MedCrave
Step into the World of Research

International Robotics & Automation Journal

Mini Review

Open Access

CrossMark

Visual attention in deep learning: a review

Abstract

Visual Attention is an essential part of human visual perception. This review describes two main classes of attention models. Benefiting from the rapid growth of deep learning, CNN-based or other deep learning-based models are capable of establish new state-of-the-art in this challenging research problem.

Keywords: deep learning, visual attention, CNN, saliency

Volume 4 Issue 3 - 2018

Xinyi Liu,[1] Mariofanna Milanova[2]

[1]System Engineering Department, University of Arkansas at Little Rock, USA

[2]Computer Science Department, University of Arkansas at Little Rock, USA

Correspondence: Mariofanna Milanova, Computer Science Department, University of Arkansas at Little Rock, USA, Tel +1 501-551-0662, Email mgmilanova@ualr.edu

Received: April 06, 2018 | **Published:** May 01, 2018

Introduction

When facing a complex visual scene, human can efficiently locate region of interest and analyze the scene by selectively processing subsets of visual input. Attention was employed to narrow down the search and speed up the process.[1] Visual attention is a hot topic in computer vision, neuroscience and deep learning area. It's widely used in object segmentation, object recognition, image caption generation[2,3] and visual question answering (VQA) etc.[4] In last few years, deep learning has been growing rapidly. Many Convolutional neural networks and recurrent neural networks have achieved much better performance in various computer vision and natural language processing tasks, compared to previous traditional methods. The visual attention models are mainly categorized into Bottom-up models and top-down models.

Bottom-up attention models

The Models are based on the image feature of the visual scene. The goal of bottom-up model is to find the fixation points, where it stands out from its surrounding and grab our attention at first glance. The main idea of bottom up visual attention models is that the attention is unconsciously driven by low level stimulus. Most of traditional bottom-up attention models use hand-designed low-level image features (e.g., color, intensity) to produce saliency map and image representation. As a classic Salient Region Detection method, histogram-based contrast (HC) and region-based contrast (RC) algorithm[5] generate saliency map by evaluating global contrast differences and spatial weighted coherence scores. CNN-based models have achieved state-of-the-art result by producing more accurate feature map. The bottom-up attention model proposed in[6] was implemented with Faster R-CNN, while spatial regions are represented as bounding boxes; provide significant improvement on VQA tasks.

Top-down attention models

The Models are driven by the observer's prior knowledge and current goal. The recurrent attention model(RAM) proposed in[7] mimicked human attention and eye movement mechanism, supposed there is a 'bandwidth' limit for each glimpse, to predict future eye movements and location to see at next time step. This method is computational effective compared to classical sliding window paradigm, which come at a high computational cost to apply convolving filter maps with entire image. The RAM model is an recurrent neural network (RNN)[8] consists of glimpse network,[7] core network, action network and location network. The glimpse representation is input for core network, which combining with the internal representation at previous time step, produces the new internal state. The location network and the action network produce the next location to look at and the action/classification using the internal state of the model. Although RAM has shown its performance on digit classification tasks, the ability of dealing with multiple objects is limited. Deep recurrent visual attention model (DRAM)[9] was proposed to expand it for multiple object recognition. It is a deep recurrent neural network to decide where to focus its computation and trained with reinforcement learning. DRAM is composed of glimpse network, recurrent network, emission network, context network and classification network. When DRAM explores the image in a sequential manner with attention mechanism, it learns to predict one object at a time. A label sequence will be generated for multiple objects. An attention-based model to generate neural image caption was proposed.[3] The form of attention described has two variants: stochastic "hard" attention and deterministic "soft" attention. The idea of this attention-based model is to focus on different attention regions of the visual input as it refined predictions and move focus in time. A salient object detection method proposed[10] and introduced short connections to the skip-layer structures within the Holistically-Nested Edge Detector (HED) architecture.[11] A Unified Framework[12] was proposed to extract richer feature representations for pixel-wise binary regression problems (e.g., salient object segmentation and edge detection). They construct a horizontal cascade, which connects a sequence of stages together to transmit the multi-level features horizontally. Enriched deep recurrent visual attention model (EDRAM)[13] combined special transformer and recurrent architecture, was used for multiple object recognition. It can perform object localization and recognition at same time. With Spatial transformer[14] used as attention mechanism, the architecture became fully differentiable, achieved superior performance on MNIST Cluttered dataset[15] and SVHN dataset.[16–18]

Conclusion

In this brief review, several widely used and state-of-the-art visual attention models were introduced. With visual attention model being introduced into image classification, object recognition and VQA tasks, the computational cost was relieved. Benefiting from the rapid growth of deep learning, CNN-based or other deep learning-based models are capable of establish new state-of-the-art result on datasets like MNIST and SVHN, it also can deal with more complicated cases compared to classical methods. We introduced them in a manner of categorizing as Top-down and Bottom-up mechanism.

Acknowledgements

The research was sponsored by University of Arkansas at Little Rock.

Conflict of interest

The author declares there is no conflict of interest.

References

1. Milanova M, Mendi E. (2012) *Attention in Image Sequences: Biology, Computational Models, and Applications.* In: Kountchev R, Nakamatsu K, editors. *Advances in Reasoning-Based Image Processing Intelligent Systems*. Intelligent Systems Reference Library. Springer, Berlin, Heidelberg; 2012:147–170.
2. Chen L, Zhang H, Xiao J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *Computer Vision and Pattern Recognition*. IEEE Conference; 2017:6298–6306.
3. Xu K, Ba J, Kiros R, et al. *Show, attend and tell: Neural image caption generation with visual attention.* International Conference on Machine Learning; 2015:2048–2057.
4. Singh J, Ying V, Nutkiewicz A. *Attention on Attention: Architectures for Visual Question Answering (VQA)*. 2018.
5. Cheng MM, Mitra NJ, Huang X, et al. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015;37(3):569–582.
6. Anderson P, He X, Buehler C, et al. *Bottom-up and top-down attention for image captioning and VQA*. 2017.
7. Mnih V, Heess N, Graves A. Recurrent models of visual attention. *Advances in neural information processing systems*. 2014: 2204–2212.
8. Lipton ZC, Berkowitz J, Elkan C. *A critical review of recurrent neural networks for sequence learning*. 2015.
9. Ba J, Mnih V, Kavukcuoglu K. *Multiple object recognition with visual attention.* 2014.
10. Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections. *Computer Vision and Pattern Recognition*. 2017 IEEE Conference. 2017:5300–5309.
11. Xie S, Tu Z. Holistically-nested edge detection. *International Journal of Computer Vision*. 2017;125(1–3): 3–18.
12. Hou Q, Liu J, Cheng MM, et al. Three birds one stone: a unified framework for salient object segmentation, edge detection and skeleton extraction. 2018.
13. Ablavatski A, Lu S, Cai J. Enriched deep recurrent visual attention model for multiple object recognition. *Applications of Computer Vision.* 2017 IEEE Winter Conference. 2017:971–978.
14. Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. *Advances in neural information processing systems*. 2015:2017–2025.
15. https://github.com/deepmind/mnist-cluttered
16. Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning. *NIPS workshop on deep learning and unsupervised feature learning*. 2011;2011(2):5.
17. Mendi E, Milanova M. Contour-based image segmentation using selective visual attention. *Journal of Software Engineering and Applications*. 2010;3(08):796.
18. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*. 2017;39(6):1137–1149.