Vision Transformers are Inherently Saliency Learners

Yasser Abdelaziz Dahou Djilali[1,2]
yasser.dahoudjilali2@mail.dcu.ie
Kevin McGuinness[1]
kevin.mcguinness@dcu.ie
Noel O'Connor[1]
Noel.OConnor@dcu.ie

[1] Dublin City University (DCU), Dublin, Ireland.
[2] Technology Innovation Insitute, Abu Dhabi, UAE.

Abstract

Training a Convolutional neural network's (CNNs) auto-encoder has been the defacto approach for visual attention modelling. Recently, (Vision) Transformer models (ViT) achieved excellent performance on various computer vision tasks. In this context, the self-attention mechanism plays a crucial role enabling early aggregation of global information, and ViT residual connections strongly propagate features from lower to higher layers. This raises two important questions: are Vision Transformers inherently learning saliency maps? Are the self-attention maps focusing on the salient regions of the input image? Analyzing the self-attention maps of a pretrained ViTs on saliency prediction datasets, we find that smoothing the internal attention maps with a small number of convolutional filters can achieve reasonable saliency maps with acceptable metric scores. We explore how this phenomenon arises, finding that self-attention promotes early aggregation of global information, then in higher layers, it associates highly attended features, compares their dependencies, and makes analogies over the recurring patterns. This suggests that ViTs first perform feature search, followed by conjunction search combining multiple features sharing higher mutual information. We study the analogies between the self-attention maps and the human generated saliency maps, and conclude with a discussion on the relationship to human visual attention such as feature integration theory.

1 Introduction

The last decade has witnessed the remarkable progress of saliency prediction, and many methods have been presented and achieved remarkable performances on the recently introduced benchmarks, especially the deep learning based methods have yielded a boost in performance. The success of deep learning on visual attention modelling has mainly relied on convolutional neural networks (CNNs) [8, 31, 54]. The convolution operation has the inductive bias of spatial equivariances, enabling impressive results on visual attention datasets. Notably, recent studies pushed the frontiers on many critical computer vision tasks (i.e. Classification [20, 46], Object Detection [14], Semantic Segmentation[59], etc) leveraging the Transformer neural networks [63] at large scale in both model and dataset size. Interestingly, the attention block

Code: https://github.com/YasserdahouML/saliency_transformers

is universal and is shared between Vision Transformers (ViT), language [8, 19], audio [3, 56], and many other modalities [70]. Different than convolution, the self-attention treats the input sequences as a fully connected dense graph, and parse information across nodes. Hence, the symmetries and structure are learned from the data, rather than explicitly incorporated using image-specific inductive biases [17, 36, 57]. Also, Hybrid ViT-CNN models has become an active area of research [14, 22, 46, 68].

This inductive-bias free nature of transformers raises a fundamental question: how does a Vision Transformer's learning paradigm relate to the feature integration theory [61] guiding human visual attention? Do they inherently assign high attention mass to salient areas? If so, are there methods to extract a saliency map from these self-attention maps? The focus of this paper is to explore and analyze the analogies of the self-attention maps with saliency maps and investigate ways in which ViTs can be used as of-the-shelf saliency models. Specifically, our contributions are:

- By analyzing how local/global spatial information is utilised, we show that the the ViT penultimate self-attention maps are similar to the saliency maps.
- Furthermore, we find that retrieving the salient attention maps using a tensor decomposition method yields a rich tensor of saliency information, and is easily mapped to a saliency map using a shallow convolutional neural network decoder.
- We study the main differences between ViT's trained on a supervised classification tasks, and a self-supervised regime, and find that the latter generalizes better to saliency, whereas the former is prone to focus on the most salient objects in the scene.

