The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Banner Click Through Rate Classification Using Deep Convolutional Neural Network

Nicolas MICHEL † ‡ Hayato SAKATA * Keita KURITA† Toshihiko YAMASAKI†

† Department of Information and Communication Engineering, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku Tokyo, 113-8656

‡ Département Informatique, Institut Mines-Télécom Atlantique, 2 rue Alfred Kastler, 44300 Nantes, France

* 2-11-1 Osaki, Shinagawa-ku, Tokyo, 141-0032 Japan, a.i lab., So-net Media Networks Corp.

E-mail: † nicolas.michel@imt-atlantique.net, ‡ {kurita, yamasaki}@hal.t.u-tokyo.ac.jp * hayato_sakata@so-netmedia.jp

ABSTRACT

In banner advertising, Click Through Rate (CTR) is one of the most important indicators to evaluate one advertisement's quality. Advertisers create massive number of banner candidates in empirical ways, then proceed to actual tests by delivering advertisement to measure each banner's effectiveness. This process is expensive and therefore our CTR prediction helps reducing online advertising costs. In this work, we propose a method to classify 'effective' and 'ineffective' advertising banners based on image processing using state-of-the-art CNN. We first focus only on images then conduct experiments including metadata (product, advertiser, etc) to increase the CTR prediction accuracy and demonstrate which metadata is the most influential. Subsequently, each approach is compared to human performance. In the second part of our work, we detect which parts of the image contribute predominantly to increase the CTR by generating heat maps for each classes. This work leads to a deeper understanding of a banner advertising success and helps making decisions on how to improve it.

1 INTRODUCTION

In the field of banner advertising, Click Through Rate (CTR) is one of the core indicators to describe one advertisement's efficiency [10]. Scoring each banner's real quality is very expensive and time consuming as advertisers rely on concrete trials over a tremendous number of banners to infer. Moreover, in some cases, promising candidates are not used, because advertisers filter out many of them without use to decrease advertising costs, and this process strongly depends on intuition of staffs. Accurate CTR prediction helps reducing online advertising costs and improving its performance.

In this work, we propose a CTR prediction method based on image processing using a state-of-the-art CNN [5, 6], and demonstrate that a machine learning technology overcomes human performance. CTR prediction using visual features such as CNN is emerging and further study is needed. In addition, current research in Deep Learning gave impressive results for Image Classification and more precisely for feature extraction [5, 6, 11]. As in banner advertisement, visual feature is highly correlated to attractiveness [4], recent studies addressed CTR prediction based on CNN as feature extractor [3, 7]. Throughout this study, a CTR classification model is proposed.

The organization of the paper is as follows. In section 2, related works are summarized. The methods and experiments adopted for this work are described in section 3. Data cleaning process is explained in subsection 3.1. In subsections 3.2 and 3.3, the architectures and the experimental results are outlined. Moreover, training details are developed in subsection 3.4. In section 4, experimental results are interpreted in regard to our knowledge. Eventually in section 5, our conclusions are established and we discuss our future work.

2 RELATED WORK

This work is related to Image Recognition using Deep Learning as well as CTR prediction. In the past few years CNN have demonstrated high capabilities to process images [5, 6, 11]. In 2012, the first CNN was used to win ILSVRC, a well-known Image Classification challenge based on ImageNet dataset. Since then, Deep Learning techniques evolved and their performances drastically improved to eventually overcome human-level performance [14]. Considering such technological advances, several studies soon started to explore CNN techniques for Online Advertising, and specifically Banner Advertising [1-3, 7].

In this regard, Fire et al. [1] experienced an ImageNet pre-trained CNN to automatically determine banner categories. Fortunately, banners included in our dataset had been manually categorized already. Nevertheless, ImageNet classification can still be of use to enhance the image description and give more impact to the visual information [12]. Thus, our work explores CTR classification using an architecture similar to Chen et al.'s study [2] and proposes an alternative by introducing ImageNet 1,000-class log-probabilities as CNN input. Moreover we use variable number of metadata and interpret their impacts, constructing a metadata-expandable architecture. Eventually we compare those results to human performance which, to the best of our knowledge, has never been conducted in others studies.

