Common Sense for Image Analysis: A Study to Decode Advertisements using Vision Transformers

Chandreen Liyanage
Department of Computer Science
Lakehead University
Thunder Bay, ON, Canada
cliyanag@lakeheadu.ca

I. Introduction

With the significant advancement of deep learning and machine learning that happened during the past years, Artificial Intelligence(AI) systems become more intelligent and less artificial. In other words, the common sense reasoning ability of AI systems has been developed and this enabled AI to be applied widely in many areas. The reasoning skill of an artificial system is described in a way that, it must be able to logically understand and act upon a situation even without prior knowledge [1]. However, the literature explains that the implementation of common sense in perceiving and understanding visuals is not fully established in AI systems [2]. For example, an automated driverless car will not be able to identify the children walking on roads dressed in fancy Halloween costumes [3]. Moreover, consider the following example in Fig 1, where a person cuts tofu in a bowl. However, the AI algorithm has detected this as a knife cutting the bowl, not tofu. This explains, even though the AI algorithms are very good at detecting the objects in an image, they are still struggling with understanding the real object-action relationships due to a lack of logical reasoning. Therefore, it is notable that AI systems are still poor in understanding things around the world as humans due to a lack of common sense.

Logically understanding and interpreting images or visual sceneries is an extremely important research area in AI cognitive skill enhancement and the findings of these studies will facilitate various applications of AI in almost all fields, including manufacturing, health, transportation, business, and education. Nevertheless, understanding the visuals containing symbolic representations or abstract objects will be even more challenging in AI application research. More specifically, automatic advertisement understanding is a similar research area, which is rarely investigated but is still important in enhancing, as it can serve many other sectors.

Advertisements act as powerful manipulators, especially in business, socio-cultural and education contexts as they can handle human choices by stimulating their thoughts. While most of the images combine objects inside a scene, advertisements convey messages to request their viewers to perform some actions by understanding the text and relations among their objects [4]. There are several applications related to advertisement analysis that are beneficial for advertisers to correctly convey their message to viewers, such as identifying the right audience, predicting user preferences by understanding their expressions, and predicting the effectiveness of advertisements by analyzing the ad content.

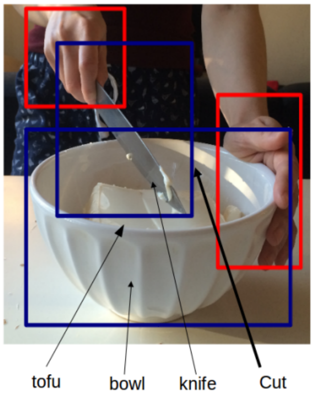


Fig. 1: The Tofu cutting example for explaining common sense: depicted from [2]

Among these, the applications that engage with advertisement content understanding are comparatively complex due to the abstract expressions and symbolic meanings they carry [5]. Moreover, as many creative advertisements convey their messages in abstract form and contain symbolic representations without texts or logos, understanding their meaning will be a critical task. However, as this decoding needs human common sense and reasoning skills, automatic implicit meaning extraction of symbolic advertisements will be an essential but challenging task in the field of computer science research.

Image understanding in computer vision is the ability to identify objects and their actions in an image automatically. Thus far, Convolutional Neural Networks(CNN) were extensively used for various image understanding tasks, including object recognition and image classification [6]. Lately, Vision Transformer, a hybrid version of CNN, has come into the stage by combining neural network layers with a special self-attention mechanism to identify relations in a broader range. [7].

Vision Transformer(ViT) is an image classification architecture derived from Transformers in Natural Language Processing(NLP). This mechanism has proved to obtain better accuracies and efficiencies on larger datasets, and the self-attention mechanisms in ViT avoid the requirement of focusing on nearby elements in an image due to its encoder-decoder architecture [7]–[9]. Therefore, the hybrid CNN version prevents the limitation with fixed-size image vectors as in other traditional neural network architectures. Furthermore, this procedure allows the ViT to collect details from the entire image, including its lowest layers by splitting the images into smaller elements called ”tokens” [10].

