M2FN: Multi-step Modality Fusion for Advertisement Image Assessment

Kyung-Wha Park
Interdisciplinary Program in Neuroscience, Seoul National University.
kwpark@bi.snu.ac.kr

Jung-Woo Ha
NAVER AI LAB, NAVER CLOVA.
jungwoo.ha@navercorp.com

JungHoon Lee
Statistics and Actuarial Science, Soongsil University.
ssutartup@gmail.com

Sunyoung Kwon
School of Biomedical Convergence Engineering, Pusan National University.
skwon@pusan.ac.kr

Kyung-Min Kim
NAVER AI LAB, NAVER CLOVA.
kyungmin.kim.ml@navercorp.com

Byoung-Tak Zhang
Interdisciplinary Program in Neuroscience, Seoul National University.
Department of Computer Science and Engineering, Seoul National University.
Surromind Robotics.
btzhang@bi.snu.ac.kr

Abstract

Assessing advertisements, specifically on the basis of user preferences and ad quality, is crucial to the marketing industry. Although recent studies have attempted to use deep neural networks for this purpose, these studies have not utilized image-related auxiliary attributes, which include embedded text frequently found in ad images. We, therefore, investigated the influence of these attributes on ad image preferences. First, we analyzed large-scale real-world ad log data and, based on our findings, proposed a novel multi-step modality fusion network (M2FN) that determines advertising images likely to appeal to user preferences. Our method utilizes auxiliary attributes through multiple steps in the network, which include conditional batch normalization-based low-level fusion and attention-based high-level fusion. We verified M2FN on the AVA dataset, which is widely used for aesthetic image assessment, and then demonstrated that M2FN can achieve state-of-the-art performance in preference prediction using a real-world ad dataset with rich auxiliary attributes.

1. Introduction

Unlike other industries of the 20th century, the advertising industry has rapidly expanded only in recent times. The global advertising market, already worth 532.5 billion
dollars in 2019, is expected to reach 769.9 billion dollars by 2024. Of this large market, nearly one third is based on online advertising, which has achieved enormous commercial success in the modern world. The market value of online advertising is projected to grow significantly from 193 billion dollars in 2017 to 236 billion dollars by 2026.

Given the immense size and scope of the modern advertising industry, much competition comes from within the industry. Businesses and advertisers, therefore, must ensure that their advertisements (hereinafter referred to as ads) are more attractive and compelling than each other’s.

A well-created ad not only attracts user attention but also leads to engagement. From the perspective of online advertisers, an effective ad would not only attract more interest but also generate more clicks, leading to increased product purchases with fewer ad displays, thereby reducing the cost of bidding on the ad schedule. Therefore, a precise quantitative evaluation of ads is important for advertisers and ad designers.

Accordingly, this study conducts an ad image assessment that involves understanding image preferences: which images are preferred and to what extent is an image preferred over others. In addition, an initial study is provided via a deep network visualization: of which part of the image might be preferred.

Existing image assessment methods typically use an image as the only input. Some studies have proposed models to quantitatively evaluate human-subjective aesthetic judgments. Additionally, various studies have been conducted in the fields of neuroscience, cognitive science, computational photography, psychology, marketing, and aesthetics to assess the emotional impressions and aesthetic qualities of images.

However, predicting image preferences for ads differs from conventional aesthetic-image assessment in two ways. First, previous aesthetic-image assessment studies did not utilize auxiliary attributes such as image annotations, time, and age of participants. Thus, they were ineffective at evaluating ad images because user preferences may depend significantly on various conditions. Second, unlike aesthetic-image assessment studies that use ratings as a metric, studies on ad image preference use the click-through rate (CTR), the ratio of the number of users who click on the ad to the number of its displays. CTRs are commonly used to measure the success of online advertising; additionally, they can be used to assess image preferences for ads. Although CTRs are not automatically indicative of user preference, they do indicate the extent to which an ad accomplishes its objective (i.e., capturing user attention and engagement), which are induced by user preferences. Most CTR prediction studies on ad image preference rely on large-scale user data, which are sensitive and expensive to be collected. Therefore, recent studies have focused on reducing dependence on user information and leveraging other attributes.

Despite the differences between aesthetic-image assessment and CTR prediction studies, multi-modal fusion, which integrates information from different inputs, has become essential for both fields to expand the application domain and enhance the performance of deep network models. However, existing CTR prediction studies on ad images that apply multi-modal fusion do not utilize image-driven auxiliary attributes, but only facilitate linguistic attributes (e.g., ad title and description) and mainly involve simple operations, such as concatenation.

Handling complex ad images using simple concatenation-based modality fusion is difficult. Therefore, we extensively analyzed large-scale, real-world, and in-house log data from online game ads, including the CTR for exposure events, to utilize auxiliary attributes more effectively in assessing ad images. We determined that auxiliary attributes interact with the visual information of ad images at different conceptual levels, depending on each attribute. For example, at a low level, the overall color of an ad image could depend on the season. At a high level, the position or content of the embedded text could vary depending on the users’ ages. We required techniques that intervene in the early vision process and integrate auxiliary attributes with the refined vision representation in the later stages.

Owing to the recent advent of vision and language fusion studies, two techniques exist that match our needs—conditional batch normalization (CBN) and attention mechanism. The CBN technique modulates the visual processing in the early stage based on other input modalities. The attention mechanism is an essential technique in recent multi-modal deep learning frameworks. Both these techniques are mostly used for fusing language modalities into visual representations. We experimentally extended their scope for visual processing by leveraging auxiliary attributes, which consist of content-related attributes such as linguistic attributes (e.g., title and dominant colors) and classical metadata (e.g., age and time).

