Saliency Prediction in 360º Videos with Transformers

Mateo Vallejo, Diego Gutiérrez, Edurne Bernal

Afiliación: Graphics and Imaging Lab (GILab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: mvallejo@unizar.es

Abstract

We present a model for saliency prediction in 360º videos based on the Transformer architecture. Our model leverages the global attention mechanism in order to represent the temporal dependencies that drive human attention. We compare our model with a current state-of-the-art model and outperform it for all metrics measured.

Introduction

Recent advances in virtual reality (VR) technologies have caused a great increase in user adoption. As such, virtual reality is undergoing a surge of interest, with an increasing amount of applications in a wide variety of fields such as education, medicine and entertainment. However, there still exist many challenges related to the creation of immersive content for virtual reality that do not exist in traditional 2D content, since users have control of the camera at all time. For this reason, designers of immersive applications for virtual reality need to develop techniques in order to guide user attention. Therefore it is of great interest to study the mechanisms by which humans perceive and explore virtual environments.

Saliency is one of the main features studied when dealing with the human visual system. Saliency is defined as the quality that a visual stimulus possess for capturing human visual attention. This manifests as gaze fixations that occur while humans view salient regions. However, these regions need to be captured by detecting the gaze fixations of a large number of human participants, which can be a costly process. As such, many works have studied the automatic generation of saliency maps, most successfully through deep learning. Inspired by the recent advances in the field of natural language processing, where the Transformer was introduced to great success by being capable of representing the global dependencies between words, we present a saliency prediction model for 360º videos based on the Transformer architecture capable of representing the temporal dependencies in videos.

Our Model

The architecture of our model is based on the Transformer [1]. More specifically, it follows a structure similar to the Video Vision Transformer [2], although there are some clear differences to the previous models due to the nature of the tasks being performed, as the model generates saliency maps for 360º videos. For every frame in the input video the model needs to generate an output saliency map which corresponds with a probability distribution of the predicted salient regions of the given frame. These salient regions can be interpreted as those which have a higher probability of receiving gaze fixations from users. Figure 1 shows the main structure of the model which can be separated into 3 main parts, token embedding and positional encoding, the transformer encoder, and lastly the output generation.

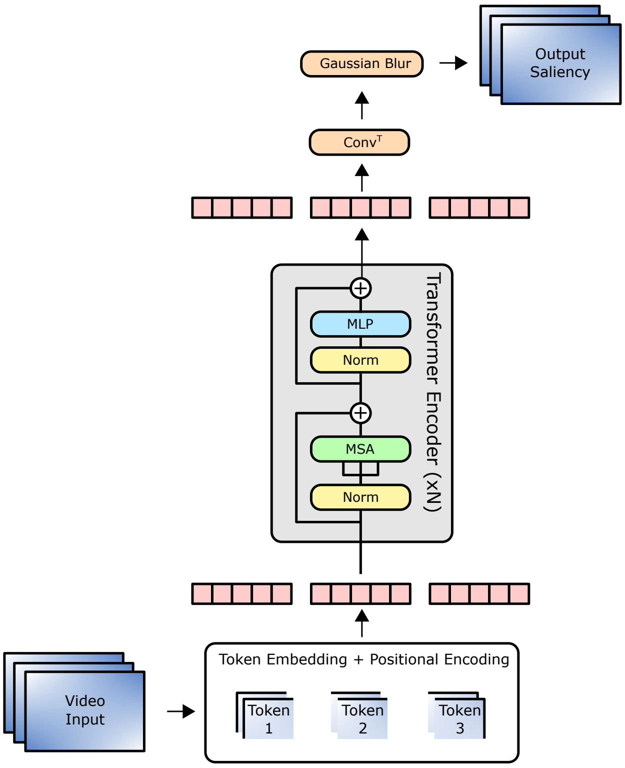

Figure 1: Overview of the proposed model. The input video is transformed into a sequence of tokens which are processed by a stack of transformer encoders which produce the corresponding output saliency maps of the video.