Although there were several studies conducted to analyze image advertisements using CNN, their results were not adequate due to the inability to capture the diverse meanings of symbolic objects and their complex relationships. Therefore, this study proposed to utilize a different technology, the Vision Transformers, to investigate their capability of increasing the accuracy of symbolic advertisement understanding. By automatically decoding such advertisements, this study will help to accurately convey messages concealed behind a set of symbolic objects in an image.

The rest of the sections in the report is organized as follows. The section II summarizes the five objectives of this proposed work. In section III, a brief review of the literature on the areas of symbolic image understanding and applications of Vision transformers in image analysis is presented. Next, the proposed solutions section explains the techniques and approaches that can be used to perform this work, indicating how the data collection and analysis can be performed. In the end, a conclusion from this proposed research study will be presented.
Word Count in Introduction: 847

II. Objectives

The main objective of this proposed study is to develop a model to decode the implicit meaning of image advertisements based on their symbolic contents using Vision Transformers. Ultimately, this study focuses on enhancing the common sense skills of AI systems in understanding and interpreting images. The specific objectives to be achieved within the study are;

1) To develop a comprehensive dataset with a list of symbols, advertisement images, and their corresponding annotations.
2) To interpret the unspoken meanings of different symbolic objects in an image advertisement.
3) To develop a model to understand the interrelation between symbolic objects and image segments to predict the topics of an image advertisement.
4) To enable the model to perform question and answering.
5) To investigate the performance of different Vision Transformer models in the symbolic image understanding task.

This proposed study will be built upon accomplishing the above five objectives, where the first target is to build a large dataset with image advertisements, both from products and services. Annotating the collected images will also need to be done to make it a labeled dataset. The second objective of the study is to understand the symbolic meanings of the objects in the image. The results of this attempt will help to identify any abstract object or objects with a different appearance from normal, that are called “atypical objects” in any type of image in other applications. Thirdly, the relationship between the objects and regions in an image will be explored. This helps to extract the meaning of the whole image so that to predict the topic of the advertisement. These results can be later applied to any other application for understanding or automatically annotating the images. The fourth research question is related to question and answering performed on the images, where the questions fed into the model, for example, in Fig1 if a user asked ”what is person cutting?” this model should be able to answer ”The person is cutting Tofu”.Finally, the results of several significant ViT models will be compared to analyze their performance in symbolic image understanding.
Word Count in Objectives: 350

III. Existing Works

Automatic image understanding is the process of identifying the content and annotating the images automatically. Although machine learning-based image understanding applications have been developed, due to a lack of knowledge on integration and common sense reasoning, the methods in the context still undergo limitations.

A. Image understanding through Neural Networks

The application of deep neural networks to improve the common sense of AI models for understanding and interpreting images is an emerging field in the research domain. A study has been conducted to improve the knowledge of spatial common sense by using a novel type of deep neural network called “Geometry-Aware Recurrent Network” [11]. This attempt enabled AI Systems to predict any dynamic visual sceneries by changing the number of objects, configurations, and appearances. The researchers of another study have combined computer vision techniques with techniques in natural language processing to enhance knowledge gathering through visuals and text [2]. They specifically used semantic parsing in NLP and common sense reasoning techniques to enhance the visual reasoning skills of an artificial system. As this study stated, the CNN and Recurrent Neural Networks(RNN) are not optimized in improving the high level of reasoning skills in visual understanding.

B. Transformers for image analysis

In recent years, the concept of Transformers was introduced through language translation. This transformer model was completely built on a self-attention mechanism by replacing the usage of recurrent and convolutional networks. Later, several studies implemented transformers in NLP applications to gain higher processing speeds than neural networks [12]. Recently, this concept was adopted to computer vision as Vision Transformers for many applications, including image retrieval, object detection, and semantic segmentation [6], [7], [10], [13]–[15].

Another study has introduced a non-convolution method called Data-efficient image Transformers(DeiT) with a distillation token and special training strategies [6]. They have performed their experiment on the ImageNet dataset and proved the unnecessity of larger datasets for training with the help of their novel distillation technique. Meanwhile, another work has presented a ViT approach for image classification where they have divided the images into patches and perform supervised learning by considering these patches similar to the tokens in NLP [10]. LeViT is another ViT model introduced very recently by researchers to increase the speed of the image feature extraction task [7]. They have used different datasets with different ViT architectures to experiment with the sense of balance between accuracy and speed. According to their paper, several other promising ViT architectures, such as Bottleneck transformers [16], Visual Transformers [17], Pyramid Vision Transformer [18], Pooling-based Vision Transformer[14] and [19] were comparatively slower than their model.

