Predicting the Appreciation of Multimodal Advertisements

Serra Sinem Tekiroğlu (serrasinem@gmail.com)
Carlo Strapparava (strappa@fbk.eu)
Gözde Özbal (gozbalde@gmail.com)
FBK - irst, Trento, Italy

Abstract

Creativity is an essential factor in successful advertising where catchy and memorable media is produced to persuade the audience. The creative elements in the visual design and in the slogan of an advertisement elevate the overall appeal providing a perceptually grounded attractive message. In this study, we propose the exploitation of creativity cues in textual and visual information for the appreciation prediction of multimodal advertising prints. Moreover, as a novel dimension space of multimodality, we propose using the human sense (i.e., sight, hearing, taste, and smell) information embedded in the language. Our findings show that sensorial information is an invaluable indication of whether the advertisement is appreciated or not. Furthermore, combining linguistic and visual models significantly improves the unimodal appreciation detection performances.

Keywords: advertising creativity; human senses; multimodal creativity

Introduction

Creativity in advertising is an entangled, multi-dimensional phenomenon that reflects the complex structure of human creativity. A catchy and memorable advertisement is coherent and captivating. It is carefully designed with a diverse range of approaches including the ways of visualizing concepts, the use of rhetorical devices, such as exaggeration, paradox, metaphor and analogy, and taking advantage of shock tactics and humour (Pricken, 2008). In case of the advertising prints, visual and textual contents are designed to have a complementary and coordinated meaning. Advertising makes use of sensory and linguistic sensorial information heavily in order to reach the customers and persuade them. Elder and Krishna (2009) propose that multi-sensory ads induce higher taste perceptions than ads focusing on taste alone. They also state that using multiple senses in slogan increases the positive thought about the advertised food product. As a way of improving advertising communication, Percy (1982) suggests the use of concrete and high imageary words and concepts to stimulate better recall, better comprehension of the advertised message leading to an easier and more accurate understanding of the ad. Another creativity infusion strategy in ad production is using sensory words especially generating linguistic synaesthesia as an imagination boosting tool (Pricken, 2008). The slogans '*The taste of a paradise*' (Bounty bar commercial), where the sense of *sight* is combined with *taste*, and '*Hear the big picture*'(CBC Radio One commercial), where *sight* and *hearing* are merged, can be considered as the examples of linguistic synaesthesia.

As a topic being on the rise in computational linguistics, multimodality is mostly exploited by adding other modalities on top of the linguistic models to perceptually ground the current tasks. For instance, semantic representations benefit from the reinforcement of linguistic modality with visual (Bruni, Tran, & Baroni, 2014) and auditory (Kiela & Clark, 2015) modalities. In the same manner, we propose devising visual modality in collaboration with linguistic modality in the appreciation detection task. To our knowledge, this is the first study aiming to identify multimodal appreciation in a computational manner. Moreover, the multimodality of the dataset stands out amongst the others since the linguistic channel of an ad is complementary to the visual channel instead of being a scene description or an image label. This study focuses on the appreciation of the advertising print by the advertising professionals and communities instead of the appreciation by the audience/customer of the advertised product.

In this paper, we investigate the appreciation level of multimodal advertising prints focusing on the creativity cues in the slogan and in the corresponding image. We use a set of fundamentally creative artworks; an advertising dataset which is composed of 4265 images and corresponding slogans. The objective of this paper is twofold: i) to capture the potency of sensorial dimension of semantics as a creativity cue in the language along with various creative properties both in visual and linguistic modalities, ii) to develop a multimodal appreciation detection model. We utilize a random forest model trained on a dense feature set extracted from the slogans, ad categories and product types for the linguistic modality. For the visual modality, we employ a fine-tuned convolutional neural network model and a random forest model trained on the observable features of the images to determine whether the appreciated images display common visual characteristics and whether these characteristics have a distinctive effect on the overall appreciation of a multimodal ad.

