Task 2

André Bastos, nº 56969
Rafaela Cruz, nº 56926

In this report, we will compare four different attention mechanism approaches proposed in papers [1] to [4] to solve different kinds of problems. We will start by briefly explaining the attention mechanism used in each paper and then we will present the advantages and disadvantages of each approach against the others.

Starting by [4], the authors created an attention-based task-driven recurrent network for visual processing, namely visual object classification. This model, called the Recurrent Attention Model (RAM), is capable of extracting information from an image or video by processing only selected regions of that image or video.

RAM has a built-in goal-directed agent, that interacts with a visual environment, but never senses this environment in full. Instead, at each point in time, the agent looks only into a small region via a bandwidth-limited sensor. In a more detailed way, at each step t, the agent receives a partial observation of the environment in the form of an image x_t and extracts information from x_t via its bandwidth-limited sensor, by focusing the sensor on some location l_{t-1} of interest and extracting a square patch centered around that location. The sensor encodes only the region around l_{t-1} at a high-resolution, resulting in a low-resolution representation called glimpse.

The glimpse is used to produce a glimpse feature vector g_t, which is combined with the internal representation of the network at previous time step h_{t-1} to update the internal state to h_t. This internal state summarizes the information extracted from the history of past observations and is used by the agent to decide the next location to focus on, l_t, and how to act (i.e., what action, a_t, it must perform based on the task at hand).

Thus, instead of processing an entire image at once, the attention mechanism selects a small location to attend to at each time step based on past information and the demands of the task it must perform, needing only a few time steps (i.e., a few glimpses) to perform the task at hand.

The authors of [3] also used an attention-based model, this time for speech recognition problems. In this work, the authors developed a recurrent neural network (RNN) capable of generating a sequence of phonemes given a sequence of speech frames.

In more detail, the authors used an attention-based sequence generator (ARSG) to generate an output sequence from an input. First, the input, x, which is a sequence of feature vectors (each vector being extracted from a small overlapping window of audio frames), is processed by an encoder, obtaining a list of representations h. Then, an attention mechanism included in the network uses h in order to obtain the output sequence, y (a sequence of phonemes).

At each time step i, an ARSG creates a vector of attention weights (the alignment, α_i) using the state of the network at the previous time step $i-1$ (s_{i-1}), the alignment at time step $i-1$ (α_{i-1}) and h. The alignment is then used to weigh all the elements in h, creating a vector g_i (the glimpse), that will thus contain information about which of the elements in h are relevant. The output y_i is then generated taking into account s_{i-1} and the relevant elements of h. This is, thus, a hybrid attention mechanism: it takes into account both the location of the focus from the previous time step, as it uses the previous alignment α_{i-1} to select a list of possible elements from h, and the content (the features of the input sequence), as it uses this list to select the relevant h for the generation of the output.

Note that, to improve their results, the authors also proposed some modifications to this attention mechanism, including sharpening, to prevent g_i from containing noisy information when the input sequence h is long, and smoothing, in order to avoid concentrating the attention on a single frame.

The authors of [2] presented a neural network model with a spatial attention mechanism to extract information from images. The model developed by the authors processes images of street views through a convolutional neural network (CNN), attentionally weighs the features obtained, and passes them into a RNN, in order to extract street names from these images.

In order to extract only the street names from the street views, the model has to focus on the important parts of the scene and ignore visual clutter. To achieve that, the input to the RNN is determined by a spatially weighted combination of image features (the context), obtained by multiplying a spatial attention mask, α_t, by the feature map f derived by passing the image x through a CNN.

In order to compute α_t, the authors combined the content from the image, via f, with the hidden state of the RNN at time t, s_t. Moreover, the authors also made the model location aware, like in [3]. However, while the approach used in [3] uses α_{t-1} for location awareness, [2] concatenates each element of the feature map f with an one-hot encoding of the spatial coordinates of that element. This allows to make a big jump to the left side of the line below, which is useful in multiline text recognition problems (which is the goal of this paper).

