All about attention

Shin Asakawa

Tokyo women's christian university
asakawa@ieee.org

Abstract

Cognitive scientists are paying attention to attention since Broadbent. Especially studies about psychological evidence, computational models, and their neural correlates of attention were contributed to the advances of these areas. Recent advances in deep learning for both image understanding and natural language processing are worth considering. Questions whether these studies are compatible might be interesting. We gave a brief survey in physiology, psychology, and computational models about attention. We also focused on the saliency map and winner-take-all (WTA) circuits and proposed that the WTA function might be implemented in the penultimate layer. Despite differences between physiology and computational modeling such as bottom-up and top-down interactions. Attention is still worth studying and attractive all the scholars who are interested in cognitive functions.

Keywords:Attention, neural networks, winner-take-all, bottom-up and top-down

1. Introduction

Recent progress of deep learing for both image and language understanding [19, 9, 6] might be influential for coginitive science. These studies might be included as

1. the multi-head attention, or self attention in the transformer [33] is bottom-up attention in cognitive psychology.
2. Attention based image and language interactions [34]

Th sensory cortex (Fig. 1) is studied is one of encoding-—the process by which stimuli are transformed into patterns of neural activity–and decoding, the process by which neural activity generates behavior. The ventral visual pathway is the most comprehensively studied sensory cascade. It consists of a series of connected cortical brain areas. PIT,CIT, AIT, RGC, LGN, Fig.1(c) are multilayer neural networks, each of whose layers are made up of a nonlinear combination of simple operations such as filtering, thresholding, pooling and normalization.

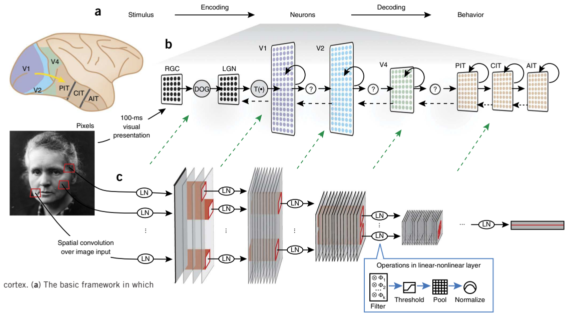

Fig. 1 A shematic correspondings between the brain areas and deep convolutional neurral network model. From [38] Fig. 1

2. Psychological evidence

Attention was studies in psychophysics, functional brain imaging, electoro-phisiology, neuropsychology, and computational modelings. We will give a brief review of psychological and computational models with respect to recent advances in deep learning. There are several important concepts such as **Spotlight (search light) metaphor** [5], **feature binding** [31], **attention bottle neck** [27], and guidance, selection, enhancement, exogeous vs endogenous, saliency map [18],WTA[18]. Among them, this paper was intended to address the following points:

1. saliency map = penultimate layer hypothesis, but multi scale attention proposed by Wang and Shen [35].
2. botom-up/top-down attention = winner-take-all = softmax hypothesis [20] It can be explained the Inhibition of Return [16]
3. In conjunction with the layer represenation [29], another possibilities might be considered for atttetion

3. Related brain areas

Most prominent areas was shown in fig. 2. Areas in blue in Fig. 2 indicate the dorsal frontoparietal network. “FEF”, frontal eye field; “IPs/SPL”, intraparietal sulcus/superior parietal lobule. Areas in orange indicate the stimulus-driven ventral frontoparietal network. TPJ, temporoparietal junction “IPL/STG”, inferior parietal lobule/superior temporal gyrus; “VFC”, ventral frontal cortex “IFg/MFg”, inferior frontal gyrus/middle frontal gyrus).

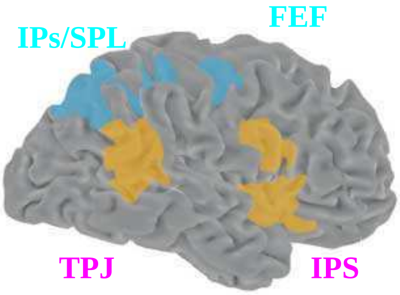

Fig. 2 Dorsal and ventral frontoparietal networks and their anatomical relationship with regions of damage in patients with unilateral neglect. From [4] Fig. 7a

A shematic diagram was shown in Fig.3. Visual information enters the primary visual cortex via the lateral geniculate nucleus (LGN), and the superior colliculus (SC). From there, visual information progresses along two parallel hierarchical streams. Cortical areas along the “dorsal stream” (including the posterior parietal cortex; PPC) are primarily concerned with spatial localization, or “where pathways” directing attention and gaze towards objects of interest in the scene. Cortical areas along the “ventral stream” including the inferotemporal cortex(IT) are mainly concerned with the recognition and identification of visual stimuli, or “what pathways”. Several higher-function areas are thought to contribute to attentional guidance, in that lesions in those areas can cause a condition of “neglect” in which patients seem unaware of parts of their visual environment.

