ADNet: A Deep Network for Detecting Adverts

Murhaf Hossari*[1], Soumyabrata Dev*[1], Matthew Nicholson[1], Killian McCabe[1], Atul Nautiyal[1], Clare Conran[1], Jian Tang[3], Wei Xu[3], and François Pitié[1,2]

[1] The ADAPT SFI Research Centre, Trinity College Dublin
[2] Department of Electronic & Electrical Engineering, Trinity College Dublin
[3] Huawei Ireland Research Center, Dublin

Abstract. Online video advertising gives content providers the ability to deliver compelling content, reach a growing audience, and generate additional revenue from online media. Recently, advertising strategies are designed to look for original advert(s) in a video frame, and replacing them with new adverts. These strategies, popularly known as product placement or embedded marketing, greatly help the marketing agencies to reach out to a wider audience. However, in the existing literature, such detection of candidate frames in a video sequence for the purpose of advert integration, is done manually. In this paper, we propose a deep-learning architecture called *ADNet*, that automatically detects the presence of advertisements in video frames. Our approach is the first of its kind that automatically detects the presence of adverts in a video frame, and achieves state-of-the-art results on a public dataset.

Keywords: deep learning · advertisements · ADNet · CNN · product placement.

1 Introduction

With the ubiquity of multimedia videos, there has been a massive interest from advertising and marketing agencies to provide targeted advertisements for customers. Such targeted adverts are useful, both from the perspectives of marketing agents and end-users or consumers. The rise of television and online videos has provided several avenues in transmitting brand information to the audience. Several studies [14, 8] have shown that there is a steady growth in the generated revenues from internet advertising over the last decade, and the growth is forecast to continually increase in the coming years.

Traditionally, in online videos, advertisement videos are inserted within the online video content. The advertisement videos are played independently, either prior the online content (pre-roll); or during the online content (mid-roll); or after the online content is viewed by the user (post-roll). The option of fast-forwarding the advertisement video is disabled. Such adverts hugely interrupt the seamless viewing experiences of the viewers [10]. Therefore, nowadays, advertisement placements are integrated directly into the scene. They are popularly

* Authors contributed equally and arranged alphabetically.

known as embedding marketing, or product placements, where specific brands or products are deliberately inserted into the scene [9]. Such placements are mostly manually done via digital editing technologies in the post-processing phase. This involves manually going through all possible frames of the video, and *identifying* candidate video frames for advert integration. The existing adverts are replaced with new adverts, in order to reach out to specific demographic population or markets.

In this paper, we automatically *identify* the frames in a video sequence that have an existing advert. Such an automatic process will greatly help during the post-processing stage, as the video editor no longer needs to parse the video frames manually. Our innovative AI-powered system allows the detection of adverts in video sequences. Such system will greatly expedite the process of post-processing, and save a huge amount of man-hours.

In this work, we focus our attention on detecting billboards in outdoor scene videos. We propose a deep learning architecture called *ADNet*, that automatically detects billboard adverts [4] in a video sequence. Our proposed model is inspired from the state-of-the-art architectures, in creating a bespoke network that can automatically detect existing billboards in a video. The rest of the paper is organized as follows: Section 2 discusses the related work. In Section 3, we describe the ADNet architecture and its layer configuration in a detailed manner. We train the ADNet model using a composite dataset consisting of positive- (images containing a billboard) and negative- (images that do not contain any billboard) examples. Section 4 describes the details of the dataset, and the benchmarking results. Finally, we conclude the paper in Section 5, and mention our future work.

2 Related Work

In the recent years, deep convolutional neural networks have proven to be very effective in the areas of image- and video- processing. More specifically, they have been widely used to tackle tasks related to image classification and object detection. The convolutional neural networks, popularly known as ConvNets are a type of feed-forward neural nets, that produce state-of-the-art results in the areas of visual recognition. There are several reasons for these massive successes in this area. Firstly, the public repository of large-scale image- and video- datasets have helped the machine learning researchers to train the deep neural networks on millions of images [5]. The ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) is a good example [15], that led to the development of several novel neural networks. Secondly, the rise of high computing machines viz. graphics processing unit (GPU) and clusters [4] helped the researchers to train deep networks on very large datasets.