Related Work

Considering that the essential focus of this study is computational creativity and multimodality, we summarize the most relevant studies conducted on these topics. Elgammal and Saleh (2015) quantify the creativity in paintings within the context of historical creativity where creative paintings adequately differ from the antecedent paintings and influence the subsequent. They present a computational framework that is

based on a creativity implication network.

Regarding the linguistic creativity, Özbal, Pighin, and Strapparava (2013) present a creative sentence generation framework, BRAINSUP, on which several semantic aspects of the output sentence can be calibrated. The syntactic information and a huge solution space are utilized to produce catchy, memorable and successful sentences. Kuznetsova, Chen, and Choi (2013) focus on identifying creativity in lexical compositions. They consider two computational strategies, first investigating the information theoretic measures and the connotation of words to find the correlates of perceived creativity and then employing supervised learning with distributional semantic vectors. Alnajjar, Kundi, Toivonen, et al. (2018) propose a methodology to automatically create slogans for a target concept and its adjectival property by first generating metaphors, based on a metaphor interpretation model. They produce a semantic space with the generated metaphors and use the semantic space to fill the slogan skeletons extracted from the existing slogans. They evaluate the slogans through crowd-sourcing with respect to the relatedness of the slogan to the concept and property, the correctness of the language, the metaphoricity, the catchiness, attractiveness and memorability, and the overall appropriation of the expression as a slogan.

Concerning the multimodality, Bruni, Boleda, Baroni, and Tran (2012) analyze the affect of different types of visual features such as SIFT and LAB on semantic relatedness task, and present a comparison of unimodal and multimodal models. Sartori et al. (2015) experiment on a complementary multimodal dataset similar to ours. They explore the influence of the metadata (i.e., titles, description and artists statement) of an abstract painting for the computational sentiment detection task. For the combination of modalities, they propose a novel joint flexible Schatten p-norm model exploiting the common patterns shared across visual and textual information. Shutova, Kiela, and Maillard (2016) exploit visual modality to improve the metaphor detection performance while Zadeh, Chen, Poria, Cambria, and Morency (2017) apply multimodal input to sentiment analysis.

Creativity in Advertising Prints

The creativity elements and dimensions in advertising have been investigated thoroughly. Ang and Low (2000) explore the influence of dimensions of creativity such as novelty (expectancy), meaningfulness (relevancy), and emotion (valence of feelings) to the effectiveness of the advertisement. While novelty could be identified as the unexpectedness and out-of-box degree of an advertisement, meaningfulness is the relevancy of the advertisement to the message aimed to be conveyed. The third dimension, emotional content, focuses on the feelings awakened in the audience. These three dimensions should manifest themselves in a creative advertising media. Smith, MacKenzie, Yang, Buchholz, and Darley (2007), on the other hand, elaborate on the divergence, which is the encapsulation of novel, different, or unusual elements, in ads proposing that the most significant characteristic of creative ads is their divergence. In addition to the above-mentioned general dimensions of advertising creativity, we specifically focus on the sensorial elements and their effect in the objective creativity level. Sensorial language makes use of multiple senses to induce higher taste perceptions (Elder & Krishna, 2009). Multiple senses in the advertising text trigger the positive thinking in the audience. Using highly imageable, in other words highly sensory words, helps to convey the advertised message better and easily. Finally, linguistic synaesthetia is a specific but a very significant way of effective advertising. Furthermore, for visual creativity in advertising, we focus on capturing the divergence factors through a transfer learning mechanism built on top of a deep learning image classification model and artistic values by taking advantage of the observable visual features in the image.

Multimodal Advertising Creativity Dataset

To investigate the appreciation of a multimodal advertising print, we first need to identify a dataset that reflects relatively upper and lower levels of appreciation from human subjects. To this respect, we chose AdsOfTheWorld[1], which has a wide range of coverage of ads considering its characteristic of being a social network that aims to inspire the advertising professionals. The members of the website can share their advertisement artwork, rate and discuss the ads created by others. The published advertising prints are diverse in terms of the level of creativity and ratings such that some ads are award-winning while some are highly disfavored by the community. We collected the ad images, their slogans and meta-data from AdsOfTheWorld[2]. The meta-data of an ad includes the average user rating, which is an integer within the range from 1 to 10, the number of raters, brand name and category.