It is not straightforward to explain user preferences toward ad images using a single pipeline of a deep neural network, because such an architecture would contain an excessive number of layers to analyze. Inspired by this finding, we proposed a multi-step modality fusion network (M2FN) that incorporates auxiliary attributes, such as user-related metadata and visual–linguistic features, to assess ad images. To integrate ad images and auxiliary attributes effectively, we performed multiple steps of modality fusion and employed CBN and a spatial attention mechanism for predicting CTR. To evaluate our approach and the performance of our model, we validated M2FN on two datasets: the aes-

https://www.imarcgroup.com/global-advertising-market
thetis visual analysis (AVA) dataset [23], which is an aesthetic image assessment dataset, to test the quality of image assessments made by M2FN; and a dataset of real-world ad images, to evaluate the model using CTR predictions. We achieved state-of-the-art (SOTA) results on both datasets. In addition, we used a visualization method for each step in M2FN to empirically determine the image regions that influence user preferences as a initial study.

The main contributions of our research are summarized below:

• We propose a novel deep neural networks model for advertisement image assessment that involves performing multiple steps of modality fusion via conditional batch normalization and a spatial attention mechanism.

• We leverage visual–linguistic attributes for both aesthetic and advertisement image assessment tasks, along with classic metadata, denoted as auxiliary attributes.

• We demonstrate state-of-the-art performance for both tasks.

• We show the contribution of this model with ablation studies and visualizations to investigate the effectiveness of each step and auxiliary attribute.

2. Related Research

Image Assessment: Image assessment conventionally refers to the evaluation of an image in terms of quality (such as enhancement and noise). Visual aesthetic assessment, a form of image assessment, focuses on art, beauty, and personal preference. Driven by the importance of aesthetic assessment, the AVA dataset, which is a large-scale image database accompanied by various metadata and rich annotations, was introduced for advanced research [23]. Using the AVA dataset, several studies have performed aesthetic assessments of images by learning universal traits that characterize aesthetic quality [30] [8] [37]. Although aesthetic assessment generally depends on individual tastes, some universal rules, such as the golden ratio, color harmonies, and rule of thirds [5] [10] [36] [31] [18], have been reported. To obtain a higher correlation with human ratings, a study [30] focused on distributions of ratings instead of aggregated mean scores using the earth mover’s distance (EMD) [38]. Most visual aesthetic assessment studies concentrate on image content and do not involve additional textual metadata.

Vision and Language: Research on visual–linguistic representation learning has become popular since the advent of visual question answering (VQA) challenges [1]. Generally, learning methods comprise two key components:

1. Attention mechanism: Specific parameters are used to train models to attend to particular regions of images or correlated words in sentences [35]. Because of the sparsity of regions that carry useful information in images or videos, most methods reduce the redundancy on spatio-temporal inputs based on other modalities, e.g., text [15]. It typically finalizes the fusing step using matrix multiplication methods between other modalities.

2. Multi-modal fusion: The fusion of multi-modal data, which provides information characterized by different properties or formats for computational models, can lead to enhanced performance. While two distinct levels of fusion, i.e., early and late, were widely used [27], recent computational models can handle various multi-modal representations; moreover, the levels of fusion have become more diverse. Various methods utilizing multi-modal representations, ranging from simple methods, such as concatenation or element-wise addition, to complex methods, such as compact bilinear pooling [11], have been proposed.

CBN is a fusion method that initially modulates vision by controlling the batch normalization parameters in layers of a visual representation model based on other inputs [7]. It trains multilayer perceptrons (MLPs), which predict the mean and variance for in-depth normalization. Our approach adopts ideas from learning methods of visual–linguistic representations. We utilized auxiliary attributes via multi-step fusion based on both CBN and attention-based fusion mechanisms. In recent studies, methodologies for effectively performing bi-modal fusion of vision and language representations have been developed [24] [26]. These methodologies search for hidden representations or structures to achieve the best performance within multi-layer networks. While these methods can be applied to bi-modality problems, the problem in this study involves more than two modes. Through extensive experiments, we attempted to determine the optimal output from layers via multi-modal fusion.

Advertisement Preference Prediction: Recently, because of the success of deep learning, neural network-based approaches for predicting CTRs have been proposed. Automatic and flexible feature learning methods (directly from raw images) have been introduced [22] as an alternative to conventional, custom-developed, feature-based approaches. For more accurate predictions, Mo et al. [22] used ad information, such as detailed item categories and display positions, in addition to raw images, and achieved improved performance. Other CTR prediction methods involve building end-to-end deep learning architectures that learn representative features [4] [24]. These models
are trained using both raw images and other related information, such as ad zones, ad groups, item categories, and user features. Whereas these approaches use additional meta-information together with images, their use of simple concatenation-based modality fusion leaves room for improvement. End-to-end architectures and single pipeline designs also make it challenging to determine the reason for the superiority of an ad image.

3. Preliminaries

In this section, we define the terms relevant to our study. First, impression is a common term that represents the number of exposures of an ad to users. After an exposure, clicks can be collected from users, which are assumed to be induced by their preference or the attractiveness of the images; note, however, that the number of clicks is not automatically indicative of user preferences. In this paper, we use the term “preference” to describe both attractiveness and user preference. Furthermore, the term ad insights refers to observed phenomena that are discovered through statistical analysis of advertising data. Examples of ad insights are presented in the sections on datasets.