2 Related works

The theory of feature integration [61], stands out as the pioneering work that identified the visual features that guide human attention. Indeed, this theory served as the foundation for many computational models of saliency, such as pioneers [32] which used center-surround differences across multi-scale image features to derive bottom-up visual saliency. The model produced conspicuity maps by linearly summing and normalizing three key features - color, orientation, and intensity - and then averaging them to generate the final saliency map. Le Meur et al. [44] proposed a more advanced bottom-up saliency approach that utilized additional HVS features, including contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions. Other static saliency models, such as e.g. [9, 25, 38, 50, 51] , are mostly cognitive based models relying on computing multiple visual features such as color, edge, and orientation at multiple spatial scales to produce a saliency map. Bayesian models were also developed, building on top of cognitive models, to incorporate prior knowledge (such as scene context or gist) using a probabilistic approach like Bayes rule for combination [23, 52, 60, 69]. These models attempted to model the human visual system in a principled manner, but their performance is still far from the "infinite humans" baseline [35]. Refer to [5, 6] for a comprehensive review.

In the past decade, the use of deep learning techniques has led to significant improvements in saliency prediction by adapting existing CNN architectures. These models are typically trained end-to-end on large-scale datasets of static scenes, as described in studies such as [7, 11, 34]. The first CNNs to be used for saliency prediction were eDN and DeepFix, introduced by the authors of [53] and [41], respectively. DeepFix incorporated VGG-16 weights

to initialize the first 5 convolution blocks, then added two Location Based Convolutional (LBC) layers to capture semantics at multiple scales. On the other hand, Pan et al. [64] used Generative Adversarial Networks (GANs) [29] to develop the SalGAN model. The generator model's weights were learned through back-propagation computed from a binary cross-entropy (BCE) loss over existing saliency maps. The resulting prediction was then processed by a discriminator network trained to solve a binary classification task between the saliency maps generated at the generative stage and the ground truth ones in a min-max game. EML-NET proposed by [33] consists of a disjoint encoder and decoder trained separately. Furthermore, the encoder can contain many networks extracting features, while the decoder learns to combine many latent variables generated by the encoder networks. These deep models achieve results closer to the human baseline results on the Salicon [34], MIT300 [10], and CAT2000 [7] datasets.

It is clear from the above review that the deep learning based approaches targeting image visual attention modeling have little in common with classical methods, and in fact diverged from the initial cognitive motivations. Authors from [62] proposed Class-wise Jensen-Shannon (JS) distance to produce an error-consistency metric that is very close to Cohen's κ [27]. This metric was used to measure whether CNN or ViTs correlate more with human vision from an error consistency point-of-view. They concluded that ViTs have higher shape bias and are largely more consistent with human errors. The work from [71] concluded that except from brain's ventral stream hierarchical correspondence that ViTs and CNNs reveal, neither CNN nor transformer is an optimal model paradigm of the human visual system. The contributions of this paper attempt to investigate whether ViT's are inherent saliency learners by exploring their analogies with saliency cognitive based models. We also study the effect of supervised vs self-supervised ViTs on the saliency task.