3 METHODS

3.1 Data Sets and Cleaning

For this study, we used images collected over two years by So-net Media Network Corporation (SMN), an advertisement delivering company, for a total of 115,250 advertisement banners. The dataset in itself is widely heterogeneous on many aspects as it was gathered from 2,585 different advertising companies, 105 product categories (rent, sale in lots, newly-built, etc), 20 product category groups (real estate, etc) and 13 different sizes of banners. These product categories and product category groups are manually defined by human. As a consequence, data-

preprocessing was essential and mainly impacted the quality of our results.

The first step was to prevent leakage issue. In banner advertising, it is common to create several near identical images to fit different displayed sizes and identify the impact of minor modifications. This process implies that a large number of banners are near-duplicate. This number is difficult to estimate and in many cases induces leakage. Indeed near-duplicates are repeatedly present in the training set and test set and usually have close CTR values to each other. In practice we sorted our sets in a way that every near-duplicate banners are in the same set. This prevents the model to be tested on a banner where near-duplicate was in the train set.

The second step was disposing of duplicate banners. Conversely, same banners can also have very different CTR depending on the commercial strategy or the device on which it is displayed. This results in admitting in our dataset several possible CTR values for the same banner. Though those information might give decisive clues to advertisers, this makes CTR prediction even more complex. Hence, duplicate banners were removed from our dataset reducing their number to 85,255. The CTR value selected among duplicate was the one corresponding to the highest number of impressions, as its value is more representative of the overall banner effectiveness.

The third step was to select images whose aspect ratios are similar. When using a deep learning algorithm, considering images of variable input size can be delicate. The common practice is to re-scale the input image, which can induce huge distortions when the banner aspect ratio differs from 1:1. In that sense, only banners being close to square were considered, with ratio between 6:5 and 1:1. A majority of the banners used for online advertisement are long-shaped, shrinking the dataset size to 35,108 banners. Every experiment described in this study has been conducted on this dataset

3.2 Network Design

Our first approach was to fine-tune a ResNet-101 architecture [6] to classify the banners. Here the final fully connected layer has simply been truncated and replaced with another fully connected layer with two output nodes in the output layer. Each node returns a value corresponding to ‘effective’ and ‘ineffective’ classes respectively. This output can be interpreted as the overall effectiveness of the banner (see Fig. 1).

Figure 1: Simple architecture using a fine-tuned ResNet-101 to extract image features.

As the CTR distribution significantly changes for each device (smartphone or PC), the device in which the banner will be displayed is later included in the model. Progressively the same process is repeated, using an increased number of metadata such as the category of the advertisement and the image size. Metadata are processed using a Multilayer Perceptron (MLP). In the last fully connected layer (FC) the ResNet-101 and MLP outputs are concatenated to compute the effectiveness (see Fig. 2).

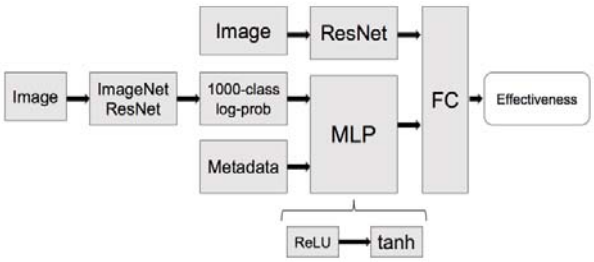

Figure 2: Architecture using fine-tuned ResNet-101 and metadata.

The last architecture uses the device, the size and the category as metadata plus the log-probability for each ImageNet label, obtained by the mean of another ResNet-101 trained on ImageNet dataset (see Fig. 3).