4. Auxiliary Attributes

Auxiliary attributes denote a superset of classical metadata, where metadata is a common term for data that describe a user or item. Currently, such information used in most studies comprises the personal data of users, which are sensitive and expensive to collect. We use the expression “auxiliary attributes” to describe expanded metadata that describe the content, which, in this study, is an ad image. We exploited several features, including visual–linguistic attributes, which are outlined in Table 1. In contrast to most existing image assessment studies [30, 17] that focus only on images for preference assessment, we used additional features as well. In several studies investigating CTR, the input data were full of metadata, such as user demographic data (e.g., gender and age) and ad-exposure event-related data (e.g., weekday and time). However, an ad image assessment model should be capable of using rich auxiliary attributes related to images, such as brands and embedded text [21], and colors [20, 16], including classic user-related metadata and ad display policy. Auxiliary attributes are crucial for assessing ad images. For example, the top row in Fig. 1 shows different ad images for the same product. Among these images, users prefer the image on the right, which uses embedded text (e.g., “please change the character’s outfit”) to request for their participation. Other attributes, such as the time of ad display, which are listed in the middle row, may also affect user engagement. Some cases may even be contrary to the overall statistical analysis results, as shown in the middle row’s right. For further demonstration, various ad images consisting of pixel arts, real photos, and embedded text are shown at the bottom row in Fig. 1. In this study, all these metadata were used together with additional data from various sources, e.g., embedded text, which is expected to express the intention of an ad. To allow computational models to learn these catchphrases, visual–linguistic attributes were used in the Real-Ad dataset. In addition, the dataset possesses an optical character recognition (OCR) attribute with titles and descriptions, which are common linguistic attributes in image datasets. Details on the utilized auxiliary attributes are explained in the datasets section.

5. M2FN: Multi-step Modality Fusion Network

M2FN consists of four main modality fusion steps as depicted in Fig. 2: i) input step, which integrates features using the conventional early-fusion method, wherein auxiliary attributes are concatenated to be processed in the next step; ii) low-level fusion step, which deals with specific features, such as dominant colors of ad images and auxiliary attributes; iii) spatial attention step, which addresses text-expression location in ad images; and iv) high-level fusion step, which involves abstracted visual features and auxiliary attributes, such as demographics, time, and semantics of linguistic information.

5.1. Low-Level Fusion Step

We hypothesized that by providing each layer of the network a hint of the complexity of the ad image, auxiliary attributes can improve prediction performance. To achieve this, we employed CBN [7] for low-level fusion. CBN modulates the layers of the network instead of performing simple concatenation. Additionally, it trains shallow neural networks that predict the scale factor parameter $\gamma$ and shift factor parameter $\beta$ in batch normalization from the input data. In VQA, for example, the question sentences are embedded
Table 1. Auxiliary attributes from the Real-Ad dataset.

| Feature name   | Description                                      | Format                                      |
|----------------|--------------------------------------------------|---------------------------------------------|
| Gender         | Gender of user                                   | {1(female), 2(male)}                       |
| Age            | Age of user                                      | {1(under 14), 2(15–19), …, 9(over 50)}    |
| Month          | Month                                            | {1(Jan), 2(Feb), …, 12(Dec)}              |
| Weekday        | Weekday                                          | {1(Mon), 2(Tue), …, 7(Sun)}               |
| Time           | Time                                             | {0, 1, 2, …, 23}                           |
| Position       | Position of ad in platform app                   | {1, 2, 3, 4}                               |
| Cate2          | Mid-category of item                             | {1(casual), 2(gambling), 3(mid-core), 4(core)} |
| Cate3          | Low-category of item                             | {1(RPG), 2(sports), 3(simulation), 4(action), 5(adventure), 6(strategy), 7(casino), 8(casual), 9(puzzle)} |
| Title          | Title of ad, embedded using BERT                 | 768-dimensional embedding vectors          |
| Desc           | Description of ad, embedded using BERT           | 768-dimensional embedding vectors          |
| OCR            | Detected from image using open API, embedded using BERT | 768-dimensional embedding vectors          |
| Domcol         | Dominant color of image labeled using K-means-based algorithm | {1(black), 2(blue), 3(brown), 4(green), 5(grey), 6(multiple), 7(pink), 8(red), 9(white), 10(yellow)} |

using long short-term memory (LSTM), and the question embedding $e_q$ is fed to the CBN module. In this study, instead of text questions, auxiliary attributes are used as input to the CBN. Before being input to the CBN, categorical attributes are encoded into a one-hot vector, whereas text information is represented as a bidirectional encoder representations from transformers (BERT)-based embedding vector.

Similar to batch normalization, CBN can be performed at any layer of a neural network. In this study, CBN was used to combine images and auxiliary attributes only after the first convolution of the early stage VGG-19, based on SOTA results.

5.2. Spatial Attention Step

The location of textual information and its expression in ad images have a crucial influence on the CTR of an ad. To address this, we introduced an attention mechanism that considers spatial relationships for modality fusion.

More specifically, image features and auxiliary embedding vectors were represented using $N_b \times C \times W \times H$ and $N_b \times dim_{aux}$ tensors, respectively, where $N_b$, $C$, $W$, and $H$ denote batch size, number of channels, width, and height, respectively. The auxiliary embedding vectors were replicated to be a $N_b \times C \times W \times H$ tensor. The auxiliary tensor was concatenated with the image feature tensor to be a $N_b \times (C + dim_{aux}) \times W \times H$ tensor. Thereafter, it was fed to a fully connected layer of an MLP. The resulting vectors $\{N_b, W \times H\}$ obtained using the SoftMax function became the attention matrix. This attention matrix was multiplied (Hadamard product) with image features to achieve a soft-attention map.

5.3. High-Level Fusion Step

The role of high-level fusion is to emphasize the effect of spatial relationships between visual features and auxiliary attributes by integrating two modalities near the output layer. Unlike in previous approaches, where a simple concatenation was used, a matrix element-wise multiplication between the output of the attention mechanism and auxiliary attributes was performed, as in [7]. To match the dimension size, affine transformation was performed using linear layers for each image feature and auxiliary attributes. The activation function of the hyperbolic tangent was then applied to both outputs, and the resulting vectors were multiplied element-wise. There was a significant difference in the regression performance with and without this fusion mechanism, which is detailed in the experiments section.

