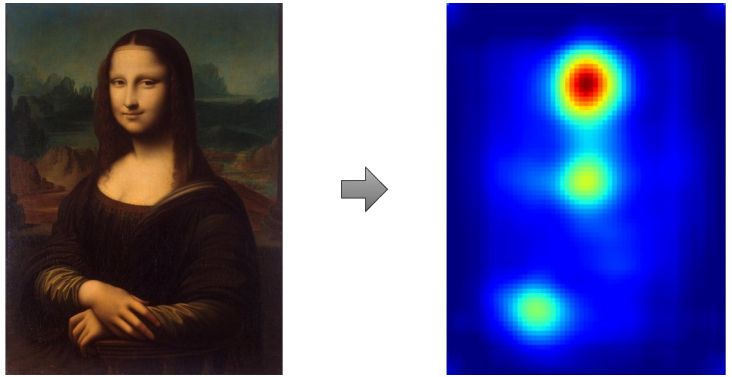
The Importance of Time in Visual Attention Models

Author: Marta Coll Pol
Advisors: Kevin Mc Guinness and Xavier Giró-i-Nieto

INTRODUCTION

Visual Attention Models

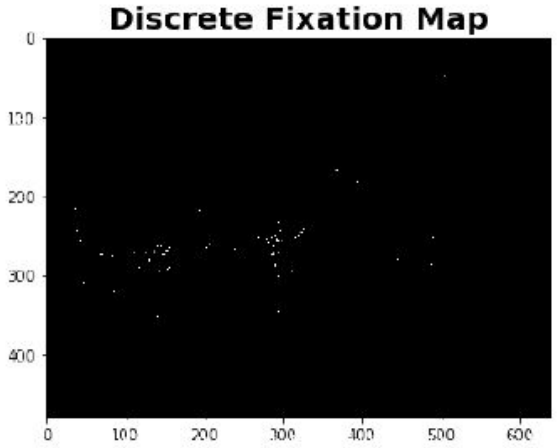


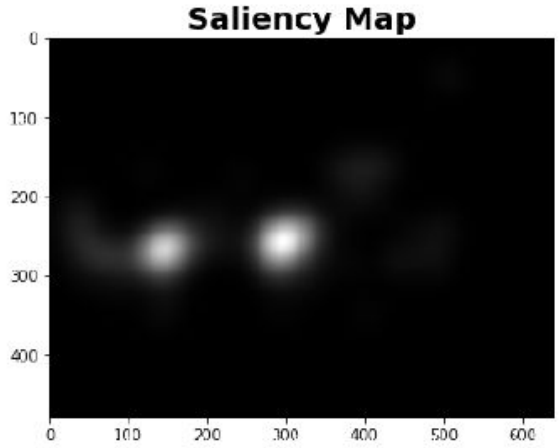
- **What is visual attention?**
  - Visual System Mechanism: allows humans to selectively process visual information.

- **What is a visual attention model?**
  - Predictive models that aim to predict regions that attract human attention.

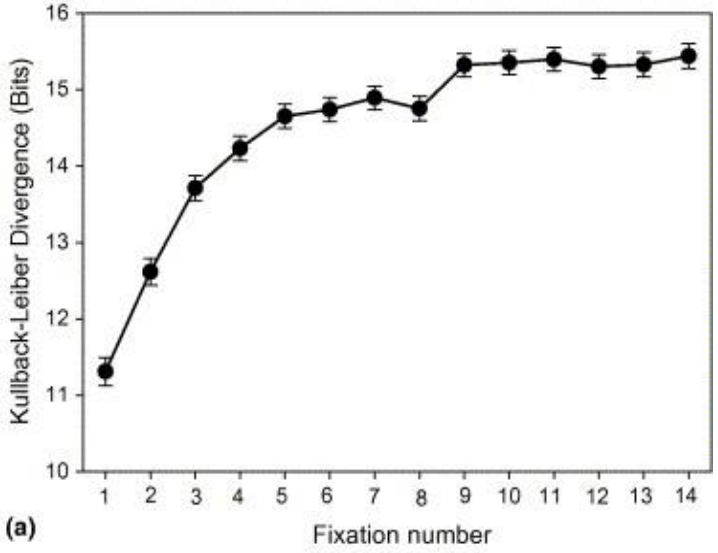
Source: https://algorithmia.com/algorithms/deeplearning/SalNet/docs

Saliency Information Representations





Motivation



Consistency in fixation selection between participants as a function of fixation number [1].

[1] Benjamin W Tatler, Roland J Baddeley, and Iain D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. Vision research, 45(5):643-659, 2005.

Motivation

- **Hypothesis I:** Visual Attention models have difficulties predicting later fixation points.

- **Hypothesis II:** Adding less weight to later fixations in the ground-truth maps used for training, could help improving visual attention models' performance.

DATASETS OF TEMPORALLY SORTED FIXATIONS

Datasets

- **iSUN:**
  - Data collection: eye-tracking [2].
- **SALICON:**
  - Data collection: mouse-tracking [3].

