This is the peer reviewd version of the followng article:

Attentive Models in Vision: Computing Saliency Maps in the Deep Learning Era / Cornia, Marcella; Abati, Davide; Baraldi, Lorenzo; Palazzi, Andrea; Calderara, Simone; Cucchiara, Rita. - 10640:(2017), pp. 387-399. (Intervento presentato al convegno 16th International Conference of the Italian Association for Artificial Intelligence tenutosi a Bari, Italy nel November 14-17, 2017) [10.1007/978-3-319-70169-1_29].

Springer
Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

24/04/2024 14:58

(Article begins on next page)

Attentive Models in Vision: Computing Saliency Maps in the Deep Learning Era

Marcella Cornia, Davide Abati, Lorenzo Baraldi,
Andrea Palazzi, Simone Calderara, and Rita Cucchiara

University of Modena and Reggio Emilia
{name.surname}@unimore.it

Abstract. Estimating the focus of attention of a person looking at an image or a video is a crucial step which can enhance many vision-based inference mechanisms: image segmentation and annotation, video captioning, autonomous driving are some examples. The early stages of the attentive behavior are typically bottom-up; reproducing the same mechanism means to find the saliency embodied in the images, i.e. which parts of an image pop out of a visual scene. This process has been studied for decades in neuroscience and in terms of computational models for reproducing the human cortical process. In the last few years, early models have been replaced by deep learning architectures, that outperform any early approach compared against public datasets. In this paper, we propose a discussion on why convolutional neural networks (CNNs) are so accurate in saliency prediction. We present our DL architectures which combine both bottom-up cues and higher-level semantics, and incorporate the concept of time in the attentional process through LSTM recurrent architectures. Eventually, we present a video-specific architecture based on the C3D network, which can extracts spatio-temporal features by means of 3D convolutions to model task-driven attentive behaviors. The merit of this work is to show how these deep networks are not mere brute-force methods tuned on massive amount of data, but represent well-defined architectures which recall very closely the early saliency models, although improved with the semantics learned by human groundthuth.

Keywords: Saliency, Human Attention, Neuroscience, Vision, Deep Learning

1 Introduction

When humans look around the world, observing an image or watching at a video sequence, attentive mechanisms drive their gazes towards salient regions. Attentional mechanisms have been studied in psychology and neuroscience since decades [17], and it is well assessed that the attentional mechanism is mainly bottom-up in its early stages, although influenced by some contextual cues, and guided by the salient points in the scene which is scanned very quickly by the eyes (in about 25-50 ms per item). If the person has a task-driven behaviour, e.g.

when one drives a car, top-down attentive process arise; they are slower (at least 200 ms of reaction in humans) and due to the learned semantics of the scene. In general, the control of attention combines some stimuli processed in different cortical areas to mix spatial localization and recognition tasks, integrating data-driven pop outs and some learned semantics. It has also a temporal evolution, since some mechanisms such as the inhibition of return and the control of eye movements allow humans to refine attention during time.

Reproducing the same attentional process in artificial vision is still an open problem. In the case of a static image, researchers have shown that salient regions can be identified by considering discontinuities in low-level visual features, such as color, texture and contrast, and high-level cues as well, like faces, text, and the horizon. When watching a video sequence, instead, static visual features have lower importance while motion gains a crucial role, motivating the need of different solutions for static images and video. In both scenarios, computational models capable of identifying salient regions can enhance many vision-based inference mechanisms, ranging from image captioning [11] to video compression [13].

Since the seminal research of Kock, Ulman and Itti [23, 18], traditional prediction models have followed biological evidences using low-level features and semantic concepts [14, 22]. With the advent of Deep Learning (DL), researchers have developed data-driven architectures capable of overcoming many of the limitations of previous hand-crafted models. This is not only due to the brute-force of DL architectures, with their capability of being trained by supervised data. This is one area where these architecture are particularly suitable since they recall precisely the neural biological models. Still, it is surprising to see how much today's models share with those early works.