5.4. Loss Function

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

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

where $w_n$, $\hat{y}_n$, and $y_n$ denote the impressed number, predicted CTR, and real CTR, respectively, of the $n$-th instance, and $N$ represents the data size.

In addition, Kullback–Leibler divergence (KLD) was used as a loss function when the preference score of the
Figure 2. Overall structure from data input phase to multi-step modality fusion network (M2FN).

The data was represented as a distribution form similar to that of the AVA dataset [23]:

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

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

6. Datasets

6.1. Real-Ad Dataset

For evaluating the model, we used a large-scale dataset of click logs of ad image impressions, collected from a global online ad service in Japan in 2018. The total number of impressions is approximately 500 million and includes 3,747 distinct ad images, as shown in Table 2. While the training dataset was being constructed, aggregation and attribute selection were performed on the data. These processes are illustrated in Fig 3. Note that this dataset cannot be released because of confidentiality issues.

6.2. Data Construction

Before the dataset was constructed, irrelevant logs were discarded to mediate the survivorship bias [19]; instances of malicious users were removed before the ad banner was read.

Table 2 outlines the specifications of the Real-Ad dataset. One instance of raw data corresponds to one exposure event of an image ad with auxiliary attributes. Each instance has a label of 1 or 0, denoting whether it was clicked or not, respectively. Each ad was viewed multiple times by
Table 2. Details on Real-Ad dataset. Each instance in raw dataset represents one exposure of an ad. Aggregated 100+ and 500+ datasets were constructed from over 100 and 500 impressions, respectively.

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

different users. Moreover, each ad was displayed at different times, ad positions, etc.

Each instance was aggregated to CTR:

\[ y_n = \frac{M_n^c}{M_n}, \]

where \( M_n^c \) denotes the number of clicks, and \( M_n \) denotes impressions for the \( n \)-th unique instance.

The aggregated data were compressed and compared with the raw data, although these aggregated data remained suitable for ad image assessment. In the raw data, more than 99% of the impressions were not clicked, leading to severe imbalance problems. Because of this imbalance, label conflicts and errors were frequent at an instance level. In contrast, aggregated CTR values were represented as real numbers with a higher resolution, which facilitated comparative analysis with existing image assessment studies \[30\].

We separated the training and test datasets. Every experimental result was obtained from unseen image data from the test set.

6.3. Statistical Analysis of Real-Ad Data: Ad Insights

Through statistical analysis methods such as analysis of variance (ANOVA) and logistic regression analysis, we investigated the influence of ad images and their auxiliary attributes on CTR. Through statistical analyses, we identified the relationship between specific CTR responses and each auxiliary attribute. Generally, these characteristic movements (ad insights) are strategically considered points identified by marketers to increase the CTRs of ads. Fig. 4(a) illustrates that the ad insights of three of these CTR-effective attributes usually affect the ad performance regardless of the category of the item. For example, the age bar plot in Fig. 4 shows that game ads have a higher probability of being clicked when they are shown to elder users. In addition, we reaffirmed that the dominant color and visual–linguistic attributes also significantly influence the effectiveness of ads \[21\].

Furthermore, the time and month bar plots in Fig. 4(a) show the influence of time-sequential attributes on CTR. In detail, ads published at mid-dawn (2 AM to 4 AM), mid-morning (9 AM to 11 AM), noontime (12 PM), and mid-afternoon (3 PM to 5 PM) exhibited better results than those published at other hours. Additionally, the beginning and end of a week exhibited higher CTRs than the other days. Finally, the lowest CTRs occurred during the third quarter of the year, i.e., July, August, and September.

In contrast, many cases exist that do not follow these ad insights. Fig. 4(b) shows the unusual cases discovered during our statistical analyses of Real-Ad, and the results of CTR prediction by our model. These results demonstrate that preference prediction on ad images is highly challenging. For example, CTR distributions for the same ad image and auxiliary attributes varied depending on different time attributes, which can be considered a covariate shift.

Our statistical analyses provided some interesting observations. In most images, the brand logo was observed in a variety of ways. The logo shape is known to have a considerable influence because it conveys the image of the brand \[32\]. Additionally, there were specific dominant colors (black, blue, brown, etc.) that correlated with higher CTR instances. Designing to represent even low-level visual parts is therefore important. According to these insights, a sophisticated model design must leverage auxiliary attributes and integrate them with ad images effectively.

6.3.1 Attribute Selection

The raw Real-Ad dataset contains approximately 40 attributes, which were collected with click logs. Some of these attributes are redundant or weakly related to CTR. Attribute selection was therefore performed to cleanse the dataset and select only the auxiliary attributes; it was not performed on the images. To select attributes that are crucial for CTR prediction, we employed ANOVA and logistic regression analysis at a significance level of 0.05. Through these analyses, we selected nine key attributes responsible for CTR. These attributes are outlined in Table 1.

6.3.2 Attribute Preprocessing

The dominant color and visual–linguistic auxiliary attributes had to be preprocessed for computational models to be trained. These auxiliary attributes included title, description, and embedded textual expression of ad images (OCR). The dominant color of an ad image was represented as an element of a predefined color set, which included 10 colors. We used \( K \)-means clustering with a minimum covariance determinant (MCD) distance to extract the intermediate dominant color from each image. After the intermediate color was obtained, it was mapped to one of the predefined dominant colors.

\[3\http://info.rocketfuel.com/\]
Figure 4. Ad insights and task introduction. (a) Ad insights and statistical analysis results for Real-Ad. (b) Counter-evidence of insights derived from (a) and CTR prediction results of M2FN. Although table contents shown in (b) differ from results in (a), M2FN exhibits reliable performance. CTR values are divided by the smallest CTR in each table because of confidentiality issues.