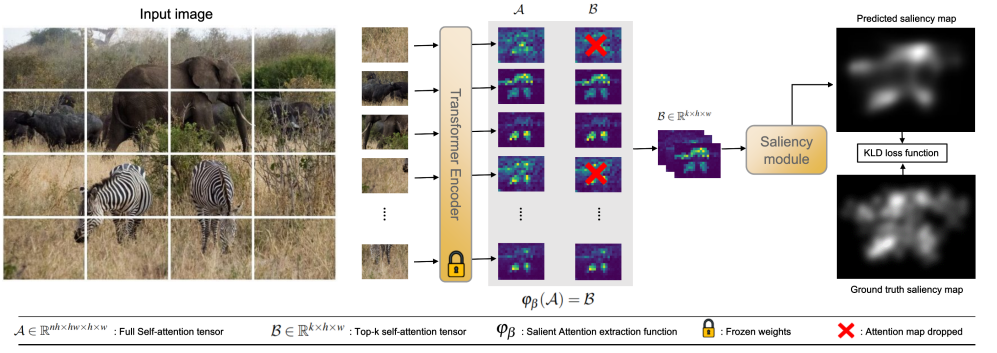
3 Vision Transformers Visualisation

This has recently become an active research topic with many methods proposed. Treating the pairwise attention values between the different patches as relevancy scores is the most commonly used visualisation [14, 15, 56]. This can be obtained over the last layer, or by combining multiple layers. This method suffers from a blurring effect, as well as over amplifying the role of irrelevant tokens. The rollout method [1] is an alternative, which reassigns all attention scores by considering the pairwise attentions and assuming that attentions are linearly combined into subsequent contexts. The method seems to improve results over the utilization of a single attention layer. Solving the issue of losing the relevancy when using a Self-attention layer is a challenge since a naive propagation rule would lead to negative contributions inducing noise in the relevancy map. The authors of [16] guided the relevancy propagation with a rule that is applicable to both positive and negative attributions. They compute the scores for all attention heads using the Layer-wise Relevance Propagation (LRP) method [4], then to remove the negative contributions, the method incorporates both relevancy and gradient information throughout the attention graph. We highlight the main methods:

Attention Rollout [1] Given the attention graph representing the flow of information between positions in different layers as a series of edges, with weights representing the proportion of information transferred. The attention rollout is calculated recursively by multiplying attention weights matrices in all the layers below, and the input attention is computed by setting the lower layer to be the input layer.

Partial LRP [64] was the first method to exploit the LRP [4], based on the observation

Figure 1: Complete pipeline for training. The input image is processed with a Vision Transformer model to get the self-attention maps tensor $\mathcal{A}$. φ_β serves as a medium to produce the Salient Attention maps tesnor $\mathcal{B}$ by selecting top-k informative modes. Then, $\mathcal{B}$ is fed to the Saliency Module to predict the visual saliency map.



that the mean attention heads is not optimal since the relevance of the attention heads in each layer can vary [64]. However, the application of LRP was limited in that no relevance scores were propagated back to the input, providing only partial information on the relevance of each head. It should be noted that the relevance scores were not directly evaluated but rather used for visualization of relative importance and for pruning less relevant attention heads.

Relevance LRP [16] propagates the relevance and gradients with respect to a specific class. The two tensors are the input feature map and weights for layer n. Their relevance propagation follows the Deep Taylor Decomposition [49], and satisfies the conservation rule in [49] across two successive layers. It is worth mentioning that the result of rollout is fixed given an input sample, regardless of the target class to be visualized. In addition, it does not consider any signal, except for the pairwise attention scores. However, the relevance LRP requires the classification logits, hence, can not be applied on self-supervised methods.

4 Salient Attention

In visual attention modelling, the salient regions in an image may correspond to the objects or regions of interest within the image. Clearly however, the most salient regions are not necessarily the objects in the image, but could rather be other features or patterns that catch the viewer's attention [6]. Existing methods for extracting the relevance map of ViTs focus on classification models, thus, they consider only the [CLS] token, which summarizes the self-attention tensor. Consequently, the relevance map is generated from the row $C_{[CLS]} \in \mathbf{R}^s$ that relates to the [CLS] token. Within this row, there is a score assigned to each token, which assesses its impact in solving the downstream classification task.

Furthermore, as shown in Figure 1, we propose a simple token-agnostic attention map approach to select the ViT's self-attention maps of the salient locations. Let's note the **Vision Transformer backbone** ($\mathbf{f}_\theta$) that maps the input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, to a representation Λ. $\mathbf{f}_\theta$ processes the input image in patches $x_p \in 1 \ldots N$, where N is the total number of patches, and $x_p \in \mathbb{R}^{C \times h \times w}$. We can extract the last layer's full self-attention tensor $\mathcal{A} \in \mathbb{R}^{nh \times hw \times h \times w}$ (i.e. the [CLS] token is excluded as the full attention tensor would be of size $h.w+1$). The aim is to select the salient positions along the second dimension, and retrieve their self-attention