Saliency ground-truth data for one observer in a given image:

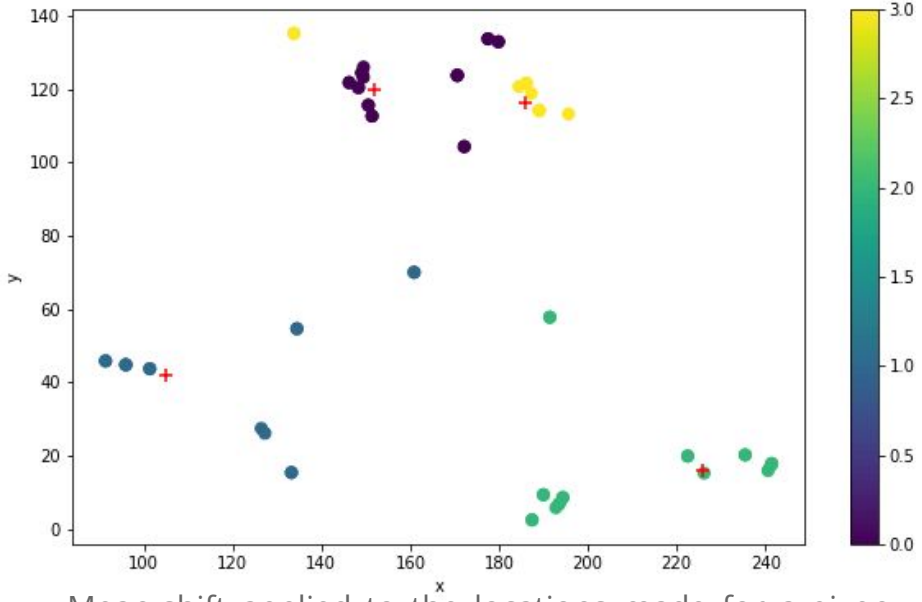
location	timestamp	fixation
88x2 dou...	*88x1 double*	[96.9873,1...
69x2 dou...	*69x1 double*	*7x2 double*
79x2 dou...	*79x1 double*	[166.6424,...

[2] Yinda Zhang, Fisher Yu, Shuran Song, Pingmei Xu, Ari Se, and Jianxiong Xiao. Large-scale scene understanding challenge: Eye tracking saliency estimation.
[3] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pages 1072-1080. IEEE, 2015.

Fixation points in order of visualization

iSUN:

- **Locations to Fixations:** Mean-shift.
- We used timestamps as a weight to retrieve fixation points in order.



Mean-shift applied to the locations made for a given image by an observer. Colored dots are locations assigned to different clusters. Red Crosses are the mean of each cluster (centroid), Fixation points.

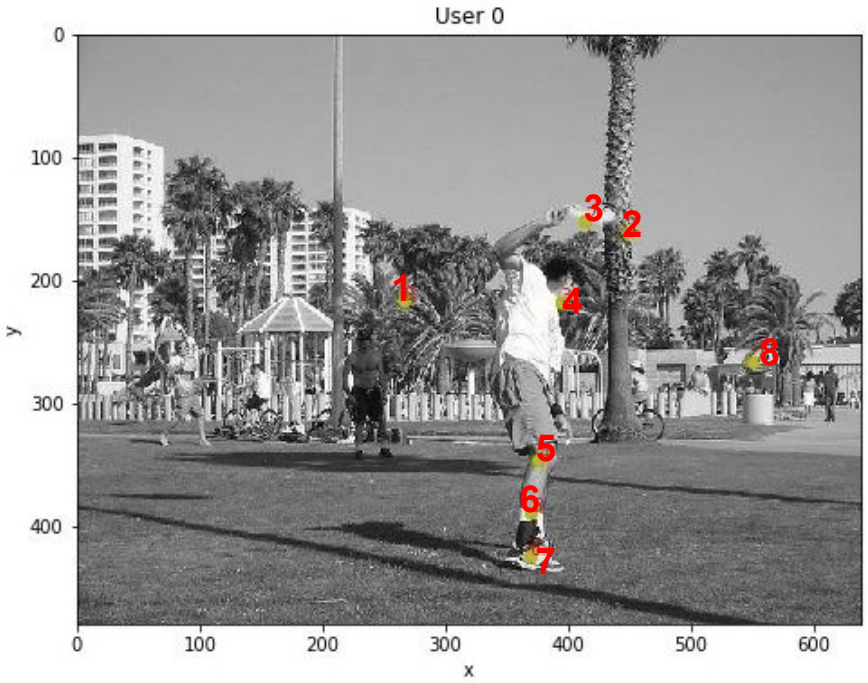
Fixation points in order of visualization

