
**UNICAMP**
**Institute of Computing**


# Attention as a leverage for Deep Learning

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

University of Campinas

July 27, 2019

**Abstract**

Attention is fundamental for intelligent beings. It is necessary for filtering the significant volumes of stimuli we constantly receive and for applying the adequate mental resources to perform tasks. Deep Learning is currently broadly applied to Artificial Intelligence. The use of Attention in Deep Learning has been increasingly frequent, resulting many times in better results. In this context, this work proposes the study and elaboration of approaches to use Attention in Deep Learning for more power and efficiency to solve problems in Artificial Intelligence. We aim at obtaining a framework generically applicable in broad problem classes such as Computer Vision, Natural Language Processing, Program Composition and others.

# Introduction

We continually receive high volumes of multimodal stimuli from both external sources – such as visual, auditive signals – and internal sources – such as proprioception and memories. It would be very inefficient or even impossible to process all the information with the same intensity once a significant portion of it is irrelevant for the task executed at the moment and considering that we have limited actuation capacity. When we read, our vision does not focus on all words equally, but instead on a small subset of the text at a time. When we are addressing a given subject (in a train of thought), it tends to mediate the focus in the memory search process, mostly retrieving memories that are useful, whereas 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. The abilities to filter and select stimuli that are relevant for a task, to keep the focus for an extended period and to adequately direct mental processes is fundamental to human beings and other sophisticated forms of life. We name this set of abilities **Attention** [5].

Attention can potentially play an essential role in Artificial Intelligence (AI). The pursue of intelligent machines is an old effort in Computer Science [17] and is still very relevant today due to the potential to radically benefit society. Although there have been significant advancements in the field of AI, it is broadly accepted that machines still cannot perform specific complex tasks nearly as efficiently as humans or some animals and the path to achieving more intelligence is still unclear, with many different proposals [13]. Part of the problem comes from the difficulty to accurately define intelligence itself, but surveys of the works on the subject [12] suggest that a reasonably accepted concept is *the ability to perform elaborate tasks in complex and dynamic environments to achieve a wide variety of goals*. From the narrow to the broader aspects of intelligence, the functionalities of Attention are of great importance – and it increases as the level of intelligence considered increases [9].

A considerable amount of advancements in AI in recent years comes from the popularization of Deep Learning (DL) [11]. As we will discuss in the following sections, the technique mostly consists of artificial neural networks architectured in a hierarchical manner. DL showed to be effective in a variety of tasks in Computer Vision [10][8], audio processing [15] and Natural Language Processing (NLP) [18], mainly due to its ability to learn what features should 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 (Figure 1) where the task consists of giving a natural language description of a given image. The work presented in [4] shows that the task benefits from sequentially focusing on different parts of the image in a sequence, through the use of an attentional component in the model. Other examples – which will be discussed in-depth in following sections – include linguistic translation [1], audio recognition [3] and neural computation [7]. These are evidence that concepts of Attention have indeed been useful for the field.

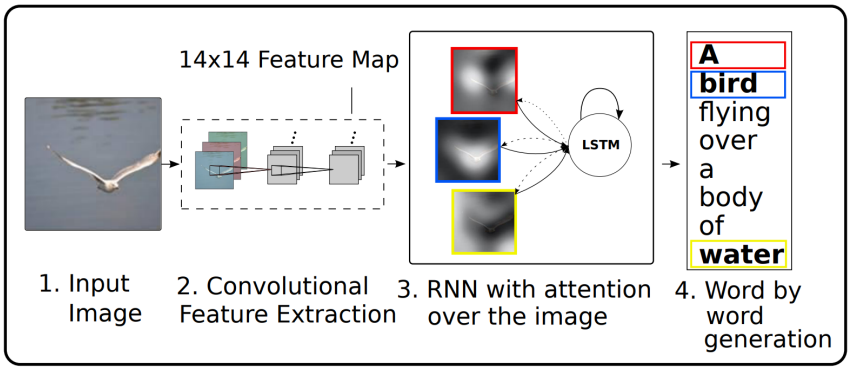


Figure 1: Diagram of natural language image description using Attention (from [4]).

### Motivation and Objectives

In spite of the recent adoption of Attention by a variety of Deep Learning models and the significant improvements it has shown, we conjecture that there are still many other tasks that are still not explored. Current works also tend to focus more on the filtering functionality of Attention, but there are other aspects – such as the allocation of mental resources over time – that can be of potential benefit (we further discuss the taxonomy of Attention in following sections). 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 [14], but we believe it is possible to go further. Therefore, the specific objectives of this work are:

- To perform an extensive literature review on the use of Attention in modern Deep Learning;
- From the perspective of a theoretical framework of Attention (from areas such as psychology and neuroscience), identify in the current literature opportunities in the form of theoretical aspects to be explored that are of potential benefit for the area;
- To identify a specific problem (in robotics, vision, natural language or other area) with improvement potential through the use of Attention, proposing and implementing a solution based on the findings of the work to validate the ideas and evaluate them in an application.

