

# Chapter 15

## The Future of Attention Models: Convergence of Deep Learning with Artificial and Human Attention



Matei Mancias , Vincent P. Ferrera, and Antoine Coutrot

This book contributes to the crucial endeavor of understanding and modeling human attention. It gives an overview of physiological and computer science models, an extensive approach to model validation, as well as new trends of attention models. It also paves the way for further investigations. Some directions for future research are discussed in the next section, in relation to the major contributions summarized above. In the second section, a perspective on issues beyond attention, such as higher-level processing and the link with the deep learning techniques, is provided. We propose that human attention can be viewed as a suite of computational strategies that are essential for autonomous behavior by agents both natural and artificial. The study of attention should go beyond filtering of sensory data to develop an understanding of how relevant and valuable information is actively gathered by agents who possess an integrated awareness of both their internal goals, needs, and abilities and external sources of sustenance or danger. This kind of awareness implies an ability to model both the environment and the self that acts within that environment. Understanding the computational mechanisms underlying active, goal-oriented attention may be a step towards autonomy and a more general and “curious” AI.

---

M. Mancias (✉)  
Numediart Institute, University of Mons, Mons, Belgium  
e-mail: [matei.mancias@umons.ac.be](mailto:matei.mancias@umons.ac.be)

V. P. Ferrera  
Zuckerman Institute on Mind Brain and Behavior, Columbia University, New York, NY, USA

A. Coutrot  
CNRS, INSA Lyon, Universite Claude Bernard Lyon 1, LIRIS (UMR5205), Lyon, France

## 15.1 From Human Attention Study to Models

### 15.1.1 Attention, Emotions, Memory, and Actions

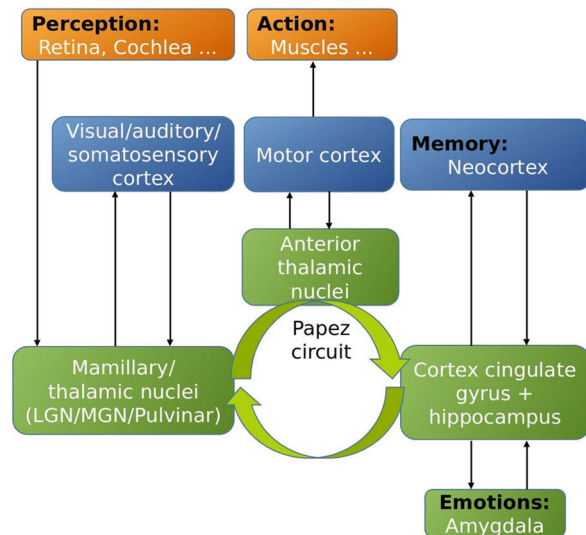
Attention and memorability are heavily interlinked. While the influence of emotions and memory on attention is obvious and this influence is part of the definition of top-down attention, in the other direction (attending towards emotions and memory) things are less clear. Nevertheless, even if the links are not as obvious as one would think, the first step towards memorizing an object may require attention as indicated by studies of memorability studies based on visual attention [1].

In the brain, a basic structure within the thalamus provides very interesting clues about a possible relationship between attention, memory, and emotions. This is the Papez circuit (Fig. 15.1) which was initially seen as a mechanism for emotions [2].

The main element of this circuit is the hippocampus which is important for episodic and spatial memory [3]. At the rostral end of the temporal lobe, a collection of nuclei called the amygdala is involved in emotions. The emotions related to the amygdala are mainly negative but positive emotions also evoke a response in this area [4]. The amygdala also responds to high interest or unusual images which attract attention [5].

