Applying Generative Adversarial Networks to perform gaze path prediction in websites

(Discussion Paper)

Gianluca Bonifazi[a], Enrico Corradini[a], Gianluca Porcino[b], Alessandro Scopelliti[c], Domenico Ursino[a] and Luca Virgili[a]

[a]DII, Polytechnic University of Marche
[b]Data Labs, Daimler AG
[c]Energy Intelligence

Abstract

In recent years, gaze path prediction has become a topic widely studied by computer scientists, who have proposed a variety of approaches to address this issue in the context of natural images. Among these approaches, the ones based on deep learning and, in particular, on Generative Adversarial Networks (GANs) have proven to be extremely accurate. When moving from natural images to websites, gaze path prediction becomes much more complex. As an evidence of this fact, no GAN-based approaches have yet been presented to solve this problem. In this paper, we aim at filling this gap by proposing two GAN-based approaches capable of predicting the gaze path of a user when looking at a website.

Keywords

Gaze path prediction, Deep learning, Generative Adversarial Networks, PathGAN, FiWI dataset

1. Introduction

With the passing of the years, more and more contents are present on the Web, leading to an increase in the difficulty of capturing the attention of a user when she visits a website. The evaluation of the attention posed by a user when visiting a website is a non-trivial problem, which depends on several factors. To cope with it, researchers have proposed a powerful tool, namely visual scanpath [1]. It represents a formal definition of the path made by a user's gaze while looking at an image. In the past, such concept was applied to natural images. However, the huge development of the web has led to an increasing interest in applying it also to websites. However, the website scenario is much more complex than the natural image one. In fact, a single web page can contain more natural images, text, logos and animations. This peculiarity makes gaze path prediction tools designed for natural images much less effective when applied

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy
g.bonifazi@univpm.it (G. Bonifazi); e.corradini@pm.univpm.it (E. Corradini); gianluca.porcino@daimler.com; (G. Porcino); alessandro.scopelliti@energyintelligence.it (A. Scopelliti); d.ursino@univpm.it (D. Ursino); l.virgili@pm.univpm.it (L. Virgili)
0000-0002-1947-8667 (G. Bonifazi); 0000-0002-1140-4209 (E. Corradini); 0000-0003-1360-8499 (D. Ursino); 0000-0003-1509-783X (L. Virgili)

to websites. In fact, the presence of several competing stimuli in a web page makes the accurate prediction of the eye fixation much more complex [2].

To address this problem, the past literature has proposed several approaches, belonging to different categories. Among them, deep learning represents one of the categories that recently has attracted a lot of interest. One of the first proofs of the effectiveness of deep learning in this area is reported in [3]. After this attempt, several others have been proposed, and many of them obtained important results. In particular, one architecture that is gaining increasing attention is Generative Adversarial Networks (hereafter, GANs) [4, 5, 6, 7]. It has proven to be extremely valuable in the context of gaze path prediction of natural images [8, 9, 10], where the corresponding approaches have achieved state-of-the-art results. However, as for websites, to the best of our knowledge, no GAN-based approaches have been proposed yet.

In this paper, we aim at filling this gap by proposing two GAN-based approaches specifically designed to operate on websites. We start from PathGAN [10], an approach that has been shown to be very effective for gaze path prediction of natural images, and propose two variants of it. They present a number of refinements derived from several observations we made during some experiments conducted "in the field".

The outline of this paper is as follows: Section 2 provides a technical description of our approaches. Section 3 discusses the experiments performed and the results obtained. Finally, Section 4 draws some conclusions and provides a look at possible future developments.

2. Characterization of the proposed approach

In the field of gaze path prediction, scientific research is still at an early stage. Among the few studies to address this problem, a GAN-based approach, called PathGAN [10], stands out for its results. It predicts the visual scanpath of people watching natural images in both normal and 360-degree formats. Its architecture is shown in Figure 1. The first part of the network is a generator that receives an image and returns a gaze path denoting the eye path of a potential user in observing that image. The generated path is a sequence of 63 fixations, each modeled by a tuple of 4 variables, namely an x-coordinate, a y-coordinate, a timestamp and an end of path probability. The first three variables indicate the position and the duration of each fixation. The fourth one guarantees the possibility of generating paths of variable length. To this end, a threshold is defined for this variable, which makes it possible to determine which fixations should not be included in the final prediction.