In addition, we supplemented the data with visual-linguistic auxiliary attributes, such as title, description, and OCR result. The title and description were collected from the ad application form. OCR results were obtained using an open-source API to identify the letters in the ad images. This API detects scene text by exploring each character and affinities between characters. Thus, it can flexibly detect complicated scene text images, such as curved or deformed texts, which frequently appear in ad images. Unique sentences were collected for each visual-linguistic auxiliary attribute, with 1,583, 2,695, and 2,695 sentences being collected for title, description, and OCR, respectively. We embedded these sentences into vectors using BERT. Other non-linguistic auxiliary attributes were represented via one-hot encoding.

While grouping and summing CTRs, we considered only the instances with more than 100 and 500 impressions for the feasibility assessment of CTR values. The levels of attributes that have less than 50,000 impressions were merged with the closest level to cope with biases in the dataset. The total number of dimensions of auxiliary attributes was 2,383, of which 2,304 were visual-linguistic attributes (with 768 dimensions for each).

### 6.4. Benchmark Dataset: AVA Dataset

M2FN can contribute to a conventional image assessment task. We evaluated our model on AVA, which is an image dataset designed for aesthetic assessment studies. This dataset comprises images and their annotations, including aesthetic, semantic, and photographic-style annotations. The photographic-style annotations are composed of 14 styles (complementary colors, duotones, HDR, etc.) that represent the camera settings. However, these annotations were not used in this study because an excessive number of cases exist without annotations.

Semantic annotations are textual tag data, e.g., nature, black and white, landscape, animal. There are approximately 200,000 images, and each image has at least one tag. The tag set size is 67 when no-tag cases for image-only CTR predictions are added. Each text tag was represented as a BERT embedding vector. In this study, embedding vectors with 768 dimensions were used to represent visual-linguistic auxiliary attributes.

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

7. Experimental Results

7.1. Experimental Setup

Evaluation Metrics: Depending on the dataset, the resulting outputs were in distribution (10 buckets) or scalar value (score) form. Performance evaluation was based on ranking. Spearman rank correlation (SPRC) and the linear correlation coefficient (LCC) were computed to rank the output scores and compare them with the ground-truth ranking. If the output score was in the form of a distribution, the SPRC and LCC were calculated and examined for both the mean and standard deviations.

Model Parameter Settings: The shape and the number of training parameters for each layer in M2FN are outlined in Table 3. Because VGG19 is widely used and well-known, the table contents are omitted, and only the total number of parameters is presented.

The benchmark and Real-Ad datasets have distinct hyperparameter 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 MLP for embedding representations inside the module. When the AVA datasets, which have relatively small numbers of auxiliary attributes, were trained, the hyperparameters of the modules were 64, 512, and 512, respectively. When the Real-Ad dataset was used, the hyperparameters were set to 256, 512, and 512. The hyperparameters are highlighted in bold in the third column of Table 3. This decision was based on over 20,000 automated parallel hyperparameter-searching experiments.

The batch size of the dataset was 128, and five P40 GPUs were used for training. All models were trained for 100 epochs.

All the experiments are implemented and performed based on NAVER Smart Machine Learning (NSML) platform [13] with AutoML [14] [25].

7.2. Comparison of Prediction Performance

We compared our M2FN with a simple concatenation-based modality fusion model and neural image assessment (NIMA) [30], which is an SOTA method for image assessments, as the baselines. Table 4 compares their 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 provides only a mean of the CTRs. Because our proposed method requires training on a CTR value, the CTR distributions were approximated based on a log-normal distribution that resembles CTRs. When converted to this distribution, it is represented as “dist”; Otherwise, it is represented in Table 4 as “regr.” Previously, the best performing model (NIMA) on the benchmark dataset used EMD as a loss function. The asterisk (*) signifies that, instead of EMD, Kullback–Leibler divergence (KLD) was applied as a loss function (for “regr.” cases, the impression-weighted MSE was used, as mentioned in a previous section). Regardless of the number of impressions (100+ or 500+), our method clearly performed better. For this dataset, KLD, compared with EMD, appears to be a better choice for the loss function.

Table 5 compares our models with baseline models using the AVA dataset to verify that our model can be effective at image assessments, with a quantitative comparison between NIMA (InceptionV2) [8], the previous SOTA model, and M2FN.

The results of the comparison based on the necessity and location of multi-modal fusion are also listed in the table. “-cat” denotes simple concatenation of both the auxiliary and image, whereas “-LLF” represents low-level fusion. If a data entry in the table is denoted by “-,” the model was trained on images only (no auxiliary attributes).

In Table 5, the number of training parameters are included to show the complexity of each model. Experiments with InceptionV4, a deeper version of the previous SOTA, confirmed that deepening the network does not always improve the task performance. This inference can also be verified using the results for “Resnet101,” which is widely used for image-related tasks [12]. Although it has numerous layers, the model is not reliable. From the “-cat” results for “InceptionV4-cat” and “VGG19-cat,” simple concatenation was inferred to be degrading the performance, compared with the corresponding image-only cases (“InceptionV4” and “VGG19”). In contrast, in the case of “Resnet101-cat,” simple concatenation enhanced the performance. Because the visual feature extraction module (“Resnet101”) is large, it overfits easily and severely degrades performance. Enhanced performance is assumed to have been achieved primarily with auxiliary attributes. “VGG19” exhibited better performance compared with the overall results, and the best performance among image-only cases. Furthermore, because we reduced the sizes of fully connected layers at the end of M2FN via hyperparameter searching, M2FN is less complex than VGG19, which is partially nested in M2FN for visual feature extraction.

Furthermore, we conducted other comparison experiments to assess how the location of low-level fusion affects performance. Because VGG19 consists of five blocks, the five numbers in the parentheses depict the location of the fusion. If fusion is performed inside the block, it is written as “1.”