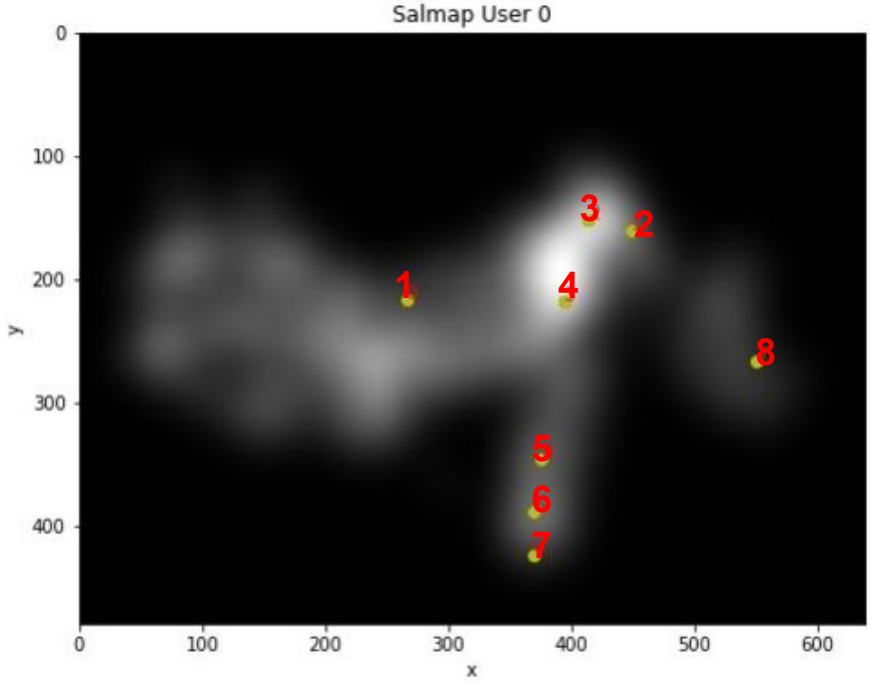
SALICON:

- **Locations to Fixations:** Samples with high mouse-moving velocity are excluded while keeping fixation points.

- An observation was made to see if fixation points are given in order of visualization in the dataset.

Fixation points in order of visualization : SALICON

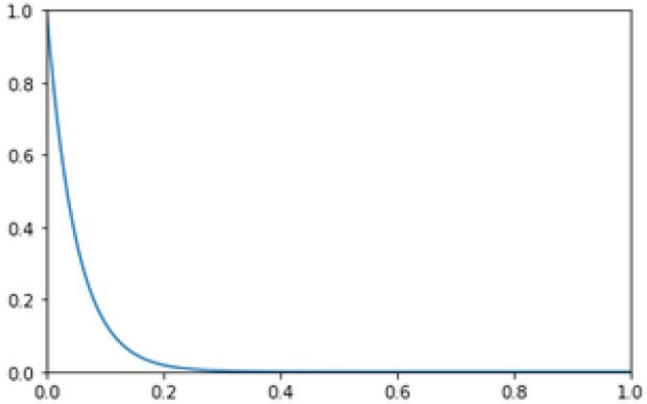




TEMPORALLY WEIGHTED SALIENCY

Temporally Weighted Saliency Maps (TWSM)

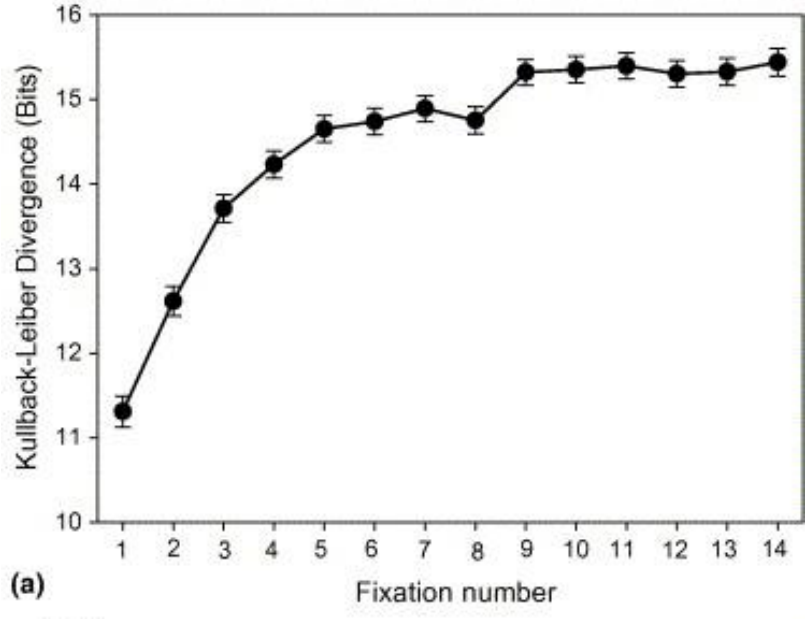
- Weighting in function of visualization order.
- Weighting function inspired in consistency graph.