On the other side of the Papez circuit one can find the mamillary body and the anterior thalamic nuclei. The mamillary body relays the output of the hippocampus and anterior thalamic nuclei. The mamillary body relays the output of the hippocampus and amygdala to the anterior thalamus and has an important role in spatial memory [6]. The anterior thalamic nuclei are linked to action and the motor cortex. Other thalamic nuclei that are important in sensory processing and attention are the lateral geniculate nucleus (relays signals from the retina to visual cortex [7]), the medial

**Fig. 15.1** A simplified schematic view of Papez circuit. In green the areas which are directly involved in the circuit. This circuit is linked to both emotions and memorability. Indeed, if impaired, new data will not go into the long-term memory, but older recollection is not affected



geniculate nucleus (auditory perception), and the pulvinar which is directly related to attention by modulating or gating sensory signals in relay nuclei [8].

It is very difficult to isolate locations in the brain which are responsible for complex functions such as attention (see Chap. 4), emotion, or memory, but the Papez circuit is of particular interest because in the limbic system, attention, memory, action, and emotions have a close anatomical proximity and are all needed in the process of memory formation. Thus, attention is heavily interconnected with emotions, memory, and action. Indeed the effects of an agent's own body on its environment are highly important in scene understanding, and it also has a crucial impact in the feeling of self-awareness and ownership which are at the basis of consciousness.

## ***15.1.2 Perspectives in Priority and Curiosity Modeling***

### **15.1.2.1 Salience and Priority in Human Vision**

The concepts of salience and priority are essential for understanding vision and eye movements in animals, including primates [9]. The theory of signal detection provides a foundation for understanding how salience and priority relate to the detection and discrimination of behaviorally relevant visual stimuli. A basic understanding of the physiological networks and computations that are associated with attention in humans and monkeys is useful when one seeks to emulate these functions in machine vision algorithms.

### **15.1.2.2 Curiosity: Uncertainty Reduction Through Guided Exploration**

When humans encounter a novel environment, one of their first priorities is to learn whom, what, and where to attend. Random, novelty-driven exploration has to be balanced with hypothesis-driven mechanisms for identifying sources of valuable or meaningful information [10]. This kind of everyday goal-directed information seeking can be called curiosity. Reward circuits in the brain respond not only to the likelihood and amount of reward, but to uncertainty [11]. Quantifying uncertainty is essential for decision-making [12]. Attention can be thought of as a mechanism to seek information that reduces uncertainty through guided exploration. Future studies should focus on the neural and computational mechanisms of attention and decision-making in environments where risk, uncertainty, and ambiguity can be controlled. An autonomous agent, such as a robot, should be able to pose questions that are relevant to its current situation and to formulate plans to seek answers to those questions.

## 15.2 Perspectives in Computational Attention Modeling

### 15.2.1 Models

#### 15.2.1.1 From Static to Dynamic Saliency Maps: Computing Eye Scan-Paths

Most saliency models take color images as input and produce saliency maps that estimate the probability distribution of the gaze in the image. The static nature of these maps could be an issue for some applications as these models do not predict the temporal sequence of human fixations (also called scan-path). What is the order of fixations? How is the image seen dynamically? This information is important in action planning and execution, or continuous learning and adaptation. Some models propose algorithms to predict the scan-path from a saliency map [13] and the field has evolved since the first edition of this book. However, this question deserves to be more deeply investigated in terms of automatic scan-path generation and especially in automatic time fixation modeling (see Chap. 7). As is the case for saliency models, most of the standard validation metrics to compare the output of attention models and human fixations are static. They do not take into account the temporal sequence of human fixations. This field has evolved to the point where it is possible to have unified metrics to compare two scan-paths [13], yet there is still room for improvement in the field (see Chap. 8).

#### 15.2.1.2 Multimodal Modeling of Attention

In the previous edition of our book we said that one of the future trends would be to aggregate results from attention algorithms on videos, but also on auditory and depth maps. This has indeed become a current trend with more and more models going into that direction. This integration of the different modalities is increasing due to real-life applications in drones, robotics, or automotive industry where the environment is (1) changing very fast and (2) a lot of sensors are able to provide different modalities from cameras to point clouds going through radars and of course audio information. These developments will drastically augment the already numerous engineering applications of attention modeling. Of course, new models will need new ground-truth and new validation techniques, but this effort is crucial to boost the attention modeling community and to augment its visibility both in other research communities and in industry.