maps. To achieve this, we propose using a tensor decomposition method to identify the most informative modes in $\mathcal{A}$. Specifically, we apply the Candecomp-Parafac decomposition via Alternating-Least Square [37], which decomposes $\mathcal{A}$ into a set of factor matrices that capture its underlying structure as shown in Equation (1):

$$\mathcal{A} \approx \sum_{r=1}^{R} a_r^{(1)} \otimes a_r^{(2)} \otimes a_r^{(3)} \otimes a_r^{(4)} \tag{1}$$

where R is the rank of the decomposition, and $a^{(1)}, a^{(2)}, a^{(3)}$, and $a^{(4)}$ are factor matrices of size $n_h \times R, h * w \times R, h \times R$, and $w \times R$, respectively. The symbol $\otimes$ denotes the outer product of two vectors.

The ALS algorithm repeats these updates for each factor matrix until convergence or a maximum number of iterations is reached. After convergence, the factor matrices can be used to reconstruct the original tensor or to extract useful information about the underlying structure of the data. By examining the factors, we can identify which modes contain the most information about $\mathcal{A}$, as these will have larger weights in the decomposition. We then select the top-k informative modes based on their weights and use them for further analysis. This approach allows us to reduce the dimensionality of the tensor and focus on the most relevant information, that are highly likely salient locations, which can lead to a class independent relevancy map. The final $\mathcal{B} \in \mathbb{R}^{k \times h \times w}$ is obtained.

4.1 Obtaining the saliency maps from ViT Attention/Relevance maps

To obtain the saliency map, we train a tiny convolutional neural network (CNN) based saliency module on top of the Attention/Relevance maps obtained using the various methods explained earlier. The saliency module takes these maps as input and learns to predict the corresponding saliency map, which highlights the most visually important regions in the input image.

Vision Transformer backbone ($\mathbf{f}_\theta$). This maps the input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, to a representation Λ. We select the ViT [20] model, which inherits a BERT-like architecture. The input is a sequence of all non-overlapping patches of size 16×16 of the input image, followed by flattening and linear layers, to produce a sequence of embedding. Similar to BERT, a classification token $[CLS]$ is appended at the beginning of the sequence and used for classification. We define (φ_β) that serves as a medium to extract the Attention/Relevancy map $\mathcal{B} \in \mathbb{R}^{k \times h \times w}$ (k is the k-modes in Salient Attention, and the number of heads for all the other methods). φ_β can be instantiated with any method from the pool of methods explained earlier: [Raw Attention, Attention Rollout, Partial LRP, Relevance LRP, Salient Attention]. These methods can be grouped in two folds: attention-maps, and relevance. Each has different properties and assumptions over the architecture and propagation of information in the network. We briefly describe each baseline in the following section and the different experiments for each domain.

The attention-map baselines are class-agnostic by definition, and include rollout [1], which produces an explanation that takes into account all the attention-maps computed along the forward-pass. A more straightforward method is raw attention, i.e. using the attention map of block 1 to extract the relevance scores. The proposed saliency based attention map is an extension of the raw attention by retrieving the attention-maps for the salient locations using the CP method, and is not necessarily tight to the salient object in the scene.

Unlike attention-map based methods, the relevance propagation methods consider the information flow through the entire network, and not just the attention maps. These baselines

Table 1: Comparative performance study on: Salicon and MIT300. Supervised denotes initializing the ViT with classification weights on ImageNet, whereas unsupervised stands for using DINO self-supervised weights. Our approach using Salient locations achieves favorable gains across different settings.