Recently, some studied have conducted application-based research using image analysis via Vision Transformers. While the usage of CNN based deep learning based architectures are very popular in many application domains, the health sector is extensively using such AI algorithms for a long time. However, with the introduction of encoder-decoder based Vision Transformers, many medical applications have eagerly shifted their experiments towards using them. The study [8] have developed a novel model for breast cancer classification using ultrasound images. They have experimented with different augmentation techniques and weighted loss function to address the common issues found in ultrasound datasets. By comparing the accuracy and area under the curve values gained through the comparison between state-of-the-art CNN models and ViT models, they verified that the performance of ViT models are much higher. The authors in study [9] have developed a model to effectively detect Tuberculosis, which is a popular infectious disease from chest X-ray images using three different machine learning models: EfficientNet, modified version of the original Vision Transformer and a new model of Hybrid EfficientNet with Vision Transformer called, ViT_Base_Efficient_B1_224. From the hybrid model, they have gained significant results of 97.72 accuracy and 100 percent of area under the curve values.

Not only applications in medical domain, Vision Transformers have gained recognition in many other sectors recently. A study has used modern Transformer based neural network models for predicting YouTube advertisement quality [24]. The authors experimented with different types of ViT models and obtained better performance compared to the feed forward neural networks. Furthermore, the study [25] has developed a vision transformer based automatic crack detection on asphalt and concrete surfaces. Besides the many studies that have experimented within similar application using CNN, such as U-Net, this study introduced novelty by using the encoder-decoder ViT architectures with transfer learning and a special loss function called ”differentiable intersection over union”. Through their results, they have proved that combination of U-Net and vision transformers are better than using U-Net models alone by improving the results by 3.8% up to around 61% of accuracy.

C. *Image advertisement understanding*

The advertisement analysis has several applications in the literature. Some of the studies have predicted the effectiveness of advertisements using several measurements, such as click-through rates [20] and the facial expressions of their viewers [21]. While very few studies have analyzed the contents of advertisements to classify them by automatically understanding their topics and meanings [5].

The study [21] has presented a method to optimize the advertisements displayed on outdoor panels by analyzing the pedestrians’ reactions to them. Here, neural networks and SVM approaches were used to analyze the emotions of the person in front of the advertisement panel. Later, these expressions were compared with the contents of the ad to understand people’s needs. The authors in [5] have presented an advertisement content analysis approach based on two frameworks; Question and Answering(Q&A) for image ads and symbolism prediction. In their opinion, both Q&A and symbolism prediction were relatively challenging, and they gained very low accuracies; 11.48% and 15.79%, respectively. For the development of these models, they have used human annotators to collect questions, answers, and topics related to images and 221 different symbols. Although they have investigated an interesting research area, many improvements can be conducted in future studies.

However, all the existing studies related to advertisement understanding and classification have been conducted using different architectures of CNNs [22], [23], and there was no study reported utilizing ViT for this task. Hence, due to the surpassing performances of ViT over CNN, there is an undiscovered research area for analyzing symbolic advertisements using ViT to increase its performance.

Word Count in Existing Work: 1043

IV. Proposed Solution

The methodology for the proposed study will be constructed step by step based on each objective. After that, the major steps of this study will be explained.