Video models that incorporate audio information define a field which addresses the key challenge of audio-video consistency which still needs to be developed. 360° images and video and the RGB-D saliency models (with depth maps) arrive as an extension with applications to VR images. 360° cameras are also being introduced in several fields such as robotics. 3D saliency on meshes or point clouds are also part of the dynamics given the 3D data production which is increasingly important along with the development of VR.

### 15.2.1.3 Towards More Natural, Diverse, and Less Biased Datasets for Model Training and Validation

The development of multimodal models is definitely one of the main future paths in computational attention along with dynamic modeling of attention through saccades and fixations. This development needs to be supported by large multimodal datasets video, audio, 360°, and depth maps and also include data from other sensors which are not yet taken into account (Lidar, radar, accelerometers, touch sensors, etc.). To avoid biases it is very important to have datasets collected in very different conditions and real-life situations. To address this issue, more databases like [14–16] with a lot of different classes have to be collected from diverse sources (paintings, drawings, websites, advertising, etc.) to understand what attracts attention when observing stimuli that differ from classical “natural images.”

It is also important to have datasets with a higher viewer diversity (much more than 20–27 years old mainly male upper social class students in computer science who seem to dominate in computer science made datasets).

Last but not least, it is important to have datasets with different precise natural tasks and not only the unconstrained free-viewing task. Attention is often studied in situations where subjects are given various cues or reinforcement. In other words, the tasks are structured to guide attention to locations, features, or objects that have been chosen by the experimenter. This begs the question of how subjects naturally deploy attention when performing everyday tasks such as driving, making a sandwich, or playing sports. Ballard, Hayhoe, and colleagues [17] have recently developed an immersive virtual reality system for recording eye, head, and hand movements when human subjects are performing simple tasks. Such systems allow experimental control over external variables like novelty, reward, and context while imposing minimal constraints on subject behavior. In the future, such systems can be combined with mobile EEG recording to map the brain activity during natural, goal-directed behavior.

### 15.2.1.4 Foundation Models of Attention and Visual Search

The arrival of foundation models in saliency [18] is maybe a new trend which will further develop especially in the framework of adding new modalities. For example, in audio-video models, the audio and visual data consistency issue is key. This issue might evolve a lot with the arrival of foundational multimodal models where video embeddings can be easily compared to audio embeddings. Indeed the gap in comparing diverse modalities (video, audio, touch, radar, lidar, etc.) might be more easily solved if embeddings from the different kinds of signals can be compared using simple distance metrics to know how similar they are. The comparison of a given current signal and memory (in terms of past datasets) will also be much easier for long-term (sustained) attention. The possibility to compare the current signal with text prompts also opens the way of visual search with saliency maps varying

based on the task provided by those prompts. Those new approaches lead to more industrial applications and more flexible and adaptable agents and go in the same direction as in Sect. 15.1.2.2.

### 15.2.1.5 More Links Between Deep-Learning and Attention?

In the first edition of the book we already said that deep learning would make important evolutions in computational attention and this was definitely the case. Most of the recent models are based on deep learning and their results are sometimes impressive. However, we see an issue in bottom-up features and deep learning. Kummerer et al. [19] showed that bottom-up attention was underestimated by deep learning models. In their experiments, a simple bottom-up model could outperform a state-of-the-art deep learning model when the images contained less top-down information. Moreover, they could not easily adapt to images in a different context from their training set, showing little adaptability to new very different stimuli. In the same way, the authors in [20] also show that deep learning-based attention models poorly explain bottom-up attention, but they capture very well top-down information about objects that are usually salient, while some bottom-up information is often present in the training set such as contrast, which is an eye-catching feature that can be learned.

