 ISSN 2362-0412

Application of Vision Transformers in Online Advertisement Identification

C.R. Liyanage [a*], M.K.S. Madushika [b], R.D. Nawarathna [c]

[a]*Department of Information and Communication Technology, Faculty of Technology, University of Ruhuna*
[b]*Department of Computer Science, Faculty of Science, University of Ruhuna*
[c]*Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya*

*Corresponding author: *ravihari@ictec.ruh.ac.lk*

ABSTRACT

Advertisements(ads) play an important role in many sectors, such as business, education and government as they can influence cultural and religious aspects of a society by disseminating important messages to people. Generally, image-based advertisements are more creative and different from other images as these contain slogans explaining the message of the ad, symbolic and atypical objects and different placements of objects within an image. Identification of advertisements from other images is important on digital media in getting customer attention or blocking them from websites. This study proposes a method to use a supervised learning approach to classify images into ads or not-ads. Another objective of this study is to verify the application of Vision Transformers (ViT) in the domain of image-based ad analysis. ViT is a novel image classification architecture derived similar to the Convolutional Neural Network (CNN), where images are divided into patches and trained using the technique called "Multi-Headed Self Attention". The experiment was conducted using 19,700 images that were labelled as ad and not-ad. Two ViT models with different patch sizes, which were pre-trained on ImageNet-21K dataset were used for image classification. These two models were trained as batches of size 10 for a maximum of 20 epochs. The dataset was split into two main parts as training and testing and set the validation split as 0.2. The highest accuracy of 82% was gained from the 32x32 patch sized model during validation. Moreover, an accuracy of 84%, precision of 85%, and recall of 84% resulted during its testing phase. The results of this study were compared with the state of the art research using CNN. The study has proved that the ViT architecture can achieve comparative results with the limited available computational resources.

Keywords: *Advertisements, Classification, Vision Transformers*