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 techniques more generically applicable in broad problem classes such as Computer Vision, Natural Language Processing, Differential Programming and others.

Introduction

We continually receive 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 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 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 (in a ''train of thought"), 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 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" [4].

Attention can potentially play an essential role in Artificial Intelligence (AI). The pursue of intelligent machines is an old effort in Computer Science [15] 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 certain complex tasks nearly as efficiently as humans or some animals and the path to achieving more intelligence is still unclear, with many different proposals [12]. Part of the problem comes from the difficulty to properly define ''intelligence" itself, but surveys of the works on the subject [11] suggest that a reasonably accepted concept is the ability to perform elaborate tasks in complex and dynamic environments in order 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) [6]. 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 [16] and Natural Language Processing (NLP) [17], 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]. These are evidence that concepts of Attention have indeed been useful for the field.

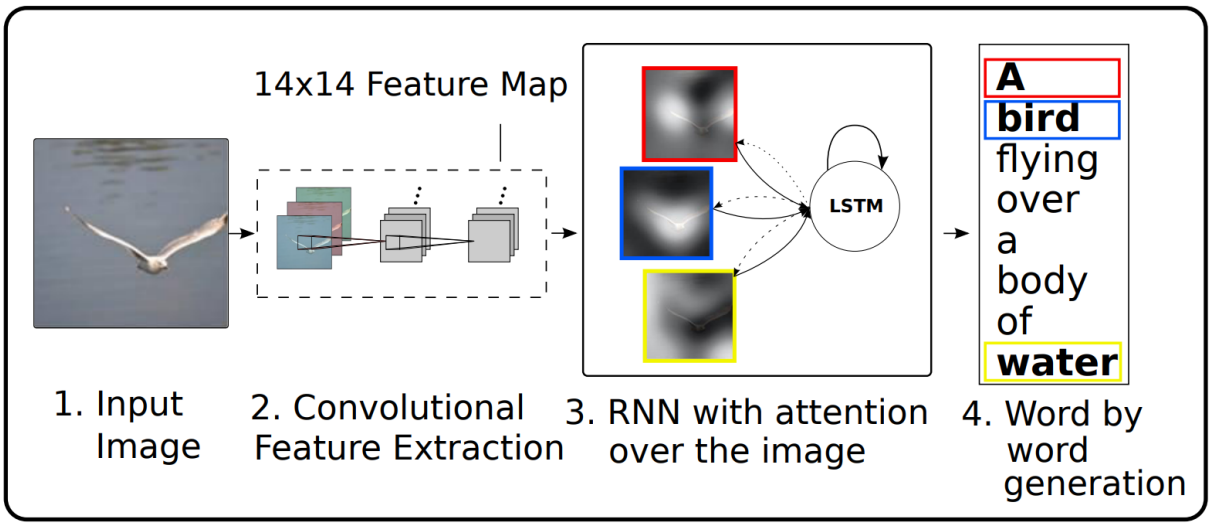


Figura 1.1: Diagram of natural language image description using Attention (from [3]).

Objectives

In spite of the recent adoption of Attention by a variety of Deep Learning models and the significant improvements it has shown, it is conjectured that there are still many other tasks that are still not very well 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 in the course of 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 techniques;
- To identify specific problems in different classes (robotics, vision, NLP, differential programming) with improvement potential by the use of Attention;
- To identify more theoretical aspects of Attention itself and their applicabilities to Deep Learning;
- To propose and implement one or more solutions based on the findings of the work in order to validate the ideas and evaluate them in an application.
- To study the viability of generalization of Attention to broader areas in AI other than Deep Learning.

Background

Attention

The interest in the concept of attention exists since a long time ago. Throughout the years, attention has been studied from various perspectives [4] such as philosophy, psychology, and neurology. There are multiple definitions of the concept. In the next items, we discuss some concrete aspects related to attention.

