Visual Saliency Prediction and Effectiveness Analysis using Metrics

Instructors - Dr. Garrett Stanley, Dr. Chethan Pandarinath

Students - Yung-An Hsieh, Nick Su, Kedar Rao (Group 7)

Georgia Institute of Technology, BMED 7610 – Quantitative Neuroscience

Abstract

Computational attention models are aimed to predict human visual attention, which are the locations where humans look at images or videos. In this project, we reviewed and implemented existing attention models, and evaluate their performance with mathematical metrics. We found that deep learning-based models outperformed the traditional models which had no training process. We then modified the best model we implemented with self-attention, features fusion, and skip connections, which are techniques that help deep learning models to better recognize objects. The modifications only slightly improved visual attention prediction, which indicated other directions of improvement are required. To identify what direction is needed, we analyzed model performance on different image categories, such as images of street view or landscape. We concluded that higher-level understandings of images, such as knowing the interactions between subjects, are needed to further improve attention models.

Chapter 1. Introduction

The human eye dedicates itself to being one of the primary senses of the body, so much so that around 80% of everything we perceive is all dependent on the human eye. This along with several other factors, such as the methods of perception, movement of the eyes, its relationship with the brain and visual cortex and ultimately how it is involved in the complex study of neuroscience is a fascinating approach to the analysis of the visual system.

Visual neuroscience is a branch of neuroscience, focused on discovering the mathematical and probabilistic as well as biological means by which living beings see things around them. It involves the study of the brain's visual cortex and its interactions with light rays projected onto the retina. This is usually studied with the use of static images or video media. The field is progressive in its nature, to a point where it hasn't been able to provide an accurate explanation to our perception of observed events, but the recent boom in experiments and studies from all over the world have begun providing solutions and methods to closely study this phenomenon.

Computational attention models are created to test visual attention. Visual attention is, in general, a set of cognitive operations or processes that perform and mediate the selection of relevant and the filtering out of irrelevant information from cluttered visual scenes. Attention models are therefore designed to extract the relevant information as seen and perceived by the observer and provide elucidated data to define the objective that the individuals conducting the study are trying to obtain.

But what assists us in making that decision of relevant and irrelevant? What one person may focus on in a static image or video may not be an object of focus for another individual. The resurgence of neuroscience in the 1980's also brought with it the interest to solve the workings of eye movement and visual attention. Since then, scientists have performed studies to find the definitive model by deriving probabilistic metrics to evaluate the efficiencies of these models (**Fig. 1**). In this project, we focused on the following objectives:

1. We brought an analysis of several existing attention models, and our aim is to evaluate their efficiency and reproduce results from previous studies using mathematical metrics.
2. We explored whether techniques that are used in deep learning-based computer vision tasks (e.g. object detection, segmentation) also improve attention models.
3. We investigated model performance on different categories of image (e.g. images with humans or with animals) and identified what is needed to further improve the models.

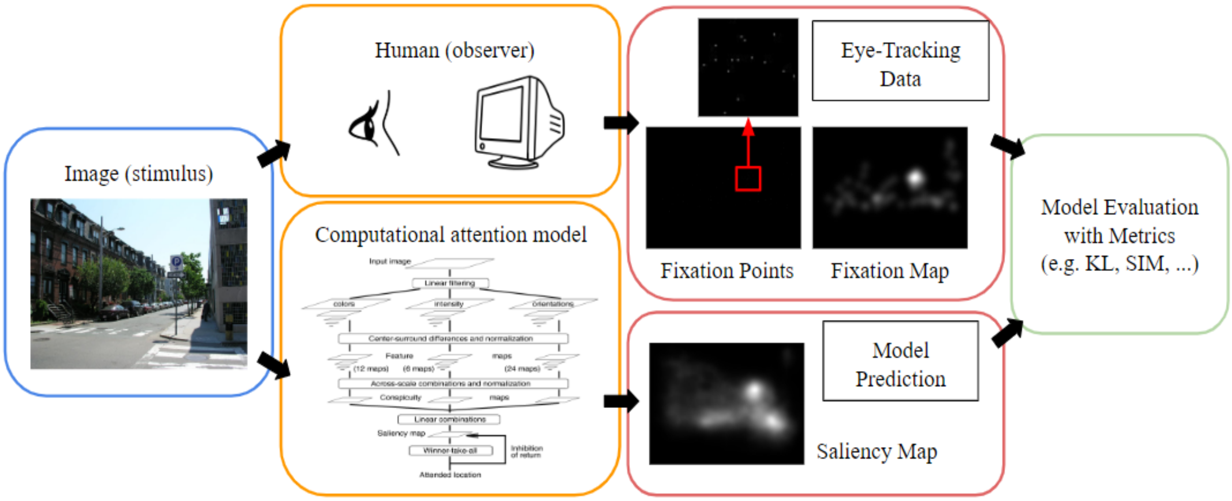


Fig. 1: The flowchart of attention models evaluation. Eye-tracking data, which is generated by eye trackers recording where humans look on images, serves as the ground truth. The attention model generates the prediction of visual attention, called the saliency map. The metrics are then used to compare the ground truth with model prediction to evaluate the model.

Chapter 2. Backgrounds

2.1 What is Saliency in Attention Models?

The most basic and fundamental way to describe saliency in neuroscience is in one word the "uniqueness" of a selection as compared to the neighboring selections. In other words, the saliency of a 'feature' is the quality by which it stands out from its neighboring elements. In visual neuroscience, the attention model would predict the 'hot-spots' on the image or video file that would seemingly pique the interest of a human or animal observer.

Attention models give us an idea of where the interesting parts of an image are and not only what they are. This is in part due to how attention works in a biological scenario. For example, we are more attuned to where a light source is coming from rather than what the light source is. How an attention model works, depends completely on the 'architecture' of the attention models. The result is however always intended to be a saliency map. A saliency map is an image that points out all the salient features or locations on the image, which usually has the same size as the original image.

2.2 Model Architecture

Based on a paper from Borji et. al. [1] computational attention models can be specifically broken down into two architectures. Or it can be said that the attention can be deployed in two ways.