Models		Salicon					MIT300				
		SIM	s-AUC	CC	NSS	KLD	SIM	s-AUC	CC	NSS	KLD
	ITTI [30]	0.37	0.61	0.20	–	–	0.46	0.13	0.44	1.11	0.95
	GBVS [30]	0.44	0.63	0.42	–	–	0.48	0.62	0.47	1.24	0.88
	Salicon [34]	–	–	–	–	–	0.51	0.73	0.56	1.70	0.78
	CASNet [24]		–	–	–	–	0.58	0.73	0.70	1.98	0.58
	EML-NET [33]	0.79	0.74	0.89	2.05	0.52	0.74	0.67	0.78	2.48	0.84
	MSI-Net [40]	0.80	0.74	0.90	2.01		0.67	0.74	0.77	2.30	0.42
	TranSalNet [48]	–	–	–	–	–	0.68	0.74	0.80	2.41	1.01
	SalGAN [32]		0.75	0.76	2.47	-	0.63	0.73	0.67	1.86	0.75
	UNISAL [21]	0.77	0.73	0.87	2.45	-	0.67	0.78	0.78	2.36	0.41
	DeepGaze [43]	–	–	–	–	–	0.66	0.77	0.77	2.33	0.42
Supervised	Relevance LRP	0.72	0.71	0.75	2.01	0.59	0.61	0.71	0.73	1.94	0.70
	LRP	0. 57	0.68	0.72	1.41	0.84	0.53	0.63	0.71	1.6	0.78
	Partial LRP	0.70	0.68	0.71	1.86	0.64	0.58	0.69	0.70	1.75	0.64
	Raw attention	0.55	0.66	0.61	1.58	0.98	0.54	0.70	0.72	1.39	0.81
	Rollout	0.58	0.69	0.67	1.45	0.80	0.55	0.72	0.73	1.42	0.75
	Salient Attention	0.71	0.69	0.73	2.12	0.57	0.62	0.72	0.70	2.14	0.72
Unsupervised	Raw attention	0.73	0.71	0.81	2.24	0.61	0.63	0.72	0.71	2.06	0.71
	Salient Attention	0.78	0.74	0.86	2.41	0.42	0.65	0.74	0.76	2.38	0.51

include the partial application of LRP that follows [2]. It is arguable whether these techniques might be practically class-agnostic, authors from [16] proved that LRP method's visualizations remain roughly constant for distinct target classes. The Relevance LRP [16] is the most effective class-specific methods as it relies on the propagation rule that considers both positive and negative contributions, hence, provides fine-grained class specific visualizations.

Saliency module (Ω_w) is a non-linear mapping $\Omega_w : \mathcal{B}^{k\times w\times h} \mapsto S_{\mathbf{p}}$ parameterised by ω, where $S_{\mathbf{p}} \in \mathbb{R}^{H\times W}$, is the predicted saliency map. The architecture of the saliency module consist of 4 blocks of $[\text{Conv2d} \rightarrow \text{ReLU} \rightarrow \text{Upsample}]$.

Saliency loss function. The saliency task can be seen as a distance measure between the predicted saliency distribution $S_{\mathbf{p}} \in [0,1]^{W\times H}$, and the continuous ground truth $S_{\mathbf{GT}} \in [0,1]^{W\times H}$. The objective function must be designed to maximise the in-variance of predictive maps and give higher weights to locations with higher fixation probability. Thus, the saliency module (Ω_w) is trained to minimize the Kullback-Leibler Divergence (KLD), widely adopted for benchmarking saliency models [13], the KLD between $S_{\mathbf{p}}$ and $S_{\mathbf{GT}}$ is given by:

$$\mathcal{L}_{\text{KLD}}(S_{GT}, S_p) = \sum_{i=1}^{W\times H} S_{GT\,i} \log\left(\varepsilon + \frac{S_{GT\,i}}{\varepsilon + S_{p_i}}\right) \tag{2}$$

5 Experiments

Training. We experiment with the ViT-base/16 [20] model trained in a supervised way on ImageNet [18], and in a self-supervised setting (i.e. DINO [15]). To evaluate the proposed framework, we train the CNN saliency module on the 10k/5k train/validation splits of the image saliency dataset Salicon [34]. We use the aforementioned methods to obtain a final

Table 2: Results on the CAT2000 validation set.