We started by applying the original PathGAN to websites; unfortunately, this task did not produce encouraging results. However, it allowed us to identify some problems to solve for improving performances.

First of all, we decided to update the weights of the generator and the discriminator with different frequencies, but the choice to make more updates on the discriminator than on the generator did not lead to performance improvements. Furthermore, assigning a very low weight to the content loss (for instance, a weight equal to 0.05) makes the discriminator much stronger than the generator. Training only the generator for the first 5 epochs was not enough to prevent this phenomenon.

The length of the gaze path to be predicted was another factor that greatly complicated the

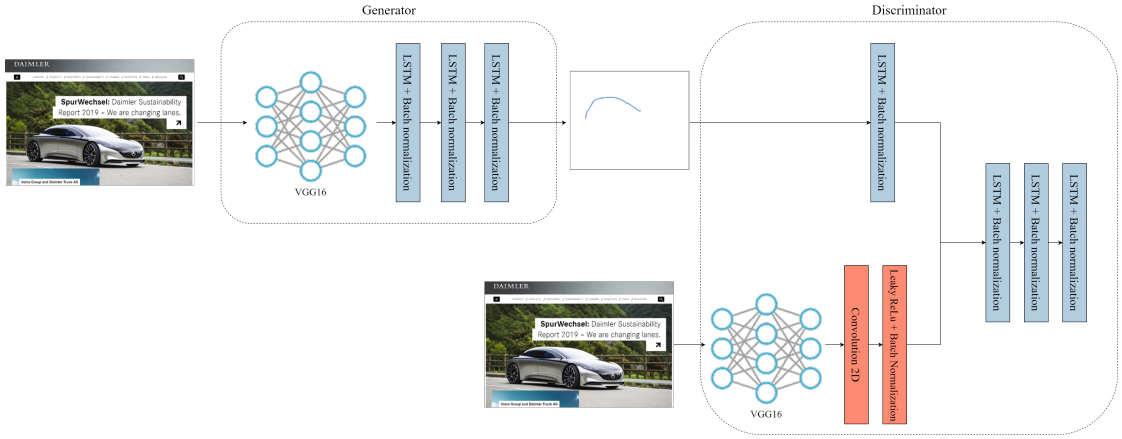

Figure 1: Original PathGAN architecture

problem. In fact, in order to handle paths of variable length, the end of path probability was included. However, the training phase proved unable to tune this variable adequately, leading to paths that were either too short or too long. Overall, we experienced completely wrong predictions that, in the long run, led to mode collapse during training. In this case, the generator tends to return gaze path predictions matching the edge of the image and completely ignoring the original ground truth.

The cause of this phenomenon was identified in the combination of several elements. The discriminator becomes too strong with respect to the generator, which can no longer make realistic predictions. The scarcity of data available does not help to solve this problem. The discriminator clearly overfits on the training data after several epochs, and this cannot be prevented by changing the update frequency of the weights of the generator and the discriminator.

The next step led us to make improvements in the way the network is trained and to change the weights assigned to content and adversarial loss. The only way to make the discriminator weaker is to update the generator weights more often. Decreasing the weight assigned to adversarial loss in the objective function produced positive effects. A higher content loss weight prevented the discriminator from taking over. The quality of the generated samples increased a lot, allowing the network training to be completed successfully. In conclusion, we decided to multiply the adversarial loss for a weight equal to 0.35 and the content loss for a weight equal to 1.

Afterwards, we tuned the correct number of updates of the weights for each part of the network. In particular, we decided to update 16 times the weights of the generator and 4 times those of the discriminator at each step. We also introduced other changes to avoid overfitting. In particular, taking inspiration from saliency prediction models, we added noise to the images passed to the discriminator.

After all these changes and updates to the original PathGAN, we obtained a new version of it called NormalGAN. It has the same architecture as PathGAN but is specifically designed to operate on websites.

