Which Ads to Show? Advertisement Image Assessment
with Auxiliary Information via Multi-step Modality Fusion

Kyung-Wha Park[1], JungHoon Lee[2], Sunyoung Kwon[3], Jung-Woo Ha[3], Kyung-Min Kim[3], Byoung-Tak Zhang[1,4,5]

[1] Interdisciplinary Program in Neuroscience, Seoul National University, [2] Statistics and Actuarial Science, Soongsil University,
[3] Clova AI Research, NAVER Corp, [4] Department of Computer Science and Engineering, Seoul National University, [5] Surromind Robotics
kwpark@bi.snu.ac.kr, ssutartup@gmail.com, {sunny.kwon, jungwoo.ha, kyungmin.kim.ml}@navercorp.com, btzhang@bi.snu.ac.kr

Abstract

Assessing aesthetic preference is a fundamental task related to human cognition. It can also contribute to various practical applications such as image creation for online advertisements. Despite crucial influences of image quality, auxiliary information of ad images such as tags and target subjects can also determine image preference. Existing studies mainly focus on images and thus are less useful for advertisement scenarios where rich auxiliary data are available. Here we propose a modality fusion-based neural network that evaluates the aesthetic preference of images with auxiliary information. Our method fully utilizes auxiliary data by introducing multi-step modality fusion using both conditional batch normalization-based low-level and attention-based high-level fusion mechanisms, inspired by the findings from statistical analyses on real advertisement data. Our approach achieved state-of-the-art performance on the AVA dataset, a widely used dataset for aesthetic assessment. Besides, the proposed method is evaluated on large-scale real-world advertisement image data with rich auxiliary attributes, providing promising preference prediction results. Through extensive experiments, we investigate how image and auxiliary information together influence click-through rate.

Introduction

Measuring image preference is both a fundamental and challenging problem because it is linked to the complex multimodal cognitive processes of humans visual perception (Palmer, Schloss, and Sammartino 2013; Braun et al. 2013). Various studies have been conducted in the fields of cognitive science, computational photography/aesthetics, and neuromarketing to assess the emotional impressions and aesthetic qualities from images (Murray, Marchesotti, and Perronnin 2012; Talebi and Milanfar 2018).

Image preference prediction can contribute to practical applications. Determining images for impressions: the number of times ads are displayed to customers, in online ads, is a prevalent application of image preference prediction. Predicting image preference for ads has two differences from the conventional aesthetic assessment. One is that an assessment model can use auxiliary information related to images, target subjects, and ad display policy. ad images are created by a designer for a purpose. It is a complex product that has undergone the creative processes of human designers. In particular, because the textual information is directly exposed to the content, the computational model must cope with both visual and language. The other difference is that an explicit metric can be used for evaluation, e.g., click-through rate (CTR).

Most existing methods on image assessment generally focused on an image as the input with respect to the quality and aesthetic. However, It has been reported that there exist other factors influencing on preference in addition to image quality, for example, colors (Mehta and Zhu 2009; Labrecque and Milne 2012), brands, and verbal components (Mitchell 1986) were referred to affect traditional advertising. Furthermore, most CTR prediction studies heavily rely on user or meta information. There has been research on multimodal fusion with other information based on content (Chen et al. 2016). However, modality fusion can be more advanced to improve preference performance. Talebi and Milanfar (Talebi and Milanfar 2018) proposed a model to quantitatively evaluate human subjective aesthetic judgments about images. However, this approach might be ineffective for evaluating the advertising images due to not utilizing auxiliary information.

Here we propose a new approach to predict the preferences of users for determining which image is impressed to be advertised. For achieving this, we statistically analyzed the CTR for impressions, exposure events of online ads. Also, we explored the neural network structure inspired by the known knowledge and notions of existing advertisement professionals. i) We approached the CTR prediction problem as a regression problem rather than a classical classification. ii) Preprocessed as in image assessment tasks. iii) Inspired by statistical analysis and cognitive science, we propose a multi-step modality fusion network (M2FN). The term: multi-step refer to between images and auxiliary information, using conditional batch normalization (CBN) (De Vries et al. 2017) and a spatial attention mechanism for predicting CTR.

To evaluate our approach and model performance, we validated M2FN on two datasets. Our model is evaluated on impression CTR predictions for real-world ad images. AVA: an aesthetic assessment dataset similar to our content, was tested on M2FN for impression assessment. We achieved both image and impression assessments state-of-the-art results. With neural network visualization, we analyzed through various and sufficient experimental results in

which area of the image influences the user's preference.