Motivated by these considerations, we present an overview of different solutions that we have developed for saliency prediction on images and video with DL, which represent now the state-of-the-art in public available benchmarks. We compare the neural network model with the early models of computational saliency map, to show similarities and differences. The main contribution of this work is a discussion on why the model of attention prediction with Deep Learning is useful. The paper will show that today's models, based on Convolutional Neural Networks (CNNs) share many of the principles of early models, while at the same time solving many of their drawbacks. Different convolutional architectures will be presented, to deal with features extracted at multiple levels, and to refine saliency maps in an iterative way. Eventually, a solution for video saliency prediction will be discussed and analyzed in the case of driver attention estimation.

2 Related Work

2.1 Saliency prediction on images

Early works on saliency prediction on images were based on the Feature Integration Theory proposed by Treisman *et al.* [32] in the eighties. Itti *et al.* [18], then,

proposed the first saliency computational model: this work, inspired by Koch and Ullman [23], computed a set of individual topographical maps representing low-level cues such as color, intensity and orientation and combined them into a global saliency map. The saliency map is a scalar map, as large as the image, where each point represents the visual saliency, irrespective of the feature dimension that makes the location salient. The *locus* of highest activity in the saliency map is the most probable eye fixation point or is the point where the focus of attention should be localized.

After this work, a large variety of methods explored the same idea of combining complementary low-level features [5, 14] and often included additional center-surround cues [38]. Other methods enriched predictions exploiting semantic classifiers for detecting higher level concepts such as faces, people, cars and horizons [22].

In the last few years, thanks to the large spread of deep learning techniques, the saliency prediction task has achieved a considerable improvement. First attempts of predicting saliency with convolutional networks mainly suffered from the absence of fine-tuning of network parameters over a saliency prediction dataset and from the lack of sufficient amount of data to train a deep saliency architecture [33, 25]. The publication of the large-scale attention dataset SALICON [20] has contributed to a big progress of deep saliency prediction models and several new architectures have been proposed.

Huang *et al.* [16] introduced a deep neural network applied at two different image scales trained by using some evaluation metrics specific for the saliency prediction task as loss functions. Kruthiventi *et al.* [24] proposed a fully convolutional network called *DeepFix* that captures features at multiple scales and takes global context into account through the use of large receptive fields. Pan *et al.* [27] instead presented a shallow and a deep convnet where the first is trained from scratch while some layers of the second are initialized with the parameters of a standard convolutional network. Finally, Jetley *et al.* [19] introduced a saliency model that formulates a map as a generalized Bernoulli distribution and they used these maps to train a CNN trying different loss functions.

2.2 Saliency prediction in video

When considering video inputs, saliency estimation is quite different with respect to still images. Indeed, motion is a key factor that strongly attracts human attention. Accordingly, some video saliency models pair bottom-up feature extraction with a further motion estimation step, that can be performed either by means of optical flow [39] or feature tracking [37]. Somehow differently, some models have been proposed to force the coherence of bottom-up features across time. In this setting, previous works address feature extraction both in a supervised [30] and unsupervised [34] fashion, whereas temporal smoothness of output maps can be achieved through optical flow motion cues [39] or explicitly conditioning the current map on information from previous frames [28].

As previously discussed for the image saliency setting, the representation capability of deep learning architectures, along with large labeled datasets, can

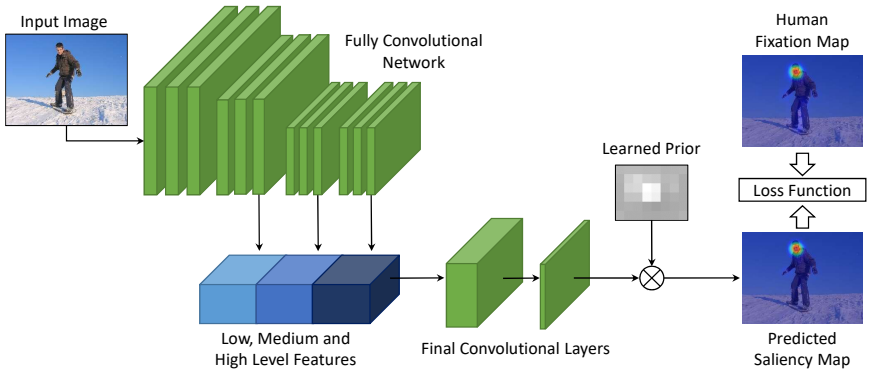

Fig. 1: Overview of our Multi-Level Network (ML-Net) [8].

