Visual Attention Should Be Modelled Using Saliency Maps

Anindya Auveek

Cognitive Science, University of Toronto

11/30/2020

ABSTRACT

Attention is the crucial process of filtering information required for perception. We have evolved to notice information that is behaviourally salient. In humans, visual attention can be identified as the sensory mode that is most important for survival. I suggest that a model called Visual Saliency Map is therefore an ideal representation of human visual attention. In order to formalise this theory I explore studies that look at the function of the human cortical regions in response to visual stimuli. The findings show that stimuli in the visual, frontal and parietal cortex of the brain are organised in a gradient of importance/salience relevant to the context, as hypothesised. In order to mechanise this process, I look into statistical and computational models of Visual Saliency maps. Lower level features are shown to be modelled by both statistics and artificial neural networks, while higher level features such as objects are currently being modelled by image classification artificial neural networks. While there remains to be more nuanced components of attention yet to be modelled such as gaze or context, research in AI shows promising results.

Attention is known as the process of filtering perceptual information for contextual relevance. Without this process it would be impossible to make sense of the world due to the excessive amount of information available. Therefore this phenomenon arises in living beings as a factor of evolution and adaptation towards the environment and the process is closely tied with natural selection. In this essay I will focus on visual attention and will detail why a saliency map model is an ideal representation of it. To do this, I will first formalize the theory of a visual saliency map using research done in cognitive neuroscience. In addition, I will outline potential mechanisms of this theory through current computational and statistical models, and therefore give it a non-homuncular basis.

The saliency map model for visual attention derives its idea from Triesman's integration theory. His equation attempts to model perception through arbitrary features reflecting properties of the environment and weights denoting their importance. The saliency map suggests a gradient of the contextual salience in the brain which denotes the importance of a stimuli (Zhaoping, L., & Zhe, L. 2015). In order to identify the properties of this gradient and how it arises, one must first realize the locations of the brain relevant to this process. According to current understanding of the human brain, the region relevant to processing visual information is known as the visual cortex, and the regions relevant to processing top-down information such as semantics are located in the parietal cortex and prefrontal cortex. The attentional stream of the brain identified through research in neuroscience directs information through the latter (frontal and parietal) regions into the visual cortex in a way similar to a cone funnelling water. When information reaches the visual cortex it is already modulated through top-down attention (Treue S. 2003). Numerous studies have been conducted on the role of the Visual Cortex (V1) in an attempt to produce a lower order (bottom-up) saliency map. As of recent the modulation of higher order (top down) information through saliency sub-maps in the parietal/frontal cortex are also being studied.

In the paper by Zhaoping, L., & Zhe, L. (2015), they demonstrated that the V1 computes a saliency map such that the saliency of a location in the visual field is shown to be represented by the maximum response from V1 neurons to this location relative to the maximum responses to other locations. If two locations or images evoke the same maximum response they are considered equally salient. They display that V1 neurons are tuned to several features including orientation, color, motion and eye of origin. In their research they solely measured the V1 response with negligible top-down attentional guidance. By the definition of evolutionary saliency, higher reaction time to a certain stimulus would mean higher importance or saliency. Zhaoping & Zhe demonstrated that the latency of attentional shift to a stimulus is shorter for a more salient region as defined by maximum neural response. Empirically, the reaction time for a feature was directly linked to its maximum evoked V1 response. This is an interesting finding because it demonstrates saliency in terms of action and evolutionary relevance. If the V1 computes a stimulus as being more important for survival, it is reacted to quicker.

While I have shown the presence of a saliency map in the visual cortex, Thompson, K. G., & Bichot, N. P. (2005) provided results from research done in the frontal eye fields (FEF) (the region responsible for planning eye movements) that it is tightly coupled with attentional saliency in the frontal eye fields. They found an enhanced gain of neurons in V4 (visual cortex) whenever the neuron's receptive field overlapped with the movement field of the FEF neuron. This research supports other studies I will outline which uses eye tracking data to represent saliency of a region. Furthermore, the localisation of the FEF as a saliency relevant region supports other research identifying this, the prefrontal cortex and the parietal cortex as being regions involved in identifying top-down saliency for visual attention. The primate visual cortex can be identified as a region that is shaped as an integrated saliency

map; a representation of the environment that weighs every input by its local feature contrast and current behavioral relevance.