One regison studied extensively is the prefrontal cortex (PFC). Areas within the PFC are bidirectionally connected to both the PPC and the IT [21].The PFC also has an important role in modulating, via feedback, the dorsal and ventral processing streams.

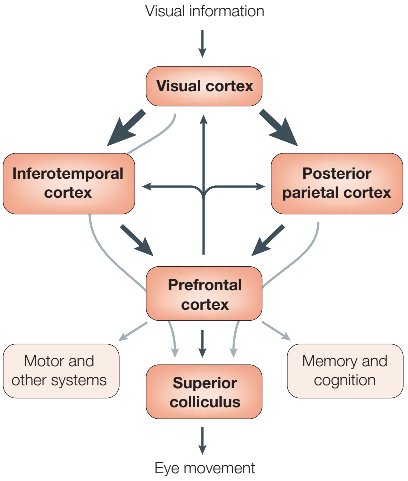

Fig. 3 A simplified overview of the main brain areas. From [13]

4. Mesurement of saliency

According to Wang and Shen [35], the terms attention, saliency, and eye fixation have the same meaning and used interchangeably here.

Given gaze data while examining targets, the strength of saliency at gaze locations is often evaluated. The normalized scan path saliency (NSS) is a measure of comparing the strength of saliency at gaze locations with the average strength of saliency in input images, which is employed in [25, 8]. Moreover, the Kullback-Leibler divergence between saliency distributions sampled from gaze locations and those sampled at random is regarded as a measure to evaluate saliency map from videos [11].

Studies that assume search tasks including visual search can employ an evaluation measure that counts the number of shifts of gaze locations to find targets, by simulating such gaze shifts based on obtained saliency maps. This measure is employed not only in the pioneer work by Itti et al. [14]

5. Top-down and bottom-up

Bottom-up: Development of computational models of attention started with the Feature Integration Theory [31], which proposed that only simple visual features are computed in a massively parallel manner over the entire visual field. Attention is then necessary to bind those early features into a united object representation, and the selected bound representation is the only part of the visual world that passes though the attentional bottleneck. Koch and

Ullman [18] extended the theory by proposing the idea of a single topographic **saliency map**, receiving inputs from the **feature maps**, as a computationally effcient representation upon which to operate the selection of where to attend next: A simple maximum detector or **winner-take-all** (WTA) neural network was proposed to simply pick the next most salient location as the next attended one, while an active **inhibition-of-return** (IOR) mechanism would later inhibit that location and thereby allow attention to shift to the next most salient location. From these ideas, a number of fully computational models started to be developed.

Another bottom-up attention was applied to the natural language processings [33, 6]. The attention in BERT 4 (self attention) might be considered as bottom-up attention.

Fig. 4 NLP based models Left:encdor-decoder based model [1], Right:Transofor model[33]

Top-down: Models that address top-down, task-dependent influences on attention are more complex, as some representations of goal and of task become necessary. In addition, top-down models typically involve some degree of cognitive reasoning, not only attending to but also recognizing objects and their context, to incrementally update the model's understanding of the scene and to plan the next most task-relevant shift of attention [23]. For example, one may consider the following information flow, aimed at rapidly extracting a task-dependent compact representation of the scene, that can be used for further reasoning and planning of top-down shifts of attention, and of action [23, 10]

Research towards understanding the mechanisms of top-down attention has given rise to two broad classes of models: models which operate on semantic content, and models which operate on raw pixels and images.

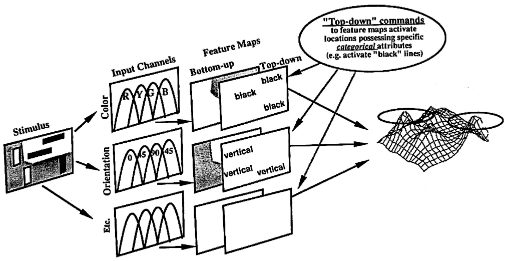

Fig. 5 The architecture of the guided search 2.0. Modified from [36] Fig. 2

One of the most probable models to account for the top-down bias is supposed to be Wolfe's Guided Search 2.0 [36](Fig. 5). Rensink [27] ellaborated these notions in terms of "Coherence theory", "gist", "attentional hand", and "tridiac stages".