Those findings show how the different mechanisms of attention such as rarity could improve classical deep learning architectures by making them aware of “surprising” data without the need of learning. Attention might be an important future mechanism letting deep learning be more reactive, especially to unseen kinds of data.

In addition to that, attention-based modules are also explicitly incorporated into deep learning, especially through transformer architectures and self-attention. The attention mechanism of self-attention compares one feature or vector  $F$  of the signal with all of the other features/vectors in the signal. It then provides more weight to parts of the signal which are closer to (or more closely aligned with)  $F$ . We can see here a clear link with the first step of a classical human model of attention based on rarity. In models such as RARE2012 [21], the same idea was already used: in an image, comparing a feature  $F$  to every other feature in the image ([21] uses a histogram on the feature map). Once this self-attention is applied, the self-information or attention  $A$  is obtained from the occurrence probability  $P$  of the feature  $F$ :  $A = -\log(P(F))$ . Thus, a “rare” feature which does not occur frequently is given more attention than a common feature with a large occurrence probability. This variant of attention has a fundamental difference with the transformer’s attention. Although they both share the starting point of self-attention, the weights of the transformer’s self-attention are trained for a specific task, whereas the RARE2012 model will find the rarest region of the input with the goal of mimicking bottom-up human attention, which focuses on surprising information. However, we can see that the basic underlying self-attention is the same leading to possible interactions between the two.

### 15.3 Links Between Human and Artificial Attention

In the last decade, artificial neural networks (ANNs) have proven very effective—and sometimes better than humans—at a variety of visual tasks, including image classifications and object recognition [22]. Several approaches have been tried to understand which visual regions ANNs “attend to” when processing images, i.e., which areas most influence the outputs of the models [23, 24]. Interdisciplinary teams including neuroscientists, cognitive scientists, and computer scientists have tried to compare how ANNs and humans visually process images when given a similar task (see Chap. 13) [25–27]. A few studies have directly compared ANN “attention” and human gaze, e.g., in Atari games [28], meal preparation [29], driving [30], and visual question answering [31]. In [32], the authors show that during a medical image classification task, as the ANN gets trained, its performance gets closer to the human experts, and the similarity between the model and human attention increases. Altogether, these studies reveal a significant overlap between ANN “attention” and human visual selectivity estimates, modulated by the task at hand and the content of the images [33].

Those findings show that human and artificial attention might converge in the framework of deep learning and application-based protocols. Object recognition and semantic understanding already show relations to saliency [34, 35] as understanding is first based on the most important parts of the signal. Thus, the convergence of deep learning architectures for scene understanding and computational attention algorithms which are directly related to human attention lead to machines becoming curious and able to prioritize data and take actions in unexpected situations. This is the most important perspective we can foresee. This convergence needs even more multi-disciplinary teams including machine learning, computational attention, cognitive psychology, and neurosciences.

### 15.4 Summary

There are many opportunities for continued development of computational strategies for taking advantage of the benefits of attention in both human and machine vision applications:

- Collection and curation of large-scale databases that integrate retinal image and eye movement information under passive and active conditions.
- Integration of visual attention models with other sensory modalities such as audition, somatosensation, vestibular (balance and head position), olfaction, etc.
- Exploring the top-down effects of memory and emotion in assigning priors and values to locations and objects in priority maps. Flexible visual search algorithms where the task is provided by text prompts for example is a very interesting development for practical applications.

- Forging closer links between the study of attention in animals (including humans) and machines.
- Working on the convergence of the different communities needed to achieve links between deep learning, artificial attention, and human or animal attention implying computer science, machine learning, computational attention, cognitive psychology, and neuroscience.