Finally, the authors of [1] developed a network model capable of predicting the bounding box of the license plate (LP) in car images and recognizing its corresponding number. The presented network, which the authors named RPnet, is divided in two modules: the detection module, a deep CNN that extracts feature maps from the input image and feeds them to three fully-connected layers that predict the bounding box of the LP; and the recognition module, that exploits regions-of-interest (ROI) in the bounding box predicted by the previous module and predicts de LP number.

The detection module works, thus, as an attention mechanism, telling the recognition module where to look to predict the LP number. Given an input image, the detection module extracts feature maps and predicts the location of the bounding box of the LP in each feature map. Then, the recognition module looks at the position of the bounding box in each feature map, extracts ROIs from these feature maps, combines them and feeds them to classifiers that predict the LP number in the bounding box. Thus, in a single forward computation, RPnet predicts the LP bounding box and the corresponding LP number at the same time.

Having understood the mechanisms presented in each paper, we can now compare them, assuming that we are interested on their application to image recognition. The mechanism proposed in [1] seems to be the approach that differs the most from the remaining ones. On one hand, it seems to be simpler, as it uses a CNN to predict a bounding box around the relevant information. Moreover, the authors claim that, by sharing feature maps between the detection module and the recognition module, the model achieves a fast recognition rate.

However, to detect the bounding box, the detection module needs to apply convolving filter maps to the entire input image, so it could become computationally expensive if the number of pixels in the input image is too high. Moreover, the fact that the attention is not based on the information collected in previous time steps, like in the other mechanisms, and is only used to create a bounding box, may not be very useful when we want to recognize long sequences like in papers [2] and [3].

The approach proposed in [4] seems to be able to circumvent the disadvantages of [1]: by using attention to focus the computational resources on small parts of a scene, fewer pixels need to be processed and the task complexity is reduced, as the object of interest can be placed in the center of the fixation and irrelevant features of the image (clutter) outside the fixated region are ignored. Thus, the amount of computation performed can be controlled independently of the size of the input images. Moreover, the model used processes the inputs sequentially, attending to different locations within the images one at a time, and incrementally combines information from these fixations to build up a dynamic internal representation of the scene. Thus, instead of processing an entire image or even bounding box at once, at each step, the model selects the next location to attend to based on past information.

The approach proposed by [3] has also many advantages. The fact that the used attention mechanism is hybrid allows the model to take into account not only the features from the previous alignments, but also the relevant representations h of the input. The authors of this paper also proposed some modifications to their attention mechanism: sharpening, to prevent the model from using noisy information when the input sequence is long; and smoothing, in order to avoid concentrating the attention on a single frame. However, this approach is not useful if we want to recognize multiline texts in images, because it does not allow to make "jumps" to the left side of the line below.

The authors of [2] also used a hybrid attention mechanism, with all the advantages presented for [3]. However, in order to make the model location-aware, in [2], the authors concatenated each element of the feature map extracted from the images with an one-hot encoding of the spatial coordinates of that element. This approach, which is different from the one used in [3], allows for the model to recognize multiline texts. This spatial attention mechanism allows, thus, to extract structured text information by reading only the interesting parts of the whole image, even if that text has multiple lines.

Note, however, that these two last approaches need to process all the input in order to apply the attention mechanism. As stated before, this is not necessary in [4], because the model proposed in this paper processes only the relevant locations in the input, which can bring advantages if the input in too large.

References

[1] Zhenbo Xu, Wei Yang, Ajin Meng Nanxue Lu, Huan Huang, Changchun Ying and Liusheng Huang (2018) *Towards End-to-End License Plate Detection and Recognition: A Large Dataset and Baseline*. Proceedings of the European Conference on Computer Vision (ECCV).

[2] Zbigniew Wojna, Alex Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li and Julian Ibarz (2017) *Attention-based Extraction of Structured Information from Street View Imagery*. ICDAR.

[3] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho and Yoshua Bengio (2015) *Attention-Based Models for Speech Recognition*. NIPS Proceedings.

[4] Volodymyr Mnih, Nicolas Heess, Alex Graves and Koray Kavukcuoglu (2014) *Recurrent Models of Visual Attention*. NIPS Proceedings.