While constituting the appreciated and unappreciated classes, namely *AP* and *UNAP*, we considered the opposite endings of the rate scale to distinguish the appreciation levels of advertisements as much as possible. We also paid regard to the number of instances that we obtained after filtering in order to have sufficient data for generalization of the *AP* and *UNAP* classes for training a classifier. In order to avoid feeding noise to our models, we empirically determined a minimum number of votes to postulate an average rating as reliable. To this end,we incorporated an ad into our final dataset if it is voted by at least 20 users and if it has an average rating in the range from 1 to 4 for *UNAP* class, or in the range from 7 to 10 for *AP* class. Finally, we eliminated the improper image styles, such as photographs of the billboards or images containing only textual content, from the dataset so that we can guarantee each image and its respective slogan contribute to the targeted message. From the final dataset that contains 4265 images-slogans, we sampled 3265 instances for training, 100 instances for development and 900 instances for testing. While the development and test sets are perfectly balanced for both classes, the training set includes

[1] http://adsoftheworld.com

[2] AdsOfTheWorld has recently changed its interface, no more showing the user ratings.

1470 *UNAP* and 1795 *AP* instances. During the sampling, we paid attention to putting the ads from the same brand into the same set since a slogan for a brand can be paired with various visual designs leading to more than one instance with the same linguistic input. As an additional meta-data, we collected the type of the products since category labels are considerably high-level. For instance, the category *House, Garden* includes a great variety of product types, such as furniture, laundry detergent, or insect killer. Utilizing product type and category labels as reference points allows us to appraise the meaningfulness of a slogan which affects its appreciation level notably. Advertisement 1[3] and Advertisement 2[4] exemplify highly appreciated and highly unappreciated samples in the final dataset, respectively. Although these advertising images seem to be very similar at the first glance with an object in the middle of the frame and in front of a blurred background, the subtle and creative details in the pictures, such as perplexing design of an octopus and sailboat made of pages of a book in Advertising 1 aims to immediately draw the attention of the audience. It holds an average rating of 10, which is the highest appreciation score, and is rated by 23 users. On the other hand, the obvious irrelevance of the main object in the image to the advertised message is a sign of an unappreciated design. Advertisement 2 with the slogan "Can't sleep?" promotes a tea brand that helps with sleeping problems by using a clearly irrelevant main object in the image. It has an average rating of 1 and its unappreciated label is trusted considering that it is rated by 171 users.

Appreciation Prediction Experiments

We design the appreciation prediction experiment of multimodal advertising prints exploiting the creativity dimension cues in the slogan and in the corresponding image, considering the studies done on the dimensions of creativity (Smith et al., 2007; Ang & Low, 2000; Elder & Krishna, 2009). To be more precise, we intend to capture surprisal, novel, meaningful, emotional, unusual and perceptual properties in an advertising slogan. Moreover, we aim to extract artistic components in the visual elements along with the latent visual descriptions and patterns.

Appreciation Detection on Slogans

For the textual model, we hypothesize that the creativity elements in the ad slogan can be mapped to features that are useful to detect the appreciation of an advertisement.

Surprisal (Self Information), as contributing to novelty/expectancy (Ang & Low, 2000) and surprisal(Smith et al., 2007) dimensions of creativity, can be interpreted as the information load of a specific outcome of an event. We calculate the self-information s of a bigram B by $s(B) = -log(p(B))$ exploiting the conditional probability distribution of bigram model trained on the corpus. We obtain the slogan self information as the average s of the bigrams extracted from the sentence.

