TDSP: A YOLO-Based Top-Down Saliency Prediction Model

Can Mızraklı
Department of Computer Engineering, TED University
can.mizrakli@tedu.edu.tr

Abstract

Visual saliency seeks to identify the most striking parts in an image. Top-down saliency prediction refines this by including higher-level task definitions. YOLO has set the standard in computer vision with its processing speed, detection accuracy and computational efficiency. Our contribution in this paper involves proposing an architecture with a pre-trained YOLO backbone for feature extraction. Coupled with a feature pyramid network for multiscale visual information capture, a Sentence-BERT-based task encoder derives the semantic embeddings from the text-based task description, which are then merged with the visual features using a transformer fusion module. It is then decoded into the resulting saliency maps with the task-relevant areas highlighted.

Keywords: Visual Saliency, Top-Down Saliency, YOLO, Feature Pyramid Network, Sentence-BERT, Transformer Fusion.

1 Introduction

Task-driven visual saliency improves the goal of visual saliency by determining image areas that relate to a high-level goal in a particular task. While the conventional bottom-up methods mark general visually striking regions, the current technique utilizes task definitions to produce task-specific saliency maps.

In this paper, we propose an architecture that incorporates a pre-trained YOLO backbone for feature extraction. To learn the visual information at different scales, we utilize a Feature Pyramid Network (FPN) that combines features at multiple spatial resolutions to conserve fine details and global context for the task context. Next, a Sentence-BERT-based task encoder is used to map the task descriptions into semantic representations. This fills the gap between vision and language. A transformer module then fuses the visual features derived by YOLO and FPN with the semantic embeddings of the task encoder. The module outputs a joint representation of visual and semantic context. The representation is decoded into saliency maps that highlight the regions based on the task, connecting model attention with high-level instructions. In our future work, we aim to improve the performance and accuracy of the model.

Although early, the architecture is promising to be a feasible alternative to current top-down approaches by combining object detection with semantic encoding.

2 Related Work & Review

For a long time, visual saliency has been a core issue in computer vision. Until recently, bottom-up approaches were the primary methods to deal with visual saliency, but later studies started to work on top-down saliency approaches. In this section, more recent work that has driven the field and applied top-down approaches to predict visual saliency is reviewed.

2.1 A Study of Visual Saliency for Free-Viewing and Task-Oriented Condition

Dilara Albayrak's *A Study of Visual Saliency for Free-Viewing and Task-Oriented Condition* analyzes the concept of saliency when varying viewing conditions are present. The thesis uses Generative Adversarial Networks (GANs) to improve the prediction of saliency maps by learning hierarchical representations that cover semantic and context-based information [1].

The data used in our project comes from the data gathered by Albayrak's work. It forms the base of our work by providing ground truth saliency maps based on four different high-level task definitions.

2.2 Visual Saliency Transformer

The Visual Saliency Transformer (VST) by Liu et al. uses a transformer-based backbone for saliency detection to overcome the limitations of conventional CNNs. VST tokenizes images and uses multihead self-attention for global interdependencies. It adopts a reverse T2T token upsampling to form a high-resolution output and a joint learning-based multitask decoder to predict saliency maps with boundary maps, utilizing task tokens and patch-task attention. Experiments demonstrate the superiority of performance compared to other state-of-the-art CNN-based architectures [2].

This paper inspired our model to integrate transformers, but in a slightly different way, by creating a transformer fusion module to merge visual features with textual task embeddings while leveraging YOLO's future extracting abilities.

2.3 Top-Down Saliency Detection Driven by Visual Classification

In *Top-Down Saliency Detection Driven by Visual Classification*, the idea of generating saliency maps through classification is introduced by the authors. They establish high-level information for tasks that can effectively guide the identification of regions most relevant for classification [3].

2.4 Visualising Image Classification Models and Saliency Maps

Simonyan et al. introduced a gradient-based method for visualizing image classification network saliency. Their work consists of calculating the derivative of the class score with respect to input pixels, which indicates which regions inside the image have the largest impact on a CNN's output. While this technique is centered on general interpretability as opposed to explicit task-oriented cues, it serves as the foundation for gradient-based methods. [4].

2.5 Discussion

Overall, the reviewed works highlight the effectiveness of incorporating high-level, task-specific definitions into saliency prediction. For example, Dilara Albayrak's thesis illustrates that incorporating task-oriented viewing conditions can enhance saliency prediction [1]. Similarly, the work on Top-Down Saliency Detection Driven by Visual Classification [3] shows how utilizing classification can guide the detection of regions most relevant for a defined task. Additionally, the Visual Saliency Transformer [2] demonstrates that transformer architectures can capture the global dependencies necessary to integrate these high-level task definitions into saliency maps.

Motivated by these findings, our TDSP model integrates a YOLO-based feature extractor with a Sentence-BERT task encoder and a transformer fusion module. This design produces saliency maps that are precisely aligned with various task requirements. Unlike methods that focus solely on generic cues or pure transformer approaches, our model combines object detection with high-level semantic encoding to capture task-driven visual attention in a balanced manner.

