Combination of 1D CNN and 2D CNN to Evaluate the Attractiveness of Display Image Advertisement and CTR Prediction

Wee Lorn Jhinn, Poo Kuan Hoong and Hiang-Kwang Chua

Abstract With the explosion of digital data nowadays, it has catapulted the usage of data analytic in the emergence of digital advertising space. One of the digital advertising giants, Facebook has accelerated the growth of this digital data volume as they are the most common used platforms for advertisers to advertise and deliver advertising messages to the mass audience online. However, this phenomenon has increased the challenges faced by advertisers to further attract audiences attentions to look at the digital advertisements when they are shown advertisements in Facebook platforms. Hence, in this paper, we proposed a method to evaluate and analyze the elements of attractiveness within the display advertisement in Facebook Advertisement platform by applying the 2D CNN on the display advertisement images while 1D CNN on the click metric data respectively. Based on our experiment results, we are able to predict the display CTR with a reasonable margin of error.

1 Introduction

In recent years, digital advertising has turned into a multi-billion dollar industry and begun to show tremendous impact in terms of generating a notable amount of revenue for ones company or raising general awareness to the public. Generally, there are mainly two types of digital advertising methods, namely: display advertising and search advertising. It is often a struggle for advertisers, particularly, to decide the placement type of the advertisement, in order to serve a better fit to the campaign objectives. Search advertising is known to provide a better chance of approaching

W. L. Jhinn (✉) · P. K. Hoong · H.-K. Chua
Axiata Digital Advertising (ADA), Level 32, Axiata Tower, Jalan Stesen Sentral 5, Kuala Lumpur Sentral, 50470 Kuala Lumpur, Wilayah Persekutuan, Malaysia
e-mail: lornjhinn.wee@ada-asia.com

P. K. Hoong
e-mail: poo.kuanhoong@ada-asia.com

H.-K. Chua
e-mail: george.chua@ada-asia.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_11

a potential customer more easily and thus, creating an opportunity of penetration throughout the emulation with multiple competitors in the digital advertising platform. Nonetheless, it is worth noticing that the search advertising only retains this customer penetration advantage under a circumstance, where a customer knows well in what category it is aiming for. On the other hand, display advertising focuses more on driving brand awareness. Since display advertisement can target specific audience segments based on the predefined demographic, locations, and the keywords within the surfed internet content, it can help to attain first impression. It has been proven that the first opportunity to see contributes 73% of the short-term sales effect of advertising [1].

Furthermore, it is forecasted by Cisco that 80% of the contents in the internet are going to be occupied by image based contents and it is strongly believed that this will be growing exponentially until year 2022 [2]. Consequently, the aesthetic design and context in an image for the digital display advertisement begin to play an essential role in capturing ones attention. Besides a well execution on an advertisement campaign, a generally correct interpretation for the display advertisement message based on the visual context is crucial in order to stimulate a series of desired post action (e.g.: conversion and brand awareness). It is believed that the image can generate a stronger impact in online advertising [3]. Moreover, the rise of big data era in digital advertising has created a revamp for the Moores law effect, where there is a massive yet swift incrementation in digital advertising data volume, in which is eagerly demanding a more sophisticated architecture design on computation hardware, specifically Graphical Processing Unit (GPU) to optimize the computation time needed by different algorithms for evaluating the advertisement campaign performance. Subsequent to that, more prospective algorithms are needed to extract a deeper insight of the data while fully leveraging the GPU processing power.

Although Human Intelligence (HI) may unfold some undiscovered attractive elements in a display image occasionally and evaluate the effectiveness of the display advertisement, there are approximately millions of historical data that are being generated simultaneously in different regions each day. With the assistance of Deep Neural Network (DNN) technique, it helps to reduce the burden of HI in evaluating over a huge amount of the same category image data to verify a similar effectiveness on the similar display advertisement. However, the challenge faced is that how optimum can the DNN technique be "taught" in order to evaluate the effectiveness of the display advertisement by extracting the attractive elements in an image data, humanly.