The results demonstrated better performances when the fusion was located closer to the input stage. Existing studies apply CBN to all blocks [7]; however, this adversely affects the task performance.
Table 3. Shape and number of training parameters of each layer in M2FN.

| Layer name       | Symbol | Shape           | # parameters |
|------------------|--------|-----------------|--------------|
| Input            |        |                 |              |
| CBN              |        |                 |              |
| VGG19 w/o classif. layers |   |                 |              |
| Attention Stack  |        |                 |              |
| High-level fusion|        |                 |              |
| Output           |        |                 |              |

Table 4. Performance comparison on the Real-Ad dataset. In the header row, m and std. represent mean and standard deviation, respectively. Dist. and Regr. denote distribution and regression, respectively.

| 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     |
| 500+ Dist.     |         |        |            |           |
| NIMA           | 0.325   | 0.343  | -          | -         |
| M2FN           | 0.384   | 0.381  | -          | -         |
| 100+ Regr.     |         |        |            |           |
| NIMA           | 0.308   | 0.249  | 0.165      | 0.166     |
| M2FN           | 0.453   | 0.448  | 0.190      | 0.190     |
| 500+ Regr.     |         |        |            |           |
| NIMA           | 0.501   | 0.451  | -          | -         |
| M2FN           | 0.561   | 0.530  | -          | -         |

As mentioned in the introduction section, ad images have a complex structure that differs from those of ordinary images. These results support our intuition behind the multi-step structure, based on the ad insights and previous studies. M2FN, which uses multi-step modality fusion, performed best on the benchmark dataset.

7.3. Ablation Study

Table 6 and Table 7 compare the performance changes depending on the presence or absence of four major modules for each dataset. The four primary modules include auxiliary attributes in the input step, low-level fusion, attention, and high-level fusion. In the four consecutive columns, “O” denotes activating, and “×” signifies deactivating. “××××” is a model trained using only images on vanilla VGG-19, the base of M2FN. In the case of “O × × ×,” where the auxiliary attributes are activated but the CBN is deactivated, the image and auxiliary attributes are learned through concatenation before the last fully connected layer.

The results in Table 7 demonstrate how essential images are to the ad image assessment task. The results of the four baseline models and M2FN were achieved using only auxiliary attributes without ad images. For the M2FN comparative result, we conducted experiments on blank images by filling the input with zeros. Better results than those of the four baseline models were exhibited, although M2FN has many parameters. Furthermore, a large difference in terms of performance between baseline models and M2FN was observed. In particular, there was a difference of approximately 0.15 compared with the result when only the image was considered (“××××” result, the sixth row in Table 7).

According to the results in the seventh and eighth rows indicating the presence and absence of low-level fusion, CBN enhances performance. The ninth row of the results also demonstrates the strength of M2FN. This row indicates the presence of high-level fusion. The best performance was achieved when all four modules were combined. The integrated results of the four modules proved that our approach is suitable for evaluating impressions. In Table 6, the row of NIMA data represents existing SOTA performance on AVA. Below the NIMA row are data on the performance of M2FN; M2FN outperforms the previous record.

The results in Table 8 are of an ablation study comparing performance changes with and without auxiliary attributes in the Real-Ad dataset. We speculate that content-related attributes (weekday, time, position, categories, and color) are more informative than other attributes, e.g., user-related attributes, for evaluating impressions. We expected better performance to be achieved even if we use only content-related attributes among auxiliary attributes. The results prove our hypothesis.

In Table 8, the performance is sorted in ascending order.
Table 5. Performance comparison on A V A dataset. In the header row, m and std. represent mean and standard deviation, respectively. “-cat” denotes simple concatenation of both auxiliary and image, which represents models without low-level fusion in existing studies. “-LLF” signifies low-level fusion. If there is no additional explanation and table entry is denoted by “-,” the model was trained only on images. M1FN is a prototype model of M2FN. Numbers in brackets refer to blocks wherein low-level fusion is either conducted (1) or not conducted (0).

| Models                | # parameters | SPRC(m) | LCC(m) | SPRC(std.) | LCC(std.) |
|-----------------------|--------------|---------|--------|------------|-----------|
| NIMA(InceptionV2)    | 20M          | 0.612   | 0.636  | 0.218      | 0.233     |
| NIMA(VGG16)          | 134M         | 0.592   | 0.610  | 0.202      | 0.205     |
| InceptionV4          | 40M          | 0.606   | 0.619  | 0.183      | 0.198     |
| InceptionV4-cat      | 40M          | 0.576   | 0.581  | 0.160      | 0.155     |
| Resnet101            | 44M          | 0.496   | 0.470  | -          | -         |
| Resnet101-cat        | 44M          | 0.549   | 0.560  | 0.192      | 0.189     |
| Resnet101-LLF        | 47M          | 0.558   | 0.561  | 0.233      | 0.230     |
| Resnet50-LLF         | 28M          | 0.544   | 0.552  | 0.225      | 0.223     |
| VGG19                | 139M         | 0.617   | 0.622  | -          | -         |
| M1FN(VGG19-cat)      | 139M         | 0.572   | 0.581  | 0.198      | 0.197     |
| M2FN 1,1,1,1,1       | 78M          | 0.552   | 0.553  | 0.183      | 0.176     |
| M2FN 0,0,0,0,1       | 65M          | 0.572   | 0.579  | 0.200      | 0.198     |
| M2FN 0,0,1,1,1       | 72M          | 0.575   | 0.573  | 0.203      | 0.196     |
| M2FN 0,0,1,0,0       | 65M          | 0.590   | 0.601  | 0.230      | 0.229     |
| M2FN 1,1,1,0,1       | 72M          | 0.612   | 0.613  | 0.253      | 0.266     |
| M2FN 1,0,0,0,0       | 65M          | 0.630   | 0.640  | 0.310      | 0.322     |

Table 6. M2FN ablation study results on A V A.