Models		CAT2000				
		SIM	s-AUC	CC	NSS	KLD
	SalGAN	0.54	0.63	0.56	1.46	0.93
	DeepGaze	0.68	0.64	0.79	1.96	0.38
Unsupervised	Raw attention	0.62	0.60	0.75	1.84	0.69
	Salient Attention	0.65	0.62	0.76	1.90	0.63

Attention/Relevancy map [1], that is fed to the CNN saliency module. It is worth mentioning that the ViT encoder weights were not fine-tuned on the saliency task; only the randomly initialized saliency module is trained on top of the frozen encoder. The motivation for this is to set a robust evaluation procedure and to prevent the encoder adapting its parameters to saliency specific requirements. See Section A.1 in supplementary for other ViT variants.

Evaluation setting. We compare against the SoTA methods listed in [65] and add newer models with available implementations [42]. Moreover, we test on the MIT300 benchmark [35] that is more challenging than the Salicon test set. As suggested in [12, 42], we use the following evaluation metrics: Similarity Metric (SIM), shuffled AUC (s-AUC), Linear Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), and the Kullback Leibler Divergence (KLD) [12]. We adopt the more recent metrics formulation from [42].

Technical details. We adopt the implementations of ViT [20] from the official DINO repository[2]. For the saliency module, we consider 4 blocks with 32 input convolution filters (59k parameters). This module is trained on the 10k images of Salicon. The saliency module is implemented in PyTorch [55] and trained using a single NVidia RTX3090 24GB GPU. All the variants are trained for 30 epochs using the AdamW [47] optimizer. We employ a warmup of 6 epochs and a cosine learning rate scheduler with maximum lr set to 10^{-3}.

5.1 Results

State-of-the-art comparison. Here we compare the proposed approach to SoTA image saliency models on both the Salicon, MIT1300 and CAT2000 validation sets. Table 1 and Table 2 shows the performance comparison in terms of the five metrics for the respective validation sets. We observe that our method performs favorably against existing approaches across different attention/relevancy extraction methods. For the supervised setting, the relevance LRP and the salient locations-based approach shows the best performance, followed by partial LRP, LRP, and Rollout. The raw attention-based approach shows the worst performance among supervised models. Surprisingly, initializing the ViT backbone with self-supervised weights, and training the saliency module on both the Raw Attention and Salient Locations maps exhibits scores on par with end-to-end methods such as SalGAN [54] and EML-NET [33]. This is likely due to the ViT encoder adapting the parameters to fit the classification task, hence, corrupting the raw attention maps, for complete comparison between supervised and unsupervised raw attention visualisation, see Section A.2 in supplementary.

Saliency for Low-level features images. SoTA saliency models capture high level features such as cars, humans, etc. However, these kinds of approaches may fail to adequately

[1] https://github.com/hila-chefer/Transformer-Explainability
[2] https://github.com/facebookresearch/dino

Table 3: Impact of the number of selected modes k. Performance comparison when k is varied. 32 modes is the optimal number for the saliency task.

Models		Salicon					MIT300				
		SIM	s-AUC	CC	NSS	KLD	SIM	s-AUC	CC	NSS	KLD
Salient Locations	k=16	0.73	0.70	0.76	2.12	0.52	0.61	0.71	0.73	1.94	0.70
	k=32	0.78	0.74	0.86	2.41	0.42	0.65	0.74	0.76	2.38	0.51
	k=64	0.70	0.69	0.72	1.90	0.60	0.57	0.69	0.70	1.75	0.64
	k=128	0.71	0.68	0.73	1.93	0.59	0.55	0.70	0.73	1.84	0.61
	k=256	0.68	0.62	0.67	1.55	0.71	0.51	0.66	0.65	1.33	0.78