yield better results. However, deep video saliency models still lack, being the work in [4] the only meaningful effort that can be found in the current literature. Such model leverages a recurrent architecture iteratively updating its hidden state over time, and emitting the saliency map at each step by means of a Gaussian Mixture Model.

3 Saliency Prediction with Deep Learning Architectures

In this section we provide a detailed discussion of different deep learning architectures for saliency prediction on images and video. We will introduce a convolutional model for images, which incorporates low and high level visual features, and which, conceptually, extends the seminal work by Itti and Koch [18] by means of a modern neural network. A discussion on the similarities and differences between these two models will follow, and forerun the presentation of a second model, in which a recurrent convolutional architecture is used to refine saliency maps in a way which is roughly similar to the human scanpath. Finally, we will present an architecture for saliency prediction on video, and show how this particular domain differs from that of images in the case of driver attention prediction.

3.1 Incorporating low-level and high-level cues in a Multi-Level Network

In [8], we proposed a Deep Multi-Level Network (ML-Net) for saliency prediction. In contrast to previous proposals, in which saliency maps were predicted from a non-linear combination of features coming from the last convolutional layer of a CNN, we effectively combined feature maps coming from three different levels of a fully convolutional network thus taking into account low, medium and high level cues. Moreover, to model the center bias present in human eye fixations, we

incorporated a learned prior map by applying it to the predicted saliency map. Fig. 1 shows the overall architecture of our ML-Net model.

More in details, the first component of our architecture is a CNN based on a standard convolutional network originally designed for image classification and then employed in several other computer vision tasks. This network, named VGG-16 [29], is composed by 13 convolutional layers, divided in 5 different blocks, and 3 fully connected layers. Since we aimed at producing a 2-dimensional map (*i.e.* the predicted saliency map), we removed the fully connected layers thus obtaining a fully convolutional architecture. Several other deep saliency models [16, 27, 19, 9] employ the VGG-16 as starting point for their architectures and almost each of them combines feature maps coming only from the last convolutional layer of the VGG-16 network differentiating from each other by designing specific saliency component or by using different training strategies. In contrast to this approach, the second component of our model took as input feature maps coming from three different levels of the VGG-16 network: the output of the third, fourth and fifth convolutional blocks. Our model effectively combined these feature maps through two specific convolutional layers that merge low, medium and high level features and then produce a temporary saliency map. Finally, we decided to incorporate an important property of human gazes in our model. In fact, when an observers looks at an image its gaze is biased toward the center of the scene. To this end, the last component of our architecture was designed to model this center bias through a learned prior map which was applied to the predicted saliency map thus giving more importance to the center of the image.

It is well known that at training time a deep learning architecture has to minimize a given loss function that, in the saliency prediction task, aims at effectively approaching the predicted saliency map to the ground-truth one obtained from human fixation points. Previous deep saliency models were trained with different strategies by using a saliency evaluation metric as loss function or, more commonly, a square error loss (such as the euclidean loss). We instead designed a specific loss function inspired by three different objectives: predicted saliency maps should be similar to ground-truth ones, therefore a square error loss was a reasonable choice. Secondly, predictions should be invariant to their maximum, and there is no point in forcing the network to produce values in a given numerical range, so predictions were normalized by their maximum. Third, the loss should give the same importance to high and low ground-truth values, even though the majority of ground-truth pixels are close to zero. For this reason, the deviation between predicted values and ground-truth values was weighted by a linear function, which tends to give more importance to pixels with high ground-truth fixation probability. The overall loss function was thus

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\left\| \frac{\phi(\mathbf{x}_i)}{\max \phi(\mathbf{x}_i)} - \mathbf{y}_i \right\|^2}{\alpha - \mathbf{y}_i} \tag{1}$$

Table 1: Comparison results on the MIT300 dataset [21].