Therefore, in this paper, we propose a DNN method by applying 2-Dimension (2D) Convolutional Neural Network (CNN) and 1-Dimension (1D) CNN on the Facebook display images click metrics dataset, respectively, to predict the Clickthrough Rate (CTR) as CTR has been one of the popular measures for evaluating the efficiency of online advertising [4]. It is noted that the proposed method is mainly focusing on optimizing the CTR regression value as the good CTR for different advertisement categories varied in the applied Facebook dataset. Thus, to avoid any induction of human biasness (predefined CTR threshold) towards a good or bad CTR, Mean Square Error (MSE) loss will be used in this study as the evaluation metric to evaluate the predicted CTR value derived by the proposed method. It is noted that a good or

bad CTR should not be defined based on whether the CTR is higher or lower than the predefined CTR threshold value in the dataset but it should follow the average CTR of different advertisement categories as a threshold respectively.

2 Related Works

There have been several proposed techniques that applied CNN to CTR prediction based on the click metrics data. In 2007, Richardson et al. showed the effectiveness of logistic regression model on predicting how likely is the search advertisement will be clicked [5]. In 2015, Liu et al. proposed a CNN based CTR prediction model, where it can handle a varied length types of input instances as each sample contains a derived one-hot encoded vector based on the click metric data [6]. Nonetheless, this leads to a high dimensionality and sparse input space for the CTR model [7]. In 2016, Qu et al. proposed a Product Neural Network (PNN) that is able to capture the high order feature interaction by including a product layer after the embedding features are generated [8]. Meanwhile, in 2018, Patrick et al. presented a random multiple sequence embedding feature vector combination to study the influences of the sequential characteristic in embedding feature towards CTR prediction [9]. Generally, these methods shared a similar model structure design, where the embedding layers are firstly adopted to derive a dense representation from sparse features and followed by attaching a Multilayer Perceptron (MLP) layer to learn the relationship between the different combinations of features [10]. Zhou et al. introduced a DIN (Deep Interest Network) and an efficient mini-batch aware regularizer where DIN serves to learn the significant representation of user interests based on the historical behaviors w.r.t specific display while the mini-batch aware regularizer technique applies the L2-norm regularization on parameters that generate non-zero features in order to speed up the training process in industry level deep networks [10].

As for the studies on display images, Cheng et al. analyzed and reported a positive impact on CTR prediction accuracy by integrating the multimedia features from display images into the historical click metrics data [11]. Subsequently, Chen et al. in 2016 proposed a DeepCTR, which is the combination of ConvNet (17 layers proposed CNN architecture model) on display image and BasicNet (one-hot-encoding MLP) to reduce the dimensionality on one-hot encoded basic click metrics data. The CombNet is said to have the ability in learning complex yet effective non-linear from these two features [12]. In recent research, inspired by Chen et al., Michel et al. proposed an alternative method with a similar CNN architecture by applying ImageNet 1000 class log probabilities as CNN input to explore the influence on CTR classification [13].

3 Proposed Methods

In this section, the architecture design of the proposed 2D CNN and 1D CNN are illustrated and explained accordingly. Following that, we introduce several types of the global image feature extraction techniques that are adopted into this study. Based

on our experimental results, the extracted global image features from the display images do help in improving the proposed CNN model CTR prediction value.

3.1 2D CNN Network Architecture

The design of the 2D CNN is mainly inspired by the neural network architecture in [12, 13]. With the similar view as [12] that the trade off in between the training duration and performance must be considered. It is noted that with the consideration of overall performance, we did not build a very deep network architecture. Besides, we found that the fine-tuned transfer learning model such as ResNet for display advertising [13] tends to have a limitation on extracting the useful high-level image features from the cartoonized display images in our Facebook dataset. This is due to the fact that most of the state-of-the-art transfer learning models were developed mainly to classify only the real-life objects. Therefore, it is imperative to build a customized 2D CNN network, which is able to help in extracting not only the real-life objects, but also partially the "cartoonized" display image features that would be able to ameliorate the learning in predicting the CTR value. Figure 1 illustrates the designed architecture.