2.2.1 Top-Down Architecture

In top-down attention, goal-driven information related to the ongoing behavior, task, or goal is selected (e.g. road lane for driving behaviors). In general terms, the top-down approach can be described as involving the perception that is not dependent on the stimulus input but is the result of the stimulus, or the internal hypothesis and interactions based on expectations.

2.2.2 Bottom-Up Architecture

The bottom-up architecture of attention is also called the stimulus-driven component of attention and it processes information in a feed-forward manner. Here the image or stimulus is fed to the model and transformations are then applied to highlight the most visually interesting, important, or "salient" locations.

The aim of the bottom-up model is to process visual information fed in the form of an image or video stimulus and to return a saliency map. These models have immense value since they can be used to understand the attention mechanism in human beings at computational and neural levels. Also, the aspect of predicting where people look, may it be in images or videos, has applications in computer vision, robotics, medicine, neuroscience, surveillance, etc.

Over the years, the bottom-up approach of modeling has proved more fruitful in saliency mapping and visual prediction. Most models being studied today are elucidated by the fact that they involve a bottom-up architecture and approach. Therefore, the focus of our study has been the use of bottom-up models for their efficiency and computational power.

2.3 What are Evaluation Metrics?

What is the best way to evaluate a saliency model's predictive ability? This is an ongoing question in the field of computer vision and visual neuroscience. There can be several attention models that we can choose from but objectively determining which one is the best requires the assistance of a metric. We derive our methods of metric evaluation from the paper by Bylinskii et. al. [2].

Metric behavior depends entirely on its nature and efficiency of working in the same environment as that of a specific model. That is, if Model A works with Metric A to provide a certain result, the compatibility of Model A with Metric B in yielding a more suitable result leaves a lot of information up to the interpretation of the individual performing the comparison, hence, leaving out the scope of finding one definitive result across all models or all metrics. It is, therefore, necessary to quantify metric behavior across several saliency models and then interpret the result of one particular saliency model with another using many different metrics.

It is observed that several metrics take a probabilistic approach to provide a distribution comparison and many others treat distributions as histograms. Metrics can also be different in how they treat the presence of false positives and false negatives. We gave more analyzation on the metrics for visual saliency in **Chapter 3.3**. The following metrics that reviewed and used in our project, detailed explanations of these metrics can be found in [2]:

1. **KL** (Kullback-Leibler Divergence): It is a non-symmetric measure of the information lost when the saliency map is used to estimate the fixation map.
2. **NSS** (Normalized Scanpath Saliency): It is measured between two saliency maps as the mean value of the normalized saliency map at fixation locations.
3. **SIM** (Similarity measure of two saliency maps): Histogram Intersection and measures the similarity between two different saliency maps when viewed as distributions.
4. **EMD** (Earth Mover's Distance): Measures the distance between two probability distributions by how much transformation one distribution would need to undergo to match another.
5. **CC** (Correlation Coefficient): Also called Pearson's Linear Coefficient between two saliency maps.
6. **AUC** (Area Under Curve): The saliency map is treated as a binary classifier to separate positive from negative samples at various thresholds. We used three different types of AUC, which are AUC_Judd [3], AUC_Borji [4], and AUC_shuff (sAUC) [5].

Chapter 3. Methodology

3.1 Models Implementation

Attention models can be classified depending on the so-called 'growth' of the study in the past. Where early models of saliency were primarily used for identifying only conspicuous regions (this was regardless of where people looked), nowadays, the saliency models have been designed to predict eye movements. This growth in the science of attention modeling can be described in the form of a few stages.

3.1.1 Pre Deep Learning Age

Stage 1: It is considered to be the earliest form of computational modeling work (Koch and Ullman [6]). They introduced the concept of a saliency map derived from bottom-up inputs from all feature maps. That means that the image would be broken down into individual feature maps, for example, based on contrast layers, color, intensity, etc. and then a computational network would select the most salient one of these features for processing.

Stage 2: Models in the second stage made it possible to specifically modify models from Stage 1 to make them respond to a variety of stimuli. (Itti et. al. [7]). This particular model led to a surge in further research and study of building more comprehensive and complex models to identify saliency in attention models.

Stage 3: Largely taking place in the years between 1999 to 2013, mathematical functionality being introduced into models became more popular. Lots of computational attention models were proposed in this period. The models in this period can be categorized into different types, such as graphical models or Bayesian models. This growth in the science of visual saliency leads to a rise in the use of convolutional neural networks (CNNs) in attention models.

3.1.2 Deep Learning Age

Stage 4: The most recent and successful saliency models are entirely due to the success of deep learning neural networks. In the study of robotics and computer vision, it has opened up new avenues to the design and implementation of highly effective saliency models. A new wave of saliency models has emerged with the resurgence of CNNs.

Deep learning models rely on layers of modules that take in non-linear input and pass it into subsequent layers, where feature maps will be generated at each layer, to receive a specific output. However, these models have to be trained to do so. Hence the term 'learning' has become popular among neuroscientists and computational researchers. CNNs follow the same architecture as mentioned above, in which each layer use convolution operation to generate output feature maps from the input. Through learning from layers within layers of their architecture, the success of these models on a large scale has brought a plethora of newer and more efficient saliency models that indubitably perform better than the traditional models as seen in the pre-deep learning age.

The models we use in our project are shown in **Fig. 2**.

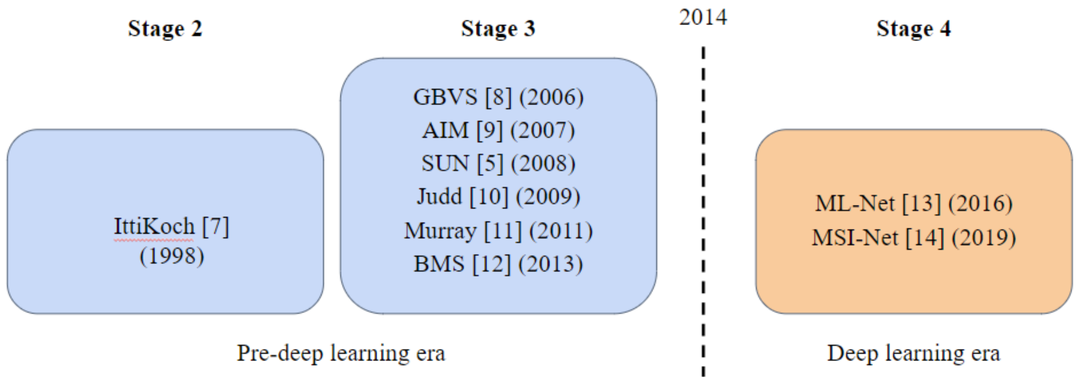


Fig. 2: The models that were implemented in this project, which we selected from stage 2 to 4 of the development of attention models.