	SIM ↑	CC ↑	sAUC ↑	AUC ↑	NSS ↑	EMD ↓
Infinite humans	1.00	1.00	0.80	0.91	3.18	0.00
ML-Net	0.59	0.67	0.70	0.85	2.05	2.63
Itti	0.44	0.37	0.63	0.75	0.97	4.26

where $\mathbf{x}_i$ are the predicted saliency maps while $\mathbf{y}_i$ are the ground-truth ones. The proposed architecture was trained with mini-batch of N samples by using the Stochastic Gradient Descent as optimizer.

3.2 Deep Learning architectures vs. the Itti and Koch's model

The first computational model for saliency prediction,and probably the most famous, was presented in a seminal paper by Itti and Koch [18]. It proposed to extract multi-scale low-level features from the input image which were linearly combined and then processed by a dynamic neural network with a winner-takes-all strategy to select attended locations in decreasing order of saliency. As we have shown in the previous section, nowadays saliency prediction is generally tackled via CNN architectures, therefore giving more importance to learning than to hand engineering of features. However, today's models share a lot with that influential work.

The model in [18] extracted three kinds of features from input images: color (as a linear combination of raw pixels in color channels), intensity (again, computed as a linear combination of color channels), and orientation, by means of oriented Gabor pyramids [12]. It should be noted that all these features can be easily extracted by a single convolutional layer, and, indeed, visualization and inversion techniques [36] showed that filters learned in the early stages of a CNN roughly extract color and gradient features. Also, the linear combinations of color channels in [18] can be computed via a single convolutional layer with channel-wise uniform weights or with a 1×1 kernel.

One detail, however, is missing in current convolutional architectures: authors of [18] extracted the same features at multiple scales, and then validated them by performing central differences between adjacent scales. In a CNN, instead, features are always computed at a single scale, even though the overall architecture extracts (different) features at different scales thanks to pooling stages. Of course the multi-scale validation of features was also motivated by the need of extracting robust features, something which comes almost for free in modern architectures. Moreover, many state of the art CNN models are multi-scale by construction, feeding a pyramid of images to the same convolutional stack. Even in our model, we combine different features extracted at different scales to form the final prediction, instead of taking only those produced by the last layer.

Conversely, the most evident characteristic that the Itti and Koch model misses with respect to today's architectures is the ability to extract higher level

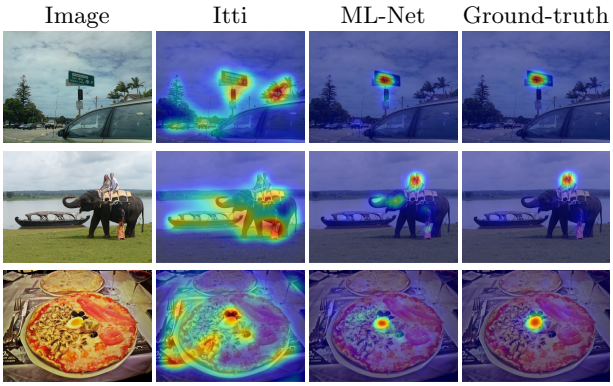

Fig. 2: Qualitative comparisons between the Itti [18] and ML-Net [8] models. Images are from the SALICON dataset [20].

features, and to detect objects and part of objects. This is achieved, in today's networks, by increasing the depth of the network (e.g. 152 layers in the ResNet model [15]). This, given the big performance gap, clearly highlights the need of high-level features for saliency prediction.