Weighting function: $y = e^{-params \cdot x}$

TWSM : Choosing a weighting function parameter

Consistency in fixation selection between participants as a function of fixation number



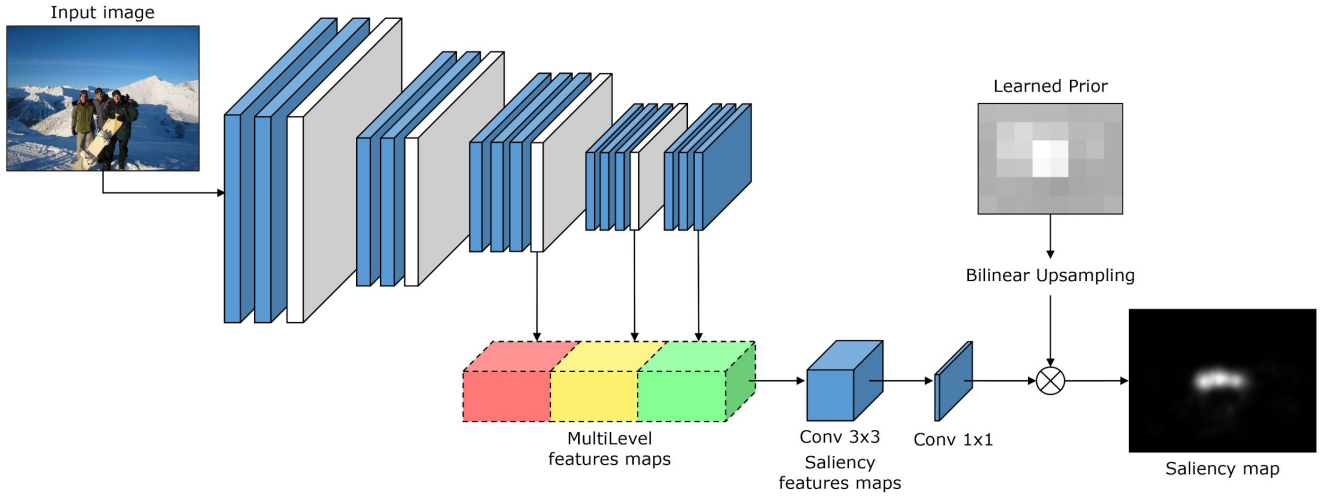
Saliency Prediction Metrics

- Area Under ROC Curve (AUC) Judd
- Kullback-Leibler Divergence (KLdiv)
- Pearson correlation coefficient (CC)

Model: MLNet

Architecture:

CNN **+** Encoding Network (produces a temporary saliency map) **+** Prior learning network **=** Final Saliency Map



[2] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. ICPR 2016.

Model: Different MLNet versions

Model	Training	Temporal Weighting
MLNet	Cornia et el [2]	× (nSM)
nMLNet (baseline)	Ours	× (nSM)
wMLNet	Ours	✓ (wSM)

[2] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. ICPR 2016.

Model: Replicating MLNet Results (nMLNet)

SALICON 2015 (Validation set)	**AUC Judd**
MLNet	**0.886**
nMLNet (baseline)	0.814

SALIENCY PREDICTION

Experiments

Experiment: Do visual attention models have difficulties predicting later fixations?

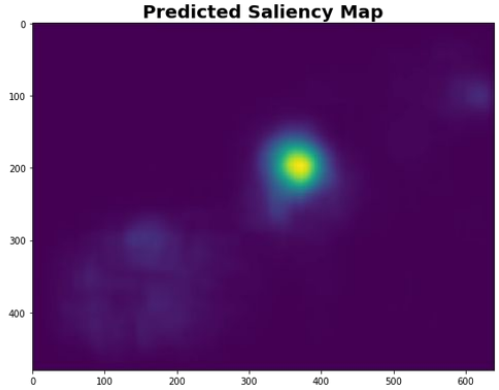
Hypothesis I: Visual attention models have difficulties predicting later fixations.

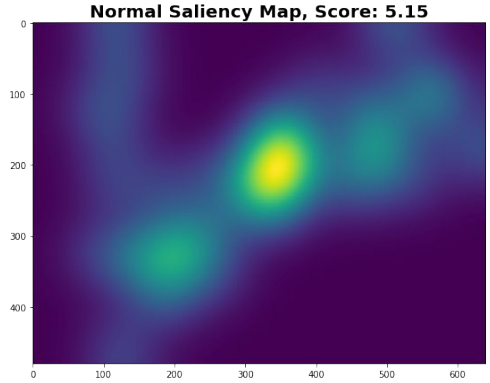
- Evaluate MLNet's predicted maps for the iSUN training set using two types of ground-truth: nSMs and wSMs.

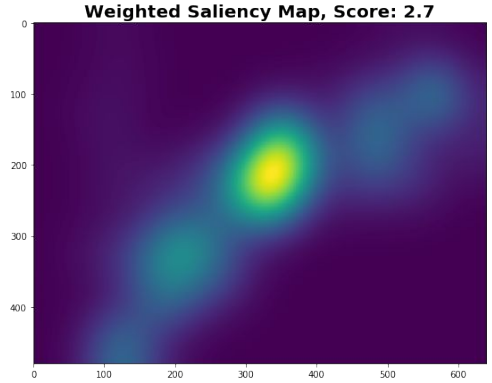
Results: Do visual attention models have difficulties predicting later fixations?