The existing works in advert detection primarily revolve around the detection of advertisement clips in a video sequence. Recently, Hussain et al. in [7] propose

[4] We will use the terms *advert* and *billboard* interchangeably throughout the paper.

a novel solution for automatically understanding the advertisement content, and analyzing the sentiments around them. Feng et al. worked on the detection of logos in television commercials [6], using a combination of audio and video features. Acoustic match profiles are also exploited in [3] to accurately locate the adverts in the video sequence. These works on advert detection do not consider identifying a particular object in the video scene, and subsequently integrating a new advert into the same scene. However, commercial companies viz. Mirriad uses patented technology to insert new objects into a video scene.

In the area of sport analytics, a few works were done to detect the billboards in the soccer fields. Watve et al. attempted to localize the position of on-field billboards using template matching techniques [18]. Cai et al. in [2] used Hough transform for advertising detection in sport TV. Aldershoff et al. used photometric invariant features for billboard track in soccer video scenes [1]. These techniques mainly used techniques from photogrammetry and traditional signal processing – neural networks were not fully exploited. Owing to the recent development in the area of artificial intelligence, we propose a AI-driven advert detection neural network in this paper.

3 ADNet architecture

Our proposed advert-detection model, ADNet (cf. Fig 1) is inspired from the VGG19 architecture. The VGG19 model uses very small convolution filters (3×3), with depth upto 19 weight layers. The VGG19 has 6 different configurations, depending on the number of weight layers. These configurations are denoted by A, A-LRN, B, C, D, and E. More details on this can be found in [16].

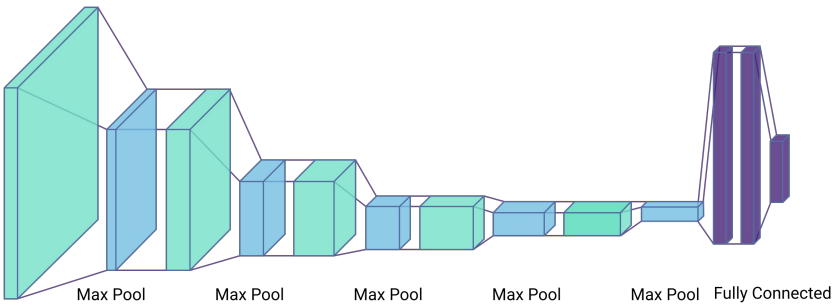

Fig. 1: Architecture of proposed ADNet.

The ADNet uses the pre-trained weights of the VGG network, trained on the ImageNet dataset. We toggle off the training for the first 5 layers of the pre-trained VGG network. We also remove the last 3 weight layers of the original VGG19 model, and add a stack of Fully-Connected (FC) layers. The first FC-layer has 1024 channels with `relu` activation function. In order to prevent the

ADNet configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 X 224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
FC-1024					
Dropout (0.5)					
FC-1024					
FC-2					

Table 1: Layer configuration of ADNet model. The ADNet model is inspired from the VGG19 model – the distinct additional layers in ADNet are indicated in blue color. The convolutional layers are indicated as conv{receptive field size}–{number of channels}. The fully connected layers are indicated as FC-{number of channels}. The dropout layer is indicated as Dropout ({rate of dropout}).

ADNet from over-fitting, we add a dropout layer with a rate of 0.5. The second FC-layer has also 1024 channels with `relu` activation function. The third and final FC-layer has 2 channels, with `softmax` activation function. The two values obtained at the end of ADNet, provide the probabilities of belonging to *billboard* or *no-billboard* classes.