As a proof of concept, in Table 1 we compare the results of the model in [18][1] with those of our method. We use the standard performance indicators for saliency: the Similarity, the Linear Correlation Coefficient (CC), the Area Under the ROC Curve (AUC) and its shuffled version (sAUC), the Normalized Scanpath Saliency (NSS) and the Earth-Mover Distance (EMD). We refer the reader to the work by Bylinskii *et al.* [7] for a detailed discussion on these metrics. It can clearly be seen that CNNs overcame that early model by a big margin, with respect to all metrics, and this experimentally confirms the need of high-level features for saliency prediction, rather than just employing low-level cues such as in [18]. To give a better insight of the performance gain, we also report some qualitative results on images randomly chosen from the SALICON dataset. We show them in Fig. 2, along with the ground-truth saliency map computed from human eye fixations. While the model of [18] tends to concentrate on color and gradient discontinuities, which often do not match with the human fixation map, our model can clearly guess most of the saliency maps in a way which is almost indistinguishable from the ground-truth. The middle image, showing a pizza, is also a good example to show the role of the center prior: when there is no a clear object which stands out in the scene, human eyes tend to fix the center of the image, as our model has learned to do. Also, predictions from our ML-Net are particularly focused on small areas, similarly to the SALICON ground-truth. This is due to the fact that, in absence of a task-driven attentive mechanism, the

[1] Numerical and qualitative results of the Itti-Koch model have been generated using the re-implementation of [14], which is also the one reported in the MIT Saliency Benchmark [6].

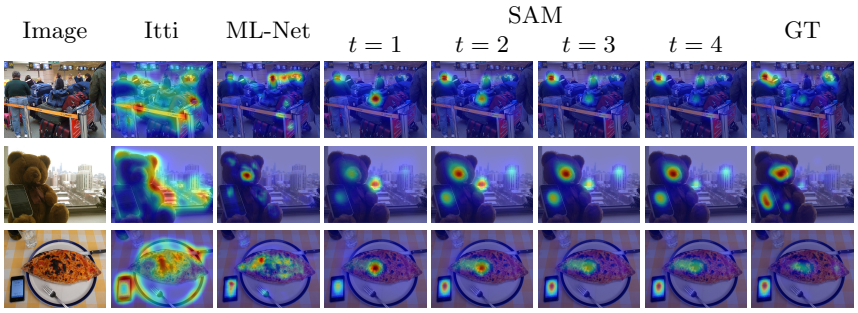

Fig. 3: Qualitative comparison between the Itti [18], ML-Net [8] and SAM [10] models on images taken from the SALICON dataset [20]. For the SAM model, we show predictions given by the recurrent attentive network at different steps.

focus tends to be directed on what is *a-priori* known, such as a person, a face, a traffic sign. The architecture, trained on similar data, does not overfit specific points, but tends to replicate the same semantic-based attentive behaviour.

3.3 Saliency map refinement via a convolutional recurrent architecture

Models for saliency prediction can also go beyond feed-forward neural networks and include recurrent components. Recurrent neural networks are usually employed to deal with time-varying input sequences, but can be used, in general, to process any kind of sequence. Following this intuition, we proposed a second model [10] in which we combined a fully convolutional network (similar to the one described in the previous sections) with a recurrent convolutional network, endowed with an attention mechanism. The recurrent network, instead of looping on a time sequence as in the case of video captioning [3], performs an iterative refinement of the saliency map by focusing on different part of the image. This behaviour is encouraged by using a spatial attentive mechanism, inspired by the machine translation literature [2]. We called the overall architecture SAM, i.e. *Saliency Attentive Model*.

Figure 3 shows, for some images taken from the SALICON dataset, the prediction from the model of Itti and Koch [18], that from our previous model [8], and the output of the attentive network at each step, for $t = 1, ..., 4$, as well as the ground-truth map. As it can be noticed, the refinement strategy carried out by the network results in a progressive improvement of the prediction, which overcomes the performance of a feed-forward neural network like the one in the ML-Net model.