Maps that scored way better when evaluated using wSM:

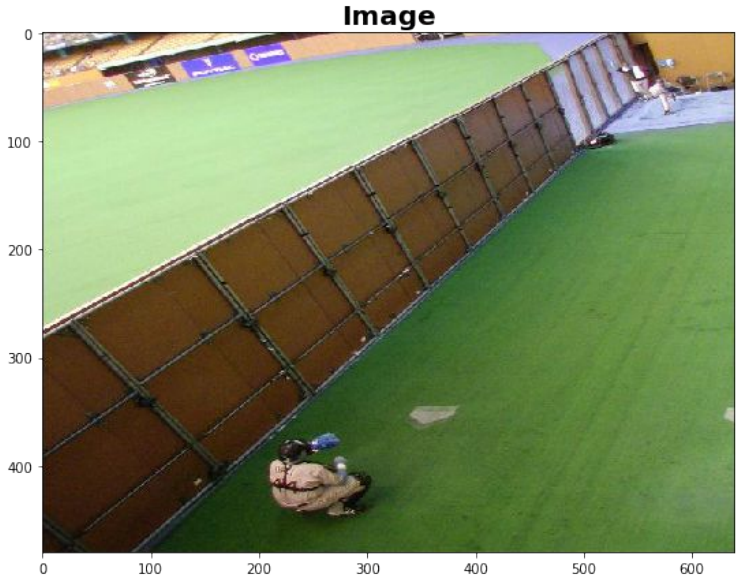
- Metric: KLdiv.

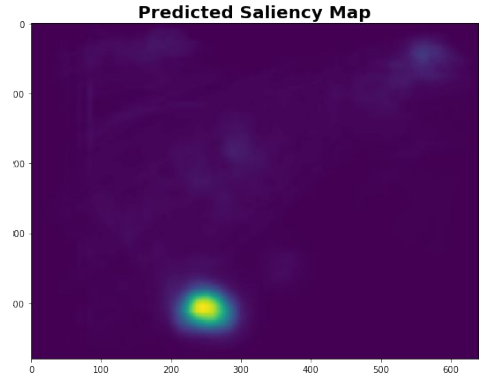


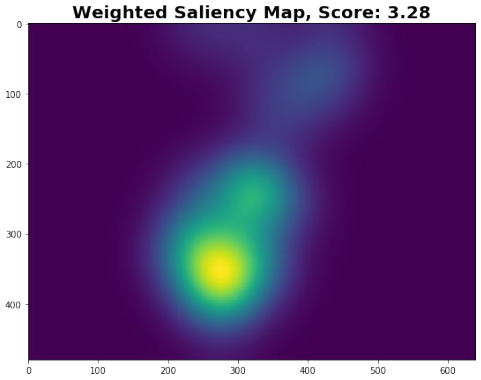




Results: Do visual attention models have difficulties predicting later fixations?



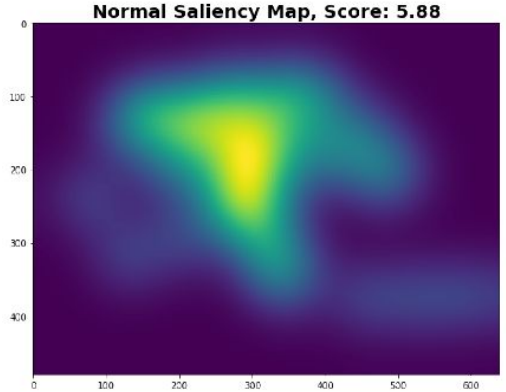


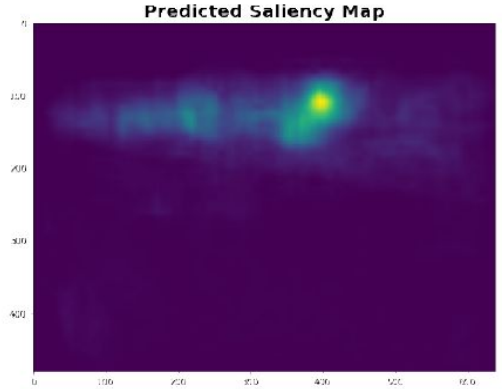


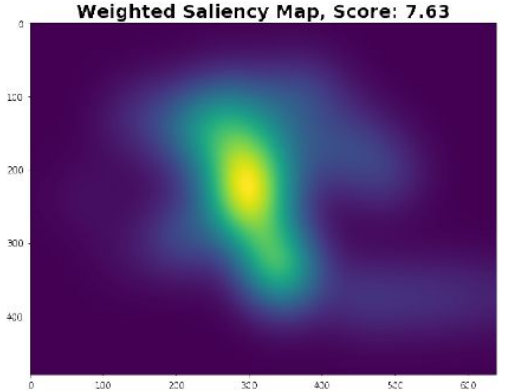
Results: Do visual attention models have difficulties predicting later fixations?

Maps that scored way better when evaluated using nSM:

- Metric: KLdiv
- Scores are in general bad for both maps, we could not extract further conclusions.







Experiment: Study on the effect on the use of Weighted Saliency Maps on a visual attention model's performance

Hypothesis II: Adding less weight to later fixations in the ground-truth maps used for training, can improve model's performance.

- **We trained** nMLNet and wMLNet, using the SALICON dataset. And **evaluated** their performance.

