Application of Reinforcement Learning in Computer Vision: A Review of Visual Attention Model

Ziyao Li [1]

Abstract

Attention Model is an ingenious invention in the areas of computer vision. It is inspired by conclusions about human attentions when processing images drawn in experimental cognition studies. The model is first raised dated back to as early as 1995, while at then it is more of an image processing techniques of selecting salient regions of an image through empirical methods. The model is incorporated with reinforcement learning and regarded as a Markovian Decision Process in a 2005 article. After development of various research work, the model reaches a high achievement of both performance and extensive abilities in an architecture of recursive neural network. Different applications are implemented from this architecture. Recently, an inclination of abusing the concept of attention arises in some areas of machine learning, such as natural language processing: an alignment techniques, firstly adopted in machine translation tasks, is misusing the name of attention model.

1. Introduction

As the heat of deep convolution neural networks' applications in computer vision areas comes a little bit down, the deficiency of convolution networks starts to emerge. The first one is the relatively high computational complexity, and the second one is its disadvantages in leveraging local features, since the parameters in the convolution kernels are trained globally. Instead of simply adjusting convolution models, an alternative starts to gain more attention and popularity: Attention Models.

Attention Model simulates the process of human decision process. It takes into consideration only a small part of the whole image, which is vividly described as a *glimpse*, and then uses optimized policies choosing next glimpse to look at. The process continues going until a maximum decision stamp is reached, or the model itself decides to quit. Final outputs are given in the end of this process. Rewards in Attention Models are defined in various ways, while the simplest one is to choose the loss of the original problem, typically an object recognition problem.

It is interesting that Attention Models are everything but newly invented. The basic ideas of adapting attentional factors in computer vision problems were come up with as early as 1998(Itti et al., 1998). It only starts to gain its recent popularity in a rapid pace after its realization with recurrent neural networks is systematically described and experimented in the 2014 Google DeepMind paper *Recurrent Models of Visual Attention* on *NIPS '14*(Mnih et al., 2014). The paper proposed a finely-designed recurrent neural network framework, which quickly became a basic baseline for coming research work. Besides computer vision, this idea of attention is extensively introduced in

[1]SID: 1500017776; Data Science and Big Data Technology Track, Yuanpei College, Peking University..

This paper is a mid-term assignment of course Reinforcement Learning, 2018 *of Prof. Zhihua Zhang, Peking University.*

various areas, such as natural language processing, especially the problems of text corresponding with images, such as visual question answering and caption generation.

2. The Origin of Visual Attention Models: an Image Processing Technique

As mentioned before, the origin of attention models in computer vision areas can be dated back to (Itti et al., 1998). In this short article, inspired by conclusions of human visual attentions drawn from experimental neural science and some naive previous explorations of visual attentions in artificial intelligence areas back in 1995(Tsotsos et al., 1995), a very delicate hierarchical, or pipeline system combining various empirical methods are proposed to simulate visual attentions. Very elegant features are selected and explored based on expert knowledge of images, which is a typical approach of machine learning at that time (in fact, these are hardly learning techniques but processing techniques, kind of similar to the Canny Edge Detector).

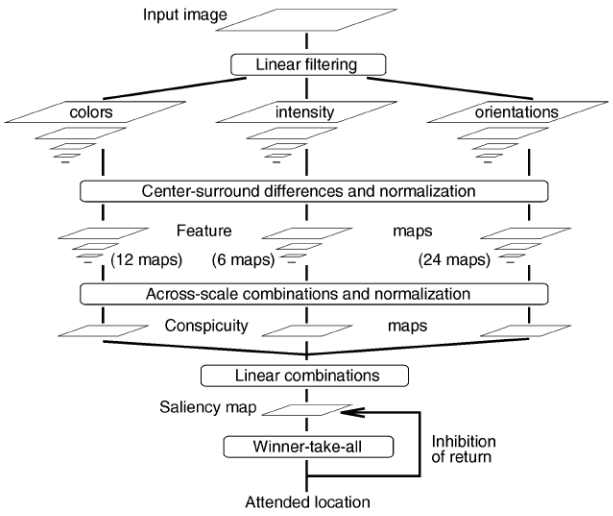

Figure 1. The 1998 model of visual attentions. The model is more similar to processing techniques than learning techniques. Areas of greater importance are marked in a salient map.