In this network, it consists of 13 convolution layers. The first convolution layer is a 5×5 convolution kernels in order to generate a higher-level feature map on the first hand. After that, there are three blocks of convolution operations, where each block contains of four convolution layers with a 3×3 convolution kernels individually. Subsequently, each convolution block operation, a 2D max-pooling is performed to down-sample the input image by summing up the activation matrix based on two hyperparameters known as: filter and stride, where the filter, F is set to have a standard value of 2 in order to sum up the feature values in each 2×2 matrix region without overly destructing the potential useful image features while stride, S with a standard value of 2 to down sample every depth slice of the activation output matrix. The updated size, $W \times H \times D$, for the image activation output matrix can be defined as:

$$W_i = \frac{W_{i-1} - F}{S} + 1 \tag{1}$$

$$H_i = \frac{H_{i-1} - F}{S} + 1 \tag{2}$$

where D remains to be the same three image dimension.

3.2 1D CNN Network Architecture

Unlike most of the state-of-the-art techniques that convert the basic click metrics data into an embedding features vector. In this paper, the proposed 1D CNN fully leverages the numeric correlation between these non-linear click metrics to extract the linear

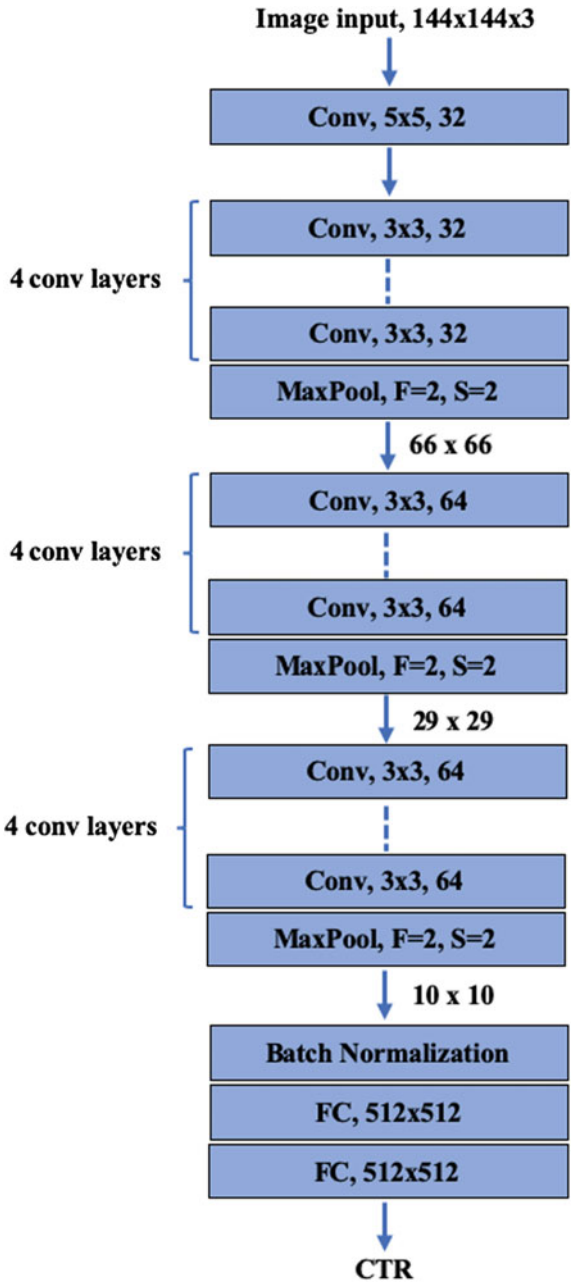

Fig. 1 The proposed architecture of the 13 convolution layers

relationship. Moreover, the striding and filtering properties of the CNN architecture contribute to a vast reduction in training parameters as compared to the general linear regression model. Consequently, the model convergence speed can be greatly accelerated, which is essential to perform a fast iteration on parameters fine-tuning, especially on a large volume dataset. It is worth mentioning that the Facebook dataset used in this study has a significant different sequential characteristic compared to the datasets in the other related work techniques. In [6, 9, 10, 12], they have the advantage of first party data, where the time sequential relationship between click metrics based on each customers historical data is apparent. However, for our Facebook dataset, which served as a second party data, contained only the accumulated value for each corresponding click metrics data, hence losing the valuable insight on sequential click metrics information of each customer. Therefore, 1D CNN can be solely acts as an alternative to predict the CTR based on the large-scale second party dataset.

Given a list of Facebook dataset input instance with n elements (metrics), denoted as $S_{i_{n=1,2,3...74}} \varepsilon R^{d \times i}$, the input instance matrix, x can be formed as below:

$$x = \begin{bmatrix} \vdots & \vdots & \vdots \\ S_i & \dots & S_n \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{3}$$

Note that the element values in the x have been pre-processed by applying a min-max normalization to produce a normalized matrix, z, in order to reduce the sparse variations in between different input instances:

$$Z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4}$$

After that, the convolution process begins by convolving with the weight matrix, $w \varepsilon R^{i \times n}$, to generate the final activation output matrix, r in a one-dimensional way. To be specific, given the receptive field, $w_f \varepsilon R^{i \times n}$ of height in between $w_f = \{32, 64, 128\}$ instead of i and the number of kernel for each convolution, $k_{l=1,2,3}$ in between $2 \leq k \leq 10$, where l represents the lth designed convolutional layer, the lth layer's activation matrix output, z_l is computed by applying the non-linear activation function known as Rectified Linear Unit (ReLU) during the convolutional operation (Fig. 2):

$$Z_l = ReLU(w_f \cdot z_{l-1}) \tag{5}$$

where ReLU function only accepts the positive z_i as the features while denoting the negative or zero value found in z_i as 0 during the convolution operation. After each convolution operation, a 1D max-pooling will be performed to down-sample the input instance. The updated output volume, V for each layer of z_l can be defined through the calculation in below:

Total Facebook Data Records	130, 000
Total Unique Image	2,557
Facebook metric data (Total 74 types of features)	account_currency, impressions, clicks, post_reach, reach, spend, page_engagement, frequency, ctr, link_click, comment etc.
Image global features	brightness, saturation, colorfulness, naturalness, grayscale
Brightness	Average: 136.611 Standard Deviation: 61.179 Max: 255 Min: 3.262
Saturation	Average: 0.392 Standard Deviation: 0.249 Max: 1 Min: 0
Colorfulness	Average: 136.611 Standard Deviation: 61.179 Max: 255 Min: 3.262
Naturalness	Average: 0.502
Grayscale	Average: 62.811

Fig. 2 Summary of the applied Facebook dataset

$$V_{update} = \left(\frac{n - filter}{stride} + 1\right) \times w_f \tag{6}$$

The overall architecture of 1D CNN is illustrated in Fig. 3.

3.3 Image Global Features

Despite of the combination in between the 2D CNN and 1D CNN to improve the learning process between the complex non-linear features, some global image feature calculations are adopted to enhance the CTR prediction by establishing a direct correlation between the image features and click features. The applied global features are listed in the following:

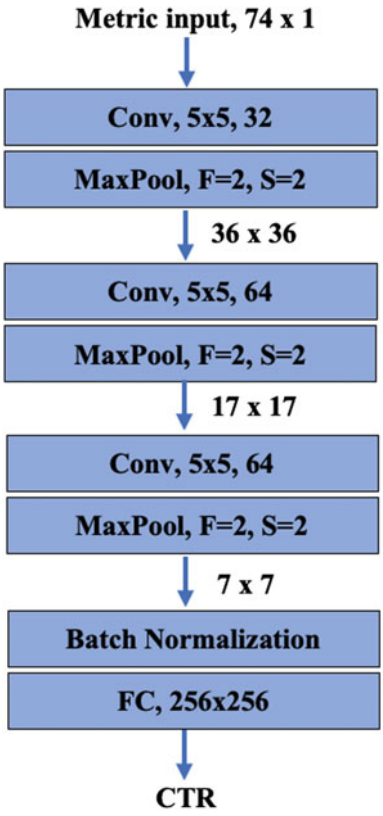

Fig. 3 The proposed architecture of the 3 convolution layers

- ***Brightness***: The image brightness value can be obtained directly from two color spaces known as: YUV and HSL, where the brightness in YUV is the luma component, denoted as Y channel among the YUV and the lightness, denoted as L in HSL. The average, standard deviation, maximum, and minimum of the brightness values of each display image are computed and added into the click metrics data.
- ***Saturation***: The image saturation indicates the vividness of an image. However, the image saturation features can be extracted directly from the HSL or HSV color space models. The average, standard deviation, maximum and minimum of the saturation values are added into the click metrics data.
- ***Colourfulness***: The image colorfulness features, C measures the differences against gray scale colours through the computation in RGB color space [14].

$$C = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{7}$$

where,

$$rg = R - G \tag{8}$$

$$yb = \frac{R + G}{2} - B \tag{9}$$

- ***Naturalness***: Naturalness denoted as N, measures the correspondence degree in between the images and human visual perception. This technique is proposed by Huang et al. [15] by grouping the pixel values based on the threshold of $20 \leq L \leq 80$ and $S > 0.1$ in the HSL color space and followed by defining the pixel values into three categories namely: (1) Skin, (2) Grass, and (3) Sky as the quantitative description. The final naturalness score is defined as:

$$N = \sum_{i} NS_i \times NP_i,\ i\ \varepsilon\ \{Skin, Grass, Sky\} \tag{10}$$

  where NP indicates the proportional pixel of the corresponding categories.
- ***Grayscale***: The grayscale value, G of the image is computed in the RGB channel by:

$$G = 0.299R + 0.587G + 0.114B \tag{11}$$

The standard deviation of the grayscale features is computed and added into the click metrics data. Javad et al. showed that this grayscale level features are very effective in predicting the CTR of advertisement [16]. Thus, there is an additional of 14 image global features are added into the click metrics data.

4 Experiments

The proposed model is evaluated in this section. Firstly, the Facebook dataset used in this study is explained progressively. Follow by illustrating the details of the experiment setup for model training on Facebook image data and evaluation metrics. Lastly, the visualization on the image salient features based on the proposed 2D CNN and the comparison results of between the proposed 1D CNN with global image features and without global image features are depicted with illustrations. The overall summary of the Facebook dataset is shown in Fig. 2.

4.1 Experimental Setup

In this study, the Facebook dataset with the total of 2,557 unique display advertisement images from the 130,000 click metrics were used. For display advertising, it is a common practice to create multiple advertisement campaigns with the same group of images in order to serve the same objective such as: raising public awareness

to certain topic or triggering post conversion for instance: click on the button or hyperlink to drive the website visitation rate, purchase of item or sign up for the campaign event. As mentioned before, the characteristic of the Facebook dataset is different compared to the other datasets in terms of the sequential data where a total of 130,000 metric data were taken for the period of 6 months ranging from August 2018 to January 2019. The selected Facebook dataset varies in regards to CTR based on the different click metrics data but with the same corresponding single display image.