Results: Study on the effect on the use of Weighted Saliency Maps on a visual attention model's performance

AUC Judd ↑

Ground-truth	nMLNet	wMLNet
nSM	0.814	**0.816**

Kullback-Leibler Divergence (KLdiv) ↓

Ground-truth	nMLNet	wMLNet
nSM	1.332	**1.039**
wSM	1.493	**1.136**

Pearson's Correlation Coecient(CC) ↑

Ground-truth	nMLNet	wMLNet
nSM	0.534	**0.539**
wSM	0.509	**0.517**

VISUAL SEARCH

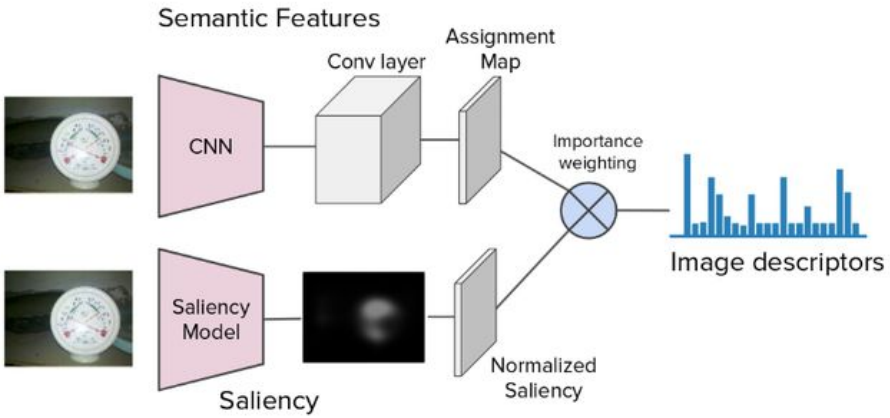
Experiments

SalBoW

- **SalBow** is a retrieval framework that uses saliency maps to weight the contribution of local convolutional representations for the instance search task.

- **The model is divided in two parts:**
  - CNN + K-means = Assignment Map (visual vocabulary)
  - Saliency prediction model

- **Model's output:** After weighting a Bag of Words is created to build the final image representation.



SalBow pipeline with saliency weighting [3].

[3] Eva Mohedano, Kevin McGuinness, Xavier Giro-i Nieto, and Noel E O'Connor. Saliency weighted convolutional features for instance search. arXiv preprint arXiv:1711.10795, 2017.

Results

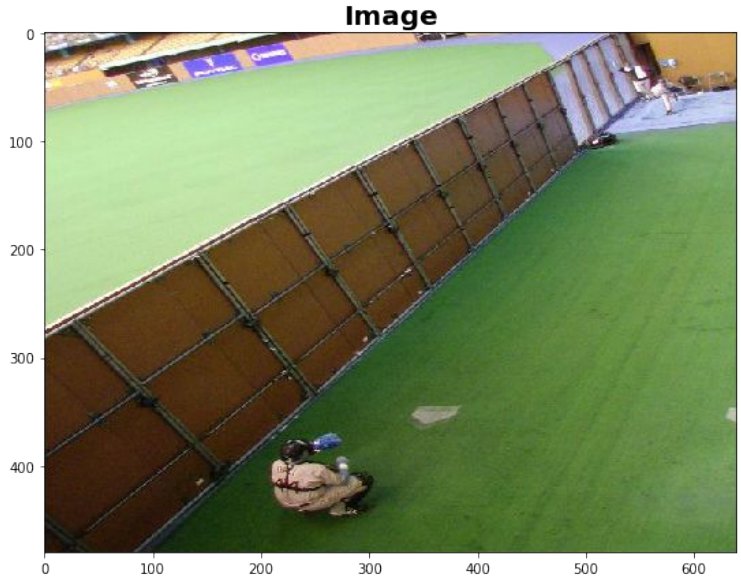
Mean Average Precision (mAP) ⬆

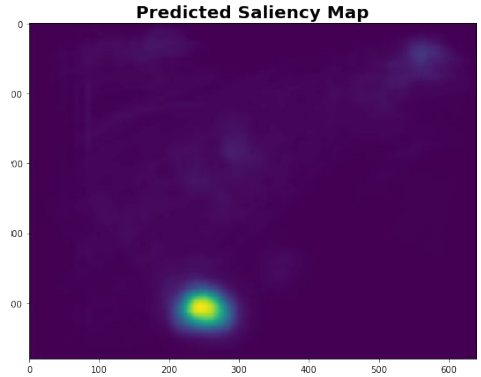
Saliency Maps type	mAP
nSM	0.6730
wSM	**0.6743**

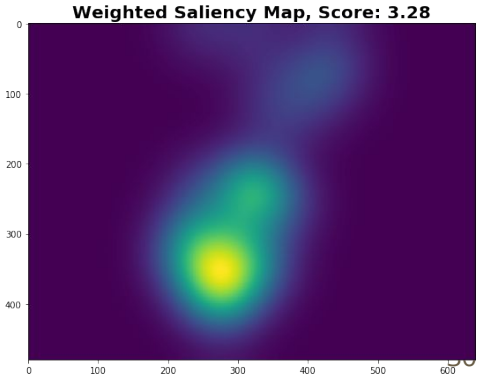
CONCLUSIONS

Conclusions

Experiment: *Do visual attention models have difficulties predicting later fixations?*







Conclusions

Experiment: Study on the effect on the use of Weighted Saliency Maps on a visual attention model's performance.