Figure 1 shows the basic structure of the proposed attention model. The model used scalars to weight the importance, or "conspicuity", of different locations on the image. This may seem as naive approaches, however, several important ideas are firstly introduced in this article, such as the concepts of Feature Map, Conspicuity Map and Salient Map. Successive work are significantly impacted by this work in its ideas. This 1998 work currently has achieved a miraculous number of citations-over 9000, and its ideas are still influencing the development of computer vision.

3. The Development: Combining Reinforcement Learning

The 1998 model, though inspiring, cannot perform any specific machine learning tasks, because its only purpose is to mark out the *Focus of Interests (FOI)*. In 2005, a model combining the idea of visual attention and object recognition task is proposed in (Paletta et al., 2005). This model firstly adopted a reinforcement learning perspective in computer vision tasks. Firstly, it adopts a local measurement of saliency to identify several (can be many) Foci of Interests, with these foci considered most likely to contain important and useful information guiding the specific learning task, i.e. the object recognition tasks. By adopting a Shannon conditional entropy of predicting the correct object,

$$H(O|f) = -\sum_k P(o_k|f)\log P(o_k|f),$$

where $P(o_k|f)$ are posteriors, and $o_k \in O, k = 1, 2, ..., |O|$, the significance a local feature is to the final object recognition problem is estimated.

After a salient map is conducted, the model considers the process of shifting FOIs to get a global view of information. The actions here to determine is which FOI to go to. By determining posteriors of object hypothesis from a histogram method (the details are skipped), a greedy method is to maximize the short-sight utility, i.e.

$$a^* = \arg\max_a \Delta H_{t+1}(s_{i,t}, a_{k,t+1}),$$

where ΔH_{t+1} is the information gained in timestamp $t+1$, i.e.

$$\Delta H_{t+1} = H_t - H_{t+1}$$

and $s_{i,t}$ is a historical state. However, it is a global optimum the agent shall achieve, that is, maximizing the expected discount reward

$$Q(s,a) = E\left[\sum_{n=0}^{\infty} \gamma^n R_{t+n}(s_{t+n}, a_{t+n})\right].$$

Therefore, instead of this greedy method, the model proposed a better and more robust Q-Learning process of finding optimal policies. The information gain ΔH_{t+1} described above is defined as the step-wise reward in this MDP. The model announced adopting the Q-Learning algorithm to estimate the transitional probabilities without providing more details.

This 2005 paper is of the pioneers aiming at specific machine learning problems using reinforcement attention mechanisms. After this paper, various research work contributes to this area. Bukto's work (Butko & Movellan, 2008) (Butko & Movellan, 2009).

The former work enumerates several disadvantages in saliency maps: atemporality, lack of integrating mechanisms, and foveation of too many locations. A Information-gathering Partially Observed Markovian Decision Process (I-POMDP) is established as an alternative in this work solving fovea problems. This POMDP model separate the state information into a hidden state s and an observation o, instead of considering them to be one variable. Using basic conclusions of POMDP, the author gives an explicit expression of the Belief States in this process, and eventually forms a $\theta-$parameterized logistic policy, or softmax function, a mapping from the calculated Belief States space to a score of action space to identify the most likely point to consist an object. While the logistic model can have explosively large scale of parameters, the work takes into account the *Shift and Rotation Invariance* and designs a weight sharing mechanism greatly reducing the model's degree of freedom, accelerating the training process 20 times than a naive policy gradient iteration.