## References

1. Mancas, M., & Le Meur, O. (2013, September). Memorability of natural scenes: The role of attention. In *2013 IEEE international conference on image processing* (pp. 196–200). IEEE.
2. Papez, J. W. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*, *38*(4), 725–743.
3. Burgess, N., Maguire, E. A., & O’Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625–641.
4. Garavan, H., et al. (2001). Amygdala response to both positively and negatively valenced stimuli. *Neuroreport*, *12*(12), 2779–2783.
5. Hamann, S. B., et al. (2002). Ecstasy and agony: Activation of the human amygdala in positive and negative emotion. *Psychological Science*, *13*(2), 135–141.
6. Vann, S. D. (2010). Re-evaluating the role of the mammillary bodies in memory. *Neuropsychologia*, *48*(8), 2316–2327.
7. O’Connor, D. H., et al. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*(11), 1203–1209.
8. Arend, I., et al. (2008). 15-the role of the human pulvinar in visual attention and action: Evidence from temporal-order judgment, saccade decision, and antisaccade tasks. *Progress in Brain Research*, *171*, 475–483.
9. Bisley, J. W., & Mirpour, K. (2019). The neural instantiation of a priority map. *Current Opinion in Psychology*, *29*, 108–112.
10. Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*(11), 585–593. <https://doi.org/10.1016/j.tics.2013.09.001>. Epub 2013 Oct 12.
11. Schultz, W., Preusschoff, K., Camerer, C., Hsu, M., Fiorillo, C. D., Tobler, P. N., & Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1511), 3801–3811. <https://doi.org/10.1098/rstb.2008.0152>
12. Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*(5), 757–763.
13. Kümmerer, M., & Bethge, M. (2021). State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*.
14. Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.
15. Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *28*(5), 802–817.
16. Chen, Y., Yang, Z., Chakraborty, S., Mondal, S., Ahn, S., Samaras, D., Hoai, M., & Zelinsky, G. (2022). Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 5031–5040).
17. Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology*, *24*(13), R622–R628. <https://doi.org/10.1016/j.cub.2014.05.020>

18. Moradi, M., Moradi, M., Rundo, F., Spampinato, C., Borji, A. & Palazzo, S. (2024). Salfom: Dynamic saliency prediction with video foundation models. <https://arxiv.org/abs/2404.03097>
19. Kummerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE international conference on computer vision* (pp. 4789–4798).
20. Kong, P., Mancas, M., Thuon, N., Kheang, S., & Gosselin, B. (2018). Do deep-learning saliency models really model saliency? In *2018 25th IEEE international conference on image processing (ICIP)* (pp. 2331–2335). IEEE.
21. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6), 642–658. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0923596513000489>
22. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
23. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.
24. Selvaraju, R. R., et al. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128, 336–359.
25. Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J., & Shao, L. (2020). Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23, 2086–2099.
26. Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2023). Explanation strategies for image classification in humans vs. current explainable AI. *arXiv preprint*, arXiv:2304.04448.
27. Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021). Human attention in fine-grained classification. In *British machine vision conference, Aberdeen*.
28. Guo, S. S., Zhang, R., Liu, B., Zhu, Y., Ballard, D., Hayhoe, M., & Stone, P. (2021). Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 34, 25370–25385.
29. Li, Y., Liu, M., & Rehg, J. M. (2020). In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 6731–6747.
30. Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, 93(2), 451–463.
31. Sood, E., Kögel, F., Strohm, F., Dhar, P., & Bulling, A. (2021). VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. *arXiv preprint*, arXiv:2109.13116.
32. Vallée, R., et al. (2024). Influence of training and expertise on deep neural network attention and human attention during a medical image classification task. *Journal of Vision*, 24(4), 1–16.
33. Langlois, T., et al. (2021). Passive attention in artificial neural networks predicts human visual selectivity. *Advances in Neural Information Processing Systems*, 34, 27094–27106.
34. Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint*, arXiv:1411.1045.
35. Zheng, T., Chen, C., Yuan, J., Li, B., & Ren, K. (2019). Pointcloud saliency maps. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1598–1606).