Related Work

Image Assessment: Visual aesthetic assessment is associated with art, beauty, and personal preference. Driven by the importance of aesthetic assessment, aesthetic visual analysis (AVA) dataset, a large-scale image database accompanied by a variety of meta-data and rich annotations, was introduced for advance research (Murray, Marchesotti, and Perronnin 2012). Using the AVA dataset, several studies have been carried out to predict aesthetic preference by reflecting the universally acceptable factors through learning (Deng, Loy, and Tang 2017; Talebi and Milanfar 2018; Yu et al. 2018); although aesthetic preference depends on individual taste, some universal rules were reported such as the golden ratio, color harmonies, and the rule of thirds (Datta et al. 2006; Dhar, Ordonez, and Berg 2011; Ke, Tang, and Jing 2006; Luo, Wang, and Tang 2011; Luo and Tang 2008). To obtain a higher correlation with human ratings, one of the papers (Talebi and Milanfar 2018) focused on the distribution of ratings instead of the aggregated mean scores using the earth mover's distance (EMD) (Zhang et al. 2018). Most visual aesthetic assessment studies mainly focus on image contents rather than using addition textual metadata together.

Vision and Language: Study on visual-linguistic representation learning has become popular thanks to the advent of Visual Question Answering (VQA) Challenge (Agrawal et al. 2015). Generally, there are two key components in the learning methods. 1) Attention mechanism. Due to the nature of image or video, most methods reduce redundancy on spatio-temporal inputs based on other modalities, e.g., text (Kim et al. 2018b). 2) Multimodal fusion. Combining vision and language of different properties is not trivial. Various methods have been proposed from simple ones such as concatenation, element-wise addition to complex ones, e.g., compact bilinear pooling (Fukui et al. 2016). CBN is a fusion method that modulates visual prior by controlling batch normalization parameters in layers of visual representation model based on other inputs (De Vries et al. 2017). Our approach adopts ideas from the visual-linguistic representation learning methods in that we utilize auxiliary information by multi-step fusion based on both CBN and attention-based fusion mechanisms.

Advertisement Preference Prediction: Recently, with the success of deep learning, neural network-based approaches have been proposed to predict CTRs, a widely used metric for evaluating ad performance. Compared to the conventional handcraft feature-based approaches, automatic and flexible feature learning directly from raw images has been introduced (Mo et al. 2015). For more accurate prediction, the paper used ad information, such as detailed category and display position, in addition to raw images and confirmed performance improvements. Another CTR prediction method built an end-to-end deep learning architecture that learns representative features (Chen et al. 2016). They are from both raw images and other related information, such as ad zone, ad group, category, and user features. Even if these

approaches use additional meta-information other than images, they have a room for improvement due to the use of simple concatenation-based modality fusion.

Advertisement Images with Auxiliary Data

Task Definition and Data

The problem addressed in this paper is to predict human preference for ad images with auxiliary information, which is formulated as a regression task. We define CTR as user preference on ad images. To solve this problem, we construct a large-scale dataset. It includes logs from ad displaying events represented with images, textual metadata, and their CTR value in a Japanese online ad service, called Real-Ad dataset. Therefore, ad images and their auxiliary information are used as input, and the CTR score is defined as the output value of regression.

Table 1: Details of Real-Ad datasets. Instances of the raw dataset represents impressions per instance, and are different from instances of aggregated datasets. Instances of the raw dataset that have more than 100 and 500 impressions are aggregated to construct train and test datasets respectively.

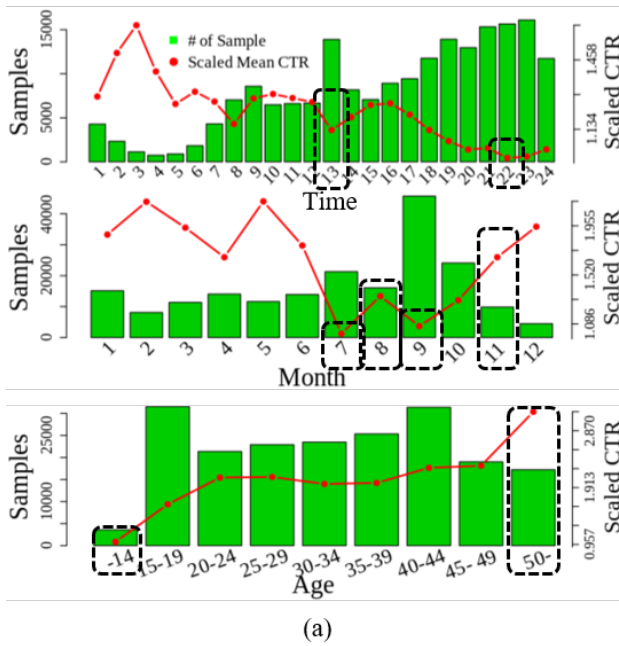| Dataset			# instances	# clicks	# ad images
	Raw(Game)		500M	20M	3,747
Aggregated	100+	Train	353,510	800K	3,045
		Test	173,248	330K	1,436
	500+	Train	47,325	350K	1,519
		Test	24,003	140K	681