In addition to this first variant of PathGAN, we designed a second one. To this end, we

Table 1
Differences between the original PathGAN, NormalGAN and WGAN

Original PathGAN	*NormalGAN*	*WGAN*
Best results with natural images	Fine-tuned for the GUI domain	Fine-tuned for the GUI domain
End of path probability	Fixed path length	Fixed path length
Content loss weight equal to 1	Content loss weight equal to 1	Content loss weight equal to 0.05
Adversarial loss equal to 0.2	Adversarial loss equal to 0.2	Adversarial loss equal to 1
Conditional GAN	Conditional GAN	Conditional Wasserstein GAN

started with the considerations that had led to NormalGAN and made additional changes. In particular, we modified the underlying architecture so that it would follow the structure of a conditional Wasserstein GAN [11] (and, for this reason, we called WGAN this new variant). We also modified the training process to comply with the characteristics of a conditional Wasserstein GAN. In particular, we updated the discriminator's weights more often than the generator's ones (i.e., 5 times compared to 1) because weight clipping was introduced. We reset the weights of the various loss terms to their original values (i.e., 0.05 for content loss and 1 for adversarial loss). These two changes allowed us to obtain a balance between the generator's and the discriminator's strength. In fact, increasing the update rate of the weights leads to a scenario where the generator and the discriminator learn too much from our dataset, causing overfitting. On the other hand, decreasing this parameter implies that the training process takes longer, or that, for the same duration, the generator and/or the discriminator cannot learn enough from the dataset.

In conclusion to these discussions, in Table 1, we provide a scheme of the differences between PathGAN, NormalGAN, and WGAN.

3. Experiments

Our approaches for gaze path prediction of users accessing websites have been implemented in a web-based tool; in the design and realization of it, we paid particular attention to user experience. We adopted Python as programming language, Keras and Tensorflow for handling GANs, and Django for managing the underlying CMS. With User Experience Design techniques in mind, we created a home page where a user can upload an image and specify the gaze path prediction technique she want to apply. Our tool returns the original image, the corresponding ground truth and the gaze path prediction. In Figure 2, we report an example of the output provided by it.

3.1. Description of the dataset

To the best of our knowledge, the only dataset available in the literature, which contains both web page layout images and gaze data is FiWI (Fixations in Webpage Images) [2]. In fact, it is used as a reference in all researches regarding gaze path prediction in the GUI domain. It was constructed by collecting data from 11 volunteers, who observed 149 websites. The limited number of volunteers makes FiWI unable to represent all the ways in which people observe websites. Also, the gender of people is not balanced in it, because 7 volunteers were female and 4 were male. All of these considerations prompted us to build a new dataset aiming at avoiding,

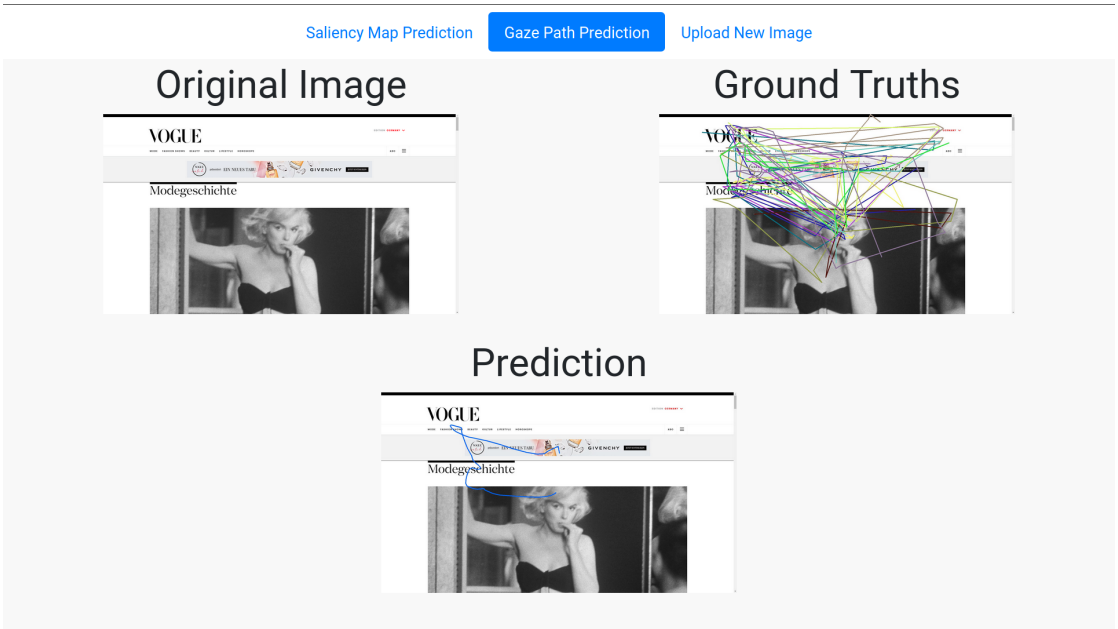

Figure 2: An example of a gaze path prediction returned by our tool

or at least mitigating, these problems.