Domain Relatedness features for the slogan are generated to address the meaningfulness (relevance) dimension (Ang & Low, 2000) of creativity. We expect that a meaningful slogan could contain words that are mapped to the same semantic domain with the product type and product category. On the other hand, a surprising effect could be achieved by injecting words from different domains. The ads dataset contains 24 categories, such as fashion or food. We obtain the domain information for each category as a noun, from WordNet Domains (Magnini, Strapparava, Pezzulo, & Gliozzo, 2002). Similarly, for each lemma-POS in the slogan and for the product type, we collected the related domains. The categories, product types and lemma-POS pairs are associated with the first sense from WordNet. In addition, we exploit a smaller set of domains that is constructed by normalization with respect to the middle level of WordNet hierarchy. The normalization of the domains provides a higher level of abstraction (Ozbal, Strapparava, Tekiroğlu, & Pighin, 2016) and could allow us to capture whether indirect concepts or ideas are employed for expressing the targeted message.

Semantic Similarity features also allow us to capture the meaningfulness dimension (Ang & Low, 2000) of ad creativity. We exploit ad category and the product type to calculate similarity scores with respect to the lemmas in the sentence. We employ 300 dimensional word representation vectors from GloVe (Pennington, Socher, & Manning, 2014) pretrained embeddings trained on Wikipedia 2014 articles and English GigaWord 5 (LDC). The similarity scores between a category/product type and a lemma are obtained by calculating the cosine similarity of their embedding vectors. The average score of a slogan is encoded as a real valued feature for category and another for product type.

Emotion as a creativity dimension focuses on the feelings awakened in the audience(Ang & Low, 2000). We also generated the emotion features as suggested by Özbal et al. (2013).

Sentiment scores are estimated and used as a part of the emotion(Ang & Low, 2000) dimension features. A word with a highly negative or positive sentiment can induce a positive or negative feeling and might alter the effectiveness and appreciation of the sentence. For instance, an environmental awareness slogan would intend to evoke negative sentiment intensifying the feeling of danger in order to be more striking. Thus, we determined the highest values of positive and negative sentiments in the slogan by checking each lemma-POS and encode them as real valued features. We use the sentiment scores of SentiWordNet (Esuli & Sebastiani, 2007).

Unusual Words contribute to the creativity as the unusual elements dimension suggested by Smith et al. (2007), also as a suprisal factor. We generated unusual words features following the study by Özbal et al. (2013).

Variety can be mapped to the flexibility dimension (Smith et al., 2007). We employ variety scores to detect whether creative and appreciated language displays a particularly different word variety than a less-creative and unappreciated lan-

[3] http://www.adsoftheworld.com/media/print/anagram_sea

[4] http://www.adsoftheworld.com/media/print/gryphon_slippers

guage in a similar way with Özbal et al. (2013).

Phonetic scores can be considered as contributing to the artistic value dimension (Elgammal & Saleh, 2015; Smith et al., 2007; Fichner-Rathus, 2011). The exploitation of phonetic features in creative and persuasive sentence analysis has been deeply explored by Özbal et al. (2013). Following them, we explore the alliteration, rhyme and plosive scores generated by using the HLT Phonetic Scorer[5].

Sensorial features are created regarding the sensory dimension of ad creativity (Elder & Krishna, 2009). A slogan aims to trigger a sensory activation in the mind of the audience. For instance, to evoke the sense of taste for an ad in the food category, certain sensorial information, such as the 'warmness' of a soup or the 'sweet aroma' of a cake, should be transmitted through the language. To identify the sensorial load of the sentences, we obtain the word-sense associations from Sensicon (Tekiroğlu, Özbal, & Strapparava, 2014) and Voted Norms, which we generated as a new set of sensory modality association norms through a voting mechanism and labeling the words with the senses that receive the majority of the votes from 4 different sensorial lexicons (Lievers & Winter, 2017; Tekiroğlu et al., 2014; Lynott & Connell, 2009, 2013; Winter, 2016). Sensicon embodies 22,684 and Voted Norms Lexicon includes 3890 English lemmas together with their part-of-speech (POS) information that have been linked to one or more of the five senses. For each sensory modality, we encode the average sensorial associations of the lemma-pos tuples in the slogan. In addition, we explore how the sensorial trait of a product interacts with the sensorial information in the slogan. Therefore, we create a binary feature indicating whether the sensorial modalities with the highest value of the product type and the slogan are identical. We also add the sensorial association relation of the product type and the slogan as a feature set by taking the mean of the slogan associations and product associations with respect to Sensicon and the Voted Norms. As another hypothesis, we expect that sensory experience ratings (Juhasz, Yap, Dicke, Taylor, & Gullick, 2011) can provide a second channel of sensorial information since *SER* resource estimates the sensory experience triggered in human mind instead of the sensorial information that one word carries. We extracted sensory experience ratings by averaging the SER values of the words in the slogan. Based on the category and sensorial modality correlations provided by Tekiroğlu et al. (2014), we propose a set of sensorial features encoding whether the sensorial information in the slogan conforms to the predetermined sensorial structure of its category.