Nevertheless, a low CTR display advertisement does not fully imply low attractiveness level of the display image as it is partially affected by the period (e.g.: weekly or monthly) of the advertisement campaign. Therefore, it will lead to a different CTR value for different campaigns, even though these campaigns served under the same objective. To address the problem of different CTR of the same image during the 2D CNN training, the varied CTR values on a single display image are averaged before passed into the 2D CNN model. On the other hand, the 130,000 metrics data with global features are passed instantly into the 1D CNN model without averaging the CTR as each metrics data has a different features value combination leading to different CTR compared to image features.

In this experiment, the Facebook dataset is divided into 90% for training and 10% for testing respectively. During the training, the Adam optimizer is adopted with a learning rate of 0.01 and batch size of 32 to prevent the overfitting in both CNN models. The predicted CTR value is evaluated based on MSE. A NVIDIA Geforce GTX 1050 Ti with Max-Q Design GPU processor that contains a 4GB memory is used to conduct these experiments.

4.2 2D CNN Experimental Results

From Fig. 4, it illustrates the saliency maps of three test display images after the proposed 2D CNN is applied. Throughout the experiment, it is showed that the saliency map of the pre-trained 2D CNN with the image size of 144 (second column), denoted as 144-saliency map, displays more robust salient features compared to the saliency map with the image size of 192 (third column), denoted as 192-saliency map that induced the image features as noise inadvertently. For instance, it can be observed that the 144-saliency map on the first image visualizes the little boys head at a superficial level while visualizing more at the table region. Meanwhile, the 192-saliency map overly visualized the region at the background yellow color wall. As for the second image, the 144-saliency map protrudes the 8 GB text and the person while the 192-saliency map specifies in terms of visualizing the “8” from the “8G”. Although the saliency maps of both image size in the third image seemed to be displayed at the similar region, it is observed that the visualization on the cartoon character in 144-saliency map tends to show a thicker focus region compared to 192-saliency map. Therefore, through this empirical test, we have concluded to train the 2D CNN with the image size of 144 in order not to extract the noise image features

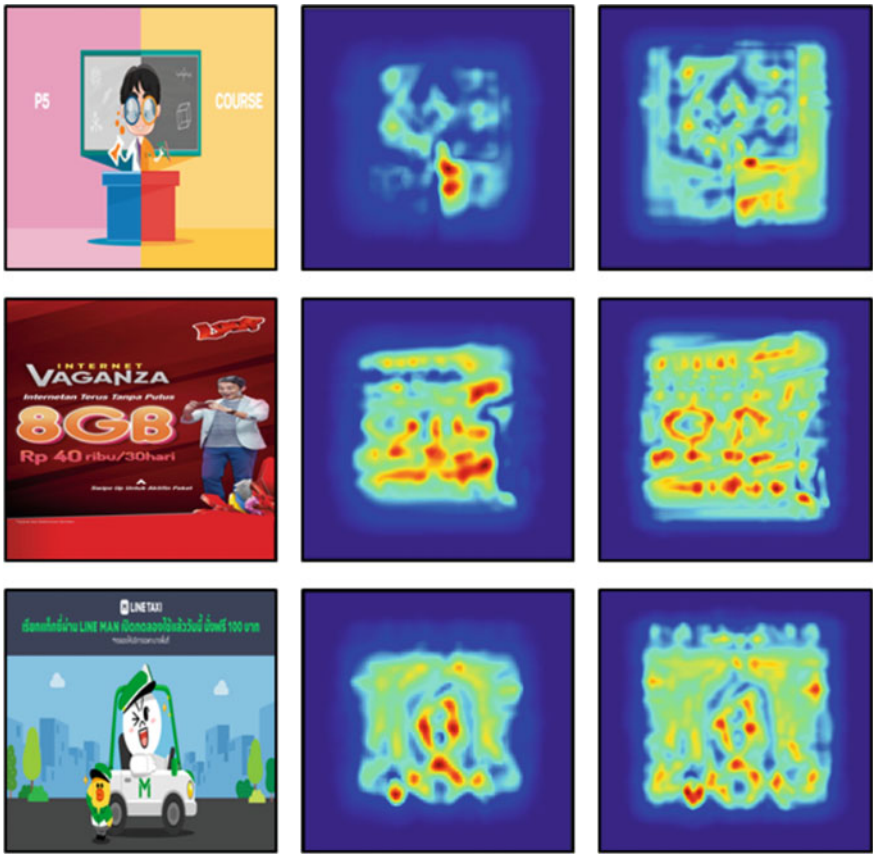

Fig. 4 The saliency map result after visualizing the activation matrix at the last convolutional layer of 2D CNN