Real-Ad Data Construction

An instance of raw CTR data is represented with one image and multiple auxiliary attributes. And it corresponds to one exposure event of ad content. The label of an instance is 1 or 0, denoting clicked or not-clicked. One instance is viewed many times with different times and target users. All instances exposed with the same auxiliary information and ad image are aggregated to be the CTR score of the unique impression:
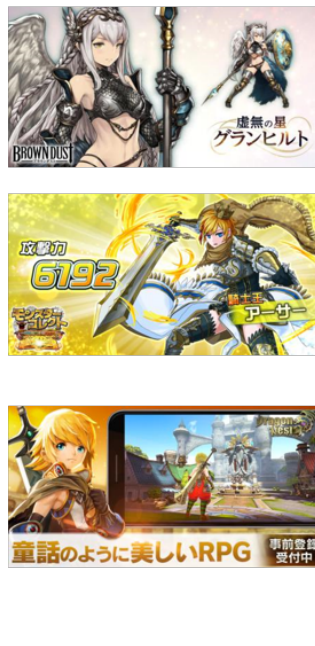
$$y_n = \frac{M_n^c}{M_n},\qquad(1)$$

where M_n^c and M_n denote the numbers of clicked and total ad displayed instances of the n-th unique exposure.

These preprocessed data are suitable for CTR prediction. The human preference for ad images may be represented as a distribution of real numbers with higher resolution. This facilitates comparative analysis with existing image assessment studies (Talebi and Milanfar 2018). As a real-world task, it makes it possible to cope with label imbalance inevitable in CTR prediction. In the raw data, more than 99% of the impressions are not clicked. The more frequent there is a conflicting label for the same instance, the worse the learning of the prediction models. Also, in real-world, these click data has an enormous scale, and thus those thousands of instances need to be compressed to increase expressiveness.

Time	Insight	Real	Pred
13	Mid	1.07	1.00
22	Low	1.35	1.53
24	Low	1.23	1.39

Month	Insight	Real	Pred
Jul.	Low	1.44	1.98
Aug.	Low	1.52	2.19
Sep.	Low	1.22	1.49
Nov.	High	1.00	1.31

Age	Insight	Real	Pred
under14	Low	2.35	2.58
25-29	Med	1.55	1.49
35-39	Med	1.66	1.57
45-49	High	1.00	1.13

(a)   (b)

Stat. analysis of Real-Ad dataset: Ad insights   The insights breaking cases and M2FN prediction result

Figure 1: Advertisement insights and task introduction. (a) shows statistical analysis results of Real-Ad, a real-world online ad CTR dataset. It supports findings of existing literature and gives insights about ad impressions. One of the insights is that older users click ads more. (b) shows counter-evidence of insights derived from (a) and impression assessment results of our proposing model: M2FN. Although the table contents shown in (b) are different from the actual statistical analysis results in (a) (the ad insights), M2FN shows reliable performance. The CTR values are divided by the smallest CTR in each table due to confidentiality.

Following Table 1 shows the specification of the Real-Ad dataset we have constructed. The datasets appear to be very compressed compared to the raw data.

Auxiliary Data

In the existing image assessment studies, only images were used for scoring. But we went further from there: using additional auxiliary data. In general CTR studies, most input data consist of metadata such as user demographical data—gender and age, and ad exposure event-related data—date and time. We use all of these metadata, as well as additional data from a variety of sources. For example, the catchphrase of the ad is expected to contain the intent of the ad. To make computational models to learn this catchphrase, we used linguistic attributes in Real-Ad dataset. The dataset also has titles and descriptions that are commonly found in image datasets for linguistics information, but above all, OCR. If the ad image contains textual expressions such as typography, it is expected to be captured by a computational model trained with OCR auxiliary attribute. The details of used auxiliary attributes are explained in the following Datasets section. Linguistic auxiliary attributes are transformed into embedding vector by BERT (Devlin et al. 2018), and the rest are given with one-hot encoding vectors. By doing this, the various forms can be taken into account.

Statistical Analysis on Real-Ad Data

We investigate how ad images and their auxiliary attributes have influences on CTR by statistical analyses such as ANOVA and Logistic Regression Analysis. Through the statistical analysis, we could figure out the specific response of CTR along with each of the auxiliary attributes. Generally, these characteristic movements (ad insights) are strategically considered points by marketers to raise the impressions of ads. Among the impression-effective attributes, Figure 1-(a) indicates the ad insights of three attributes usually affect the performance of ads regardless of the category of the item being advertised. For example, Figure 1-(a)-the age bar plot shows those game ads have higher chances of being clicked when they are exposed to the older. In addition, we found that dominant color and linguistic attributes are also significant for ad images.

On Top of that, the time and month bar plots in Figure 1-(a) show how time-sequential attributes have an impact on CTR. In detail, ads published at dawn, morning, lunch, and day time recorded better result compare to other times. The beginning and the end of a week also showed higher CTR than other days of the week, shown in the supplementary materials. Lastly, the lowest CTRs occur in the third quarter of the year: July, August, and September, corresponding to the summer season.