From the beginning we thought that it was necessary to increase the number of websites present in the dataset and add new forms of layout than those considered by FiWI, because the latter did not fully represent the current variety of web. In particular, the original layouts of FiWI were three, namely *(i)* Pictorial, *(ii)* Text, and *(iii)* Mixed. First, we have extended the number of websites belonging to each layout, while keeping the corresponding fractions balanced. Furthermore, we added a fourth layout, i.e., *Business*, which presents analytical layouts (in particular, dashboards) and Daimler intranet layouts. Finally, we considered that the principles of website design have changed over time. For this reason, we decided to put our dataset both the original and the updated layout of the pages already present in FiWI. Finally, we removed from it the FiWI websites without an updated version available (for example, web pages of companies that no longer exist) to get a balanced set of old and new layouts.

Starting with this original core of 149 images, we reached a total of 262 web layouts. Finally, we felt it necessary to engage more volunteers than FiWI, to include more nuances of how different people look at images, as well as to balance their gender. For this reason, we selected 100 volunteers, 50 males and 50 females. Each of the 262 images was seen by 11 or 12 volunteers, in such a way as to collect 3000 gaze paths. Each volunteer was asked to watch 30 images. If compared with FiWI, the number of available gaze paths is more than twice. Table 2 shows several information on FiWI and our dataset.

Table 2
Characteristics comparison between FiWI and our dataset

	FiWI [2]	*Our dataset*
Number of subjects	11 (4 males, 7 females)	100 (50 males, 50 females)
Age range of subjects	21 - 25	15 - 70
Number of web pages	149	262
Time necessary to display a web page	5 seconds	5 seconds
Screen resolution	1360×768	1920×1080
Number of gaze paths	1,639	3,000

3.2. Results of the experiments

In this section, we apply PathGAN, NormalGAN and WGAN to the images of the dataset described in Section 3.1 to verify whether our two variants actually perform better. To carry out this task, we relied on Jarodzka metrics [12]. These define scanpaths as a series of geometric vectors, called saccade vectors, and compare them along the following dimensions: *(i) Vector shape*, denoting the difference in shape between saccade vectors; *(ii) Vector direction*, indicating the difference in direction (i.e., angle) between saccade vectors; *(iii) Vector length*, representing the difference in amplitude between saccade vectors; *(iv) Vector position*, denoting the distance between fixations; *(v) Fixation duration*, indicating the difference in duration between fixations. All these measures range in the real interval $[0, 1]$; the higher their value, the closer saccade vectors and the better the performance of the approach. Specifically, the first three measures are normalized against the screen diagonal, the fourth against π, and the last against the maximum value of the two durations being compared.

Since each website in our dataset was watched by 11 or 12 different users (see Section 3.1), there is no single ground truth for it, but the gaze path of each user watching it was considered as a ground truth. Therefore, the corresponding gaze path prediction returned by the approach to evaluate was compared with each ground truth and, next, the average performance for that website was computed. Finally, the performance associated with the various websites was averaged to obtain the evaluation of the approach with respect to the overall dataset. Table 3 shows the results obtained.

Table 3
Original PathGAN, NormalGAN and WGAN performances

	Shape	*Direction*	*Length*	*Position*	*Duration*
Original PathGAN	0.652	0.421	0.850	0.435	0.295
NormalGAN	0.992	0.693	0.991	0.836	0.290
WGAN	0.993	0.699	0.992	0.840	0.310

From the analysis of this table we can see that the original PathGAN returns results much lower than NormalGAN and WGAN for all the metrics adopted. As we know, this is due to the fact that it was designed for natural images and not for websites. If we compare NormalGAN and WGAN, we can observe that both of them predict vector shape and path length very well. The position of fixation is also high for both approaches. Direction similarity decreases significantly for both of them, although the values obtained still remain acceptable. Instead, both variants

show a low performance in predicting the duration of each fixation. This parameter is by far the main weakness of our approaches, but also of the original PathGAN, from which they inherit this issue.