6. What and where circuits

"saliency map" and "penultimate layer" The bottom-up and top-down cues are important to understand attention. Triesman and her collegues [31, 30] proposed the Feature Integration Theory. They also mentioned about the "search asymmetry" [32], "pop out". Those are hypothesized the several "feature maps" encoding each feature such as color, orientation of lines, brightness, motion disparity.

Fig. 6 shows fast R-CNN[7] shows that both information about "what" and "where" could be representated at a penultimate layer.

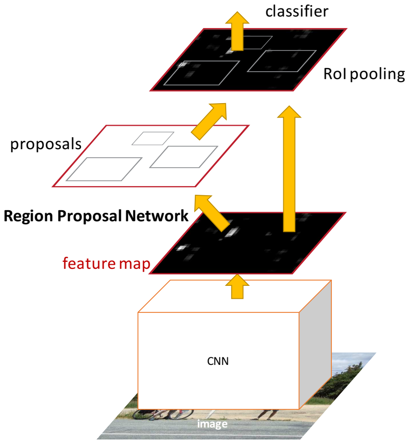

Fig. 6 A schematic diagram of Fast R-CNN [7]

Here, we propose the penultimate layer=saliency mapy hypothesis. and it is the place of attention operating with WTA=softmax function(Fig. 5). The softmax operation was also designated in the sequence-to-sequence model [1] (Fig. ??) for translation.

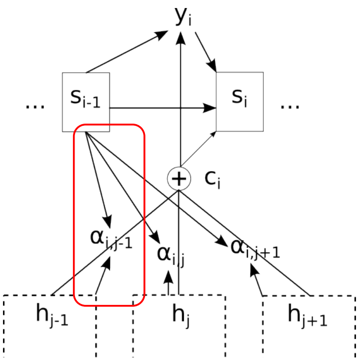

Fig. 7 Attention in natural laguage model [1]

This framework suggests that subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues.

On the other hand, physiologists found that dorsal and ventral pathways were seperated and might play different roles each other [24, 22, 15]. This discrepancy between phsiology and deep learning models might be considered more.

7. Saliency

Fig.8 shows predicted class of test images. The maps were extracted using a sing back-propagation pass thhrough a classification ConvNet.

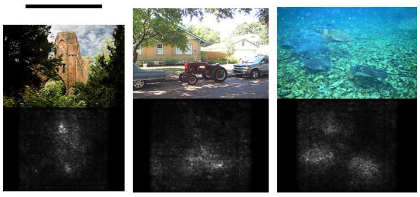

Fig. 8 Image-specific class saliency map for top-1 predicted class in ILSVRC-2013 test images. From [29] Fig. 2

Fig. 9 shows examples of weakly supervised object segmentation vis ConvNets. 9 left indicates images from the test set of ILSVRC-2013. Left-middle: the corresponding saliency maps for the top-1 predicted class. Right-middle: thresholded saliency maps: blue shows the areas used to compute the foreground colour model, cyan – background colour model, pixels shown in red are not used for color model estimation. Right: the resulting foreground segmentation masks.

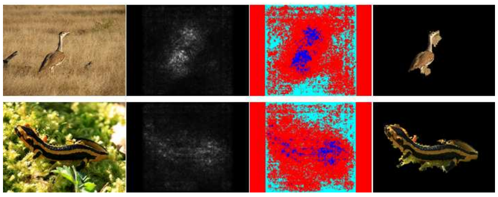

Fig. 9 Weakly supervised object segmentation using ConvNets. From [29] Fig. 3

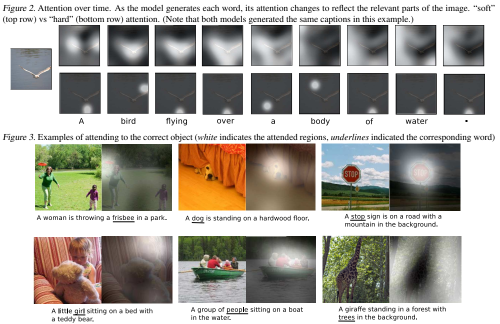

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. “soft” (top row) vs “hard” (bottom row) attention. (Note that both models generated the same captions in this example.)