A definition

We can define attention as *the act of applying mental resources to selected stimuli following an allocation policy specific to a particular goal.* This rather broad definition captures well the main concepts related to attention: in a world with virtually infinite *stimuli* to select from the environment, agents with otherwise *finite processing resources* (but with a variety of options of *mental processes* to perform) must choose what their actions will be (and in which stimuli) in a *correct sequential manner* and in *sensible time*. As mentioned before, other works may define attention in a different manner that is perhaps even conflicting with ours but these are the terms that we choose our work to be based on – noting that they reasonably capture common concepts of interest by us and other works. [9]

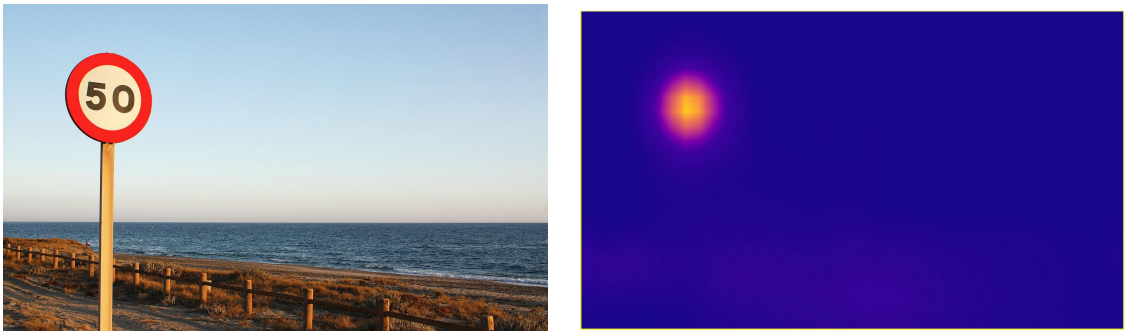
Functionalities of attention

Attention can be manifested in different manners depending on the goal. The most notable functionalities shown in intelligent beings are:

- **To select stimuli** such as looking at only a relevent portion of an image – to efficiently use resources on relevant information.
- **To sustain focus** on a specific semantic element for a period of time in order to complete a task.
- **To guide processing** in a sequential manner that is relevant for a task.
- **To orient resources** to new important stimuli – such as an abrupt noise comming from somewhere – or even in alternating the focus to multiple tasks at the same time.

Bottom-up and Top-down attention

Focus may emerge in two fundamentally different manners [4] [5]. 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



(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).

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.

Soft and Hard attention

In recent years, there has been an useful distinction between soft and hard attention [**?**]. 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.

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.

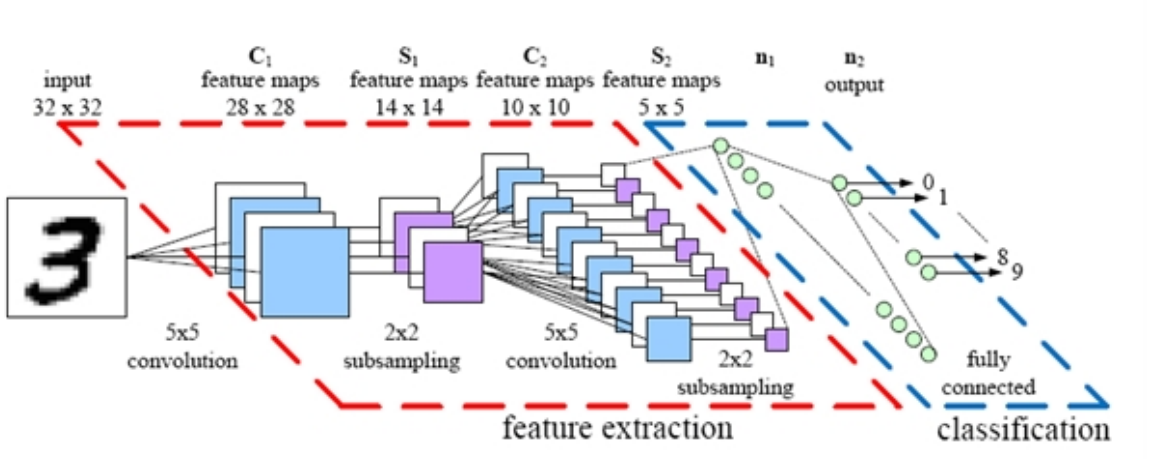


Figura 2.2: Diagram of a convolutional neural network. Learned filters extract features in an increasingly hierarchical manner.