Table 3 also shows that WGAN is the best of the three approaches, because it achieves the highest score in all five metrics. NormalGAN also shows a very good performance, lower than WGAN only by a few decimal points. Finally, this table also highlights that both approaches have the same strengths and weaknesses, because they perform well and poorly on the same metrics.

As a further test, we carried out an "absolute" evaluation of WGAN (i.e., the "winner" of the previous comparison) to verify if it was adequate. In this task, we adopted the One human baseline technique [13]. Due to space limitations, we cannot illustrate this experiment here. We can only say that obtained results were extremely satisfying.

4. Conclusion

In this paper, we have proposed NormalGAN and WGAN, i.e., two variants of PathGAN conceived to predict the gaze path of a user watching a website. First, we have explained the motivations underlying our approach. Then, we have illustrated the reasons leading us to define the two variants and have described their technical features. Afterwards, we have illustrated our dataset, which represents an advance with respect to FiWI that is the dataset adopted currently in this field. Finally, we have proposed an experiment showing that our two variants outperform PathGAN when operating on websites.

Our work should not be considered as an ending point, but the starting point for further research efforts. For instance, it could be combined with an approach for saliency map prediction in such a way as that the latter guides the former in obtaining more accurate results. Furthermore, it could be interesting to apply reinforcement learning in this scenario with the ultimate goal of defining a reward function able to highlight the correct aspects of web interfaces and, therefore, to ensure an appropriate training of the model.

Acknowledgments

This work was partially funded by the Marche Region under the project "Human Digital Flexible Factory of the Future Laboratory (HDSFIab) - POR MARCHE FESR 2014-2020 - CUP B16H18000050007". The authors thank the Daimler AG for supporting them in all their testing activities by providing its laboratories and encouraging its employees to participate as volunteers.

References

[1] J. Goldberg, J. Helfman, Visual scanpath representation, in: Proc. of the Symposium on Eye-Tracking Research & Applications (ETMA'10), Austin, Texas, USA, 2010, pp. 203–210. ACM.

[2] C. Shen, Q. Zhao, Webpage saliency, in: Proc. of the European Conference on Computer Vision (ECCV'14), Zurich, Switzerland, 2014, pp. 33–46. Springer.
[3] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, Proc. of the International Conference on Learning Representations (ICLR'14) (2013). ICLR Press.
[4] V. Singh, H. Rashwan, S. Romani, F. Akram, N. Pandey, M. Sarker, A. Saleh, M. Arenas, M. Arquez, D. Puig, Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network, Expert Systems with Applications 139 (2020) 112855. Elsevier.
[5] R. He, X. Li, G. Chen, G. Chen, Y. Liu, Generative adversarial network-based semi-supervised learning for real-time risk warning of process industries, Expert Systems with Applications 150 (2020) 113244. Elsevier.
[6] J. Oh, J. Hong, J. Baek, Oversampling method using outlier detectable generative adversarial network, Expert Systems with Applications 133 (2019) 1–8. Elsevier.
[7] G. Douzas, F. Bacao, Effective data generation for imbalanced learning using conditional generative adversarial networks, Expert Systems with applications 91 (2018) 464–471. Elsevier.
[8] J. Pan, C. C. Ferrer, K. McGuinness, N. O'Connor, J. Torres, E. Sayrol, X. G. i Nieto, Salgan: Visual saliency prediction with generative adversarial networks, arXiv preprint arXiv:1701.01081 (2017).
[9] Y. Li, Y. Zhang, Webpage Saliency Prediction with Two-Stage Generative Adversarial Networks, arXiv preprint arXiv:1805.11374 (2018).
[10] M. Assens, X. G. i Nieto, K. McGuinness, N. O'Connor, PathGAN: visual scanpath prediction with generative adversarial networks, in: Proc. of the European Conference on Computer Vision (ECCV'18), Munich, Germany, 2018, pp. 406–422.
[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved Training of Wasserstein GANs, 2017. arXiv:1704.00028.
[12] H. Jarodzka, K. Holmqvist, M. Nyström, A vector-based, multidimensional scanpath similarity measure, in: Proc. of the Symposium on Eye Tracking Research & Applications (ETRA'10), Austin, TX, USA, 2010, pp. 211–218. ACM.
[13] A. Borji, Saliency prediction in the deep learning era: Successes and limitations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 679–700. IEEE.