We generated category conformity scores utilizing Voted Norms. For each sense, we set a binary flag indicating if the average association value of the slogan and the sensorial value of the category are both positive. As an example, the feature set of an ad from the food category contains the binary features *conforms_taste=1* and *conforms_hearing=1* if the average sensorial association value of the slogan for these modalities are over 0.0. In addition, we checked the sensorial association peak of the slogan which shows the modality of the highest sensorial association among the lemma-POS tuples in the slogan. We created a binary feature if the peak modality of the slogan and its category are identical. Contrary to our hypothetical assumption, the peak sensorial conformity is observed to be an indicator of a lower level of appreciation (Mann-Whitney $p < 0.001$) in the training set. A possible explanation for this can be that the unexpected sensorial elevation contributes to the appreciation level of a slogan and a less-appreciated slogan is associated to the senses in a more conventional manner. For instance, a less-appreciated toothpaste slogan "For brighter smiles"[6], which is from the *health category*, has the sensorial peak conformity since both the sensorial peak, i.e. *brighter*, and the ad category are associated with the sense of sight. Using an overly well-known effect of the product to describe a stereotypical metonymic replacement, i.e. *brighter smile* for *whiter teeth*, might be one of the causes of a lower level of appreciation of the ad.

Model	# Feat	Training F1	Testing F1
L	62	0.573	0.577
$L \setminus Sensorial$	34	$*$0.542	$\star$0.496
$L \setminus V \cup U \cup SI$	62	0.585	0.558
$L \setminus Similarity$	61	0.573	0.560
$L \setminus Domain$	55	0.580	$*$0.548
$L \setminus Phonetic$	59	0.573	0.561
$L \setminus Emotion \cup Sentiment$	45	0.568	$^{\#}$0.552

Table 1: The linguistic modality ablation study results. $\star$ denotes $p < 0.001$, $*$ denotes $p < 0.01$, $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between L and ablated models.

Linguistic Experiment Results

We investigated the performances of linguistic features with a classification task employing Random Forest algorithm implemented within the *scikit-learn* package. To fine tune the hyper-parameters of the classifier, we perform a grid search over the number of the generated trees (between 100 and 500, with a step size of 100), the maximum depth of the tree (as [5, 10, 20]) using 10-fold cross validation on the training data. To guarantee the same slogan being only in the training folds or only in the validation fold, we divided the training set into 10 folds by taking into consideration the brand information. Since the training data is unbalanced, we selected the best model by using the weighted average of F1 values.

The results of the full model with all the implemented features and the ablation study are summarized in Table 1. The first row labeled 'L' shows the micro F1 scores for the cross validation and test phases using all the linguistic features. Each row in the rest of the table shows the ablation of the indicated feature. We marked statistical significance in terms of the drop of the performance during ablation in comparison to all features L according to McNemar's test.