3.2 Deep Learning Model Modification

To achieve our second objective, which is to improve the attention model with deep learning-based computer vision techniques, we applied the following modifications to MSI-Net [14], which is a CNN-based attention model. We chose MSI-Net because it performed the best among the models we implemented, which was shown in **Chapter 5**. In brief, MSI-Net composed of an encoder part, which encodes the input image into multi-channel feature maps, and a decoder part, in which the feature maps are decoded to generate the saliency map of the input image.

3.2.1 Self-attention Block

Zhang et. al. [15] proposed the self-attention (SA) block for CNNs, which is adopted from the non-local model [16], to efficiently model relationships between widely separated spatial regions of an image. In this project, we implemented SA blocks in three different layers of MSI-Net, which the layers are all before the max-pooling operations. (**Fig. 3**)

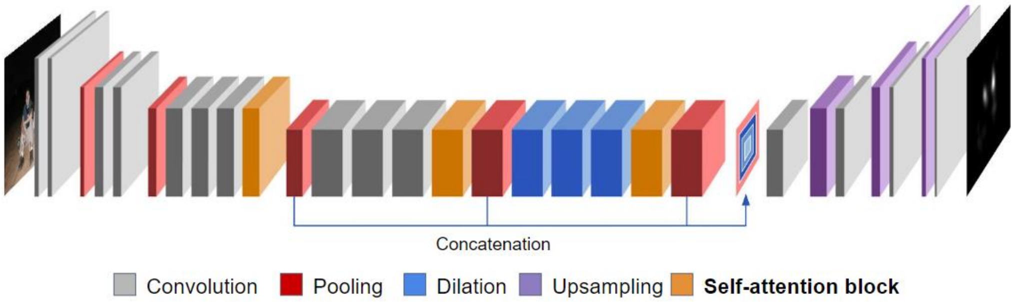


Fig. 3: MSI-Net with three self-attention blocks.

3.2.2 Features Fusion

In MSI-Net, three feature maps from different layers are concatenated and fed to the Atrous Spatial Pyramid Pooling (ASPP) layer [17]. To fuse the information of these feature maps before the ASPP operation, we implemented a convolutional layer with a kernel size of 1x1 after the feature maps are concatenated (**Fig. 4**). The depth of this convolutional layer is set to 1280, which is the sum of the three concatenated feature maps.

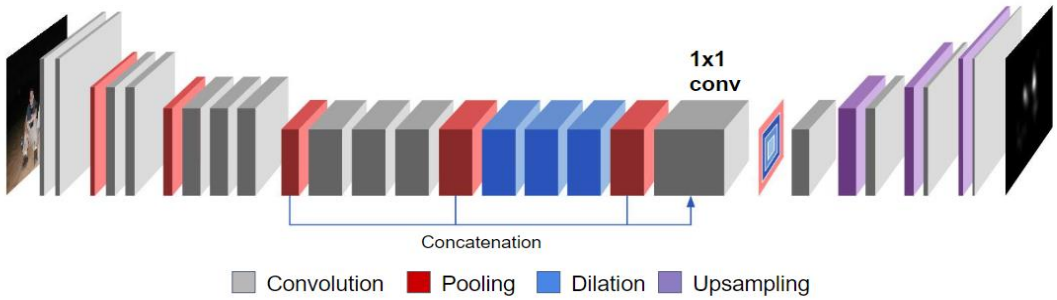


Fig. 4: MSI-Net with a features fusion layer, which is labeled as “1x1 conv” in the figure.

3.2.3 Skip Connections

Skip connections as in UNet [18] were implemented, in which the feature maps from the encoder part are concatenated to the decoder feature maps with the same size. Skip connections help to deliver information that was captured in the initial layers and was required for reconstruction in the decoder part. The skip connections are implemented in three different layers (**Fig. 5**)

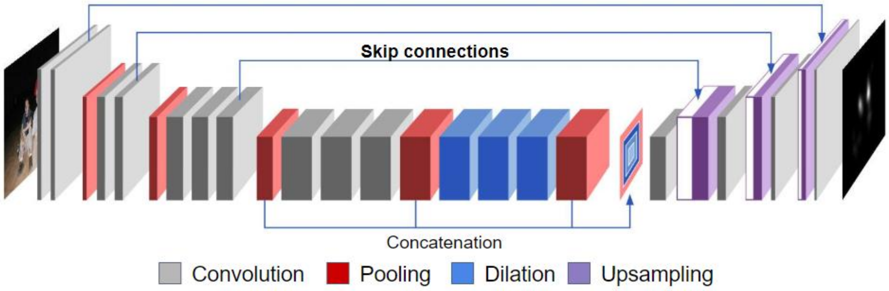


Fig. 5: MSI-Net with three skip connections.

3.3 Metrics

The 7 metrics we used, which were mentioned in **Chapter 2.4**, can be categorized as shown in **Table. 1**. These metrics were classified according to their properties. Distribution based metrics considered ground truth fixation maps and saliency maps as continuous distributions. Location-based metrics interpreted saliency maps as discrete fixation locations. Good saliency models should have high scores for similarity metrics and low scores for dissimilarity metrics.