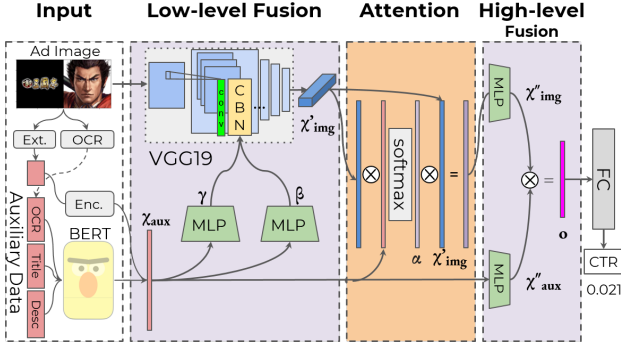On the other hand, there exist many cases which are

Figure 2: Overall structure of Multistep Modality Fusion Network (M2FN)

not following these ad insights. Figure 1-(b) shows unusual cases found during statistical analysis of Real-Ad and the results of CTR prediction using M2FN. These analyses results show the preference prediction on ad images is very challenging. For example, it is expected that the CTR distribution will be different because of the same ad image and auxiliary attributes but different time attribute. This expecting phenomenon is similar to a covariate shift.

For addressing these issues, we need to design a model to effectively utilize auxiliary attributes and integrate them with ad images based on the found ad insights.

M2FN: Multistep Modality Fusion Network

We describe our model (M2FN) for predicting user preference from ad images and auxiliary information, inspired by the ad insights through Real-Ad data analysis. For fully utilizing these findings, M2FN consists of three main modality fusion steps. i) low-level fusion to deal with specific features such as dominant colors and auxiliary information. ii) spatial attention to address text expression location in ad images. iii) high-level fusion to consider abstracted visual features and auxiliary attributes such as demographics, time, and semantics of linguistic information.

Low-level Fusion

We conjecture that it can improve prediction performance to give each layer of the network a hint of the complexity of the ad image, auxiliary information. For achieving this, we employ CBN (De Vries et al. 2017). It modulates the layers of the network instead of simple concatenation. CBN trains shallow neural networks which predict the scale factor parameters γ and shift factor parameters β as in batch normalization from the input data. Since the existing literature dealt with the VQA task, a question sentence is embedded using LSTM, and the question embedding e_q was fed into the CBN module. Instead of text question, the auxiliary information is used as the input of the CBN. To be given to the CBN, categorical attributes are encoded with a one-hot vector, and text information is represented with a BERT-based embedding vector.

CBN can be located at any layer of neural networks like Batch Normalization. In this paper, CBN was applied to fuse images and auxiliary information only after the first convolution of the early stage VGG-19, according to the best-performing experimental results.

Attention Mechanism

It is known that the location of textual information, as well as its expression in ad images, has a crucial influence on ad performance. To address this, we introduce an attention mechanism considering spatial relationships for modality fusion.

In specific, image features and auxiliary embedding vectors are represented with an $N_b \times C \times W \times H$ and an $N_b \times dim_{aux}$ tensors, where N_b, C, W, H denote batch size, the number of channels, width, and height, respectively. The auxiliary embedding vectors are replicated to be a $N_b \times dim_{aux} \times W \times H$ tensor. The auxiliary tensor is concatenated with the image feature tensor, to be a $N_b \times (C + dim_{aux}) \times W \times H$ tensor. After that, it is fed into a fully-connected layer (MLP). The matrix obtained by softmax function the resulting vector $\{N_b, W \times H\}$ becomes the attention matrix. The attention matrix was multiplied with image features to achieve a soft-attention map.

High-level Fusion

The role of High-level fusion is to emphasize the effect of spatial relationships between visual feature and auxiliary information by integrating two modalities near output layer. Dissimilar to previous approaches, where a simple concatenation is used, matrix element-wise multiplication of the output of the attention mechanism and auxiliary information is performed as in (De Vries et al. 2017). An affine transformation is performed to match the dimension size by linear layers for each of the image feature and auxiliary information. Then, the vectors from the activation function using hyperbolic tangent are element-wisely multiplied. There was a significant difference in regression performance with and without this fusion mechanism. It is shown in detail in the experiments section below.

Loss Function

Our model uses the impression weighted mean squared error as a loss function:

$$L = \frac{1}{N} \sum_{n=1}^{N} w_n \cdot (\hat{y}_n - y_n)^2, \qquad (2)$$

where $w_n, \hat{y}_n$, and y_n denote the impressed number, the predicted CTR, and the real CTR of the n-th sample. N is the data size. In addition, Kullback–Leibler divergence (KLD) is used as a loss function when the preference score of data is represented as a distribution form like AVA:

$$L_{KLD} = \frac{1}{N} \sum_{n=1}^{N} p(y_n) \cdot \log \frac{p(y_n)}{p(\hat{y}_n)}, \qquad (3)$$

where $p(y_n)$ and $p(\hat{y}_n)$ refer to the distributions of the real CTR and the CTR predicted by softmax function of the n-th sample, respectively.