immoderately in the applied Facebook dataset and able to achieve the lowest MSE of 0.68 by applying equation 12.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{12}$$

where the y_i indicates the actual CTR value while $\tilde{y}_i$ indicates the predicted CTR value through the proposed 2D CNN.

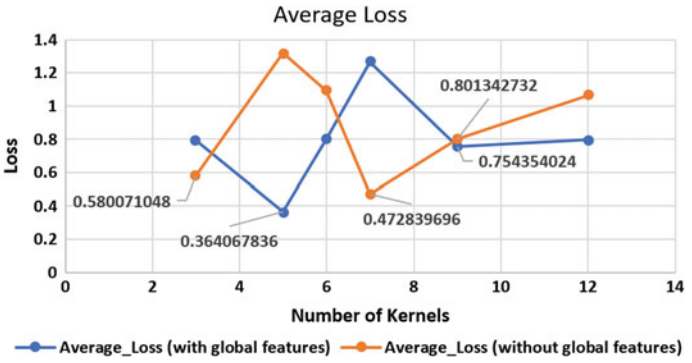

Fig. 5 The relationship of between the number of kernel and MSE Loss

4.3 1D CNN Experimental Results

In this section, the experiments of comparing the MSE between the click metrics data with global features, represented as C_{gf}, and without global features, C_{wgf} are conducted. First, the average loss result after applying the proposed 1D CNN on both metrics data as depicted in Fig. 5 will be discussed.

In Fig. 5, it is observed that the C_{gf} has achieved the lowest MSE of 0.364 when the number of kernels, K is 5. Subsequently, the average MSE begins to rise starting at $K = 6$. In contrast, C_{wgf} shows a continuous decrease of average MSE value starting from $K = 5$ and achieved the best average MSE value of 0.473 with $K = 7$. Nonetheless, C_{gf} and C_{wgf} show similar MSE value at $K = 9$. In overall, it is observed that with C_{gf}, it is able to achieve a low MSE value faster without requiring a larger kernel size like C_{wgf} in order to achieve a new lower MSE value from 0.580 at $K = 3$ to 0.473 at $K = 7$. Moreover, it also indicates that C_{gf} can achieve a much lower MSE as seen at $K = 5$. Therefore, to examine $K = 5$ is an optimum kernel size 1D CNN, the experiment of generating the standard deviation loss, Root Mean Square Error (RMSE), is computed to verify the proposed 1D CNN learning capability when $K = 5$ as shown in Fig. 6.

Figure 6 shows that the C_{wgf} at $K = 3$ has the lowest RMSE value of 0.153. However, recall that the MSE value of C_{wgf} at $K = 3$ is higher than the C_{gf} MSE value at $K = 5$. Although C_{wgf} at $K = 3$ shows a better learning capability compared to C_{gf} at $K = 5$ as it has a more rigid confidence interval, the natural characteristic of the metric data without the global features is not able to help in further reducing the MSE value. On the other hand, the second lower RMSE occurred at C_{gf} when $K = 5$ with the value of 0.189 has yield the lowest MSE. Therefore, it is proven that the proposed 1D CNN works best with $K = 5$ with the best MSE of 0.364 ± 0.189. Note that these average loss and standard deviation loss results for each number of kernels are generated and averaged by conducting the experiments for 5 times