3.4 Estimating task-driven saliency in videos

In [26], we described a model devised for predicting saliency on the `DR(eye)VE` dataset [1], and capable of replicating human attentional behavior while driv-

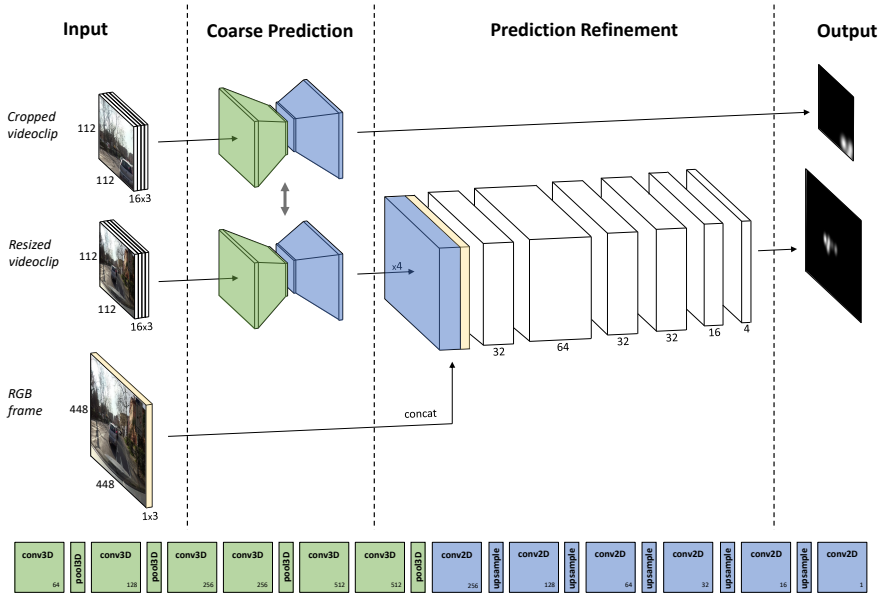

Fig. 4: Illustration of the `COARSE+FINE` model depicting the both streams guiding the optimization during training. Please note that in test stage the cropped stream is not used. At the bottom, the architecture of the `COARSE` module is illustrated.

ing. The need for a different model tailored for this specific context is twofold: first, as anticipated, objects motion in videos tends to capture human attention. Moreover, fixations recorded during the dataset acquisition in [1] are strongly related to the driving activity, and call for a task-driven model and training procedure.

Motivated by the insight that a small temporal window holds sufficient information meaningful for the task of driving, our model captures short-term correlations by means of 3D convolutions, which also stride along time axis. Accordingly, it takes as input samples holding 16 consecutive frames (called *clips* from now on) and provides a dense saliency probability map for the last (current) frame of the clip. The network is jointly trained with two input streams (Fig. 4), in order to tackle the central bias that usually affects saliency benchmarks in general, and is even more noticeable in the driving task. Both streams rely on the same backbone encoder, that we name `COARSE` module as provides a rough, harsh saliency estimate. Such model is based on the work by Tran *et al.* [31] and employs their C3D architecture to map pixels into a 512-dimensional encoding space. Being interested in spatially coherent feature maps, we drop the top fully connected classification module. Moreover, we discard the deepest convolutional layer, which encodings are strongly tailored to the original action recognition task, retaining only the most general features provided by previous

Table 2: Evaluation of the proposed models against central baselines, both on the test and attentive sequences of DR(eye)VE.

	Test seq		Att. seq	
	$CC \uparrow$	$D_{KL} \downarrow$	$CC \uparrow$	$D_{KL} \downarrow$
Baseline (gaussian)	0.33	2.50	0.22	2.70
Baseline (mean train GT)	0.48	1.65	0.17	2.85
Wang *et al.* [35]	0.08	3.77	–	–
Wang *et al.* [34]	0.03	4.24	–	–
ML-Net	0.41	2.05	0.29	2.49
COARSE	0.44	1.73	0.19	2.74
COARSE+FINE	0.55	1.42	0.30	2.24