Datasets

Real-Ad Dataset

For evaluation, we use a large-scale dataset of click logs for ad image impressions, collected from a global online ad service in Japan during 2018. The total number of impressions is approximately 500 million, which include distinct 3,747 ad images. As constructing training data, we followed two steps such as aggregation, explained in the previous section, and attribute selection.

Attribute selection Raw Real-Ad contains about 40 attributes which were collected with click logs. Some attributes have weak effects on CTR, and there exist redundant attributes. To select attributes crucial for CTR prediction, we employed Analysis of Variance (ANOVA) and Logistic Regression Analysis under a significant level of 0.05 and thus selected nine attributes. The attributes selected for M2FN include *gender, age, month, weekday, time, position (the position of ad in the displayed app page), category2 (mid-category of game, e.g. casual, hard-core, etc), category3 (low-category of game, e.g. role-playing, action, gambling, etc), and dominant color*. More details can be found in the supplementary material.

Attribute preprocessing Dominant color and linguistical auxiliary attributes are required to be preprocessed for learning, including title, description, and embedded textual expression of ad images (OCR). The dominant color of an ad image is represented as an element of the predefined color set, including ten colors. For achieving this, we use K-means clustering with the minimum covariance determinant (MCD) distance for extracting intermediate dominant color from the image. After obtaining the intermediate color, we mapped it to one of the predefined dominant colors.

Also, we supplemented linguistical auxiliary attributes such as title, description, and OCR result. The title and description were collected from the content introduction, such as smartphone app markets. OCR result is obtained by the open API[1] (Baek et al. 2019b; Baek et al. 2019a) to identify the letters in the ad images. Unique sentences collected for each of these attributes: The numbers of sentences are 1,583, 2,695, and 2,695, respectively. We embedded them into vectors using BERT[2] (Devlin et al. 2018).

While grouping and summing CTRs, we only considered instances which have more than 100 and 500 impressions for the feasibility of CTR value. We merged levels of the attribute, which has less than 50,000 impressions to the closest level for coping with being biased in the dataset.

In Real-Ad dataset, the total dimension of auxiliary attributes is 2,383, and the linguistics are majority: 2,304 (768 dimensions for each).

Benchmark dataset: AVA Dataset

M2FN can contribute to a conventional image assessment task. We evaluate our model on AVA (Murray, Marchesotti, and Perronnin 2012), which is an image dataset designed for

aesthetic preference study. This dataset consists of images and their annotations-aesthetic, semantic, and photographic style annotations. The photographic style annotations are comprised of 14 styles (complementary colors, duotones, HDR, etc.) to represent the camera settings. However, these annotations are not used in this paper due to too many no-annotation cases.

Semantic annotations are textual tag data such as *nature, black and white, landscape, still life, macro, animal*, etc. The number of images is approximately 200k, and each image has at least one tag. Tag set size is 67 by adding no-tag cases for image-only CTR prediction. A text tag is represented as an embedding vector by BERT. These embedding vectors with 768 dimensions were used for the experiment as linguistical auxiliary attributes.

Aesthetic annotations are rating data per image from hundreds of amateur and professional photographers. These data are characterized by the histogram distribution ranging from 1 to 10, which are used as score labels. The dataset was divided into the training, and the test sets with an 8:2 ratio for benchmark experiments.

Experimental Results

Explanation of terms: The resulting outputs may be in the form of distribution (10 buckets) or a scalar value (a score) depending on the dataset. Performance evaluation is based on ranking. Spearman rank correlation (SPRC) and linear correlation coefficient (LCC) are computed to rank the output scores and compare them to the ranking of the ground truth. If the output score is in the form of distribution, SPRC and LCC are calculated and examined for both mean and standard deviation.

The benchmark and Real-Ad datasets have distinct hyper-parameter settings for M2FN. Among the four major modules, CBN (low-level fusion), attention, and high-level fusion should determine the hidden layer size of the multi-layer perceptron (MLP) for embedding representations inside the module. When training AVA datasets which have relatively small auxiliary information, the hyperparameters of the modules are 64, 512, and 512, respectively. When training the Real-Ad dataset, we decided to 256, 512, and 1,024. This decision was made based on experiments. The batch size of the dataset was 128, and five P40 GPUs were used for training. All models are trained for 100 epochs. All the experiments are implemented and performed based on NAVER Smart Machine Learning (NSML) platform (Kim et al. 2018a; Sung et al. 2017).

Table 2 compares the CTR prediction performance on the Real-Ad dataset. Unlike AVA, which includes rating counts in the form of ten buckets for each image, Real-Ad only provides a mean of CTRs. Since our proposed method requires training on a CTR value, the CTR distributions are approximated by a log-normal distribution which resembles CTRs. When converted to a distribution, it is represented as "dist."; otherwise, it is described in Table 2 as "regr.". Previously, the best performing model (NIMA) on the benchmark dataset used EMD as a loss function. The star (*) means that the KLD is applied as a loss function instead of the EMD

Table 2: Performance comparison on Real-Ad. m and std. in the header row respectively represent mean and standard deviation.