AUC Judd ↑

Ground-truth	nMLNet	wMLNet
nSM	0.814	**0.816**

Kullback-Leibler Divergence (KLdiv) ↓

Ground-truth	nMLNet	wMLNet
nSM	1.332	**1.039**
wSM	1.493	**1.136**

Pearson's Correlation Coecient(CC) ↑

Ground-truth	nMLNet	wMLNet
nSM	0.534	**0.539**
wSM	0.509	**0.517**

Conclusions

Experiment: Visual search task.

Mean Average Precision (mAP) ⬆

Saliency Maps type	mAP
nSM	0.6730
wSM	**0.6743**

Conclusions : Future Work

New evaluation metric:

- Take into account what visual attention models can predict.

Ways of improving Temporally Weighted Saliency Maps:

- Weighting in function of order and the time spend on each fixation point.
- Looking for those fixation points that are common between observers for each specific image.

Project's Code

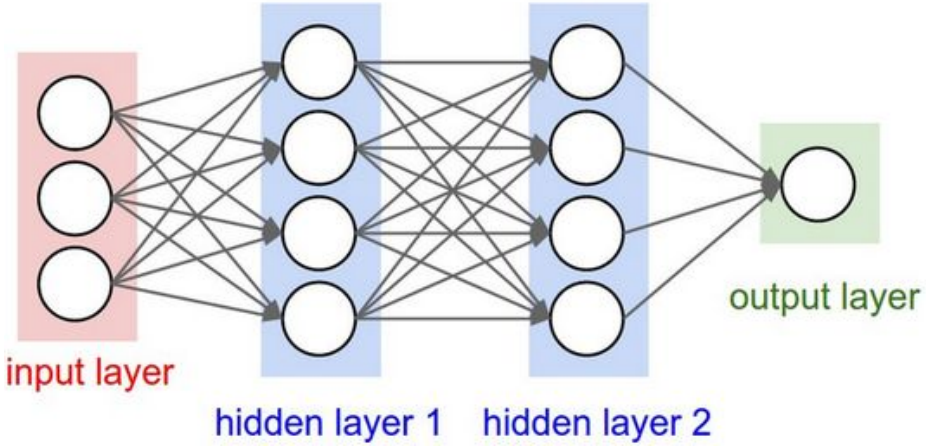
- The code for this project can be found in : https://github.com/imatge-upc/saliency-2018-timeweight
- The model and the not-owned code used in this project, is open-source.

Any Questions?

Quick Introduction to Deep Learning

Neural Networks:

- Forward propagation
- Loss function : metric to evaluate a model's performance.
- Back propagation
- The goal is to minimize the loss function.

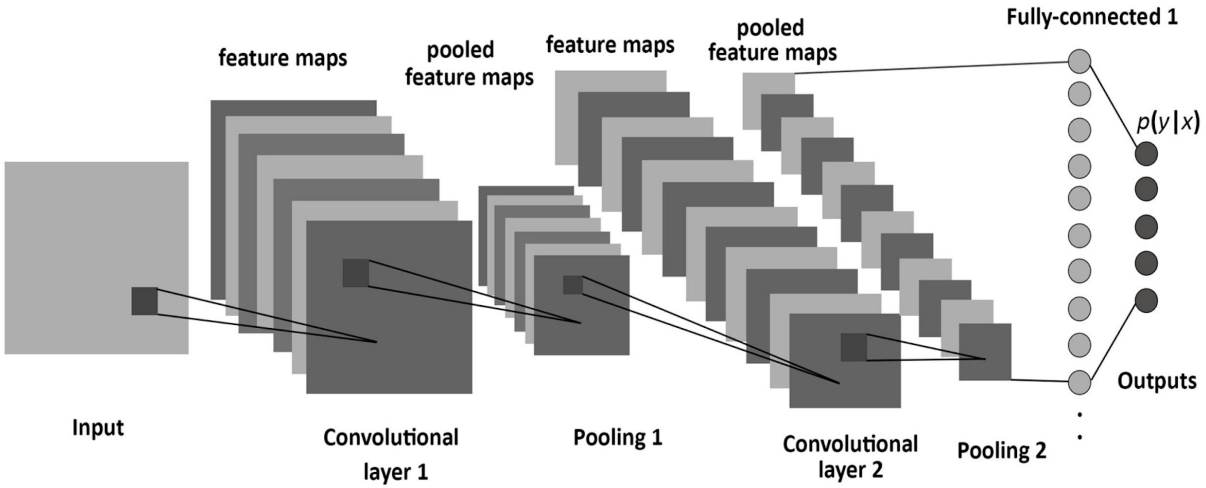


Three-layer neural network (two hidden layers of four neurons each and a single output), with three inputs.