Metrics	Distribution-based	Location-based
Similarity	SIM, CC	AUC, sAUC, NSS
Dissimilarity	EMD, KL	

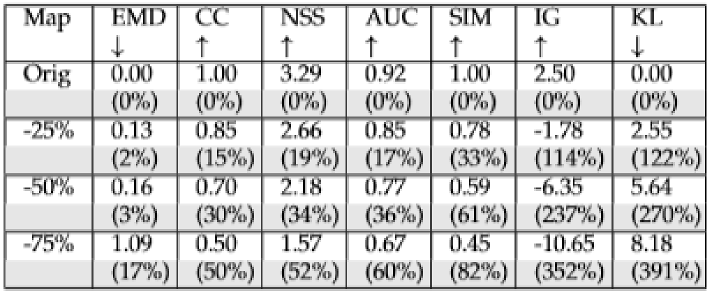
Table. 1: The classification of metrics based on their properties.

3.3.1 Metrics Behavior

In order to have a better prediction, saliency models would enhance false negatives to have a better performance. However, every metric had different sensitivity to false negatives. Based on Bylinskii at. al. [2], to analyze which metric penalized false negatives, different amounts of salient pixels were removed from the ground truth fixation map. The pixel would be uniformly and randomly selected and set to zero if its saliency value was above the mean map value. Then compared the original ground truth map with the resulting map to measure the difference in scores with 25%, 50% and 75% false negatives. The results of each metrics were presented in **Table. 2**. KL and SIM were most sensitive to false negatives. AUC ignores low-values false positives. NSS and CC were equally affected by false positives and negatives.

Center-bias also played an important role in saliency models. Because most images had a higher density of fixations in the center of the image compared to the periphery, this feature greatly enhanced the efficiency of saliency models. However, Bylinskii [2] stated that sAUC metric limited many images correspond to sampling negatives from a central Gaussian. For an image with a strong center bias, both positive and false negatives would be sampled from the same region, and a correct prediction would be unpredictable. Therefore, we had to consider that sAUC metric would penalize models that include center bias. On the other hand, EMD spatially hedged its bets.

If an image was fixated in multiple spots, instead of capturing a subset of the fixated locations, EMD would predict them in spatially distance.



Map	EMD ↓	CC ↑	NSS ↑	AUC ↑	SIM ↑	IG ↑	KL ↓
Orig	0.00 (0%)	1.00 (0%)	3.29 (0%)	0.92 (0%)	1.00 (0%)	2.50 (0%)	0.00 (0%)
-25%	0.13 (2%)	0.85 (15%)	2.66 (19%)	0.85 (17%)	0.78 (33%)	-1.78 (114%)	2.55 (122%)
-50%	0.16 (3%)	0.70 (30%)	2.18 (34%)	0.77 (36%)	0.59 (61%)	-6.35 (237%)	5.64 (270%)
-75%	1.09 (17%)	0.50 (50%)	1.57 (52%)	0.67 (60%)	0.45 (82%)	-10.65 (352%)	8.18 (391%)

Table. 2: The relation between metrics and sensitivity of false negatives [2].

3.3.2 Metrics Selection

Based on **Chapter 3.3.1**, knowing these metrics properties was important for evaluating a model. In Bylinskii [2], the behaviors of inputs would affect metrics differently: whether center bias was applied; how ground truth was evaluated; whether the input model was probabilistic; if spatial deviations occur between the fixation maps and predicting results. Besides, the cost, local, differentiable of metric computation would also influence whether a metric is suitable for model improvement. Here we concluded that for probabilistic models, the KL metric was a good method. Assertion for why it might be a better option to analyze and define probabilistic models could be found in [19], [20]. For image retargeting, compression and progressive, metrics NSS and SIM can be used. To have a fair comparison, metrics NSS and CC were the preferable decision. Both metrics made a limited assumption for input formats. Besides, they treat false positive and false negative equally.

3.4 Data Sets

In this project, we used the SALICON [21] and MIT1003 [22] datasets for training and evaluating the implemented attention models. Both datasets consist of natural images, which are used as the inputs of the attention models. For each image, fixation points are generated by using an eye-tracker to track where humans looked, and a fixation map is generated by blurring those fixation points (**Fig. 6**). Both fixation points and fixation maps are used as ground truth when calculating the metrics.

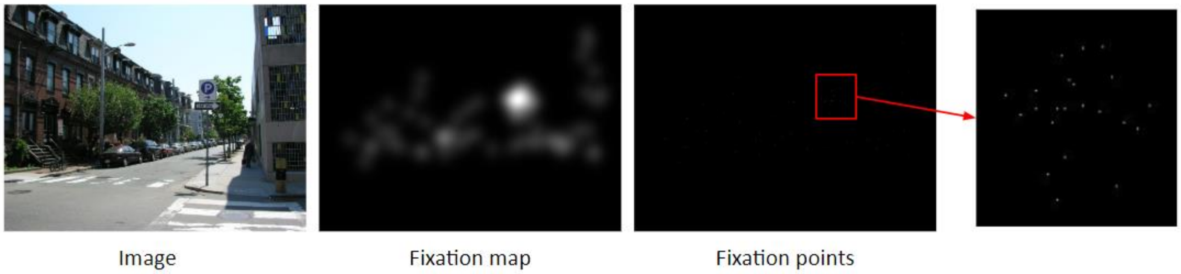


Fig. 6: The datasets we used consisted of images (left) as the inputs of models. Fixation points and fixation maps were used as the ground truth visual saliency.

3.4.1 Training - SALICON

For training the deep learning models, we used the Saliency in Context (SALICON) dataset. The dataset contains 15000 images, which we used 10000 images for training and 500 for validation. The fixation maps are used as the ground truth for training.

3.4.2 Evaluation - MIT1003

The dataset contains eye-tracking data of 15 viewers on 1003 images and it is a challenging dataset for saliency models, as images are highly varied and natural. In our project, we used this dataset as testing data to evaluate the performance of models.

To achieve our third objective, which is to analyze the model performance on different categories of images, we divided the images into the following categories:

1. **Objects**: Images having main subjects that are easily identified, which can be further divided into the categories of human, animal, text, gaze or action, and building (**Fig. 7**). Note that the gaze or action category is images with humans looking at or interacting with other humans or objects.
2. **Scenes**: Images having no main subject and contain more complex information, which can be further divided into the categories of street view, landscape, and indoor (**Fig. 8**).