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Fig. 10 Attention for neural image captioning [37]

8. Compulatation model

We show recent computational models of focal visual attention, with emphasis on the bottom-up, saliency of attentional deployment. We highlight five important trends that have emerged from the computational literature:

1. The perceptual saliency of stimuli critically depends on surrounding context; that is, a same object may or may not appear salient depending on the nature and arrangement of other objects in the scene. Computationally, this means that contextual influences, such as non-classical surround interactions, must be included in models.
2. a unique "saliency map" topographically encoding for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Many successful models are based on such architecture, and electrophysiological as well as psychophysical studies have recently supported the idea that saliency is explicitly encoded in the brain.
3. inhibition-of-return (IOR), the process by which the currently attended location is prevented from being attended again, is a critical element of attentional deployment. Without IOR, indeed, attention would endlessly be attracted towards the most salient stimulus. IOR thus implements a memory of recently visited locations, and allows attention to thoroughly scan our visual environment.
4. attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. Understanding the interaction between overt and covert attention is particularly important for models concerned with visual search.
5. scene understanding and object recognition strongly constrain the selection of attended locations. Although several models have approached, in an information-theoretical sense, the problem of optimally deploying attention to analyse a scene, biologically plausible implementations of such a computational strategy remain to be developed.

9. Summary

We gave a brief survey of 1) physiological or imaging studies, 2) psychological evidence, 3) computational model. Attention for both image and natual language processing recently advanced employed attenion mechanisms. Although physiology and brain imaging studies insisted these mechanisms must be processed in seperate pathways (what and where pathways), recent computational models deal these information in the same (peneultimate) layer. The discrepancy between physiological and computational models must be considered deeply for further understanding.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *Proceedings in the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[2] Narcisse P. Bichot, Matthew T. Heard, Ellen M. DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88:832–844, 11 2015.

[3] Donald E. Broadbent. *Perception and Communication*. Pergamon, Oxford,UK, 1958.

[4] Maurizio Corbetta and Gordon L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3:201–215, 3 2002.

[5] Francis Crick. Function of the thalamic reticular complex+ the search light hypothesis. *Proceedings of the National Academy of Sciences*, 81:4586–4590, 1984.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018.

[7] Ross Girshick. Fast R-CNN. *arXiv:1504.08083*, 2015.

[8] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Proc. Conference on Neural Information Processing Systems (NIPS)*, 545–552, 2007.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.033835*, 2015.

[10] Laurent Itti and Michael A. Arbib. Attention and the minimal subscene. In Michael A. Arbib, editor, *Action to Language via the Mirror Neuron System*, pages 289–346. Cambridge University Press., Cambridge, U.K., 2005.

[11] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49:1295–1306, 2009.

[12] Laurent Itti and Ali Borji. Computational models: Bottom-up and top-down aspects. In Anna C. Nobre and Sabine Kastner, editors, *The Oxford Handbook of Attention*, chapter 38, pages 1122–1158. Oxford University Press, 2014.

[13] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:1–11, February 2001.

[14] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[15] EveLynn McGuinness John Allman, Francis Miezin. Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neurosciece*, 8:407–430, 1985.

[16] Raymond M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4(4):138–147, 2000.

[17] Eric I. Knudsen. Fundamental components of attention. *Annual Revivew of Neuroscience*, 30:57–78, 2007.

[18] Christoh Koch and Simon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[19] 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, *in Advances in Neural Information Processing Systems 25*, Montréal, Canada, 2012.

[20] D. K. Lee, L. Itti, Christoph Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, pages 375–381, 1999.

[21] E. K. Miller. The prefrontal cortex and cognitive control. *Nature reviews Neuroscience*, 1:59–65, 2000.

[22] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–284, 1985.

[23] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45:205–231, 2005.

[24] Ernst Niebur and Christof Koch. Control of selective visual attention: Modeling the "where" pathway. In *Neural Information Processing Systems*, volume 8, pages 802–808, 1996.

[25] Robert J. Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 18–23, Minneapolis, Minnesota, USA, 6 2007.

[26] Steven E. Petersen and Michael I. Posner. The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, 35:73–89, 2012.

[27] Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3):17–42, 2000.

[28] Tim Shallice, Paul W. Burgess, Frederick Schon, and Doreen M. Baxter. The origins of utilization behaviour. *Brain*, 112:1587–1598, 1989.

[29] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint*, cs.CV(arXiv:1312.6034v2), 2014.

[30] Ann Treisman. Feature and objects: The fourteenth bartlett memorial lecture. *The quarterly Journal of Experimental Psychology*, 40A:201–237, 1988.

[31] Ann Treisman and George Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[32] Ann Treisman and J. Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *JEP:General*, 114(3):285–310, 1985.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Łukasz Kaiser. Attention is all you need. *arXiv preprint*, 2017.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015.

[35] Wenguan Wang and Jianbin Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2018.

[36] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin and Review*, 1(2):202–238, 1994.

[37] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*, 2015.

[38] Daniel L. K. Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 3 2016.