Bogler, C., Bode, S., & Haynes, J. D. (2011) conducted supporting studies demonstrating that the correlation of low-level saliency and high-level concepts does not mean that attention is directed mainly through high level concepts rather than behavioural saliency. This was done using a study that investigated patients with visual object agnosia (inability to recognize objects) that showed the presence of visual saliency detection in V1 regardless of their ability to identify objects (high-level concept). Results support a computational bottom-up saliency model and associate different anatomical regions to different computational stages of the model. Information about graded saliency is gathered from the higher order regions which is further modulated by the Visual Cortex in order to yield a representation of the most salient stimulus in the visual field.

Treue, S (2003) addressed that all visual information in the brain is shaped by top-down attentional influences. He presented that research done in the visual cortex of monkeys demonstrates a gradient of spatial attention specified by a cell's firing rate. Specifically, Treue showed that directing attention to a stimulus matching the cell's preference will tend to increase responses. In addition, he stated with evidence from other studies that feature-based, non-spatial effects also exist (eg. size, shape, etc). This latter feature based effect brings some problems to this bottom-up saliency model, as it is difficult to distinguish from object-based attention. To mitigate this issue, he provided a potential merging of bottom-up and top-down saliency.

As I stated previously, like a filtering tunnel, the information in the visual cortex is already filtered by top-down attentional guidance which has its own saliency map. After this filtering is done, the visual cortex uses bottom-up features to map the saliency of stimulus. Trueue (2003) suggested that multiple representations of a distributed saliency map could

exist in the visual and frontal/parietal cortex, through which perception would be based on the activity in the area whose neuronal properties are best matched to the current perceptual task. In accordance, he provided results from other research that higher cognition areas can model salience with objects such as houses, faces, etc.

To summarise from neuroscience research, current studies support the saliency map model of visual attention. It is suggested that visual information is filtered in the human brain with the convergence of top-down and bottom-up saliency from adjacent brain regions. In order to mechanise this theory I will provide statistical and computational bases for this function both from research done using neural data and from models in computer vision science.

For mechanization of the visual saliency model, the neural response to visual sensation can be modelled statistically so that this process can be applied in a computational model. Duan, H., & Wang, X. (2015) explained that the distribution probability of neuronal responses of the V1 region can be described using Gaussian (normal) Distribution Model. Focussing only on bottom up saliency, they demonstrated that objects that differ from the rest of scene maximally are regular focuses of attention in a free view. According to their hypothesis, the computation of a saliency map can be demonstrated using statistical probabilities of the neuron responses, where the salience map is computed as a combination of values from all neurons in the region. Furthermore, it was shown how our prior knowledge can prime the neuronal response such that previously experienced salient stimulus can influence the outcome of future saliency map. This was done using the Bayes theorem of conditional probability. Their resultant saliency model from V1 neural responses were sub-maps of different color spaces (RGB) that combined into a final saliency map, described as a probability map. Using this understanding, one can attempt to build a mechanized computational model using statistics.

Li (2018) presented mechanisms of these saliency sub maps with a computer generated saliency model. He first described a previous model of bottom up saliency of regions in an image using a statistic called KL divergence; which measures the difference between a value compared to others i.e the difference between a location and its surrounding area. This is similar to the probability map explained by Duan, H., & Wang, X (2015). Li suggested that a more accurate model would count size in addition to visual location in order to simulate the dynamic attentional process in humans; where size of a region in the visual field plays a big role. This is demonstrated from human eye tracking data where the focus length of our eye is adjusting from short to long continuously i.e from large to small. His model proposes that producing a model using non-salient regions which mutually inhibit each other can more accurately identify salient regions, and therefore be a better representation of visual saliency. He also used a gaussian filter (normal distribution) to model saliency like previously.

The problem with these statistical models is that only basic features can be taken into account such as colour, sharpness of image, etc. In order to resolve this issue I will outline a study conducted using a statistical process called PCA which identifies important features in data intrinsically. In correspondence, Bruce, N. D., & Tsotsos, J. K. (2009) presented a model which has an output qualitatively similar to human density maps (eye tracking). The outputs were from both video data and static images and highlighted the importance of specific neural code in the determination of visual saliency as similar regions were identified as being salient to a similar degree.