layers. Eventually, we modify the last pooling layer to cover the whole time axis, and therefore squeeze out the temporal dimension from the output features. The resulting map, which is reduced by a 16x factor along spatial dimension and lacks the temporal axis due to pooling layers, is then processed to produce a saliency estimate as big as the original image and featuring a single probability channel. This is achieved by means of a series of upsampling followed by convolutions.
During training, the model is fed with two streams. The first stream encourages the model to learn saliency estimation given visual cues rather than prior spatial bias, and feeds the COARSE model with random crops. Cropping is also employed in the original C3D training process. Indeed, in [31] authors perform a tensor resize to 128×128 and then a random 112×112 crop. In our experience, this cropping policy is too polite, and yields models strongly biased towards the image center since ground-truth maps still suffer a poor variety. The policy we employ is immoderate, and features a 256×256 resize before the crop. This way, samples cover a small portion of the input tensor and allow variety in prediction targets, at the cost of a wider attentional area. Intuitively, the smaller crops are, the larger the attentional map will appear. Thus, the trained model was able to escape the bias when required, but unfortunately provided over-rough estimates. To address this issue, we feed the COARSE model with a second stream providing images resized to match the crop size. The prediction, after being resized and concatenated with the last frame of the clip, than undergoes a further block of convolutional layers (FINE module) that refine the map. Estimates from both streams are modeled as a probability density P over pixels, and optimized jointly against a ground-truth map Q by means of the Kullback-Leibler divergence:

$$D_{KL}(P, Q) = \sum_i Q_i \ \log\left(\epsilon + \frac{Q_i}{\epsilon + P_i}\right) \tag{2}$$

where the summation index i spans across image pixels and ϵ is a regularization constant.

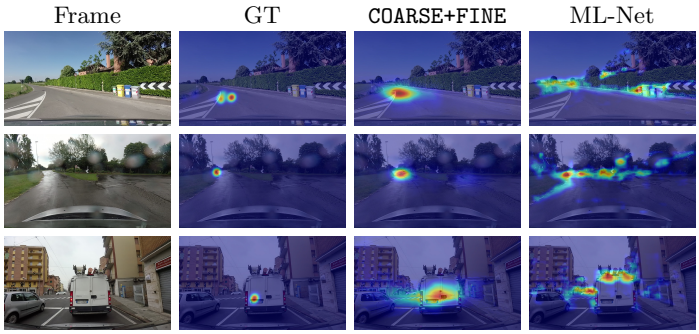

Fig. 5: Representation of differences in the video saliency estimation. This qualitative assessment indicates the suitability of the COARSE+FINE model in encoding temporal information. On the other hand, the ML-Net model processes still images and is more influenced by low-level non temporal features.

Evaluation Here we discuss the experiments performed in order to assess the design choices of our architecture for video saliency. As common in public benchmarks, we first compare our model against two central baselines. The first one represents the central bias as a Gaussian $\mathcal{N}(\mu, \Sigma)$, being μ the image center and Σ a diagonal covariance matrix whose variances are coherent with the image aspect ratio. A more precise, task-driven baseline is obtained by averaging all training ground-truth maps, and two unsupervised state-of-the-art video saliency models [35, 34] are also included in the comparison. The evaluation has been carried out comparing the shift between predicted and ground-truth maps both in terms of Pearson's correlation coefficient (CC) and Kullback-Leibler divergence (D_{KL}). We report such measures evaluated both in the whole test set and in the attentive subsequences only[2] in Tab. 2. Moreover, we report the results of the ML-Net model, that was originally proposed for image saliency and has been properly trained from scratch on the DR(eye)VE dataset.
Several conclusions can be drawn from this evaluation. Firstly, from the poor performances of unsupervised models emerges the peculiar nature of the driving context, that demands for task-driven supervision. Moreover, it can be noticed that the attentive subset of samples is crucial for the evaluation, as simple input-agnostic baselines perform positively overall. Finally, an important remark is revealed by the superior performance of the proposed model w.r.t ML-Net. The gap in performance is due to the temporal nature of video data: indeed, COARSE+FINE profitably learned to extract temporal features that are meaningful for video saliency prediction, whereas the design of ML-Net cannot capture such precious dependencies. A qualitative illustration of the difference in predictions is illustrated in Fig. 5.