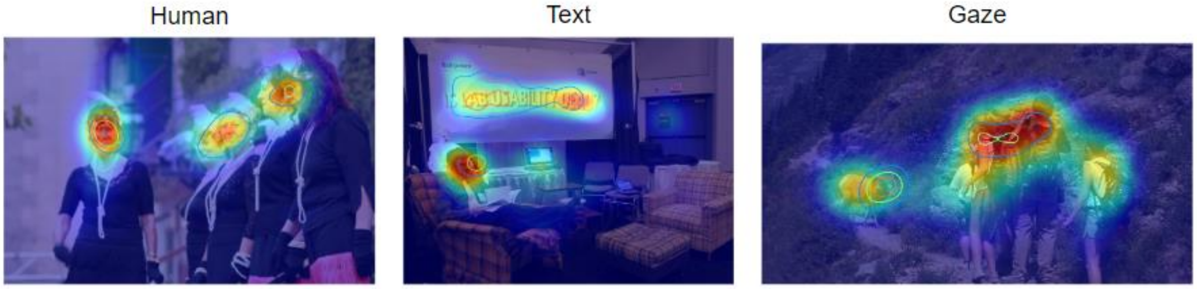


Fig. 7: Examples of the categories of objects. Both the "saliency maps" generated by an attention model and the ground truth "fixation maps" are overlapped with the images. The continuous color maps are the saliency map and the contours are the fixation map.

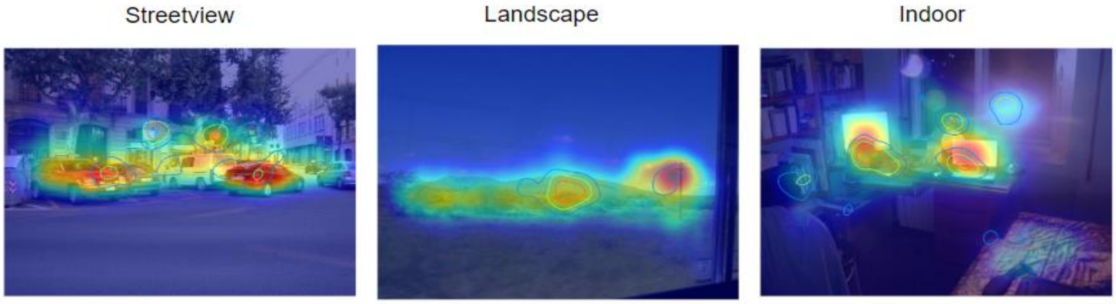


Fig. 8: Examples of the categories of scenes.

Chapter 4. Experiment Setup

4.1 Models Implementation

For non-deep learning models, we used Matlab for implementation. Each of the models generates saliency maps, which are predictions of where humans look at, for each image in MIT1003 dataset. These saliency maps are then used to compute the scores of the metrics.

We implemented deep learning models, ML-Net and MSI-Net, with Tensorflow. Both models were trained and validated with SALICON dataset, and the trained models were then used to generate saliency maps of MIT1003 dataset. For hyper-parameters, we used KL divergence as loss function, Adam optimizer, a batch size of 8, and trained the model with 10 epochs.

4.2 Metrics Computation

The flowchart (**Fig. 9**) below explains how we evaluated a model's performance with an input image. After getting a saliency map from the attention model and the associated fixation map and fixation points, we used them as input for metrics to get the scores.

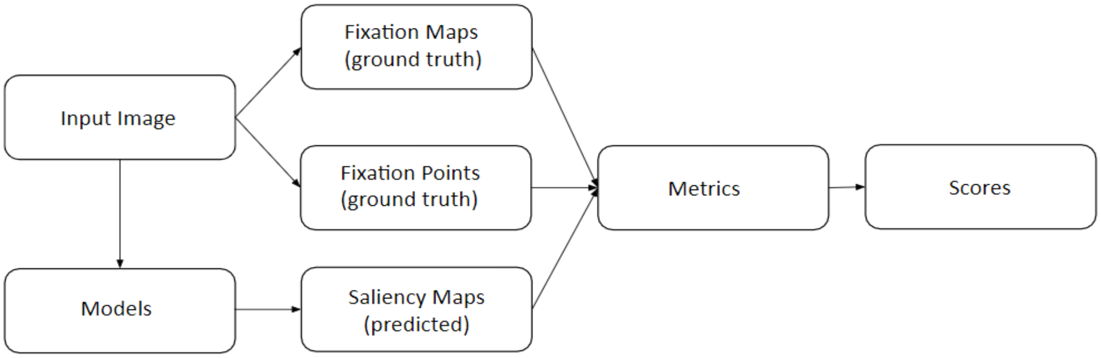
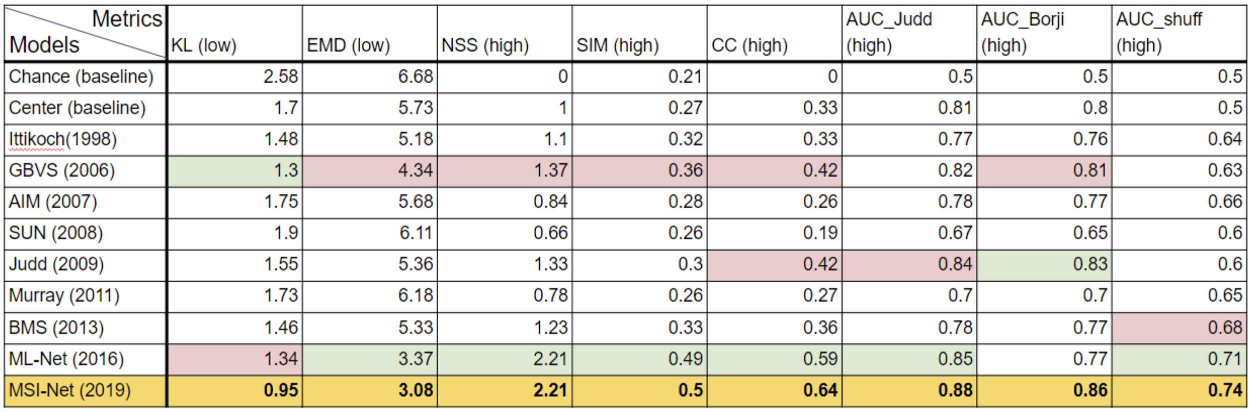


Fig 9: The procedure of evaluating models with metrics.