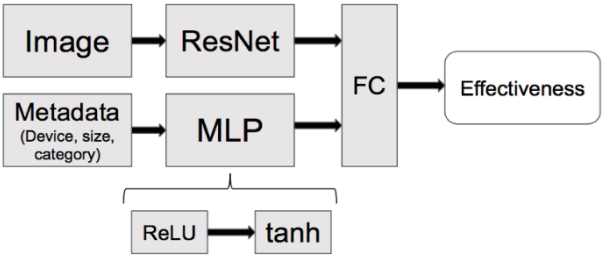

Figure 3: Architecture using fine-tuned ResNet-101, metadata and ImageNet-trained ResNet-101 1,000 log-probabilities.

For comparison and to understand the impact of metadata for our task, we also considered an architecture composed solely of metadata and ImageNet 1,000-class log-probabilities (see Fig. 4).

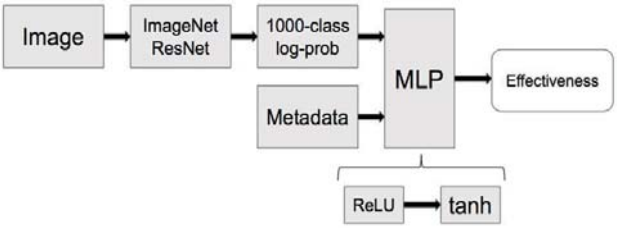

Figure 4: Architecture using image metadata and ImageNet 1,000 log-probabilities only.

4 EXPERIMENTS AND TRAINING DETAILS

4.1 Experiments

As aforementioned, the main objective of this work is to help advertisers and banner advertising experts having a hint concerning a banner’s quality. We used a state-of-the-art ResNet architecture pre-trained on ImageNet dataset. Therefore, we focused on a simpler task consisting in predicting a banner as ‘effective’ or ‘ineffective’. We define a banner as ‘effective’ when admitting a CTR above a specific threshold and ‘ineffective’ when under a second specific threshold. Practically, we set those threshold as to divide our dataset between the 30% most effective and 30% most ineffective banners. To prove

superiority of machine learning over human, we asked seven members of SMN to try classifying manually few banners. In that sense, we train several architectures and measure the impact of .adding data others than image features to the model. Our intuition here is that an image itself cannot contain enough information to make the prediction accurate, additional metadata are needed. In this regard, we gradually include metadata in the model and evaluate the performances thereafter. Our latest experiment includes ImageNet labels as log-probabilities. Further we use grad-CAM technique [8] to generate heat-maps corresponding to each 'effective' and 'ineffective' classes. Thus, we visualize which part of the network is mostly decisive in the last convolutional layer of our network. This technique is applied to the entire test set. However, due to copyright issues only one sample can be displayed in this paper.

4.2 Training Details

For each architectures we used a dataset composed of 35,108 banners as described in subsection 3.1. Therefore we divided our dataset in two splits of which represents 80% and 20% of the dataset for the train set and the test set respectively.

Furthermore, to transform our problem into a classification problem two CTR threshold values are determined on the train set such that two different classes are obtained: the 30% highest and lowest CTR values, referred as top30 and bottom30 classes. As a consequence, the dataset size is reduced to 60% of its original size corresponding to 21,064 banners.

For the training itself, we implemented the model using PyTorch framework and used one TITAN Xp as GPU. Moreover, we used SGD with momentum of 0.9 as optimizer and Cross Entropy as loss. We used a learning rate of 1e-4, with weight decay of 1e-3 and a batch size of 32 to try reducing overfitting [9]. Regarding the implementation, we incorporate metadata in the form of a one-hot vector as input of the MLP.

5 RESULTS AND DISCUSSION

5.1 Results

In the following section we present the results obtained after training different architectures described in the above section. The results show the accuracy obtained after the training of several models on our dataset, considering top30 and bottom30 classes. Each architecture's tag described in Table 1 corresponds to the data used during training. During this study, we trained 4 different architectures described in subsection 3.2, gradually increasing the amount of metadata. As a result, 61.9% accuracy is obtained using metadata and log-probabilities compared to 64.1% when using image features only.