[2] *attentive subsequences* in DR(eye)VE are clips in which the driver is looking far from the image center due to a peculiar maneuver he is performing. We refer the reader to [26] for details.

4 Conclusions

In this work we presented different deep learning architectures for saliency prediction on images and video, showing the importance of multi-level features and the ability of recurrent architectures to enhance saliency prediction results. We also shown, with experiments on a driving dataset, that dealing with video sequences requires ad-hoc architectures due to the need of extracting motion features. The comparison between today's models and the early model by Itti and Koch [18] revealed several similarities in the way feature are extracted, and motivated the gap in performances with current models, which is not merely due to the their brute-force nature, but also to their ability to recall very closely early saliency and biological models, although improved with the semantics learned on the ground-thuth.

References

1. Alletto, S., Palazzi, A., Solera, F., Calderara, S., Cucchiara, R.: DR(eye)VE: a Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving. In: CVPR Workshops (2016)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Baraldi, L., Grana, C., Cucchiara, R.: Hierarchical boundary-aware neural encoder for video captioning. CVPR (2017)
4. Bazzani, L., Larochelle, H., Torresani, L.: Recurrent mixture density network for spatiotemporal visual attention. In: ICLR (2017)
5. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: ANIPS. pp. 155–162 (2005)
6. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: Mit saliency benchmark. http://saliency.mit.edu/
7. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)
8. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A Deep Multi-Level Network for Saliency Prediction. In: ICPR (2016)
9. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Multi-level Net: A Visual Saliency Prediction Model. In: ECCV Workshops. pp. 302–315 (2016)
10. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. arXiv preprint arXiv:1611.09571 (2017)
11. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Visual Saliency for Image Captioning in New Multimedia Services. In: ICME Workshops (2017)
12. Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., Anderson, C.H.: Overcomplete steerable pyramid filters and rotation invariance. In: CVPR (1994)
13. Hadizadeh, H., Baji, I.V.: Saliency-aware video compression. IEEE Trans. Image Process. 23(1), 19–33 (2014)
14. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: ANIPS. pp. 545–552 (2006)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In: ICCV (2015)
17. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience 2(3), 194–203 (2001)
18. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI 20(11), 1254–1259 (1998)
19. Jetley, S., Murray, N., Vig, E.: End-to-end saliency mapping via probability distribution prediction. In: CVPR (2016)
20. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: CVPR (2015)
21. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. In: MIT Technical Report (2012)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
23. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence, pp. 115–141. Springer (1987)
24. Kruthiventi, S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. arXiv preprint arXiv:1510.02927 (2015)
25. Kümmerer, M., Theis, L., Bethge, M.: DeepGaze I: Boosting saliency prediction with feature maps trained on ImageNet. In: ICLR Workshops (2015)
26. Palazzi, A., Solera, F., Calderara, S., Alletto, S., Cucchiara, R.: Learning to attend like a human driver. In: Intelligent Vehicles Symposium (2017)
27. Pan, J., McGuinness, K., E., S., O'Connor, N., Giró-i Nieto, X.: Shallow and Deep Convolutional Networks for Saliency Prediction. In: CVPR (2016)
28. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning video saliency from human gaze using candidate selection. In: CVPR (2013)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
30. Stefan Mathe, C.S.: Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE TPAMI 37 (2015)
31. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
32. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology 12(1), 97–136 (1980)
33. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: CVPR (2014)
34. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR (2015)
35. Wang, W., Shen, J., Shao, L.: Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans. Image Process. 24(11), 4185–4196 (2015)
36. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
37. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: ACM MM (2006)
38. Zhang, J., Sclaroff, S.: Saliency detection: A boolean map approach. In: ICCV (2013)
39. Zhong, S.h., Liu, Y., Ren, F., Zhang, J., Ren, T.: Video saliency detection via dynamic consistent spatio-temporal attention modelling. In: AAAI (2013)