Chapter 5. Results

5.1 Model Comparisons - Quantitative

In our project, we simulated all models with different metrics. **Table. 3** was the quantitative results of each model. The top 3 best scoring maps were highlighted. Color yellow represented the best score. Light green and light red were second and third place respectively. The GBVS model, ML-Net, and MSI-Net had better results than other models. We can observe that the deep learning models performed better than non-deep learning models. This can be explained by the trainable features in deep learning help models to better "recognize" the objects in the images, and the recognition of objects is essential for predicting saliency.



Metrics Models	KL (low)	EMD (low)	NSS (high)	SIM (high)	CC (high)	AUC_Judd (high)	AUC_Borji (high)	AUC_shuff (high)
Chance (baseline)	2.58	6.68	0	0.21	0	0.5	0.5	0.5
Center (baseline)	1.7	5.73	1	0.27	0.33	0.81	0.8	0.5
Ittikoch(1998)	1.48	5.18	1.1	0.32	0.33	0.77	0.76	0.64
GBVS (2006)	1.3	4.34	1.37	0.36	0.42	0.82	0.81	0.63
AIM (2007)	1.75	5.68	0.84	0.28	0.26	0.78	0.77	0.66
SUN (2008)	1.9	6.11	0.66	0.26	0.19	0.67	0.65	0.6
Judd (2009)	1.55	5.36	1.33	0.3	0.42	0.84	0.83	0.6
Murray (2011)	1.73	6.18	0.78	0.26	0.27	0.7	0.7	0.65
BMS (2013)	1.46	5.33	1.23	0.33	0.36	0.78	0.77	0.68
ML-Net (2016)	1.34	3.37	2.21	0.49	0.59	0.85	0.77	0.71
MSI-Net (2019)	**0.95**	**3.08**	**2.21**	**0.5**	**0.64**	**0.88**	**0.86**	**0.74**

Table. 3: Quantitative results of different models in this project.

In **Fig. 10**, we showed that with given images, how the saliency maps looked like when we applied different models. Also, we could notice that the results of ML-Net and MSI-Net are highly close to fixation maps which meant they precisely predicted human attention.

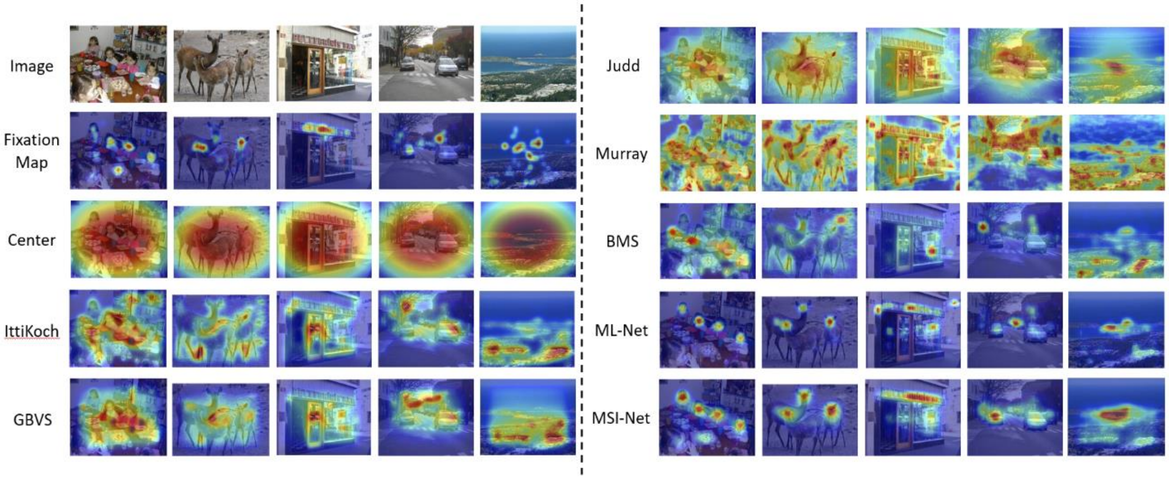


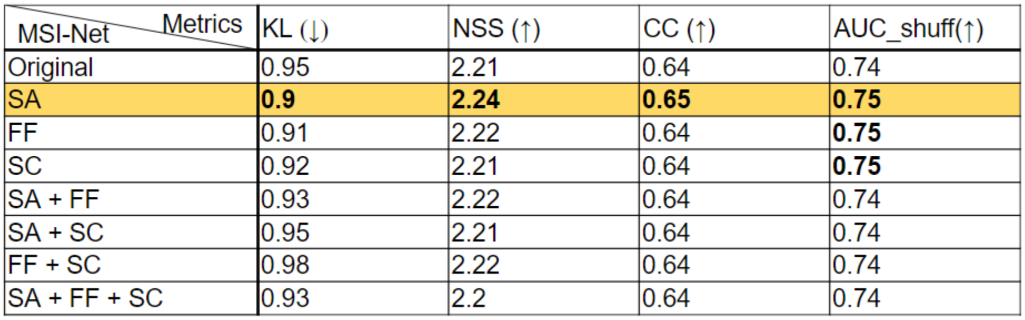
Fig. 10: Qualitative results of models. ML-Net & MSI-Net exhibit accurate saliency prediction.

5.2 Deep Learning Attention Model Improvement

With MSI-Net being the best model among the models we implemented, we applied the model modifications mentioned in **Chapter 3.2** to MSI-Net. **Table. 4** demonstrated the performance of the modified MSI-Net. We selected the four metrics based on the discussion of **Chapter 3.3**.

