Universidade Estadual de Campinas
Instituto de Computação

Attentional models and Deep Learning

Erik de Godoy Perillo

Supervisor: Prof.a. Dr.a. Esther Luna Colombini

Research Project
6 de agosto de 2018

Abstract

Attention is a fundamental mechanism in intelligent beings. It is necessary for filtering the significant and constant volumes of stimuli we receive and for selecting information that is essential for a particular task. Deep Learning is currently broadly applied to Artificial Intelligence. The use of attention and Deep Learning has been increasingly frequent, resulting many times in better results for the task addressed. In this context, this work proposes the elaboration of attentional models based on Deep Learning for problems in Artificial Intelligence. We aim at obtaining frameworks more generically applicable in broad problem classes such as Computer Vision, Natural Language Processing, Differential Programming and others.

1 Introduction

We are continually 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 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 cognitive 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, it tends to mediate the focus in the memory search process, essentially 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 ability to filter and select stimuli that are relevant for a task, keeping the focus for an extended period and directing it to new stimuli when appropriate is fundamental to human beings and other sophisticated forms of life. We name this ability ''attention" [4].

Attention can potentially play an essential role in the development of Artificial Intelligence (AI). Areas such as computer vision often involve a significant quantity of data and most of the time only part of an image is relevant to the task. 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 is necessary for real-time applications. Furthermore, paying attention to abrupt changes in the environment that may affect the robot's navigation is vital for the success, robustness, and safety of the application. Computational models of attention have been elaborated for years. A classic example is VOCUS [5], a model proposed to simulate the visual attention process in humans. Many of its mechanisms are based on concepts from psychology.

In recent years, there have been significant improvements in AI due to the popularization of Deep Learning (DL) [6]. As we will discuss in the following sections, the technique consists of artificial neural networks architectured in a hierarchical manner. DL showed

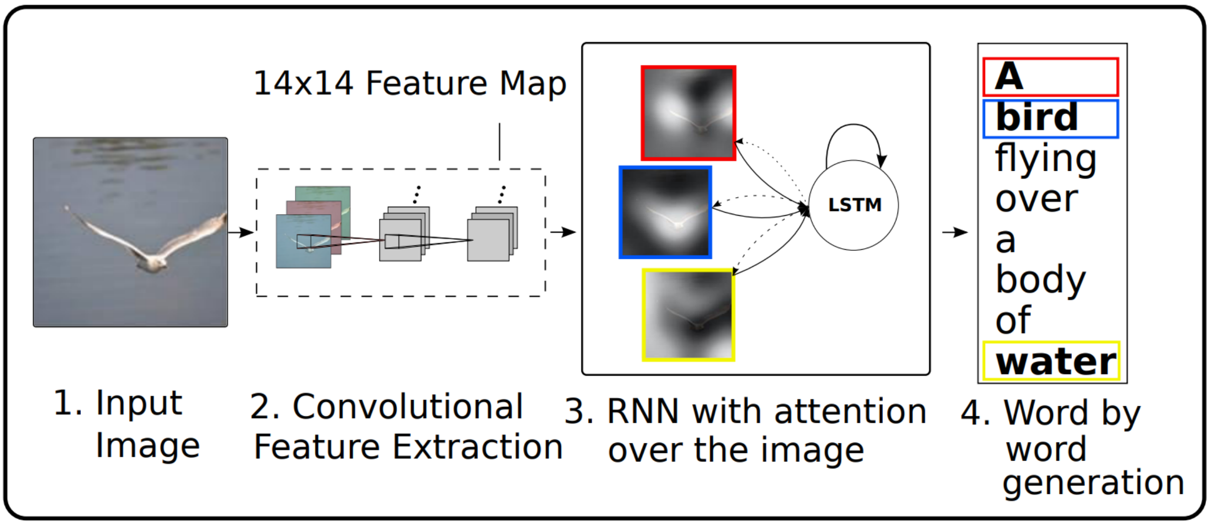

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figura 1.1: Diagram of natural language image description using attention (from [3]).

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 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 1.1) where the task consists of giving a natural language description of a given image. The work presented in [3] 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 [2] and neural computation [7].

1.1 Objectives

Attention might be fundamental for AI in general. 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 do not 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. 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).