Further, when using devices as metadata and image features at the same time an even higher accuracy of 66.5% is achieved. Next, we increase the number of metadata, including the device, size and category group of the image. In this scenario, the metadata input size is 35 and the accuracy reaches its peak of 69.3%. The last experiment includes the maximum number of information: size, device, category group, ImageNet log-probabilities as MLP input, as well as the image features. In this case, the accuracy falls to 64.6% (see Fig. 5). In addition we conducted several experiments using marginal thresholds values (see Table 1).

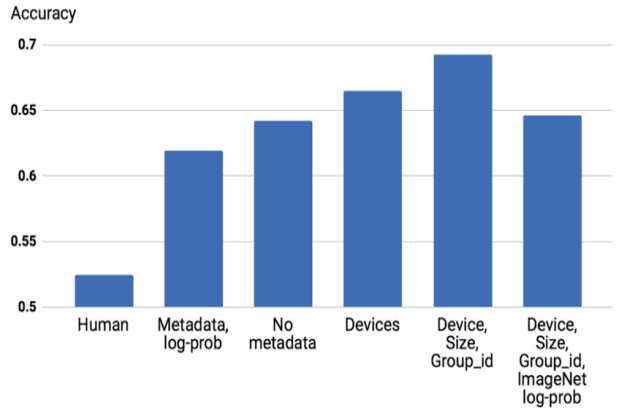

Figure 5: Accuracy obtained with different architectures.

Threshold (%)	Accuracy (%)
50	58.691
40	58.604
30	69.258
20	69.211
10	66.272

Table 1: Accuracy obtained for different threshold values.

Once the previous results obtained, we generate two heat maps using grad-CAM technique [8] for each banners of the test set. Each heat map highlights the part of the image that contributes to either being 'effective' of 'ineffective'. On Fig. 6, the part of the banner where it is written "Machine Learning" in Japanese is highlighted as contributing to make the CTR higher. Same process demonstrate that the part where "Full-scratch" in Japanese is highlighted as contributing to make the CTR lower.

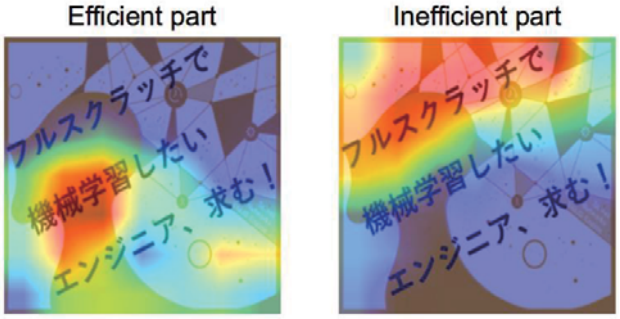

Figure 6: grad-CAM application.

Finally, manual classification were directed by seven members of SMN among 100 randomly chosen pictures from the test set. Three of those members have been Artificial Intelligence Laboratory members for 3 years, two have been working in SMN for around 1.5 year and the last two members are working in "Advertising Banner Team" for less than a year. Manual classification resulted in 52.7% accuracy.

5.2 Discussion

In this subsection we discuss the results obtained in the preceding one. First, we noted that the accuracy using only metadata is lower than when using only image features. This suggests that

the banner itself contains more decisive information than advertisement metadata and ImageNet labels.

Moreover, Fig. 6 demonstrates the significance of metadata for CTR classification as the accuracy steadily increases with the quantity of metadata. Even though some redundant information can be found in the image features and the metadata, the accuracy obtained using both combined demonstrates that each feature is separately decisive and must be included.

However, the last model including the most important amount of data as input shows an accuracy of 64.6 % slightly higher than 64.1% obtained with image features only. Our explanation is that the information contained within the 1,000 log-probabilities is somehow redundant with the feature-extracted information. Adding those information increased overfitting. Better regularization could help solving this problem, unfortunately this was not conducted in this study due to time constraints.