The main contribution of the work proposed we wish to accomplish is to *establish a theoretical framework of Attention as a series of components and its applicabilities to Deep Learning.* Recent works show that the effort on establishing more general concepts and frameworks for Deep Learning design has been broadly useful. Examples include the ideas of *Curriculum Learning* [2] and *Generative Adversarial Networks* [6].

# Activities

The work can be summarized in three main activities (or **phases**):

- **1. Literature Review**: an extensive survey on current uses of Attention in modern Deep Learning.
- **2. Proposal of an Attention framework for Deep Learning**: defining a component set of Attention elements currently used in Deep Learning design from results of the previous phase and further survey.

- **3. Implementation and validation of Attention in Deep Learning**: proposing a model with components of Attention and evaluating it on a set of tasks.

More specifically, the activities to be executed are:

- **A1.1 - Theoretical definition of Attention and its components:** From a variety of previous works [9][5], we establish a theoretical framework of Attention on which all later work will be based. It is worth noting that this theoretical framework is not necessarily the same as the framework we propose to produce specifically for Deep Learning in phase 2. It will work as the base that will be used to organize and classify the current literature.
- **A1.2 - Elaboration of survey:** Exploration of selected work under the point of view of the theoretical framework established in **A1.1**. For each work, we identify the main components of Attention the authors use, the consequences for the performance in the application domain and elaborate a critical evaluation.
- **A2.1 - Establishment of Attention components for specific Deep Learning domains:** From the theoretical framework obtained in **A1.1** and the exploration of current uses and results in **A1.2**, we devise sets of useful components of Attention for specific main problem domains in which Deep Learning is broadly used, such as image classification, text-to-speech, language translation, image segmentation.
- **A2.2 - Establishment of Attention framework for Deep Learning:** From the theoretical framework obtained in **A1.1**, exploration of current uses and results in **A1.2** and results from **A2.1**, we elaborate a set of components of Attention under a single framework to be applied to more general areas of use of Deep Learning, such as Computer Vision, Sequence Processing, Program Composition.
- **A3.1 - Arrangement of experiments:** From the framework obtained in phase 2, we select a set of problem domains (such as text-to-speech), Deep Learning models to use, components of Attention to implement and metrics to evaluate the task. The activity aims at selecting all main devised components from phase 2 in order to evaluate the real consequences of their adoption against what was predicted.
- **A3.2 - Execution of experiments:** We implement and execute the planned experiments following a pre-defined protocol that pays particular attention to reproducibility.
- **A3.3 - Evaluation of experimental results:** We evaluate the results using established metrics for each experiment, elaborating discussions that include exciting aspects of the results in general and comparisons between the theoretical predictions and tangible outcomes. It is worth noting that the metrics we'll use will vary depending on the specific problem, but they will always be selected to reflect the improvement of the models with the use of attention.

# Use of results

The findings of the work may be of benefit for the Deep Learning community. On a previous and related work [16], the authors built a fully convolutional network for the prediction of visual saliency maps, achieving top-10 results in a variety of metrics on the MIT300 benchmark. The results of the work (model weights/code) were made available to the public. In a similar fashion, we intend to make all findings public, as well as make the data/code obtained openly available – along with documentation and tutorials.

# 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] Yoshua Bengio et al. "Curriculum learning". en. In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. Montreal, Quebec, Canada: ACM Press, 2009, pp. 1–8. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553380. URL: http://portal.acm.org/citation.cfm?doid=1553374.1553380 (visited on 09/21/2018).

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

[4] 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.

[5] 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.

[6] Ian J. Goodfellow et al. "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]* (June 2014). arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 09/21/2018).

[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] Helgi Helgason. "General Attention Mechanism for Artificial Intelligence Systems". In: (May 2013).

[10] 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.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". en. In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539 (visited on 09/21/2018).

[12] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *CoRR* abs/0712.3329 (2007). arXiv: 0712.3329. URL: http://arxiv.org/abs/0712.3329.

[13] Tomas Mikolov, Armand Joulin, and Marco Baroni. "A Roadmap towards Machine Intelligence". In: *CoRR* abs/1511.08130 (2015). arXiv: 1511.08130. URL: http://arxiv.org/abs/1511.08130.

[14] Volodymyr Mnih et al. "Recurrent Models of Visual Attention". In: *arXiv:1406.6247 [cs, stat]* (June 24, 2014). arXiv: 1406.6247. URL: http://arxiv.org/abs/1406.6247 (visited on 09/11/2018).

[15] 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.

[16] Erik Perillo and Esther Colombini. “Efficient Visual Attention with Deep Learning”. en. In: *IEEE International Conference on Systems, Man, and Cybernetics*. To be published. Oct. 2018. URL: https://goo.gl/6DTfcL.

[17] Alan M. Turing. “Computing Machinery and Intelligence”. In: *Mind* (1950).

[18] 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.