3 Proposed Method

Our goal is to create a top-down saliency prediction model that identifies visually striking regions to the human eye based on a specific task. By integrating a pre-trained YOLO-based architecture for feature extraction with a Sentence-BERT-based task encoder, our methodology bridges the gap between visual features and semantic task definitions. The YOLO backbone performs well at extracting visual cues, and the Sentence-BERT encoder transforms high-level text-based task descriptions into semantic embeddings. Then the transformer-based fusion module processes these two components together to create a connection between extracted features based on tasks. Consequently, the model benefits from the strengths of both object detection and contextual understanding, resulting in more accurate task-specific saliency maps.

3.1 Dataset & Preprocessing

The dataset for this project has been collected by Dilara Albayrak for her work in *A Study of Visual Saliency for Free-Viewing and Task-Oriented Condition* [1]. The dataset consists of five folders: one folder for original stimuli images and four folders for the results of each task-defined saliency map collected. Task folders contain two subfolders: “fdm” and “heatmap.” The first folder, “fdm”, contains grayscale saliency maps that have been directly collected with an eye tracker, and the “heatmap” folder contains images in which these grayscale maps have been converted to RGB and overlaid on the original stimuli images to enhance data interpretability.

For the preprocessing, both the stimuli images and the associated saliency maps are resized to a fixed dimension (384×384) in our preprocessing pipeline for uniformity during training. The image is then normalized and transformed into a tensor for fast operation in the deep learning framework. Also, we strengthen the model by paired data augmentation. In particular, we perform a paired random horizontal flip and a paired random rotation (which will apply exactly the same transformations to both the image and the corresponding saliency map). The joint augmentation approach retains spatial correspondence between stimuli and ground truth, which is important to maintain the motion consistency required for salient region prediction.

3.2 Model Architecture

As mentioned earlier, the architecture consists of a YOLO-based backbone for visual feature extraction, an FPN to aggregate multi-scale information, a task encoder to generate semantic embeddings, a transformer fusion module that combines visual and semantic features, and a final saliency decoder that produces the output saliency maps. An overview of the complete architecture is presented in Figure 1.

3.2.1 YOLO Backbone

The backbone contains a pre-trained YOLO model to extract high-level features from the images. The backbone is responsible for capturing edges, textures, and object parts that are fundamental for scene understanding. We focused on using the earlier layers of YOLO, which are optimized for object detection, helping our model benefit from its pre-trained general feature extraction capabilities.

3.2.2 Feature Pyramid Network (FPN)

The FPN combines multi-scale visual data by processing features at different resolutions to combine fine details and coarse contextual cues to make the model recognize small and large objects. It minimizes the dimensionality of high-channel feature maps into smaller ones (e.g., 512 to 128 channels) to ensure that the next modules get an effective and well-proportioned set of features with needed spatial information and decreased computational load.

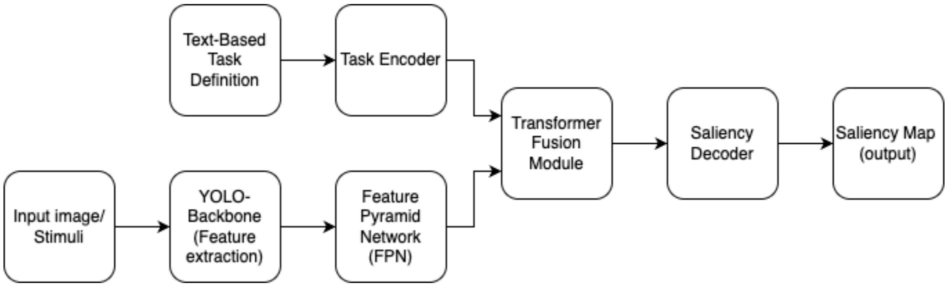
Text-Based Task Definition
Task Encoder
Transformer Fusion Module
Saliency Decoder
Saliency Map (output)
Input image/ Stimuli
YOLO-Backbone (Feature extraction)
Feature Pyramid Network (FPN)

Figure 1: Overview of the model architecture.

3.2.3 Task Encoder

To utilize task-specific information, the task encoder represents task descriptions in dense semantic embeddings. Using Sentence-BERT, the model maps text-based descriptions to vector representations. This is further fine-tuned through a linear transformation to condense the information into a compact representation that can be easily blended with visual features. This bridges the language-vision gap so that the model can recognize and focus on regions in the picture that are meaningful to the given task.

3.2.4 Transformer Fusion Module

This module's purpose is fusing the visual and semantic modalities produced before. First, feature maps outputted from FPN are transformed into a sequence of tokens, flattening the spatial dimensions. The encoded task embeddings, which represent the task definitions, are added at the end of this sequence as additional tokens. The transformer encoder processes these combined tokens, which allows it to learn complex dependencies between visual features and task definitions. The end product is a representation that emphasizes the areas of the image that are critical for that specific task. This fused sequence is then reshaped back into spatial format for further processing.

3.2.5 Saliency Decoder

The saliency decoder translates the fused feature map into a single-channel saliency map. It uses a series of convolutional layers that reconstruct and refine the spatial information. A sigmoid activation function at the output normalizes the resulting map between 0 and 1, representing the probability of each pixel being salient. This transforms the fused representation into an actual output that can be directly interpreted as a task-driven saliency map prediction.

References

[1] D. Albayrak, “A study of visual saliency for free-viewing and task-oriented condition,” Master’s thesis, TED University, 2020.

[2] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, “Visual saliency transformer,” in *Proceedings of ICCV*, 2021.

[3] F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler, “Top-down saliency detection driven by visual classification,” in *Proceedings of the Conference*, 2018.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2014.