The latter work is more of an implementation of the former I-POMDP algorithm in the precise fast object recognition problem. In this work, a Multinomial I-POMDP model is applied to construct an *off-the-shelf object detector*, such as a face detector. The location of the search target (in the grid cells, to make the problem discrete) is considered to be the state of the model; the current center of fixation is the action; the observations are defined as the total number of objects returned by the object detector in each grid cell after searching all pyramids of Image Patches (IPs); the transitional probabilities are defined the same as in the I-POMDP model. In experiments, this proposed method shows great advantages over running time than the baseline over face datasets, at a slight cost of detecting precision.

(Larochelle & Hinton, 2010) proposed a Third-Order Boltzmann Machine to extract a global view from several *glimpse* over the whole image. A glimpse have high resolution around the center (*the fovea*) and lower resolution as the eccentricity increases (*the periphery*). Combination of glimpses is the main issue in this work, with image classification problems as instances. An energy function is defined to combine the glimpses together,

$$\begin{aligned} E(y, x_{1:K}, h) &= \sum_{k=1}^{K}(-h^T W^{(i_k,j_k)} x_k - b^T x_k) \\ &\quad - c^T h - d^T y - h^T U y. \end{aligned}$$

The novelty of this model focuses on $W^{(i_k,j_k)}$. Unlike general fixed parameter matrices, this W now depends on the position of the fixation (i_k, j_k). Such connections are called third-order because they connect hidden units, input units and implicit position units (one for each possible value of positions (i_k, j_k)). It is impossible to train independent parameters for all possible positions, so $W^{(i_k,j_k)}$

is defined as

$$W^{(i_k,j_k)} = Pdiag(z(i_k, j_k))F$$

where $P \in R^{H\times D}$, $F \in R^{D\times R}$ and z is a D-dimensional vector associated with coordination (i_k, j_k). As D can be arbitrarily selected, the scale of parameters can be controlled via it. Two mechanisms are also designed to achieve two goals: integrating glimpses for global, especially shape information, and learning where to look.

$$\mathcal{C}_{hybrid} = -\log p(y^t|x^t_{1:K}) - \alpha \log p(y^t, x^t_{1:K}),$$

$$p_{controller}((i_k, j_k)|x^t_{1:k-1}) \propto exp(f(s_k, (i_k, j_k))),$$

where f is a scoring function that should maximize the probability of correct classification. An algorithm of combining two losses is proposed.

Since then, neural methods are gaining more and more popularity, and conducting attention models into neural network algorithms is so trivial that countless publications are issued on the subject, including (Denil et al., 2012) and (Ranzato, 2014). These methods and algorithms are usually very problem-specific while alike in between, with somehow lack of academic inspirations, so brief discussions are skipped here.

4. The Pearl of Neural Attention Model: *Recurrent Models of Visual Attention*

While countless trivial neural methods in Attention Models are published everywhere, *Recurrent Models of Visual Attention* (Mnih et al., 2014) remains one of the most remarkable one. The model proposed a recurrent neural network framework in conducting the idea of image attentions.

Although it is almost instinctive for researchers to associate the sequential decision process with recurrent neural networks, i.e. the *Deep Reinforcement Learning Framework*, and the idea behind this is rather simple, achieving perfect performance through these kinds of neural networks, however, is not as easy as it may sound. (Mnih et al., 2014) is successful not because of its ideas, although they may seem astonishingly fancy for amateurs in computer vision. It is its finely-grained network structures and optimization techniques.