Source: http://cs231n.github.io/neural-networks-1/

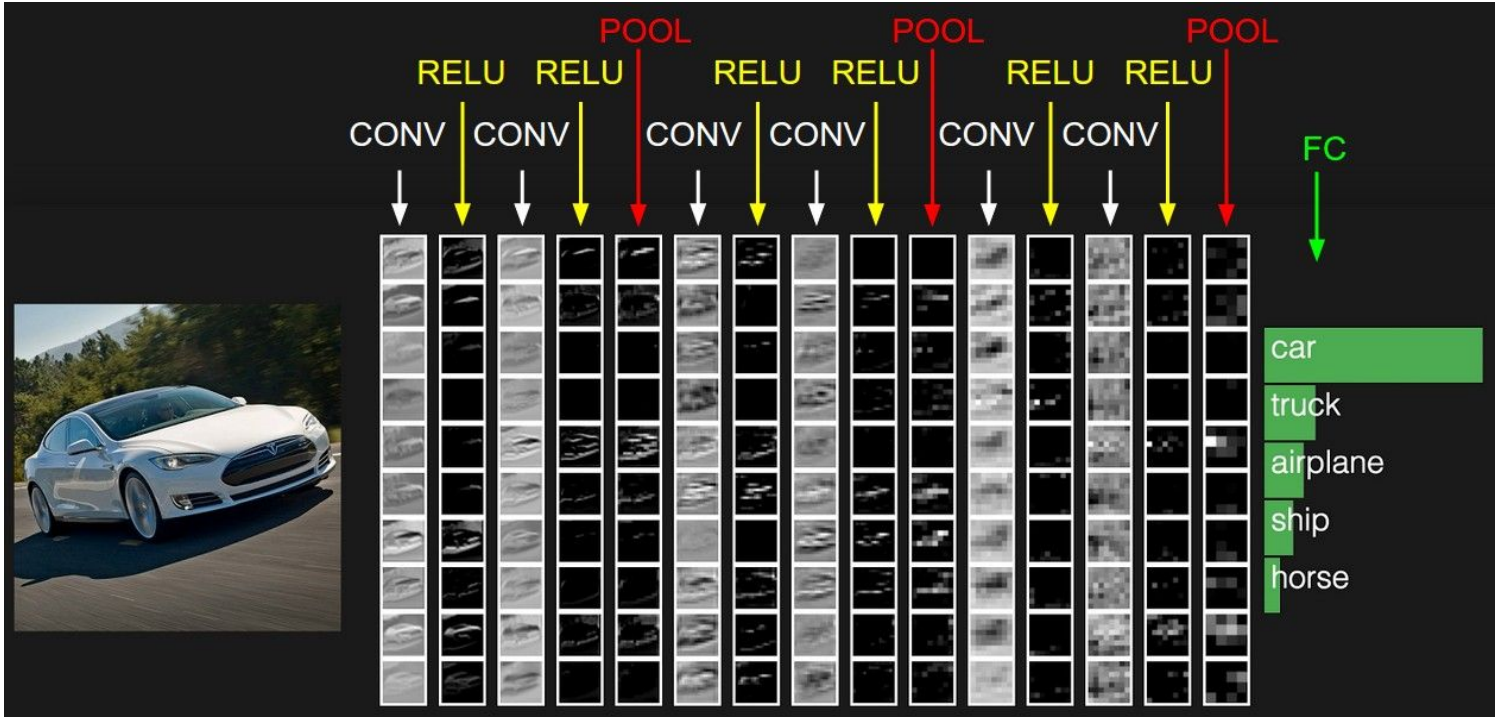
Convolutional Neural Networks

Convolutional Neural Network (CNN) architectures receive a multi-channel image as an input and are usually structured with a set of Convolutional layers each followed by a non-linear operation (usually RELU) and sometimes by a pooling layer (usually max-pooling). At the end of the network, Fully Connected layers are usually used.



Source: http://www.mdpi.com/1099-4300/19/6/242

Convolutional Neural Networks



2) Source : http://cs231n.github.io/convolutional-networks/

BUDGET

Budget

- The resources used on this thesis were of 1 GPU with 11GB of GDDR SDRAM (Graphics Double Data Rate Synchronous Dynamic RAM) and about 30GB of regular RAM. The most similar resources in Amazon Web Services can be found in EC2 instance; p2.xlarge. This service provides 1 GPU, 12GB GDDR SDRAM and 1 CPU with 61GB of RAM. The cost of this service for the duration of the project ascents to 1647€.
- Open-source software.
- No maintenance costs due to the project being a comparative study.

	Amount	**Wage/hour**	**Dedication**	**Total**
Undergraduate Research Assistant	1	8,00 €/h	30 h/week	4080 €
Research Assistant	1	20,00 €/h	8 h	160 €
Senior Engineer	2	40,00 €/h	3 h/week	4080 €
Computational Resources				1647 €
			Total	9967 €

Results: Do visual attention models have difficulties predicting later fixations?

- Evaluation performed using KLdiv metric.
- We computed the distance for each image between both evaluations' scores using the absolute difference |a-b|.