 (2017). arXiv: 1703.06870. URL: http://arxiv.org/abs/1703.06870.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-

classification - with - deep - convolutional - neural-networks.pdf.

[10] Volodymyr Mnih et al. “Recurrent Models of Visual Attention”. In: *CoRR* abs/1406.6247 (2014). arXiv: 1406.6247. URL: http://arxiv.org/abs/1406.6247.

[11] Aäron van den Oord et al. “WaveNet: A Generative Model for Raw Audio”. In: *CoRR* abs/1609.03499 (2016). arXiv: 1609.03499. URL: http://arxiv.org/abs/1609.03499.

[12] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.