Tecnológico
de Monterrey

"Where we look"

Approach to Predicting Visual Attention in Images using Deep Learning

Advanced AI for Data Science II

Adrián Galván Díaz
School of Engineering and Science
Instituto Tecnológico y de Estudios Superiores de Monterrey
Querétaro, México
A01704076@tec.mx

***Abstract* — Visual saliency prediction, particularly the ability to anticipate where humans will look in images, remains a significant challenge in computer vision and human-computer interaction. In this project, we present a deep learning approach for eye fixation prediction using a hybrid architecture that combines a U-Net structure with a ResNet50 backbone. This model was trained and evaluated on the CAT2000 dataset, which contains eye-tracking data from 120 observers across 4000 images spanning 20 different categories [1].**

I propose a modified U-Net architecture that leverages the feature extraction capabilities of ResNet50 while maintaining high-resolution spatial information through skip connections and other techniques further shown. This approach addresses the challenges of accurate fixation map prediction while preserving fine-grained details essential for understanding human visual attention patterns with low or moderate computational cost.

Experimental results demonstrate that this model effectively predicts eye fixation patterns across diverse image categories. Using standard evaluation metrics including NSS (Normalized Scanpath Saliency), CC (Correlation Coefficient), and sAUC (Shuffled Area Under Curve), our model achieves competitive performance compared to existing approaches. Furthermore, our analysis reveals particularly strong performance in [specific categories or scenarios, que podemos completar una vez tenga los resultados].

This work contributes to the field by [contribuciones específicas], providing insights into the relationship between deep neural network architectures and human visual attention mechanisms.

***Keywords* —**

I. Introduction

Understanding where people naturally direct their attention within an image offers profound insights into human perception and can reshape the way industries approach design, media, and user engagement. By predicting fixation points, deep learning models enable new ways to interpret human attention and interest, creating valuable applications across a range of industries. Knowing where a person will focus their gaze can provide critical insights for marketing, where advertisements can be optimized to capture attention; in film and photography, where scenes and compositions can be crafted to guide the viewer's experience; and in healthcare, where understanding visual attention can aid in diagnosing conditions related to cognition, eye health, and neurological disorders.

Eye-tracking devices are pivotal in this field. These devices capture eye movements, allowing researchers to record where and how long a person looks at specific elements within a visual scene. Typically, eye trackers employ infrared light and high-speed cameras to precisely map gaze positions on a screen or environment. By studying these gaze patterns, it is possible to reveal subconscious attention behaviors, which can be highly informative for personality studies, education (to analyze focus in learning environments), and even safety research, such as tracking driver attention in autonomous vehicles. For instance, atypical gaze behaviors can indicate conditions such as autism or ADHD, while specific gaze patterns help in understanding the progression of neurodegenerative diseases like Alzheimer's. In the context of deep learning, these datasets of eye-tracking data serve as ground truth for training models to predict visual saliency maps — heat maps indicating likely areas of focus.

By integrating data from eye-tracking devices into deep learning models, we can create predictive systems that simulate human gaze, empowering diverse fields to tailor visual content more effectively. This project explores these possibilities through the development of a deep learning model aimed at predicting visual saliency, showcasing the transformative potential of gaze prediction in both digital and physical domains.

II. Related Work

Research in visual saliency prediction has evolved significantly, driven by its utility in fields like computer vision and human-computer interaction. Early models, such as Itti and Koch's [2], employed handcrafted features based on color, contrast, and orientation to approximate human attention. However, these approaches were limited in adaptability and struggled with complex image types.

The introduction of deep learning spurred the development of more accurate models. SalGAN is one prominent example [3], using a VGG-based encoder-decoder architecture with skip connections to improve pixel-wise saliency map predictions. It leverages adversarial training, similar to generative adversarial networks (GANs), which helps the model generate saliency maps that resemble true human attention distributions, adding a layer of realism beyond conventional binary cross-entropy loss [4].

The introduction of deep learning led to a paradigm shift, enabling models to automatically learn features directly from data. SalGAN [3], one of the pioneering models in this field, employs an encoder-decoder architecture based on VGG networks, with skip connections that allow the encoder's feature maps to bypass certain layers and connect directly with the decoder. This design preserves finer details from the encoding layers, which improves the model's pixel-wise saliency predictions. SalGAN introduced an innovative use of adversarial training, similar to GANs, where a discriminator is trained alongside the generator to distinguish between real and generated saliency maps. This adversarial setup, beyond conventional binary cross-entropy loss, encourages the generator to produce maps that not only align pixel-by-pixel with the ground truth but also mimic the broader distribution of human gaze, making the predictions more realistic and naturally aligned with human attention patterns [4].