capture a number of other crucial features that describe aspects of human visual attention that have been extensively investigated in psychology and neuroscience. Visual search, often couched in relation to Feature Integration Theory (FIT), is one of the most prominent processes shaping human attention [39, 61]. This is where a subject's brain parallel processes regions that differ significantly in one feature dimension i.e. Color, Intensity, Orientation. These correspond to low level features, that operate as the basic mechanisms of the Human Visual System. We conducted evaluations of the performance of UNISAL, and our ViT based approaches on samples of low level attention using images from a recently proposed dataset [39]. The aim is to understand the main differences on how saliency exploration is performed when the self-attention mechanism promotes global connectivity between the image patches. See Section A.3 in supplementary for more details.

As shown in Figure 2, Unisal produces high quality saliency maps consistent with the ground truth maps for natural images on Salicon, high-level features such as: human faces, bus, monument, etc; are dominant in these images (3^{rd},4^{th},5^{th} rows in the Figure 2). The human visual system combines the bottom-up with top-down features to solve the attention task. This behaviour might not be reflected in the fixation/saliency datasets. Hence, End-to-end deep learning based models might learn a good saliency, but violates its subtle definition. Early computational approaches for the visual human system e.g. [9, 25, 26, 28, 38, 50, 51, 58] were mostly cognitive based models relying on computing multiple visual features such as color, edge, and orientation at multiple spatial scales to produce a saliency map. Moreover, could self-supervised ViTs bridge this gap, by reasoning over the reccuring features and drawing dependencies/similarities?

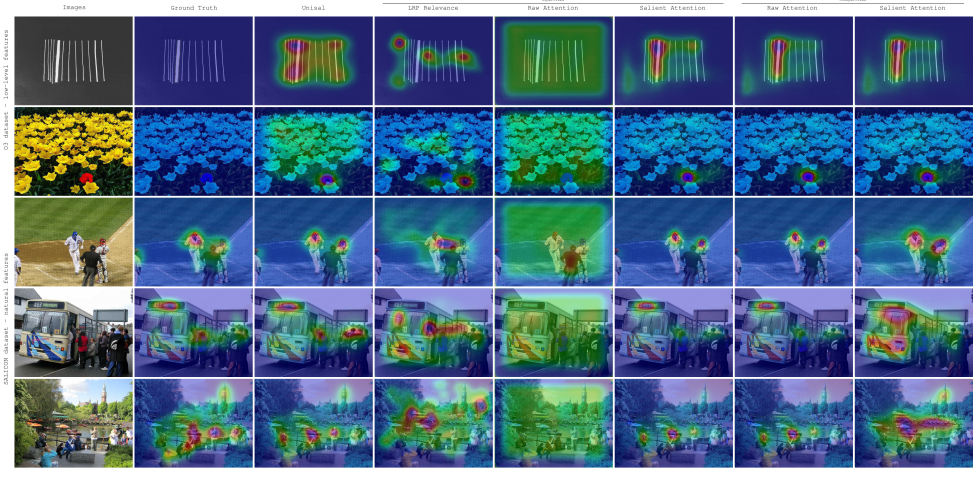
In fact, Unisal fails to respond to simple features. For example, considering colour (2^{nd} row in Figure 2), Unisal [21] did not capture the red flower as the most salient object, whereas the ViT succeeded in doing so, as this pattern is solved with the global nature of the self-attention mechanism. Also, Unisal do not discriminate the larger shape (1^{st} row in figure). Furthermore, a differently shaped object to others should capture the viewer's attention, but all of the approaches fail to do so. This suggests that ViTs do not only correlate with the gaze data, but also incorporate characteristics of the visual system as important priors induced by the self-attention mechanism.

5.2 Ablation study

In this section we justify the choices by ablating key features of the procedure.

The effects of the number of Salient Locations. Table presents the performance of the Salient Locations based framework when the number of top k-modes is varied. We observe that 32 is the optimal value and higher values degrade the performance since the added modes do not carry any useful information that and may act as noise.