We describe the layer configuration of the ADNet architecture in Table 1 in a detailed manner. Similar to VGG model in [16], we experiment with various configurations of ADNet viz. A, A-LRN, B, C, D, E. As expected, we get the best performance for advert detection with the configuration E, consisting of 19 weight layers. This makes sense, as the neural network becomes more discriminative with deeper layers, and provides better classification performance. In the

subsequent discussion, the default configuration of ADNet is configuration E, with 19 weight layers.

4 Experiments and Results

4.1 Dataset

We train our ADNet model on a *composite* dataset that consists of both positive- and negative- examples of billboard detection. The positive examples are those outdoor-scene images that contain a billboard. The billboard object should cover a *significant* proportion of the image. We set the threshold as follows – the billboard should cover more than 10% of the total image area. The images from this composite dataset are collected from two sources:

- Mapillary Vistas Dataset [13]
- Microsoft COCO (MS-COCO) Dataset [11]

Our composite dataset has a total of 18945 images. Out of these images, 9141 images contain billboard images; and the remaining 9804 images do not contain any billboard. Most of the positive examples are images from the Mapillary Vistas Dataset. We include those images from the dataset, that contain billboard(s) in them. Any images that contain billboard(s), covering less than 10% of the total image area are ignored. We also exclude images, where the billboards are partially off-screen. Most of the negative examples (i.e. images with no billboards) are selected from Microsoft COCO Dataset. We randomly selected 9395 images from MS-COCO that do not contain any adverts.

We divide the entire 18945 images of the composite dataset into training- and testing- sets. Table 2 describes the distribution of images into training- and testing- sets.

Set	**Positive**		**Negative**		**Total**
	Mapillary	MS-COCO	Mapillary	MS-COCO	
Training	6380	-	303	6617	13300
Testing	2761	-	106	2778	5645
Total	9141	-	409	9395	18945

Table 2: Distribution of training- and testing- sets, with respect to the two datasets.

4.2 Subjective Evaluation

In this section, we provide a subjective evaluation of the advert detection using ADNet model. Our large-scale composite dataset contains a balanced number of positive- and negative- billboard examples. Figure 2 shows a few visual results.

(a) Positive examples of adverts correctly identified by ADNet.

(b) Negative examples of adverts correctly ignored by ADNet.

(c) Examples of misclassification errors.

Fig. 2: A few visual illustrations of ADNet performance.

Figure 2(a) shows a few sample results that are correctly classified by the ADNet model. All of them are outdoor scenes that contain billboard(s) at prominent location of the scene. Figure 2(b) are a few images that do not contain adverts, and are correctly ignored by ADNet. Finally, we show a few misclassification errors of ADNet in Fig 2(c). Most of these images contain objects that are similar in shape to a regular four-sided billboard. We observe that the back of the truck, the solar cell stand and the road signage has similar shape and appearance to a regular billboard. Therefore, the ADNet model predicts that these images contain regular billboard in them.

4.3 Objective Evaluation

In order to provide an objective evaluation of our model, we benchmark the performance of ADNet model with the Inception model. We used Inception-v3 and used the pre-trained weights that were obtained by training the model for ImageNet Large Visual Recognition Challenge 2012. This recognition challenge involved classifying the images into 1000 predefined classes. The Inception-v3 network consists of several blocks of 'inception' layers, which are a combination of multiple convolutional and pooling layers [17]. We use the keras-implementation of Inception-v3 that uses 311 layers. We froze the training for the first 172 layers, and toggled the training 'on' for the remaining layers. We also added an average pooling layer, and a fully connected layer with 1024 channels (FC-1024). Finally, we added a fully connected layer with 2 channels (FC-2) with `softmax` activation. Similar to ADNet, this Inception-v3 model provides us two values that indicate of probability of *billboard* or *no-billboard* classes.