In the linguistic experiment, we found out that all the fea-

[5] hlt-nlp.fbk.eu/technologies/hlt-phonetic-scorer

[6] http://www.adsoftheworld.com/media/print/colgate_hide_and_seek

tures contribute to the performance of the final linguistic model even if they cause a slight increase in the F1 scores. By utilizing all the features, we obtain an average training cross-validation F1 score of 0.573 and testing F1 score of 0.577. The linguistic model without *Sensorial Information* yields an F1 test score of 0.496 that is significantly lower ($p < 0.001$) than the model *L*. We obtain a significantly lower ($p < 0.01$) F1 score of 0.548 on the absence of *Domain* features in the linguistic feature set. Removing the *Emotion* and *Sentiment* features from the model decreases the score down to 0.552 on the test set causing a statistically significant loss of performance ($p < 0.05$). The contributions of the strongest features point out that the relevance of a slogan to the product category and type, the positive or the negative feeling that a slogan induces and most importantly the sensorial structure of the slogan and its sensory impact in the audience are indeed essential for a creative and *AP* slogan which is in line with the creativity dimension analysis.

Appreciation Detection on Images

The message of an advertising print is conveyed through both linguistic and visual channels. In this experiment,we utilize the raw sensory input in the form of embedded representations of the image and visual surface features.

Transfer Learning (CNN) Deep learning approaches are proven to be successful in multimodality tasks yielding the state of the art performances on Computer Vision studies such as image classification (Krizhevsky et al., 2012) or object detection (Ren et al., 2015). Considering the promising strength of the convolutional neural networks in image recognition, we hypothesize that certain characteristics of an image, such as objects and patterns, can tamper with its appreciation level as a creative artwork. For instance, marketing images mostly encode cultural and historical stereotypes of masculinity and femininity in order to invoke the feeling of gender identity in the customers (Schroeder & Zwick, 2004). We conjecture that such patterns can be utilized to predict the appreciation level of an ad image if we can capture them automatically.

Since our dataset is not large enough to train a deep network from scratch, we employ transfer learning where we fine-tune Inception V3 image recognition model(Szegedy et al., 2016) as an appreciation predictor. Inception V3 is a deep convolutional neural network that significantly improves the state of the art ILSVRC 2012 1000-class ImageNet classification benchmark. It is trained using stochastic gradient on Tensorflow. Although, the object classification on ImageNet and the appreciation classification task on carefully designed ads are fundamentally dissimilar, Yosinski et al. (2014) state that transferring features from a distant task is still better than randomly initialized variables. Therefore, it would be feasible to boost a network by transferring deeply trained features to overcome the scarcity of the advertising data.

While conducting a transfer learning on Inception V3, first, we only retrained the top 2 layers, labeled as Inception-V3/Logits and Inception-V3/AuxLogits, and kept the earlier layers frozen. In this phase, we obtained a checkpoint after 1000 steps. We exploited Tensorflow Slim[7] implementation to train the new layers and we set the *learning rate* to 0.01. The last layer of the network is a softmax layer that provides posterior probabilities as normalized prediction values for *AP* and *UNAP* classes. In the second phase, we fine-tune all trainable weights in the whole network only for 500 steps and with a learning rate of 0.001. We keep the learning rate small in order to protect the powerful weights of the original Inception V3 model from changing too quickly and loosing their representation ability.

Observable Visual Features (OVF) Together with the implicit properties and patterns belonging to *AP* and *UNAP* classes, we also utilize the explicit elements such as lines and their properties found in the images. In addition, we seek the impact of the color information per se in the creativity detection by encoding the dominant colors as another feature. These features are mostly related to the artistic dimension of the creativity. We extract top 10 dominant colors in the images by k-means clustering on the color values of the pixels and map the center values of the clusters to 16 colors. In addition, we extracted lines in an image through Hough Transform (Duda & Hart, 1972). This feature set contains the normalized length of the longest line and average line length; 3 binary flags indicating whether the longest line is horizontal, vertical or diagonal. We also encoded an interpretation of the "Rule of Thirds", which is a well-known rule of photographic composition[8] and mainly states that the center of interest in images should be on the intersection points or along the lines when an image is divided into 9 equal sections by 2 horizontal and 2 vertical lines. We generated a binary feature indicating if the longest line in the image starts from an outer area and crosses over only 2 sections horizontally and/or 2 sections vertically. Our intuition is that cutting the continuum of the line close to "Rule of Thirds" interest areas can guide the eye of the viewer to the center of focus and contribute to the message and aesthetic value of the image.