Token embedding is the process by which spatio-temporal patches are selected and projected in order to transform the input video into a sequence of tokens. The Transformer encoder employs stack of six Transformer encoders in order to compute the global attention between all tokens. Each encoder is formed by a Multi-Head-Self-Attention (MSA), a Multi-Layer-Perceptron (MLP) and two residual conections. Once the output tokens are obtained, an inverse convolution is performed in order to transform the tokens into saliency maps for each frame. Each frame is further processed with a Gaussian blur in order to smooth the discontinuities present at the edges of the patches.

Our model has been trained with the VR-EyeTracking dataset [3], applying data augmentation in order to increase the total amount of videos available for training, with Mean-Square-Error (MSE) as the loss function for training.

Results

We employ 4 different metrics commonly used to evaluate saliency models: Similarity (SIM), Pearson Correlation Coefficient (CC), Kullback-Leibler Divergence (KLD), Normalized Scanpath Saliency (NSS) and Receiver Operator Curve (ROC).

All metrics were obtained with test videos from the Sports-360 dataset [4]. We compare the results of our model with those obtained by the CP-360 model, a current state of the art saliency prediction model for 360º videos based on convolutional long short-term memory neural networks [5]. Table 1 shows the comparison between both models. As can be seen, our model obtains better results for all metrics.

Table 1: Comparison of saliency prediction metrics between our model and a current state of the art model, where our model obtains better results in every measured metric.

	SIM↑	CC↑	KLD↓	NSS↑
Our Model	**0.3375**	**0.3048**	**6.3080**	**1.387**
CP-360 [5]	0.2761	0.2338	8.3600	0.9515

Figure 2 shows ROC curves of the ground truth, our model and CP-360. As shown, the curve that corresponds to our model is relatively close to the curve that corresponds to the ground truth, converging to a value of ROC=1 at a selection of around 50% of most salient regions, while the curve that corresponds to the CP-360 model is further away to the ground truth model and converges to a value of 1 at a larger percentage.

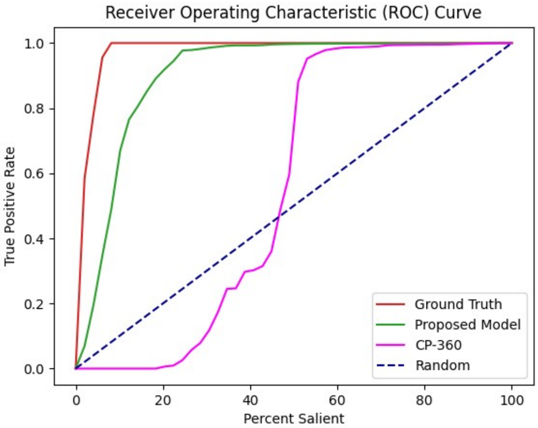

Figure 2: ROC curves of the ground truth saliency maps, predicted saliency maps of our model, predicted saliency maps of CP-360 and random saliency map distributions. The faster a curve converges to a value of 1 the better it is at predicting the ground truth gaze points.

REFERENCES

[1]. VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I.. Attention Is All You Need. *Proceedings of the International Conference on Neural Information Processing Systems.* 2017

[2]. ARNAB A., DEHGHANI M., HEIGOLD G., SUN C., LUČIĆ M., SCHMID C.: ViViT: A Video Vision Transformer. *IEEE/CVF International Conference on Computer Vision (ICCV).* 2021

[3]. XU Y., DONG Y., WU J., SUN Z., SHI Z., YU J., GAO S.: Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017

[4]. HU H., LIN Y., LIU M., CHENG H., HANG Y., SUN M.: Gaze prediction in dynamic 360° immersive videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2018

[5]. CHENG, H., CHAO, C., DONG, J., WEN, H.K., LIU, T.L., and SUN, M.: Cube Padding for Weakly-Supervised Saliency Prediction in 360º Videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2018