Even so, there remain problems with these mechanizations of saliency maps. To start the statistical models have a lack of context explanation. While pixel probabilities can provide functional results, it cannot produce a qualitative saliency map. To add, high level (Top down) saliency is yet to be modelled by the previously shown statistics. Also, the

performance of a mechanized model cannot be measured accurately. While it can be measured by fixations of the eye, the accuracy of its convergence with neural maps in the brain is very complicated to calculate (Boccignone, G., et al. 2019).

As of recent, the mechanisms of human neuron networks have been modelled using artificial neural networks. These networks are trained similarly to how neurons in the brain are by adjusting weights of features according to the data. These models can extract higher level features from data rather than the basics identified previously. Using these artificial neural networks, the top-down saliency which was previously avoided can be modelled. While these features are hard to label semantically, some are identified as symmetry, surprise, texts, and signs. Further features such as social and action cues are yet to be modelled (Borji, A., 2019).

While research in mechanizing top-down visual saliency is currently ongoing, recent findings with neural networks show optimistic results. Artificial networks used for image/object recognition, have been used to model saliency, and can more accurately predict human saliency, calculated using eye tracking data (Oyama, T., & Takao,Y., 2018). The issue of modelling higher order salience in the primate parietal and frontal cortex can be mitigated. Nevertheless, many of these features used in the networks are semantically difficult to identify, and nuanced features such as the gaze of an individual in the image (or vision) and the context of an image are still yet to be modelled. Furthermore, current mechanizations of saliency maps do not account for temporal dynamics, i.e the relevance of a previous image to the current. Although Bayesian statistics model prior probabilities, these are long term learned rather than continuously in a dynamic process (Bruce, N. et al., 2015).

In summary, I have shown that visual attention in the human brain is mapped by it's behavioural salience, whether it is by higher level properties in the frontal and parietal cortex or by lower level in the visual cortex. The intuition behind saliency based models is derived

from natural selection and behavioural importance, and therefore is in line with current understanding of biology. In addition using statistical and AI models I have outlined how this can be mechanised in a non-homuncular fashion such that it passes the turing test. Although there are still more complicated parts of attention in humans that have yet to be mechanized (eg. context, knowledge, memory, etc.), research in artificial intelligence shows promising results. With further work done in the field, neural networks should be able to model neural saliency maps of visual attention more accurately. As the higher level processes of the human brain can be broken down semantically both in Neuroscience and AI, the saliency map model can be better formalised and mechanised, further strengthening the theory.

References

Boccignone, G., et al. (2019). Problems with saliency maps. Lecture Notes in Computer Science 558. https://link.springer.com/bookseries/558

Bogler, C., Bode, S., & Haynes, J. D. (2011). Decoding successive computational stages of saliency processing. Current biology : CB, 21(19), 1667–1671. https://doi.org/10.1016/j.cub.2011.08.039

Borji, A., (2019). Saliency Prediction in the Deep Learning Era: Successes, Limitations, and Future Challenges. IEEE Transaction of Pattern Analysis and Machine Intelligence. 1-44. http://arxiv.org/abs/1810.03716.

Bruce, N. et al. (2015). On computational modeling of visual saliency: Examining what's right, and what's left. Vision Research, 166. 95-122. https://doi.org/10.1016/j.visres.2015.01.010

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: an information theoretic approach. Journal of vision, 9(3), 1–24. https://doi.org/10.1167/9.3.5

Duan, H., & Wang, X. (2015). Visual attention model based on statistical properties of neuron responses. Scientific reports, 5, 8873. https://doi.org/10.1038/srep08873

Li, J., (2018). Visual Attention Is Beyond One Single Saliency Map. CoRR 1811-02650. http://arxiv.org/abs/1811.02650.

Oyama, T., & Takao,Y., (2018). Influence of Image Classification Accuracy on Saliency Map Estimation. CoRR 1807-10657. http://arxiv.org/abs/1807.10657.

Thompson, K. G., & Bichot, N. P. (2005). A visual salience map in the primate frontal eye field. Progress in brain research, 147, 251–262. https://doi.org/10.1016/S0079-6123(04)47019-8

Treue S. (2003). Visual attention: the where, what, how and why of saliency. Current opinion in neurobiology, 13(4), 428–432. https://doi.org/10.1016/s0959-4388(03)00105-3

Zhaoping, L., & Zhe, L. (2015). Primary Visual Cortex as a Saliency Map: A Parameter-Free Prediction and Its Test by Behavioral Data. PLoS computational biology, 11(10), e1004375. https://doi.org/10.1371/journal.pcbi.1004375