| Module          | Mean       | LCC       | SPRC       | LCC       | SPRC       | LCC       |
|-----------------|------------|-----------|------------|-----------|------------|-----------|
| NIMA(Inception-v2) | 0.612  | 0.636  | 0.218  | 0.233  |
| NIMA(VGG16)     | 0.592 | 0.610 | 0.202 | 0.205 |
| InceptionV4     | 0.606 | 0.619 | 0.183 | 0.198 |
| InceptionV4-cat | 0.576 | 0.581 | 0.160 | 0.155 |

If an attribute is excluded and the performance drop is significant, it indicates that the attribute is essential. Therefore, we observe the following: i) The time-sequential attribute (weekday) is vital. ii) According to the bottom of the table (-(title + ocr)), the description has negligible effect when used alone. However, description has a significant impact when it is integrated with other attributes. iii) As previous studies have asserted, the color attribute proved essential for advertising content in the table. iv) The user attribute is not crucial, even though it has been extensively used in previous studies. v) However, when the user attribute is combined with others (“text aux” in the table), performance degradation seems to be substantial. User attribute is assumed to enhance performance when integrated with other attributes.

Table 7. M2FN ablation study results on Real-Ad.

| Module                     | Mean       | LCC       | SPRC       | LCC       | SPRC       | LCC       |
|---------------------------|------------|-----------|------------|-----------|------------|-----------|
| Only Linear regression    | 0.239  | 0.195  | 0.166  | 0.156  |
| Only SVM regression (RBF) | 0.102  | 0.160  | 0.183  | 0.126  |
| Only SVM regression (poly)| 0.242  | 0.177  | 0.201  | 0.197  |
| Only MLP                  | 0.241  | 0.193  | 0.240  | 0.217  |

7.4. Advertisement Image-Assessment Results

Fig. 5 shows ad image-assessment results for M2FN. Results that have a small difference between the true CTRs and predicted CTRs were selected from the test set. We chose time, age, and month as attributes to determine how this model works according to ad insights. The results are outlined in each row. M2FN is observed to perform better in most cases.

In contrast, the model is robust even under exceptional circumstances that do not follow the insights. There are unusual cases that counter previous statistical evidence, as in
Figure 5. Ad image assessment results of usual cases that follow ad insights of single condition or combination. CTR values are divided by the smallest CTR in each table because of confidentiality issues.

| Time   | CTR  | Pred |
|--------|------|------|
| 4      | 1.352| 1.333|
| 18     | 1.241| 1.204|
| 23     | 1    | 1    |

| Age    | CTR  | Pred |
|--------|------|------|
| 15-19  | 1.286| 1    |
| 20-24  | 2.643| 2.786|

| Month  | CTR  | Pred |
|--------|------|------|
| Jun.   | 2.516| 2.654|
| Jul.   | 1.032| 1    |

| Multiple | CTR  | Pred |
|----------|------|------|
| Sat. 9   | 2.112| 1.959|
| Sat. 12  | 1    | 1.375|
| Sun. 8   | 3.272| 3.289|
| Sun. 22  | 1.539| 1.509|

| Week    | Time | CTR  | Pred |
|---------|------|------|------|
| 20-24   | Thu. | 1.071| 1    |
| 25-29   | Sun. | 1.418| 1.497|

| Age | Week | CTR  | Pred |
|-----|------|------|------|
| 15-19 | 8    | 1.098| 1.19 |
| 15-19 | 18   | 1.058| 1    |
| 20-24 | 18   | 1.756| 1.492|

Fig. 6 The results in the first row suggest that CTR increases with age. Other noticeable results are highlighted.
Figure 6. Ad image assessment results of unusual cases that do not follow ad insights of single condition or combination. Dashed rectangle indicates that results inside boxes are of same images with different visual-linguistic auxiliary attributes (title, desc, and OCR). CTR values are divided by the smallest CTR in each table because of confidentiality issues.
with dashed rectangles. These are results for identical images but different visual–linguistic auxiliary attributes, such as title, description, and OCR. For example, two ads with identical images may have different catchphrase texts, such as “this game is a current big hit!” and “we payback 50% of your purchases.” When the visual–linguistic information displayed to the users is different, the resulting CTRs are also affected. From the results, M2FN is proven capable of handling these unusual cases.

7.5. Further Analysis

The results of experiments and statistical analyses provide some insights that can be used to assess the attractiveness of ad images created by human designers. Based on the Real-Ad statistical analyses (ad insights), we reaffirm the results of previous studies, i.e., a single or a combination of auxiliary attributes influences performance. As listed in Table 8, four attributes are particularly influential on performance. In addition, with the combined attributes, the fused representation in M2FN fires like an activation function in neurons (e.g., they fire, united but not divided). This would be interesting from the perspective of cognitive science and marketing research.

In Fig. 7, each layer of M2FN is leveraged using Grad-CAM [28] to visualize the operation of each layer in the model. Grad-CAM is a tool that shows which part of an image the neural network sees and how it makes a decision on a particular label. It facilitates the layer in understanding the importance of each neuron with the gradient information. In Fig. 7, heatmaps are drawn, with red indicating high importance for an 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). Thus, in the attention map (fourth column), a visualization result also shows that partial salience with characters and embedded text is obtained. Based on an analysis of these results, our model demonstrably learns where the visual–linguistic elements that humans consider salient in advertising images are located.

The visualization results also demonstrate the purpose and structural benefits of multi-step fusion, effectively fusing different levels of ad images by each step, and aiding in in-depth analyses of the ad images. The model provides a hint regarding the visual–spatial saliency of humans toward the ads. This supports the results of existing literary, cognitive, and marketing science studies. These findings show that M2FN has achieved preliminary success in learning human aesthetic preferences in advertising images.