Therefore, using grad-CAM technique [8] to understand deeper the model decision-making [13] appears promising. Here the result is coherent with actual job market. As machine learning is currently popular, job-seekers might be appealed by those words. Nonetheless, implementing machine learning from scratch is not common as most user prefer utilizing framework and other ready-to-use tools.

Furthermore, Table 2 shows that the accuracy is higher for smaller threshold values. This is expected as a lower threshold implies important visual differences between banners. The exception is the accuracy observed for top10 and bottom10. In this scenario, only 20% of the total dataset is used and smaller dataset have been proven to impact unfavorably deep learning models.

Overall, every result obtained is significantly higher than human performance. This situation expresses the actual difficulty of the task and even though our model still has a considerable margin of progression, it seems to surpass human capabilities.

6 CONCLUSIONS AND FUTURE WORK

Through this study, we addressed a key point of banner advertisement that CTR prediction is. In this field, the common way to estimate any banner efficiency without actually delivering them is empirical. As it is, this process in time-consuming and expensive for advertisers. Thus, offering decisive clues to dissociate different banners is primordial. While many factors influence an image's CTR, our experiments expose the adversity of CTR prediction. As results subsection 5.1 highlights, advertisement-related professionals cannot produce a good estimation off-hand. Concerning our best classifier, an accuracy approaching 70% can be achieved and outperforms considerably human performance. The impact of metadata have been demonstrated, improving by 7% point the prediction accuracy. Additionally, metadata by itself can be used to achieve 61,9% accuracy compared to 62,8% when using image features only. This points out that even though the image itself contains more decisive information than metadata, both have separated impact and need to be taken into account simultaneously.

Furthermore we believe that our results can fairly be improved using other visual recognition technologies such as Optical Character Recognition (OCR). All the banners used were provided by Japanese advertisement companies. This implies that content information can be extracted as well as visual information using OCR. Eventually, we have the intention to forthwith conduct real-world experiments to validate our model.

ACKNOWLEDGMENTS

We thank So-net Media Corp. for providing the ad image dataset.

REFERENCES

[1] Michael Fire and Jonahtan Schler. 2015. Exploring Online Ad Images Using a Deep Convolutional Neural Network Approach, arXiv:1509.00568. Retrieved from https://arxiv.org/abs/1509.00568.

[2] Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu and Xian-Sheng Hua. 2016. Deep CTR prediction in display advertising. arXiv:1609.06018. Retrieved from https://arxiv.org/abs/1609.06018.

[3] Kaixiang Mo, Bo Liu, Lei Xiao, Yong Li and Jie Jiang. 2015. Image Feature Learning for Cold Start Problem in Display Advertising. In IJCAI'15 *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, Buenos Aires, Argentina, 3728–3734.

[4] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam. 2012. 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*. ACM Press, 777–785.

[5] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc., Lake Tahoe, Nevada, 1097-1105.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385. Retrieved from https://arxiv.org/abs/1512.03385.

[7] Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huiming Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, Pengtao Yi, Sui Huang, Zhiqiang Zhang, Xiaoqiang Zhu, Yu Zhang, Kun Gai. 2018. Image Matters: Visually modeling user behaviors using Advanced Model Server. arXiv:1711.06505v2. Retrieved from https://arxiv.org/abs/1711.06505.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv:1610.02391. Retrieved from https://arxiv.org/abs/1610.02391.

[9] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp. arXiv:1609.04836v2. Retrieved from https://arxiv.org/abs/1609.04836.

[10] Jun Wang, Weinan Zhang, Shuai Yuan. 2017. Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting. arXiv:1610.03013v2. Retrieved from https://arxiv.org/abs/1610.03013.

[11] Karen Simonyan, Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from https://arxiv.org/abs/1409.1556.

[12] Corey Lynch, Kamelia Aryafar, Josh Attenberg. 2015. Images Don't Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank. arXiv:1511.06746v1. Retrieved from https://arxiv.org/abs/1511.06746.

[13] Matthew D Zeiler, Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901. Retrieved from https://arxiv.org/abs/1311.2901.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv:1502.01852v1. Retrieved from https://arxiv.org/abs/1502.01852.