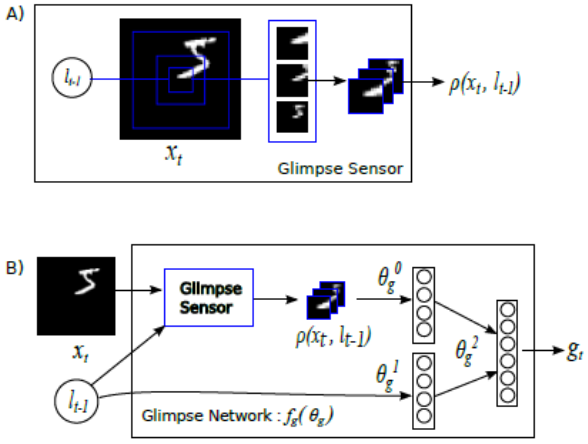

Figure 2. Recurrent neural networks designed in (Mnih et al., 2014) (1). **A) Glimpse Sensor:** Given the coordinates of the glimpse and an input image, the sensor extracts a retina-like representation $\rho(x_t, l_{t-1})$ centered at l_{t-1} that contains multiple resolution patches. **B) Glimpse Network:** Given the location (l_{t-1}) and input image (x_t), uses the glimpse sensor to extract retina representation $\rho(x_t, l_{t-1})$. The retina representation and glimpse location is then mapped into a hidden space using independent linear layers parameterized by θ_g^0 and θ_g^1 respectively using rectified units followed by another linear layer θ_g^2 to combine the information from both components. The glimpse network $f_g(\cdot; \{\theta_g^0, \theta_g^1, \theta_g^2\})$ defines a trainable bandwidth limited sensor for the attention network producing the glimpse representation g_t.

Fig 2 and Fig 3 shows the basic network structures of this recurrent network. A **Glimpse Sensor**, a **Glimpse Network** and a **Recurrent Network Architecture** are proposed in this work to achieve an end-to-end model of various image tasks (corresponding with different reward designs). In the Glimpse Sensor, different hierarchies of range/resolution windows are extracted to simulate a retina sensor. In the Glimpse Network, both image data and location information are combined to extract the feature of the glimpse. In the final architecture, a simulation of MDP process in recurrent neural networks is implemented: functions as state transition probabilities, action

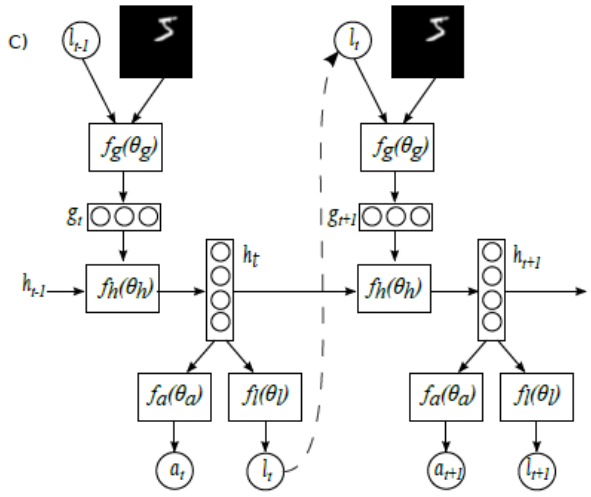

Figure 3. Recurrent neural networks designed in (Mnih et al., 2014) (2). **C) Model Architecture:** Overall, the model is an RNN. The core network of the model $f_h(\cdot;\theta_h)$ takes the glimpse representation g_t as input and combining with the internal representation at previous time step h_{t-1}, produces the new internal state of the model h_t. The location network $f_l(\cdot;\theta_l)$ and the action network $fa(\cdot;\theta_a)$ use the internal state h_t of the model to produce the next location to attend to lt and the action/classification at respectively. This basic RNN iteration is repeated for a variable number of steps.

selecting policies and location movement (or environmental effects) are all simulated with neural layers.

What's more, this work provides training strategies to improve complexity and converging speed. First of all, a *Reinforce Rule* is used to sample for the expectation in the loss function, as is usually the case in traditional reinforcement learning; secondly, a variance reduction technique is adopted to lower the sampling variance at the cost of bias by separating the reward into action-invariant reward and action-dependent reward; thirdly, a hybrid supervise loss is recommend in this model. The model significantly outperforms state-of-the-art results, especially in cluttered image corpora. This is due to the characteristics of automatic feature selection brought by Attention Models.