We trained our ADNet model on the composite dataset of 18945 images, for 50 epochs with stochastic gradient descent optimizer. We set the learning rate as 0.0001, with a batch size of 16, and using categorical cross-entropy loss function. We used the same configuration file to train the Inception-v3 model on the composite image dataset.

In this paper, we report the classification accuracy of ADNet. Suppose TP, TN, FP and FN denote the true positive, true negative, false positive and false negative samples of our binary classification of billboard detection. We define the classification accuracy for this task as:

$$\text{Classification Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

Table 3 describes the performance evaluation of our proposed model, with the state-of-the-art Inception model.

Approach	**Classification Accuracy**
Inception-v3	0.56
Proposed model ADNet	0.94

Table 3: Performance evaluation of our proposed model in detecting adverts.

The ADNet model has a competitive classification accuracy of 0.94, as compared to the benchmarking approaches. Moreover, it has a competitive processing speed on video frames. In a Linux computer with a GPU configuration of Nvidia GTX1080Xtreme D5X 8GB, the computational time of ADNet during testing stage is 57ms for every frame in the video. Such competitive results on real-life video frames, opens up the possibility of deploying such detection model in a real-time advert creation system [12].

5 Conclusion

In this paper, we have proposed a deep neural network called ADNet, that can automatically parse the frames of a video, and identify the frames that contain an advert. This is useful for the advertisement agencies and marketers, to provide targeted advertisements for specific markets and demographics. The ADNet can help the user, in identifying the key frames in the video where new adverts can be integrated. Our approach is the first of its kind that uses deep neural networks for detecting adverts in video frames. In our future work, we plan to extend ADNet by relaxing the criterion of outdoor scene images. We will explore its possibility of detecting adverts, for indoor-scene images viz. movie and reality television shows.

Acknowledgement

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

1. Aldershoff, F., Gevers, T.: Visual tracking and localization of billboards in streamed soccer matches. In: Storage and Retrieval Methods and Applications for Multimedia 2004. vol. 5307, pp. 408–417. International Society for Optics and Photonics (2003)
2. Cai, G., Chen, L., Li, J.: Billboard advertising detection in sport tv. In: Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on. vol. 1, pp. 537–540. IEEE (2003)
3. Covell, M., Baluja, S., Fink, M.: Advertisement detection and replacement using acoustic and visual repetition. In: Multimedia Signal Processing, 2006 IEEE 8th workshop on. pp. 461–466. IEEE (2006)
4. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Senior, A., Tucker, P., Yang, K., Le, Q.V., et al.: Large scale distributed deep networks. In: Advances in neural information processing systems. pp. 1223–1231 (2012)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
6. Feng, Z., Neumann, J.: Real time commercial detection in videos (2013)

7. Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Kovashka, A.: Automatic understanding of image and video advertisements. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1100–1110. IEEE (2017)
8. IAB, P.: Iab internet advertising revenue report 2011 full-year results. Tech. rep., Market research report, Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PwC). http://www. iab. net/media/file/IAB_Internet_Advertising _Revenue_Report_FY_2011. pdf (2012)
9. Karniouchina, E.V., Uslay, C., Erenburg, G.: Do marketing media have life cycles? the case of product placement in movies. Journal of Marketing **75**(3), 27–48 (2011)
10. Li, H., Lo, H.Y.: Do you recognize its brand? the effectiveness of online in-stream video advertisements. Journal of Advertising **44**(3), 208–218 (2015)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
12. Nautiyal, A., McCabe, K., Hossari, M., Dev, S., Nicholson, M., Conran, C., McKibben, D., Tang, J., Wei, X., Pitié, F.: An advert creation system for next-gen publicity. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) (2018)
13. Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV. pp. 5000–5009 (2017)
14. PricewaterhouseCoopers, L.: Iab internet advertising revenue report-2012 full year results. Bericht, Interactive Advertising Bureau (2013)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
18. Watve, A., Sural, S.: Soccer video processing for the detection of advertisement billboards. Pattern Recognition Letters **29**(7), 994–1006 (2008)