2 Background

2.1 Attention

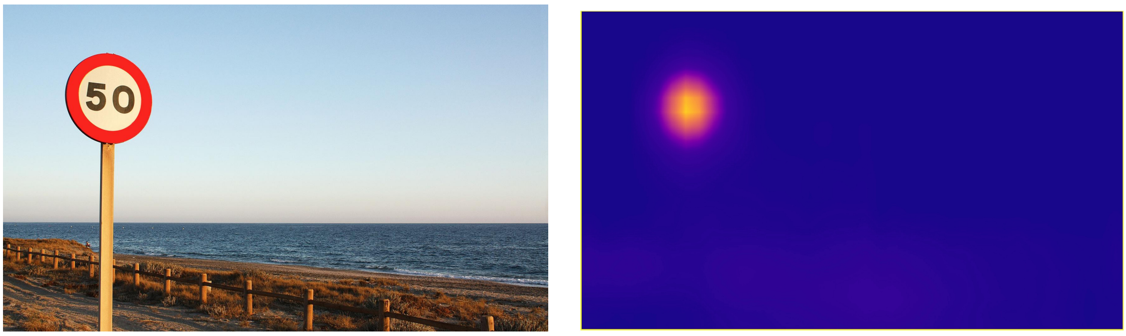
The interest in the concept of attention exists since a long time ago. Throughout the years, attention has been studied from various perspectives (c) such as philosophy, psychology, and neurology. Thus, there are multiple definitions of the concept. In broad terms, we can define attention as *the act of guiding the processing of information according to a critical evaluation method that is specific to a particular task at a given time*. In the next items, we discuss some concepts related to attention.

2.1.1 Functionalities of attention

Attention can be manifested in different manners depending on the goal (c). The most common is the act of selecting a set of stimuli over others (selective attention), such as looking at only a portion of an image. Another component is the act of guiding how one's focus moves over time (oriented attention), such as the act of looking at the words in a sequential manner when reading. Keeping the focus on a specific semantic element is important in some tasks and is known as sustained attention.

2.1.2 Bottom-up and Top-down attention

Focus may emerge in two fundamentally different manners (c). In bottom-up attention, the act of focusing is involuntarily started and guided by (usually) external and conspicuous stimuli, such as a shattering glass that tends to make us immediately turn our heads towards where the noise came from. Another example is visual saliency (figure 2.1): a glowing red ball suddenly appearing in your field of vision will probably grab your focus. In top-down attention, focus is voluntarily guided by cognition and goals. If we are talking to someone in a crowded party, for example, we focus on what the specific person is saying – ignoring other people's words – in order to maintain the conversation.



(a)

(b)

Figura 2.1: Exemple of visual saliency. b) is the saliency map where higher intensity pixels represent regions that are more salient to humans than original image a).

2.1.3 Soft and Hard attention

In recent years, there has been an useful distinction between soft and hard attention (c). Soft attention regards defining a continuous distribution of importance across all elements of information for some task. In the example of visual saliency, one can determine a saliency map M to a given image I where each pixel will have a value in $[0, 1]$ regarding its saliency. Hard attention regards determining a discrete subset of important information elements. Using again the problem of visual saliency as an example, one might want to determine a specific location (i, j) of the image to be used as center of a small patch of the image that is the most relevant to be further processed.

2.2 Deep Learning

Deep Learning (DL) is a trend in modern AI (c). Although DL started being broadly adopted around x years ago, some of its concepts date to much earlier than that (c): foundations of artificial neural networks were already discussed in the 1950s, backpropagation was introduced in the 1970s and many other key concepts that are popular mostly in the last decade or less were introduced more than 30 years ago. Many fields of AI witnessed a major shift in paradigm in the last years: models applying DL concepts now achieve state-of-the-art results in different problems regarding computer vision (c)(c)(c), audio processing (c), NLP (c), neural computation (c) among others. DL used used both in supervised and unsupervised learning (c).

One of the key concepts of DL is that of hierarchy of features (c): A deep sequence of layers apply non-linear transformations to the data in such a way that many models learn to extract features of hierarchical levels of abstraction. For this reason, DL is also regarded as Representation Learning. This characteristic enables such models to learn latent structure in intrinsically unstructured data such as images, text and audio signals. Another advantage is that of transfer learning: models that are primarily trained for a given task can be used and adapted for another task while using at least part of the representations learned. We discuss some concepts related to DL in following items.

2.2.1 Artificial Neural Networks (ANNs)

ANNs are usually adopted to prediction learning problems by means of learning a non-linear function aproximation. The ideas used in ANNs date to more than 50 years ago (c) and many of them are inspired from observed mechanisms of the human brain (x). Most of DL models are a variation of one of the families of ANNs that will be briefly discussed here.