The first objective of the study was to design and develop a complete dataset with advertisement images, a list of symbols, and their corresponding annotations. Firstly, the datasets available from existing studies will be collected. Then, more image advertisements with symbolic meanings must be collected through internet browsing. When filtering the images, the images without slogans must be prioritized, as this study is not focusing on the understanding of the text on ads, which is too straightforward in recognizing the meanings. Once done with the image ad collection, the collected set of images must be cross-checked with the images collected from the existing studies for avoiding duplicate images. Next, the images will be annotated, in terms of the topic of the ads, for example food, education, sports, beverages etc. These annotations will help to achieve the third objective. Moreover, identifying the

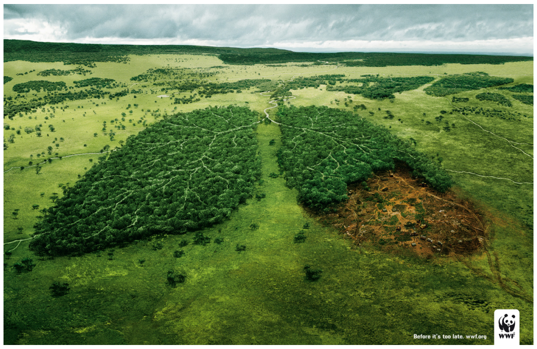


Fig. 2: The famous lungs-forest advertisement

objects and labeling them with their symbolic meanings will need to be done for accomplishing the second objective.

To interpret the unspoken meanings of different symbolic objects in an image advertisement, which is the second objective of the study, a list of objects and their symbolic meaning annotations will be utilized. These annotations, need to be passed into the ViT models together with the image advertisements. As a result of the model, the symbolic meaning of all the objects for a given image should be retrieved.

The third objective, which is to develop a model to predict the topics of an image advertisement, the list of topic annotations must be provided with the images as the input. For achieving this objective, the relationship between the identified symbolic objects and image segments must be identified and that will be automatically done by the ViT models.

The fourth objective of this study is to improve the developed model to answer the questions regarding a particular advertisement. Compared to other simple images, as advertisements carry hidden meanings, providing a way to improve the user experience regarding the ad is critical. Here, the models will be trained using a set of dummy questions gathered from people, such as regarding Fig2, what is the shape of the forest?, why the forest got the shape of lungs?, what is represented by the destroyed lungs? Moreover, the corresponding answers for each question must be collected. Since the size of the dataset must be considerably larger for training a machine learning model, for each image many number of question and answer pairs will be gathered.

Before developing the question and answering model using vision transformers, the performance of existing models for understanding ads will be evaluated. The models used in study [5]; the two layer LSTM to encode the questions will be re-evaluated with the new new dataset.

For the last objective, which is the investigation of the performance of different Vision Transformer models in the symbolic image understanding task, this study plans to conduct experiments with several Vision Transfomer architectures and finalize the model with the most suitable technique by considering both the accuracy and the efficiency. The state-of-the-art ViT models, such as the models listed below will be used.

- Data-efficient image Transformers(DeiT)
- LeViT
- Bottleneck transformers
- Visual Transformers
- Pyramid Vision Transformer
- Pooling-based Vision Transformer
- CvT

Moreover, these models will be trained and tested on different hyperparameter settings, such as the number of image tokens, batch size, epochs, optimizer, loss function and learning rate. The performance of different models will be recorded using confusion matrices and graphs. The basic steps of the proposed research study will be explained hereafter.

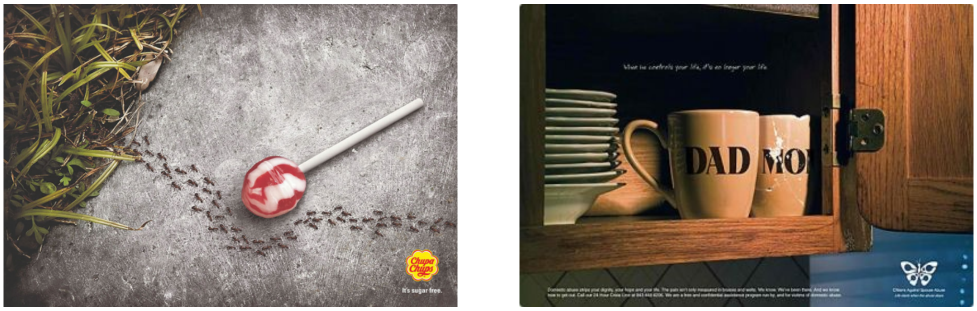
- Step 1: Preliminary Literature Review Firstly, a thorough background survey will be conducted to learn the existing studies related to image classification, advertisement image understanding, modern deep learning techniques for computer vision and Vision Transformers. Simultaneously, the planning of a robust methodology and the collection of the dataset will be conducted.