		Models	SPRC(m)	LCC(m)	SPRC(std.)	LCC(std.)
100+	Dist.	NIMA	0.110	0.121	0.146	0.142
		NIMA*	0.289	0.290	0.139	0.143
		M2FN	0.344	0.367	0.172	0.175
	Regr.	NIMA	0.325	0.343	-	-
		M2FN	**0.384**	**0.381**	-	-
500+	Dist.	NIMA	0.308	0.249	0.165	0.166
		NIMA*	0.453	0.448	0.190	0.190
		M2FN	0.484	0.478	0.216	0.190
	Regr.	NIMA	0.501	0.451	-	-
		M2FN	**0.561**	**0.530**	-	-

Table 3: Performance comparison on Real-Ad. m and std. in the header row respectively represent mean and standard deviation.

Models	SPRC(m)	LCC(m)	SPRC(std.)	LCC(std.)
NIMA(Inception-v2)	0.612	0.636	0.218	0.233
M1FN(VGG19-cat)	0.572	0.581	0.198	0.197
M2FN	**0.630**	**0.640**	**0.310**	**0.322**

(for "regr." cases, the impression weighted MSE is used as mentioned above section).

Regardless of the number of impressions (100+ or 500+), it can be seen that our method performs better. In our dataset, KLD appears to be a better choice as a loss function than the EMD.

In Table 3, we compare our models to NIMA (Deng, Loy, and Tang 2017) on the AVA dataset for verifying that our model can be applied to image assessments., with a quantitative comparison between NIMA (Inception-v2), the previous SOTA model, and our M2FN.

VGG19-cat in Table 3 describes the preliminary method we have implemented for fusing the auxiliary information: "-cat" stands for simple concatenating of both the auxiliary and image. As a result, the model adopts a simple fusion (concatenating) deteriorates performance. M2FN, which is using multi-step modality fusion performed best in the benchmark dataset.

Table 4 compares the performance changes according to the presence or absence of four major modules. The four primary modules are auxiliary data, low-level fusion, attention, and high-level fusion. In the four consecutive columns, "O" stands for activating, and "×" stands for deactivating. "× × ××" is a model trained only images on vanilla VGG-19, the basis of M2FN. In the case of "O × × ×" where the auxiliary data is activated, but the CBN is deactivated, it means that the image and auxiliary information are learned by concatenating before the last fully connected layer.

As can be seen in the results of the second and third rows indicating the presence and absence of low-level fusion, we

Table 4: M2FN ablation study results on Real-Ad

Module				500+		100+	
Aux	Low	Att	High	SPRC	LCC	SPRC	LCC
×	×	×	×	0.456	0.435	0.315	0.342
O	×	×	×	0.437	0.412	0.297	0.320
O	O	×	×	0.496	0.498	0.367	0.371
O	×	O	×	0.450	0.463	0.341	0.325
O	×	×	O	0.475	0.467	0.321	0.293
O	O	O	×	0.506	0.480	0.356	0.334
O	O	×	O	0.554	0.528	0.361	0.333
O	O	O	O	**0.561**	**0.530**	**0.384**	**0.381**

can see that CBN boosts performance. It also supports the health of M2FN once again in the results of the fourth row. It indicates the presence of high-level fusion. The best performance is achieved when all four modules are combined. The four modules integrated result proved that our approach is suitable for evaluating impressions.

Table 5: M2FN ablation study of auxiliary information using Real-Ad. The attributes start with upper cases represent consist of multiple attributes.

Used auxiliary attributes	SPRC	LCC
None	0.456	0.435
ALL[1]	**0.561**	**0.530**
— date	<u>0.454</u>	<u>0.413</u>
— description	<u>0.467</u>	<u>0.453</u>
— color	<u>0.503</u>	<u>0.470</u>
— (User+Text)[2]	0.513	0.507
— ocr	0.515	0.491
— gender	0.532	0.470
— age	0.535	0.486
— User[3]	0.539	0.514
Text	0.422	0.369
— (title + ocr)	0.453	0.411
User	0.437	0.397

[1] ALL aux: Text, User, date, time, position, categories, color
[2] remove Text (title, desc, ocr) with User(gender, age) auxes.
[3] remove User (gender, age) from ALL auxes.