Figure 2: Qualitative results of the different models on sample images from Salicon and O3 datasets. It can be observed that the proposed approach is able to handle various challenging scenes well and produces consistent saliency maps



Is the saliency module required? Research in cognitive science (e.g. [45, 67]) indicates that low-level saliency in both humans and animals happens early in the primary visual cortex. Clearly however, SoTA approaches mostly follow a supervised learning paradigm. We attempt to produce the saliency map from the [CLS] token raw-attention of the self-supervised ViT. Furthermore, the raw attention map is converted into a discrete fixation map $\mathcal{F}$ with a threshold $\lambda = 0.9$ to keep the most relevant attention nodes. Then, we smooth it with a Gaussian-filter, where each pixel value is replaced by a weighted average of its neighboring pixels according to the Gaussian distribution. The amount of smoothing is controlled by the value of σ, to obtain the predicted continuous saliency map G_p. This fully unsupervised approach resulted in a visually appealing saliency maps, however the scores were still far from the baselines on the Salicon validation set [KLD: 1.12, NSS: 1.46].

6 Discussion

Limitations: This study still need supervision to learn the saliency module. While we have achieved baseline scores across the metrics, more fine-grained methods may be able to solve the task in a fully unsupervised way. Similar to the Relevance LRP [16], this will require designing propagation rules, and conservation laws specific for the saliency task.

Conclusion: We investigate how the emerging properties of Vision Transformers could serve for Visual Attention Modelling with little supervision. We examine the existing visualisation methods for ViTs, and introduce a class-agnostic approach for selecting the ViT's self-attention maps of the salient locations, which allows the identification of informative modes within the attention tensor. The qualitative and quantitative results have demonstrated the competitiveness of the approach. We believe these findings may uncover the analogies between ViT's learning dynamics and the human visual system.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[4] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016.

[5] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):679–700, 2019.

[6] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012.

[7] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

[10] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015.

[11] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015.

[12] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[13] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[16] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[17] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020.

[22] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[23] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009.

[24] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7521–7531, 2018.

[25] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Advances in neural information processing systems*, pages 481–488, 2005.

[26] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil. Decorrelation and distinctiveness provide with human-like saliency. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 343–354. Springer, 2009.

[27] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020.

[28] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1915–1926, 2012.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[30] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006.

[31] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

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

[33] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, page 103887, 2020.

[34] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2015.

[35] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[36] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[37] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[38] Gert Kootstra, Arco Nederveen, and Bart De Boer. Paying attention to symmetry. In *British Machine Vision Conference (BMVC2008)*, pages 1115–1125. The British Machine Vision Association and Society for Pattern Recognition, 2008.

[39] Iuliia Kotseruba, Calden Wloka, Amir Rasouli, and John K Tsotsos. Do saliency models detect odd-one-out targets? new datasets and evaluations. *arXiv preprint arXiv:2005.06583*, 2020.

[40] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

[41] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

[42] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing.

[43] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 4789–4798, 2017.

[44] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):802–817, 2006.

[45] Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6 (1):9–16, 2002. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(00)01817-9.

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[48] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 2022.

[49] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[50] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*, pages 433–440. IEEE, 2011.

[51] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2049–2056. IEEE, 2006.

[52] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003.

[53] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016.

[54] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, 2019.

[56] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[57] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[58] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.

[59] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[60] Antonio Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003.

[61] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[62] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[64] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

[65] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4894–4903, 2018.

[66] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[67] Yin Yan, Li Zhaoping, and Wu Li. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, 115(41):10499–10504, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1803854115.

[68] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

[69] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008.

[70] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[71] Qiongyi Zhou, Changde Du, and Huiguang He. Exploring the brain-like properties of deep neural networks: A neural encoding perspective. *Machine Intelligence Research*, 19(5):439–455, 2022.