In Fig. 8, there are several similar ad images for the same item. M2FN works robustly against transformations such as changing the position of the objects or changing colors. As the multi-step structure copes with various conceptual levels, these results indirectly prove that the model performs well with low-level fine-grained controlled images, e.g., changing the angle, font, etc.

Although we achieved excellent performance for the ad image-assessment task, our model is limited to handling only regression problems. It is possible to predict the number of CTRs that would be achieved under a certain condition, but it cannot generate an image that will record a high CTR. Accordingly, we are currently developing a generative model that produces preferable advertising images and a service that provides automated textual advice on the design of ad images and ad displays.

8. Conclusion

In this study, we proposed a model for predicting user preference of ad images. The model was tested on an in-house large-scale dataset, containing ad images and auxiliary attributes, called the Real-Ad dataset. We statistically explored the Real-Ad dataset, focusing on the effect of images and auxiliary attributes on human preferences represented as CTR. Inspired by ad insights determined via the analyses, we proposed the novel multi-step modality fusion network (M2FN) model and employed extensive comparison experiments to verify its design. The M2FN effectively integrates ad images and their auxiliary attributes to predict CTR. We evaluated M2FN on the Real-Ad dataset and subsequently validated our method on a benchmark image-assessment dataset and the AVA dataset to verify whether
our approach can be applied to conventional image assessment. M2FN achieved better performance on both datasets compared with those of previous SOTA models. With an extensive ablation study, we investigated how each modal-
ity fusion works and which auxiliary attributes largely influence user preferences.

9. Acknowledgments

The authors thank NAVER AI LAB, NAVER CLOVA for constructive discussion and LINE Corp. for preparing data. We would like to thank Min-Whoo Lee, Chung-Yeon Lee, Taehyeong Kim, Dharani Punithan, Christina Baek, Chris Hickey, Seonil Son, Hwiyeol Jo, and Dong-Sig Han for editing and reviewing this manuscript for English language.

Funding: This research was partly supported by the Institute for Information & Communications Technology Promotion (2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-DMI, 2019-0-01367-BabyMind) and Korea Institute for Advancement Technology (P0006720-GENKO) grant funded by the Korea government. This research was also partly supported by the ICT at Seoul National University.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 2425–2433, Santiago, Chile, Dec 2015.

[2] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4714–4722, Seoul, Korea (South), Oct 2019.

[3] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9357–9366, Long Beach, CA, June 2019.

[4] Junxuan Chen, Baigui Sun, Hao Li, Hongtuo Lu, and Xian-Sheng Hua. Deep ctr prediction in display advertising. In Proceedings of the 24th ACM International Conference on Multimedia, MM ’16, page 811–820, Amsterdam, The Netherlands, 2016.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In Computer Vision – ECCV 2006, pages 288–301, Graz, Austria, 2006.

[6] Terry Daugherty, Ernest Hoffman, and Kathleen Kennedy. Research in reverse: Ad testing using an inductive consumer neuroscience approach. Journal of Business Research, 69(8):3168–3176, 2016.

[7] Harm de Vries, Florian Strub, Jeremie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In Advances in Neural Information Processing Systems 30, NIPS 2017, pages 6594–6604, Long Beach, California, Dec. 2017.

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

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019.

[10] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In CVPR 2011, pages 1657–1664, Providence, RI, June 2011.

[11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 457–468, Austin, Texas, Nov. 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, Las Vegas, NV, June 2016.

[13] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwon Kim, et al. Nsmi: Meet the mlaas platform with a real-world case study. arXiv preprint arXiv:1810.09957, 2018.

[14] Jinwoong Kim, Minkyu Kim, Heungseok Park, Ernar Kusdavletov, Dongjun Lee, Adrian Kim, Ji-Hoon Kim, Jung-Woo Ha, and Nako Sung. Chopt: Automated hyperparameter optimization framework for cloud-based machine learning platforms. arXiv preprint arXiv:1810.03527, 2018.

[15] Kyung-Min Kim, Seong-Ho Choi, Jinhwa Kim, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. In Computer Vision – ECCV 2018, pages 698–713, Munich, Germany, 2018.

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

[17] J. Lee and C. Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1191–1200, Seoul, Korea (South), Oct 2019.

[18] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In Computer Vision – ECCV 2018, pages 698–713, Munich, Germany, 2018.

[19] Marc Mangel and Francisco J Samaniego. Abraham wald’s work on aircraft survivability. Journal of the American Statistical Association, 79(386):259–267, 1984.

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

[21] Andrew A Mitchell. 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, 1986.

[22] Kaixiang Mo, Bo Liu, Lei Xiao, Yong Li, and Jie Jiang. Image feature learning for cold start problem in display advertising. In *Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI’15*, page 3728–3734, Buenos Aires, Argentina, July 2015.

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

[24] D. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, Salt Lake City, UT, June 2018.

[25] Heungseok Park, Jinwoong Kim, Minkyu Kim, Ji-Hoon Kim, Jaegul Choo, Jung-Woo Ha, and Nako Sung. Visual-hyper tuner: Visual analytics for user-driven hyperparameter tuning of deep neural networks. In *Demo at SysML Conference*, 2019.

[26] J. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie. MFAS: Multimodal fusion architecture search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6959–6968, Long Beach, CA, June 2019.

[27] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98 – 125, 2017.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, Italy, Oct 2017.

[29] Nako Sung, Minkyu Kim, Hyunwoo Jo, Youngil Yang, Jing-woong Kim, Leonard Lausen, Youngkwon Kim, Gayoung Lee, Donghyun Kwak, Jung-Woo Ha, et al. Nsml: A machine learning platform that enables you to focus on your models. *arXiv preprint arXiv:1712.05902*, 2017.

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

[31] X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, Dec 2013.

[32] Michael Francis Walsh. *Consumer response to logo shape redesign: the influence of brand commitment*. PhD thesis, University of Pittsburgh, 2006.

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Hugging-face’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.