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:

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

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 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.

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.

Related Work

The topic of integrating Attention concepts into Deep Learning has been increasingly frequent in the community (c). Augmenting the capabilities of neural network architectures with attention has shown promising results in problems from a variety of fields in which Deep Learning is currently being applied to, such as Computer Vision, Natural Language Processing and Differential Programming. In this section, we highlight some recent works.

Recurrent Attention Model (RAM)

The work [13] considers a commonly known problem in computer vision: it is usually expensive to perform processing on images and widely used current models such as CNNs tend to require computational resources proportional to the number of pixels in the image. The work proposes a *Recurrent Attention Model (RAM)*, a recurrent neural network augmented by an attentional component regarded as *Glimpse Module* that is trained via Reinforcement Learning. The Glimpse Module enables the network to select a point in the image from which it extracts ''glimpses" – patches of the image at different resolutions but with the same dimensions. These glimpes and the selected location are encoded and given as input to produce the new hidden state of the core RNN architecture. The dimensionality of the glimpes is much smaller than that of the image and furthermore does not depend on the dimensions of the input image. The authors evaluate the model for classification tasks in the MNIST (c) dataset and variations in which the input images are filled more background pixels (resulting in a larger image) and clutter. The proposed

model outperforms a convolutional neural network baseline. Furthermore, the attentional module in the model enables it to perform the same amount of computation regardless of the input size of the image and to focus sequentially only at the relevant parts of the image, which reduces the adversary effect of clutter.

Attention-based Encoder-Decoder Networks

Encoder-decoder networks are a general framework used generally for mapping from input to outputs that both are of highly-dimensional (often unstructured) data (c), having being successfully used for tasks such as machine translation (c). One drawback of such architecture is that the encoded feature vector is of fixed size and structure – regardless of the input – and not necessarily preserves spatial/temporal structure from the input. The work in [**?**] proposes the usage of an attentional module in between encoder and decoder. The proposed model's encoder produces feature vectors that have a explicit spatio-temporal structure (*context set*) of the input and the attentional module uses a relevance evaluation method to select a subset of the outputs – either by soft attention or hard attention. This allows the encoder-decoder for more flexibility to select the components of the input that are of more relevance. The authors implemented and evaluated the method for several applications:

- *Image Caption Generation*: The goal of the task is to provide a natural language description of an input image. The proposed model uses a CNN as encoder and RNN as decoder – with the attentional model in between. The model was ranked third in *MS COCO Captioning Challenge* and provided highly interpretable results regarding the importance of the regions of the image to each component of the sentence (see figure 1.1).

- *Neural Machine Translation*: The authors proposed a RNN architecture augmented with the attention module, which provided relative improvement of roughly 60% when compared to the same model without attention. The model also performs better than state of the art in some languages. It was also possible to obtain a weight matrix that maps the importance of input to output words since the context set provides structural information of the input.

- *Neural Speech Recognition*: The goal of the task is to translate audio to text sentences by using fully neural networks. The proposed model uses RNNs between the attention module and the model achieved state-of-the art results in the TIMIT corpus (c) and the outputs provide attention weights from the input signal to produced phonemes.

Overall, the proposed technique – besides achieving state of the art results – produces a semantic mapping from the input space to the output space even when they are of different nature – without explicitly being supervised to produce this mapping.

Neural Turing Machines

Adaptive computation time for RNNs

The Transformer architecture

Neural Programmer

Methodology

TODO:

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

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] Helgi Helgason. General attention mechanism for artificial intelligence systems. 05 2013.

[10] 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.

[11] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *CoRR*, abs/0712.3329, 2007.

[12] Tomas Mikolov, Armand Joulin, and Marco Baroni. A roadmap towards machine intelligence. *CoRR*, abs/1511.08130, 2015.

[13] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention.

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

[15] Alan M. Turing. Computing machinery and intelligence. *Mind*, 1950.

[16] 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.

[17] 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.