Visual Modality Experiment Results

We trained the CNN models on 3265 images from the training set. Using the trained models, we performed the testing on 900 images from the test set. At the end of the test phase, we obtained *AP* and *UNAP* scores for each test image. We employed a straightforward decision process with a 0.5 cut-off where the labeling is conducted by finding the higher value among the *AP* and *UNAP* scores. The performances of the CNN models are summarized in Table 2. The validation accuracy shown in the table is calculated by evaluating the model over the randomly sampled 265 images as the validation set during the training phase. We observe that fine-tuning all the network after the retraining the last 2 layers provide a clear boost to the classifier performance of CNN-1K. The F1 score on the test set significantly ($p < 0.05$) increases from 0.561 to 0.596.

[7] http://github.com/tensorflow/models/tree/master/slim

[8] http://en.wikipedia.org/wiki/Rule_of_thirds

Visual Model	Validation Acc	Testing F1
CNN-1K	0.608	0.561
CNN-1K+500	0.626	$^{\#}$0.596

Table 2: The CNN visual modality experiment results. $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between the models.

Model	#Features	Training F1	Testing F1
OVF	22	0.547	0.560
Colors	17	$\star$0.517	$*$.0.515
Lines	5	0.533	$^{\#}$ 0.508

Table 3: The OVF model visual modality experiment results. $\star$ denotes $p < 0.001$, $*$ denotes $p < 0.01$, $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between *OVF* and Colors or Lines.

For the *OVF* model, we performed the same training strategy that we employed in the linguistic experiment. The model yields relatively poor training (F1:0.547) and testing (F1:0.560) results as they are shown in Table 3. During the ablation study, we observed that a significant performance change occurred when we removed the *Lines* (F1:0.517, $p < 0.001$) for the training cross-validation results. Through the analysis on testing results, we found out that both *Colors* (F1:0.515, $pval < 0.05$) and *Lines* (F1:0.508, $p < 0.01$) contribute to the final model significantly. Although the results of the whole *OVF* model suggest that these aesthetic features are indeed indicating factors of image creativity, we can imply that the overall visual appreciation of an ad is affected by more subtle properties to be discovered than the aesthetic features that we implemented.

Multimodal Fusion

We embraced the late fusion (score level) strategy (Kiela & Clark, 2015) to obtain the multimodal appreciation score. To combine the scores from each model for a class by soft voting approach, we employed Equation 1 where s_n denotes the appreciation score for the ad x by the model m_k and α_n denotes the weight for the model m_k.

$$ms(x) = \sum_{n:m_k} \alpha_n \times s_n \quad (1)$$

We obtained the peak points for α values by running a grid search on the multimodal fusion of model outputs for development set. In this search, all α values are positive and their sum equals to 1.0. After calculating the multimodal appreciation scores, namely *ms*, we labeled the instance *a* by finding the maximum value among the class scores. We evaluate the multimodal experiment results by averaging (Equal α) and soft voting in terms of the F1 scores on the test set and we present the results in Table 4. The α values that we employ during the fusion are shown in the last column. While calculating the multimodal fusion results, we employ the unimodal models; *L*, *OVF* and *CNN-1K+500*. We chose to use the model *CNN-1K+500* since it yields the highest validation accuracy and has a significant improvement for the testing in comparison to its predecessor. Regarding the uni-modal results of linguistic and visual models, the lowest performance is obtained by using OVF while the highest F1 score is yielded by the CNN model. As shown in Table 4, the *ALL* model significantly outperforms the linguistic and observable visual features models. *ALL* model surpasses *CNN*, which is the best unimodal model, by increasing the performance from 0.596 to 0.625 for the soft fusion and to 0.620 for the equal α fusion. This outcome can be considered as conforming with our initial anticipation that the different modalities play complementary roles in expressing the creativity and appeal of an advertising print. The highest contributor of the complete model *ALL* is the *CNN* model and when we remove it from the fusion, the Equal α F1 score drops to 0.587.