DeepGaze I took a simpler [5], transfer learning approach by leveraging a pre-trained AlexNet model, which had been trained on the large-scale ImageNet dataset. Rather than retraining the model end-to-end, DeepGaze I extracts features from each layer of AlexNet and resizes these feature maps for consistency. By combining these layers into a single set of features, DeepGaze I applies logistic regression to output a final saliency map. This straightforward design, complemented by a center bias assumption (where human attention is often drawn towards the image center), proved effective and achieved competitive results on the MIT saliency benchmark [6]. This model demonstrated the feasibility of using pre-trained networks in saliency prediction without extensive additional training

Building on these insights, multi-resolution approaches like MLNet were developed to better capture attention patterns at multiple spatial levels [7]. Unlike DeepGaze I, which uses a linear combination of pre-trained features, multi-resolution models train end-to-end and use convolutional layers for richer feature learning. These models typically integrate a learned center bias directly from training data, reflecting more accurate natural gaze tendencies without imposing a hard-coded bias. Some architectures even employ multiplicative operations instead of addition to merge feature maps from different levels, capturing more complex feature interactions that enhance spatial detail and gaze accuracy. By adapting their center biases dynamically based on data distributions, these models are highly responsive to the actual patterns in human gaze data.

SALICON, another multi-scale model, incorporates this concept by employing the same network architecture at multiple resolutions [8]. SALICON combines feature maps via concatenation to refine saliency prediction across scales. It innovatively incorporates saliency evaluation metrics—such as Kullback-Leibler (KL) divergence, Normalized Scanpath Saliency (NSS), and correlation coefficient—as loss functions during training. By directly embedding these evaluation metrics, SALICON aligns its learning process with the criteria by which its performance will be judged. Although the effectiveness of this method in enhancing predictive accuracy remains an area of ongoing research, it represents a novel attempt to integrate evaluation criteria directly into the model's optimization process.

DeepFix, one of the more advanced models, modifies the VGG-16 architecture to improve spatial context handling [9]. By removing the max pooling layers typically found in VGG-16, DeepFix avoids excessive downsampling, retaining essential spatial information. In place of max pooling, it incorporates dilated convolutions, which expand the receptive field without sacrificing resolution, a key feature for saliency prediction tasks that demand both local detail and a broad field of view. Furthermore, DeepFix introduces "location-based convolutions," which provide explicit spatial cues by adding location images to the convolutional layers. This innovation enables the model to recognize important regions, such as the center, where attention is often focused. These adaptations allow DeepFix to maintain the integrity of learned VGG-16 features while enhancing the model's sensitivity to spatial layout and context.

More recent developments have explored domain-specific augmentations, such as frequency-domain information, to enhance model performance. For example, a model presented at AAAI 2019 named Sal-DCNN introduced Fourier-transformed inputs alongside the standard RGB channels [10]. By adding frequency-based features, this model captures patterns that might be less apparent in spatial data alone, potentially enhancing saliency prediction in scenes with complex textures or repetitive elements. In parallel, complex convolutions with non-linear activations have been utilized to exploit these Fourier features, yielding deeper representations that respond to high-frequency variations and patterns in the data.

The progression in visual saliency prediction has shown a clear shift from basic handcrafted feature models to complex deep learning architectures, each iteration aiming for greater precision and adaptability. As the field evolved, a trend emerged toward models capable of handling spatial detail while capturing broad contextual information. Techniques like multi-resolution architectures and location-based convolutions helped preserve fine details and adapt to natural gaze biases. Recent models even incorporate Fourier transforms to capture frequency-based information, further refining saliency prediction for images with complex textures and structures [4].

However, these advancements often require high computational resources, extensive training data, and considerable processing power, which can be a barrier to practical deployment, especially in limited-resource environments. My approach seeks to address this gap by leveraging a U-Net architecture with a ResNet50 backbone, augmentation techniques and most importantly, quality data for training.

By adopting this architecture and techniques, I aim to retain competitive accuracy in predicting visual saliency without the high demands of multi-stage training or complex frequency-domain processing. This balance allows us to capture essential attention patterns, creating a more accessible model that can be trained and deployed with limited computational resources. This focus not only addresses practical concerns but also aligns with broader trends toward efficient and scalable deep learning solutions in saliency prediction.

III. Dataset

IV. Methodology

V. Results

VI. Conclusion

Anexos

1. Repositorio de GitHub:

References

[1] Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581.

[2] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(11):1254– 1259, 1998.

[3] Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giro-i-Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081.

[4] Image Processing Group - UPC/BarcelonaTECH. (2020, 30 January). Visual Saliency Prediction with Deep Learning - Kevin McGuinness - UPC TelecomBCN Barcelona 2019 [Vídeo]. YouTube. https://www.youtube.com/watch?v=L0-Sm8ERvVU

[5] Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045.

[6] MIT Saliency Benchmark. (s. f.). http://saliency.mit.edu/results_mit300.html

[7] Cornia, Marcella & Baraldi, Lorenzo & Serra, Giuseppe & Cucchiara, Rita. (2016). A Deep Multi-Level Network for Saliency Prediction. 10.1109/ICPR.2016.7900174.

[8] Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 262-270.

[9] Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 26(9), 4446-4456.

[10] Jiang, L., Wang, Z., Xu, M., & Wang, Z. (2019). Image Saliency Prediction in Transformed Domain: A Deep Complex Neural Network Method. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 8521-8528. https://doi.org/10.1609/aaai.v33i01.33018521