- Step 2: Design and Develop the Dataset Mainly, there will be two types of advertisements; product ads and public service announcements. Fig 3 shows two examples of symbolic advertisements in these two categories. Hence, as the second step, the study will collect both types of advertisement images with symbolism mainly from repositories of previous studies. The dataset will be further expanded by collecting symbolic advertisement images from social media platforms and other websites. The collected dataset will be a combination of advertisement images and their annotations.

- Step 3: Analyze, Process, and Interpret the Symbolic Objects in Advertisements Expert support will be taken to prepare the annotations for these images. The set of annotations will be; what are the existing symbols, what are their meanings, what are the relations between these symbols and what is the topic of the advertisement etc. Later, a list of symbols and their meanings will be collected with the help of the literature and other reliable sources.

- Step 4: Data Pre-processing After gathering the dataset, the data pre-processing must be done. Through pre-processing, similar image removal, text and logo deletion, noise removal, object enhancement, and image resizing will be done.

- Step 5: Model Development During the fifth step of the study, system development will be conducted. Here, several steps will be implemented to perform different tasks, such as detecting symbols/ objects within the advertisement, identifying their sole meanings and extracting relationships between them. A recent model called



(a)

(b)

Fig. 3: Two examples for product and public service symbolic ads. The left is for a sugar-free product and the right is to interpret domestic violence.

"LeViT" [7] will be used as the initial ViT architecture to develop the model. This model was built on the Visual Transformer [10] architecture and DeiT [6] training model by integrating convolutional networks and self-attention mechanisms. Fig 4 gives the model development steps in brief.

- Step 6: Validation and Evaluation After training and testing the model, the final system will be able to present two types of outputs; the predicted topic of the advertisement and the distinctive meaning of each symbol, objects and their relations within the image. Finally, the model outputs will be validated with the help of experts.
- Step 7: Experiments and Improvements The proposed study can be further expanded to perform the sentiment analysis on identifying how the viewers feel about an advertisement. For this work, together with each ad image, the possible list of sentiments, such as amazed, cheerful, angry, disgust must be collected. This approach can help to improve the user experience

Word Count in Proposed Solution: 1110

V. Conclusion

Improving the common sense of AI systems in interpreting images is an emerging research area in computer science. Moreover, creative advertisement understanding is an important area in the domain where researchers can develop the reasoning capacities of AI models. However, this is a challenging task due to several reasons: Advertisements contain symbolic objects that can carry varying information, different combinations of symbols can represent different meanings, disseminate complex ambiguous meanings, and need common sense and reasoning skills to understand. Due to these reasons, this area of research has not been adequately investigated and the results of the existing works are also not satisfactory. Therefore, the proposed study attempts to address these issues to improve the performance of automatic symbolic image advertisement understanding. Moreover, the finding of this study will be novel as the area of analyzing symbolic images using Vision Transformers is untouched. One of the main advantages of using Vision Transformers in advertisement analysis is that it has more generalization power in identifying objects and symbols in an image compared to state-of-the-art CNNs. The outcome of this research; automatic advertisement classification will be beneficial for business, educational and cultural domains. Understanding ambiguous symbolic representations automatically will help to identify the appropriate audience for advertisements. Moreover, this study will help to convey critical messages indirectly to society through creative advertisements. On the other hand, the user experience of viewers can be increased by describing the implicit meanings of symbols that appeared in advertisements. Finally, the outcomes of this study will help to enrich symbolic image reasoning in the field of Artificial Intelligence and Computer Vision.

Word Count in Conclusion: 270

References

[1] Persaud, Priya, A. Varde, and Ian Drake. "Common sense knowledge, humanoid robots and human rights." IEEE MIT URTC (2016).

[2] Aditya, Somak, Yezhou Yang, and Chitta Baral. "Integrating knowledge and reasoning in image understanding." arXiv preprint arXiv:1906.09954 (2019).

[3] Wilson, James, Daugherty Paul and Davenport Chase. "The Future of AI Will Be About Less." Harvard Business Review (2019): 53-61.