The results in Table 5 are ablation study comparing performance changes with or without auxiliary attributes in the Real-Ad dataset. Table 5 shows only a part, and the entire experimental results can be found in supplementary materials. Because we conjectured content-related attributes (date, time, position, categories, and color) are more information than others, e.g., user-related attributes, for evaluating impressions. We expected that excellent performance would be achieved even if we used only content-related attributes among auxiliary attributes. As we expected, the result proves our conjecture.
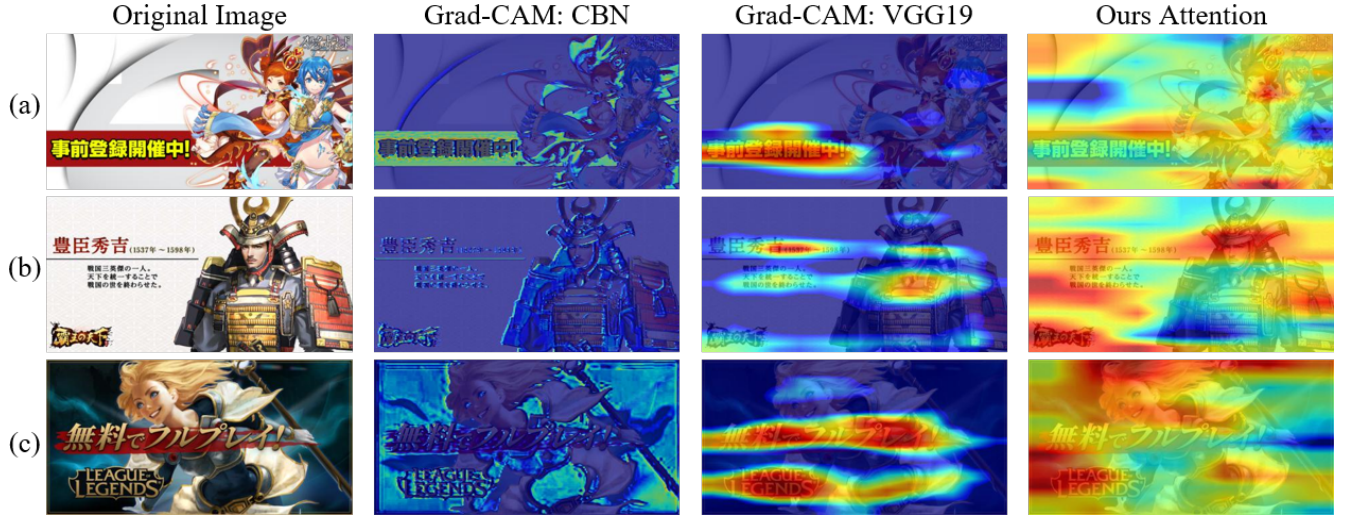
Figure 3: Visualization results using Grad-CAM. Each column displays heatmaps that depict weight vectors of specified layers. The red and blue colors respectively represent the highest value and the lowest value. Best viewed in color.

We sorted the performance in Table 5 in ascending order. If an attribute is excluded and the performance drop is significant, it means that the attribute is essential. As a result, we can see the following: i) The time-sequential attribute (date) is vital. ii) According to the bottom of the table (-(title + ocr)), description has little effect when used alone. But it is having a significant impact when it is integrated with other auxes. iii) As the previous studies asserted, the color attribute proved itself essential for advertising content in the table again. iv) The user attribute is not crucial, even though it heavily used in previous studies. v) However, when the user attribute is combined with other (text aux in the table), the performance degradation seems to be substantial. It is assumed that the user attribute has its potential when integrated.

The above five findings indicate that there is a complex action between auxiliary attributes. Therefore our hypothesizes were correct.

Further Analysis

Statistical and experimental results provide some insights, which can be used for assessing the attractiveness of ad images induced by a human designer. Based on the statistical analysis of the Real-Ad, we reaffirmed previous studies: Ad insights of single or combination of the attributes such as time, month, date, and age. As listed in Table 5, we show that the four attributes are very influential. The model was robust even under exceptional circumstances that did not follow the insights. Detailed case studies can be found in the supplementary material. Also, with the combination of the attributes, the fused representation in M2FN fires like an activation function in neurons: e.g., united they fires, divided they don't. All of this would be very interesting from the perspective of cognitive science and marketing research.

In Fig 3, each layer of M2FN was leveraged using Grad-CAM (Selvaraju et al. 2017) to visualize the operation of each layer in the model. Grad-CAM is a tool that allows us to see which part of an image the neural network sees and makes a decision on a particular label. It facilitates the layer to understand the importance of each neuron using the gradient information. In Fig 3, heatmaps are drawn, which is shown in red as a vital area. Interestingly, M2FN seems to be heavily influenced by the text on the image after the extraction of visual features (third column in the figure). As a result, in the attention map (fourth column), a visualization result also shows having a salience in part with characters and typography was obtained. Analyzing these results, our model learned well where the linguistic elements that humans consider necessary in advertising images are located. The model gives a hint about human's visual-spatial saliency toward the ad. This proves the results of existing literary and cognitive or marketing science studies, and at the same time, M2FN has achieved preliminary success in learning the human aesthetics toward advertising.