As shown in **Table. 4**, applying the proposed modifications improved the model performance, and applying only the SA blocks gave the best results. However, these modifications only lead to a slight improvement. The possible reason is that these techniques are used to help improve models' ability to "recognize" the objects, e.g. humans or animals in an image. Although recognizing objects is essential for saliency detection, existing deep learning models have already performed well on this task. Unlike object detection or instance segmentation tasks, saliency detection doesn't require models to recognize objects with very high accuracy. Therefore, applying techniques to further improve the models' ability to recognize objects doesn't effectively improve the performance. In the next section, we identified the directions to further improve the attention models.

Another thing to point out from **Table. 4** is that the model performs better when applying a single modification. A possible explanation is that the increase of model complexity causes overfitting, in which the model performs better on training data but worse on testing data.



Metrics MSI-Net	KL (↓)	NSS (↑)	CC (↑)	AUC_shuff(↑)
Original	0.95	2.21	0.64	0.74
SA	**0.9**	**2.24**	**0.65**	**0.75**
FF	0.91	2.22	0.64	**0.75**
SC	0.92	2.21	0.64	**0.75**
SA + FF	0.93	2.22	0.64	0.74
SA + SC	0.95	2.21	0.64	0.74
FF + SC	0.98	2.22	0.64	0.74
SA + FF + SC	0.93	2.2	0.64	0.74

Table. 4: Quantitative results of the modified MSI-Net.

5.3 Image Categories Analyzation

In this section, we used MSI-Net with SA blocks to generate the saliency maps for different image categories mentioned in **Chapter 3.4** and analyze the outcomes. **Fig. 11** showed the model performance on each image category.

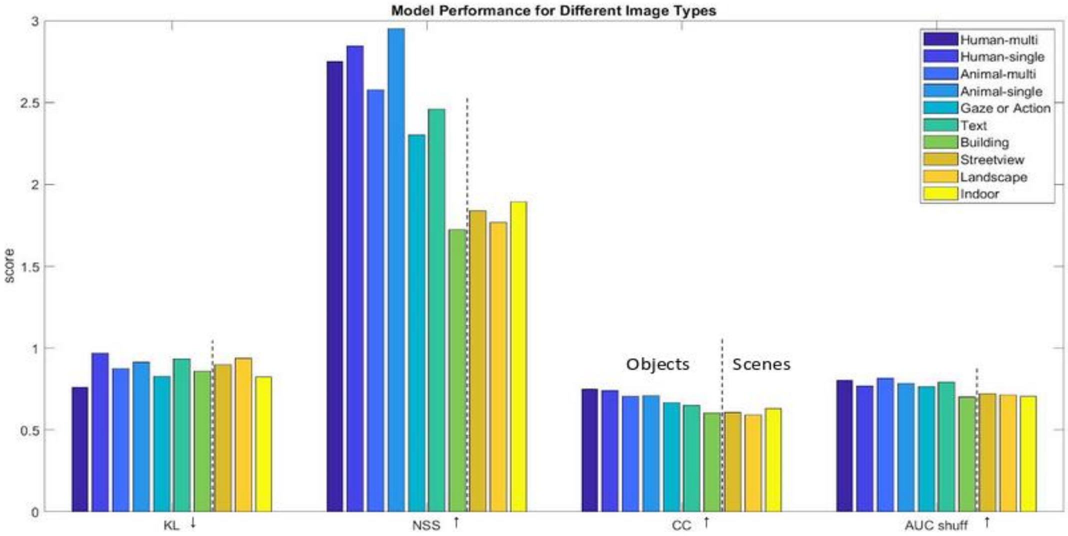
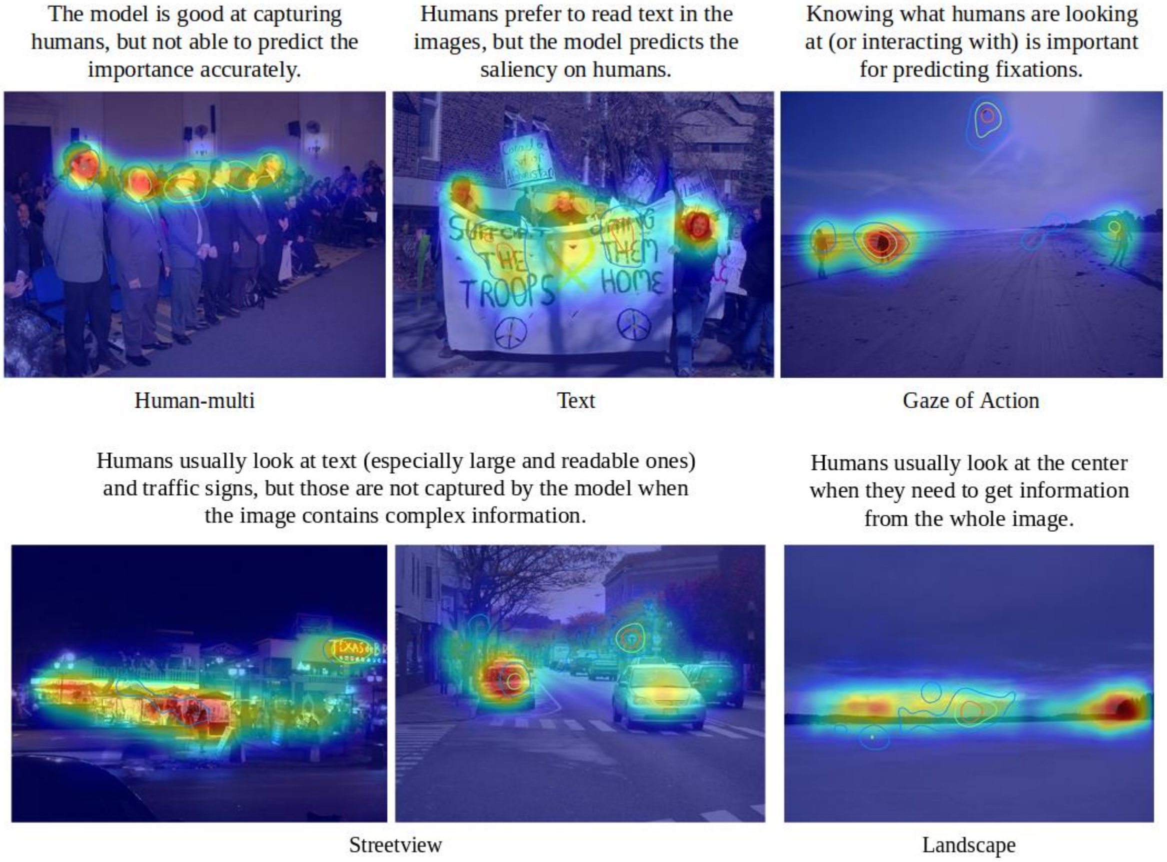


Fig. 11: Model performance on different categories of images.