[4] Barra, Silvio, et al. "Visual question answering: Which investigated applications?." Pattern Recognition Letters 151 (2021): 325-331.

[5] Hussain, Zaeem, et al. "Automatic understanding of image and video advertisements." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[6] Touvron, Hugo, et al. "Training data-efficient image transformers distillation through attention." International Conference on Machine Learning. PMLR, 2021.

[7] Graham, Benjamin, et al. "Levit: a vision transformer in convnet's clothing for faster inference." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[8] Gheflati, Behnaz, and Hassan Rivaz. "Vision transformers for classification of breast ultrasound images." 2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). IEEE, 2022.

[9] Duong, Linh T., et al. "Detection of tuberculosis from chest X-ray images: boosting the performance with vision transformer and transfer learning." Expert Systems with Applications 184 (2021): 115519.

[10] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

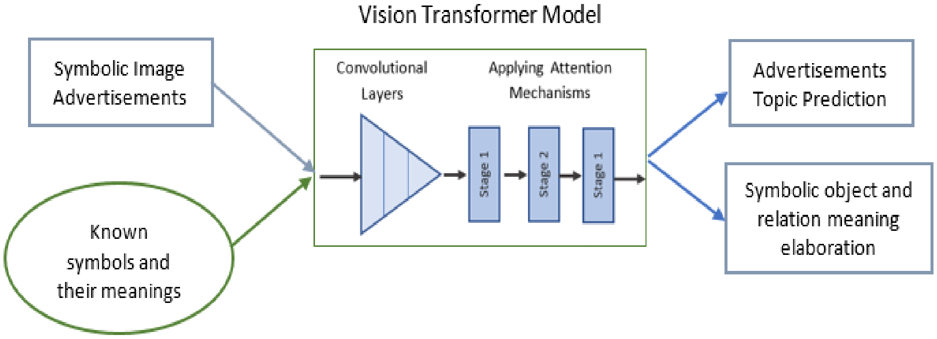

Vision Transformer Model
Symbolic Image Advertisements
Known symbols and their meanings
Convolutional Layers
Applying Attention Mechanisms
Stage 1
Stage 2
Stage 1
Advertisements Topic Prediction
Symbolic object and relation meaning elaboration

Fig. 4: Model development steps in brief

[11] Tung, Hsiao-Yu Fish, Ricson Cheng, and Katerina Fragkiadaki. ”Learning spatial common sense with geometry-aware recurrent networks.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[12] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

[13] El-Nouby, Alaaeldin, et al. ”Training vision transformers for image retrieval.” arXiv preprint arXiv:2102.05644 (2021).

[14] Beal, Josh, et al. ”Toward transformer-based object detection.” arXiv preprint arXiv:2012.09958 (2020).

[15] Zheng, Sixiao, et al. ”Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[16] Srinivas, Aravind, et al. ”Bottleneck transformers for visual recognition.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[17] Wu, Bichen, et al. ”Visual transformers: Token-based image representation and processing for computer vision.” arXiv preprint arXiv:2006.03677 (2020).

[18] Wang, Wenhai, et al. ”Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[19] Wu, Haiping, et al. ”Cvt: Introducing convolutions to vision transformers.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[20] Azimi, Javad, et al. ”Visual appearance of display ads and its effect on click through rate.” Proceedings of the 21st ACM international conference on Information and knowledge management. 2012.

[21] Costache, Alexandru, et al. ”Target Audience Response Analysis in Out-of-home Advertising Using Computer Vision.” 2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, 2020.

[22] Hou, Sujuan, et al. ”Classifying advertising video by topicalizing high-level semantic concepts.” Multimedia Tools and Applications 77.19 (2018): 25475-25511.

[23] Vo, An Tien, Hai Son Tran, and Thai Hoang Le. ”Advertisement image classification using convolutional neural network.” 2017 9th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2017.

[24] Rayavarapu, Vijaya Teja, et al. ”Multimodal Transformers for Detecting Bad Quality Ads on YouTube.” (2022).

[25] Shamsabadi, Elyas Asadi, et al. ”Vision transformer-based autonomous crack detection on asphalt and concrete surfaces.” Automation in Construction 140 (2022): 104316.