Conclusion

In this paper, we propose a model for predicting user preference of ad images. We collect a large-scale dataset, including ad images and auxiliary data, called Real-Ad dataset. Then, we statistically explore Real-Ad dataset, focusing on the influence of images and auxiliary attributes on human preference represented as CTR. Inspired by ad insights found from the analyses, we design a new multi-step modality fusion network (M2FN). M2FN is to effectively integrate ad images and their auxiliary attributes to predict CTR. We evaluate M2FN on Real-Ad dataset. Besides, we validate our method on a benchmark image assessment dataset, AVA dataset for verifying whether our approach can be applied to conventional image assessment. M2FN achieved better performance on both datasets compared to the previous state-of-the model. With extensive ablation study, we investigate how each modality fusion works and which auxiliary attributes

largely influences user preference.

As further works, we will implement a generative model that uses M2FN as a discriminator. We consider to generate ad images that can achieve higher CTR.

References

[Agrawal et al. 2015] Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

[Baek et al. 2019a] Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S. J.; and Lee, H. 2019a. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *International Conference on Computer Vision (ICCV)*. to appear.

[Baek et al. 2019b] Baek, Y.; Lee, B.; Han, D.; Yun, S.; and Lee, H. 2019b. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9365–9374.

[Braun et al. 2013] Braun, J.; Amirshahi, S. A.; Denzler, J.; and Redies, C. 2013. Statistical image properties of print advertisements, visual artworks and images of architecture. *Frontiers in Psychology* 4:808.

[Chen et al. 2016] Chen, J.; Sun, B.; Li, H.; Lu, H.; and Hua, X.-S. 2016. Deep ctr prediction in display advertising. In *Proceedings of the 24th ACM international conference on Multimedia*, 811–820. ACM.

[Datta et al. 2006] Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, 288–301. Springer.

[De Vries et al. 2017] De Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. C. 2017. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, 6594–6604.

[Deng, Loy, and Tang 2017] Deng, Y.; Loy, C. C.; and Tang, X. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34(4):80–106.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dhar, Ordonez, and Berg 2011] Dhar, S.; Ordonez, V.; and Berg, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, 1657–1664. IEEE.

[Fukui et al. 2016] Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

[Ke, Tang, and Jing 2006] Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, 419–426. IEEE.

[Kim et al. 2018a] Kim, H.; Kim, M.; Seo, D.; Kim, J.; Park, H.; Park, S.; Jo, H.; Kim, K.; Yang, Y.; Kim, Y.; et al. 2018a. Nsml: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*.

[Kim et al. 2018b] Kim, K.-M.; Choi, S.-H.; Kim, J.-H.; and Zhang, B.-T. 2018b. Multimodal dual attention memory for video story question answering. In *ECCV*.

[Labrecque and Milne 2012] Labrecque, L. I., and Milne, G. R. 2012. Exciting red and competent blue: the importance of color in marketing. *Journal of the Academy of Marketing Science* 40(5):711–727.

[Luo and Tang 2008] Luo, Y., and Tang, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, 386–399. Springer.

[Luo, Wang, and Tang 2011] Luo, W.; Wang, X.; and Tang, X. 2011. Content-based photo quality assessment. In *2011 International Conference on Computer Vision*, 2206–2213. IEEE.

[Mehta and Zhu 2009] Mehta, R., and Zhu, R. J. 2009. Blue or red? exploring the effect of color on cognitive task performances. *Science* 323(5918):1226–1229.

[Mitchell 1986] Mitchell, A. A. 1986. The effect of verbal and visual components of advertisements on brand attitudes and attitude toward the advertisement. *Journal of consumer research* 13(1):12–24.

[Mo et al. 2015] Mo, K.; Liu, B.; Xiao, L.; Li, Y.; and Jiang, J. 2015. Image feature learning for cold start problem in display advertising. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[Murray, Marchesotti, and Perronnin 2012] Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415. IEEE.

[Palmer, Schloss, and Sammartino 2013] Palmer, S. E.; Schloss, K. B.; and Sammartino, J. 2013. Visual aesthetics and human preference. *Annual review of psychology* 64:77–107.

[Selvaraju et al. 2017] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Gradcam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

[Sung et al. 2017] Sung, N.; Kim, M.; Jo, H.; Yang, Y.; Kim, J.; Lausen, L.; Kim, Y.; Lee, G.; Kwak, D.; Ha, J.-W.; et al. 2017. Nsml: A machine learning platform that enables you to focus on your models. *arXiv preprint arXiv:1712.05902*.

[Talebi and Milanfar 2018] Talebi, H., and Milanfar, P. 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing* 27(8):3998–4011.

[Yu et al. 2018] Yu, W.; Zhang, H.; He, X.; Chen, X.; Xiong, L.; and Qin, Z. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 World Wide Web Conference*, 649–658. International World Wide Web Conferences Steering Committee.

[Zhang et al. 2018] Zhang, R.; Isola, P.; Efros, A. A.; Shecht-
man, E.; and Wang, O. 2018. The unreasonable effectiveness
of deep features as a perceptual metric. In *Proceedings of the
IEEE Conference on Computer Vision and Pattern Recogni-
tion*, 586–595.