5. The Further Extension: the *Abuse* of Attention?

After the work of Google DeepMind (Mnih et al., 2014), continuous contributions over Neural Attention Models still keeps merging. (Xu et al., 2015) implements the recurrent visual attention model in image caption generation tasks. (Bahdanau et al., 2016) proposed a novel design of supervision over the whole reinforcement training process by first evaluating an abstract "value" over each decision token. (Ren & Zemel, 2017) implements the idea of visual attentions in the image segmentation tasks with a novel and rather complex network system.

However, it is the idea "Attention" itself that is more interesting. It is now no longer a limited and reinforcement-learning-related terminology in the field of computer vision. Started with (Bahdanau et al., 2015), "Attention Model" becomes a most popular topic in Natural Language Processing. Although instead of attention, (Bahdanau et al., 2015) is actually talking more about alignments of words from origin language to target language in machine translation tasks, and even no intention of so-called "attention" is ever mentioned in this article, the idea is soon interpreted and used as "attention mechanism". This kind of attention soon gained its popularity, in spite of that it hardly has anything to do with real attention models. (**?**) proposed an "Attention Mechanism" in Visual Question Answering, although it is actually an alignment of the global image features instead of real "attentions" discussed in this article. What's more, (Zhang et al., 2018) and (Nie et al., 2018) both claim that attention model is leveraged in their models. However, such usage is nothing but alignment: the former uses this simple alignment mechanism aligning words to convolution outputs from a global VGGNet to conduct Name Entity Recognition tasks; the latter uses this mechanism to align words in mention context and words in entity context to conduct Entity Linking tasks. These "abuses" of the terminology could be rather confusing and misleading. It is

probably caused by some misinterpretation of the idea "attention", or some researchers intentionally did so to attract more attention and to make their model sounds more fancy. This leads to a result that almost EVERY newly published article about neural network applications acclaims that this kind of "attentions" are adopted, because it is just so simple to be added to all kinds of models.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Processings of the International Conference on Learning Representations, ICLR '15*, 2015.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. An actor-critic algorithm for structured prediction. In *Processings of the International Conference on Learning Representations, ICLR '16*, 2016.

Butko, N. J. and Movellan, J. R. I-pomdp: An infomax model of eye movement. In *Proceedings of the 7th IEEE International Conference on Development and Learning, ICDL '08*, pp. 139–144, 2008.

Butko, N. J. and Movellan, J. R. Optimal scanning for faster object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '09*, 2009.

Denil, M., Bazzani, L., Larochelle, H., and Freitas, N. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 24(8):2151–2184, 2012.

Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

Larochelle, H. and Hinton, G. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Processings of the Annual Conference on Neural Information Processing Systems, NIPS '10*, 2010.

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. Recurrent models of visual attention. In *Processings of the Annual Conference on Neural Information Processing Systems, NIPS '14*, 2014.

Nie, F., Cao, Y., Wang, J., Lin, C., and Pan, R. Mention and entity description co-attention for entity disambiguation. In *Proceedings of the Association for the Advancement of Artificial Intelligence, AAAI '18*, 2018.

Paletta, L., Fritz, G., and Seifert, C. Q-learning of sequential attention for visual object recognition from informative local descriptors. In *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, 2005.

Ranzato, M. A. On learning where to look. *Computer Science*, 2014.

Ren, M. and Zemel, R. S. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '17*, 2017.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., and Nuflo, F. Modelling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.

Xu, K., Ba, J. L., Kiros, R., K. Cho, A. Courville, R. Salakhutdinov, Zemel, R. S., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML '15*, 2015.

Zhang, Q., Fu, J., Liu, X., and Huang, X. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Association for the Advancement of Artificial Intelligence, AAAI '18*, 2018.