One of the most basic examples is that of Multi Layer Perceprons (MLPs). The main caracteristic of this model is the use of hidden layers and neurons are a linear combination of previous layers followed by a non-linear activation. Each layer l_k (with n neurons) is connected to the previous layers l_{k-1} (with m neurons) and the neuron l_k^i, $1 \leq i \leq n$ is given value:

$$l_k^i = h\left(\sum_{j=1}^{m} l_{k-1}^j w_k^j + b_k^j\right)$$

Where $h(x) : R \mapsto R$ is a non-linear activation funcion. Commonly commonly used functions are the sigmoid hyperbolic tangent and the Rectified Linear Unit (ReLU):

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

ReLU is a currently broadly adopted due to its high efficiency and training speed (c).

Convolutional Neural Networks (CNNs) are widely used in computer vision tasks such as image classification, localization and semantic segmentation. CNNs use the fact that images tend to have correlated pixels and use convolution filters in an hierarchical manner (figure 2.2) to learn features in increasing abstraction. For a certain layer, the i-th feature

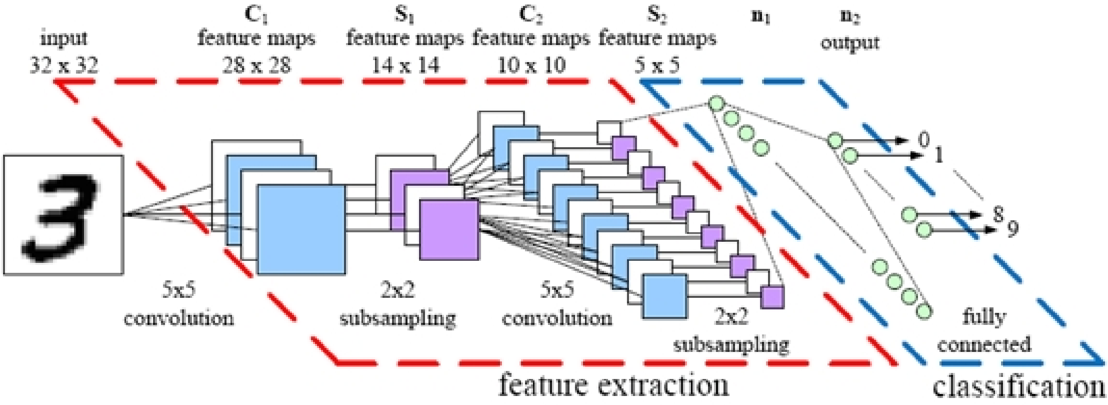


Figura 2.2: Diagram of a convolutional neural network. Learned filters extract features in an increasingly hierarchical manner.

map m_i is, given filter weights W_i, bias b_i and nonlinearity function $h(x)$, obtained as:

$$m_i = h\left(W_i * x + b_i\right)$$

with $*$ as the convolution operation.

Recurrent Neural Networks (RNNs) are characterized by a recursive architecture that uses the input of the current step and the output of the previous step to compute the predictions. The hidden state h_t at time step t, given input x_t, weight matrix W, previous state h_{t-1}, hidden-state-to-hidden-state matrix U and non-linearity $f(x)$ is given by:

$$h_t = f\left(Wx_t + Uh_{t-1}\right)$$

These architectures are widely used in NLP tasks (c) such as machine translation (c). Some variations over the original basic architecture such as LSTMs (c) are also broadly adopted.

2.2.2 Learning process

The act of learning the appropriate weights of a given model is usually obtained by the minimization of a differentiable loss function that is based on the cost function $L(y, \hat{y})$ that characterizes the error between the true value y and the predicted value $\hat{y}$. Backpropagation (c) plays an important role in DL because it's used to adjust the weights θ of models that have a differentiable cost function. A typical training process is composed of a forward-propagation step which computes the predictions over a set of input samples and a backpropagation step which computes the loss function and adjusts the weights of the model. In DL, acommon such adjustment methods include Stochastic Gradient Descent (SGD) which, for a given minibatch, adjusts weights according to:

$$\theta_{i+1} = \theta_i - \alpha \frac{\partial J}{\partial \theta}$$

where α is the learning rate.

3 Related Work

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

4 Methodology

TODO:

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

4.1 Schedule

TODO: the schedule.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[2] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell. *CoRR*, abs/1508.01211, 2015.

[3] KyungHyun Cho, Aaron C. Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *CoRR*, abs/1507.01053, 2015.

[4] E.L. Colombini, A. da Silva Simoes, and C.H. Costa Ribeiro. An attentional model for autonomous mobile robots. *IEEE Systems*, (99):1–12, 2016.

[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*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[7] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[10] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *CoRR*, abs/1406.6247, 2014.

[11] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.