Overall, the model performs better on images with objects than images of scenes. This makes sense as images of scenes contain more complex information such that humans need to have a higher understanding of the image, rather than just recognizing objects, to know where to look at. Among the images with objects, the model performs best on the "human" and "animal" categories. On the other hand, images of the landscape are the category that the model has the worst performance.

To identify what is still needed to improve the attention model, we analyzed the saliency maps of each category and extracted the cases that the model generated erroneous outcomes. **Fig. 12** demonstrates some examples of the extracted cases. In summary, to further improve the attention models, they are required to have higher understandings of the images listed below:

1. Understanding the interactions between humans and/or objects in an image.
2. Reasoning the relative importance of humans, objects, or background.
3. Understanding the text in an image.



The model is good at capturing humans, but not able to predict the importance accurately.

Humans prefer to read text in the images, but the model predicts the saliency on humans.

Knowing what humans are looking at (or interacting with) is important for predicting fixations.

Human-multi

Text

Gaze of Action

Humans usually look at text (especially large and readable ones) and traffic signs, but those are not captured by the model when the image contains complex information.

Humans usually look at the center when they need to get information from the whole image.

Streetview

Landscape

Fig. 12: Cases that attention model generates erroneous saliency maps and possible explanations. The meaning of color map and contours was explained in **Fig. 7**.

Chapter 6: Conclusions

In this project, we reviewed and implemented several attention models for saliency prediction. The implemented models expanded from the very early stage of attention models [7] to the recent deep learning-based models [14]. To quantitatively compare the models, we reviewed and implemented the metrics used for saliency prediction. We concluded that with the trainable features that help models to better recognize the objects in the images, deep learning models performed much better than the traditional non-deep learning models.

To improve the saliency prediction, we implemented three modifications, self-attention, features fusion, and skip connections, on MSI-Net, a deep learning attention model. These techniques were already proven to improve the performance of deep learning models on several computer vision tasks, such as object detection. However, the modifications only slightly improved the saliency detection. This indicated that further improving on the models' ability to recognize objects is not effective to design a better attention model. Therefore, other directions for improvement need to be explored.

We divided the images into different categories to identify the cases that the attention model was not performing well. We found that the attention model was struggling on predicting the saliency maps when the higher-level understandings of the images is required. We concluded that the higher-level understandings include understanding the interactions of subjects, the relative importance of subjects, and the text. To train the models on these understandings, image captioning or reinforcement learning models can be explored for visual saliency prediction. In addition, we can also take temporal information into consideration, as humans continuously gazing at different locations within an image. This temporal visual attention can be explored by utilizing recurrent neural networks (RNNs).

References

[1] Itti, Laurent, and Ali Borji. "Computational models: Bottom-up and top-down aspects." *arXiv preprint arXiv:1510.07748* (2015).

[2] Bylinskii, Zoya, et al. "What do different evaluation metrics tell us about saliency models?." *IEEE transactions on pattern analysis and machine intelligence* 41.3 (2018): 740-757.

[3] Riche, Nicolas, et al. "Saliency and human fixations: State-of-the-art and study of comparison metrics." *Proceedings of the IEEE international conference on computer vision*. 2013.

[4] Borji, Ali, et al. "Analysis of scores, datasets, and models in visual saliency prediction." *Proceedings of the IEEE international conference on computer vision*. 2013.

[5] Zhang, Lingyun, et al. "SUN: A Bayesian framework for saliency using natural statistics." *Journal of Vision* 8.7 (2008): 32-32.

[6] Koch, Christof, and Shimon Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry." *Matters of intelligence*. Springer, Dordrecht, 1987. 115-141.

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

[8] Harel, Jonathan, Christof Koch, and Pietro Perona. "Graph-based visual saliency." *Advances in neural information processing systems*. 2007.

[9] Bruce, Neil, and John Tsotsos. "Attention based on information maximization." *Journal of Vision* 7.9 (2007): 950-950.

[10] Judd, Tilke, et al. "Learning to predict where humans look." *2009 IEEE 12th international conference on computer vision*. IEEE, 2009.

[11] Murray, Naila, et al. "Saliency estimation using a non-parametric low-level vision model." *CVPR 2011*. IEEE, 2011.

[12] Zhang, Jianming, and Stan Sclaroff. "Saliency detection: A boolean map approach."

Proceedings of the IEEE international conference on computer vision. 2013.

[13] Cornia, Marcella, et al. "A deep multi-level network for saliency prediction." *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016.

[14] Kroner, Alexander, et al. "Contextual Encoder-Decoder Network for Visual Saliency Prediction." *arXiv preprint arXiv:1902.06634* (2019).

[15] Zhang, Han, et al. "Self-attention generative adversarial networks." *arXiv preprint arXiv:1805.08318* (2018).

[16] Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[17] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017): 834-848.

[18] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

[19] M. K¨ummerer, T. Wallis, and M. Bethge. "How close are we to understanding image-based saliency?" *arXiv preprint arXiv:*1409.7686, 2014.

[20] M. K¨ummerer, T. Wallis, and M. Bethge. "Information-theoretic model comparison unifies saliency metrics." PNAS, 112(52):16054–16059, 2015.

[21] Jiang, Ming, et al. "Salicon: Saliency in context." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[22] Judd, Tilke, et al. "Learning to predict where humans look." *2009 IEEE 12th international conference on computer vision*. IEEE, 2009.