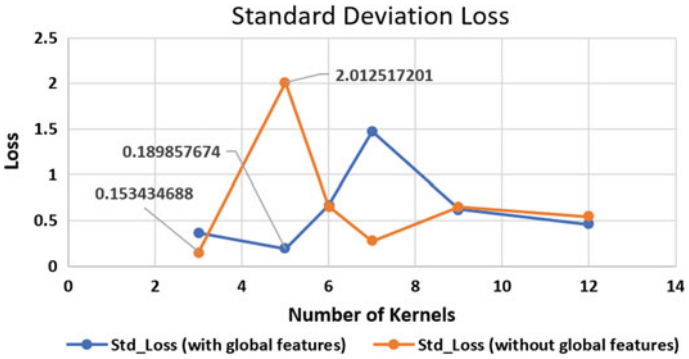

Fig. 6 The relationship between the number of kernels and RMSE

repeatedly as the stochastic learning in neural network will generate a different MSE value each time.

5 Conclusion

In this paper, a 2D CNN for extracting the salient features from the image and 1D CNN for predicting the CTR based on click metrics data were proposed. Furthermore, the experimental results showed that the loss of the proposed model can be further reduced after adding in the global image features. Nonetheless, it is noted that the experimental results showed in this paper were preliminary results as more display image data will be collected in the near future in order to verify and improve the performance of the proposed 2D CNN model. As for our future work, we will combine both CNN architecture models where the correlation between the feature maps of image and the corresponding click metric can be fused and thus, generating a more robust non-linear features to heighten the prediction of display CTR value.

References

1. Jones, J.P.: When Ads Work: new Proof that Advertising Triggers Sales. ME Sharpe (2006)
2. cisco.com: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper 2019. https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-738429.html#_Toc953325. Accessed 22 Mar 2019
3. Mei, T., Li, L., Hua, X.S., et al.: ImageSense: towards contextual image advertising. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **8**(1), 6 (2012)
4. Wang, J., Zhang, W., Yuan, S.: Display advertising with real-time bidding (RTB) and behavioural targeting. Found. Trends Inf. Retr. **11**(4–5), 297–435 (2017)

5. Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: estimating the click-through rate for new ads. In: Proceedings of the 16th International Conference on World Wide Web, pp. 521–530 (2007)
6. Liu, Q., Yu, F., Wu, S., et al.: A convolutional click prediction model. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1743–1746 (2015)
7. Guo, H., Tang, R., Ye, Y., et al.: DeepFM: a factorization-machine based neural network for CTR prediction. arXiv preprint arXiv:1703.04247 (2017)
8. Qu, Y., Cai, H., Ren, K., et al.: Product-based neural networks for user response prediction. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 1149–1154 (2016)
9. Chan, P.P., Hu, X., Zhao, L., et al.: Convolutional neural networks based click-through rate prediction with multiple feature sequences. In: IJCAI 2007–2013 (2018)
10. Zhou, G., Zhu, X., Song, C., et al.: Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1059–1068 (2018)
11. Cheng, H., Zwol, R.V., Azimi, J., et al.: Multimedia features for click prediction of new ads in display advertising. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 777–785 (2012)
12. Chen, J., Sun, B., Li, H., et al.: Deep CTR prediction in display advertising. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 811–820 (2016)
13. Dahlen, M.: Banner advertisements through a new lens. J. Advert. Res. **41**(4), 23–30 (2001)
14. Hasler, D., Suesstrunk, S.E.: Measuring colorfulness in natural images. In: Human vision and electronic imaging VIII, vol. 5007, pp. 87–96. International Society for Optics and Photonics (2003)
15. Huang, K.Q., Wang, Q., Wu, Z.Y.: Natural color image enhancement and evaluation algorithm based on human visual system. Comput. Vis. Image Underst. **103**(1), 52–63 (2006)
16. Azimi, J., Zhang, R., Zhou, Y., et al.: The impact of visual appearance on user response in online display advertising. In: Proceedings of the 21st International Conference on World Wide Web, pp. 457–458 (2012)