Model	Eq. α F1	Soft α F1	α values
ALL	0.620	0.625	L:0.18,C:0.24,O:0.58
$L \cup OVF$	0.587	$*$0.573	L:0.11, O:0.89
$L \cup CNN$	0.605	0.606	L:0.74, C:0.26
$OVF \cup CNN$	0.618	0.612	C:0.25, O:0.75
CNN	0.596	0.596	C:1.0
OVF	$*$ 0.560	$\star$0.560	O:1.0
L	$^{\#}$0.577	$^{\#}$0.577	L:1.0

Table 4: Multimodal fusion results and comparisons to the uni-modal experiments. $\star$ denotes $p < 0.001$, $*$ denotes $p < 0.01$, $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between *L* and ablated models.

Discussion and Conclusion

For the example in `www.adsoftheworld.com/media/print/act_tv_numbers_insects_vs_frog`, which is an *AP* sample resolved by the model *ALL* but not by the visual models, although the visual channel is highly expressive too, the lack of straight lines and dull color palette decreases the prediction performance of *OVF* model. *CNN* model also mislabels the image with a very low confidence since it possibly fails to recognize the peculiar focus elements. Therefore, a wider range of training samples for advertising images would be necessary to identify the style marks of creative and appreciated compositions.

Our quantitative results show that the sensorial structure of the relation between the slogan, and the product category/type is a strong indicator of the creativity appreciation level. When we analyze the feature importance of the final model *L*, we detected that especially the sensorial relation features between the product type and slogan become prominent among the implemented sensorial features. To better illustrate the contribution of the sensorial information to the final model, we fused the visual models with the model $L \setminus Sensorial$ in which we removed all sensorial features from the linguistic model. The Equal-α F1 score of the fusion model decreases to 0.594 without the sensorial features. In the ad `www.adsoftheworld.com/media/print/febreze_french_fries`, we show an *AP* test sample resolved by the contribution of sensorial features. In fact, the example epitomizes the usage of the olfactory disadvantage of the language as a creativity inducing tool. We believe

that smell related words, such as the word "odor" in the example, possibly contributes to the surprisal dimension of the creative and *AP* advertising since olfactory words tend to be less expected by the audience. Indeed, our analysis on the training set reveals that the smell association of the words are inclined to be higher in the *AP* samples in comparison to the *UNAP* samples (Mann-Whitney $p < 0.001$). On the other hand, *taste* association tends to denote the opposite behaviour in our training set (Mann-Whitney $p < 0.001$) while we cannot observe any significant difference for the other senses w.r.t. the Sensicon association values.

Although the automatic assessment of the appreciation level of advertising is a substantially compelling challenge, our findings suggest that sensorial information along with the other linguistic, semantic, cognitive and, finally visual aspects establish a starting point to tackle its complexity.

References

Alnajjar, K., Kundi, H., Toivonen, H., et al. (2018). Talent, skill and support. In *Proceedings of the ninth international conference on computational creativity, salamanca, spain, june 25-29, 2018.*

Ang, S. H., & Low, S. Y. (2000). Exploring the dimensions of ad creativity. *Psychology & Marketing*, *17*(10), 835–854.

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of ACL 2012*.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, *49*.

Duda, R. O., & Hart, P. E. (1972). Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, *15*(1), 11–15.

Elder, R. S., & Krishna, A. (2009). The effects of advertising copy on sensory thoughts and perceived taste. *Journal of consumer research*, *36*(5), 748–756.

Elgammal, A., & Saleh, B. (2015). Quantifying creativity in art networks. In *Proceedings of ICCC 2015.*

Esuli, A., & Sebastiani, F. (2007). Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation*.

Fichner-Rathus, L. (2011). *Foundations of art and design: An enhanced media edition*. Cengage Learning.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *The Quarterly Journal of Experimental Psychology*, *64*(9).

Kiela, D., & Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of EMNLP 2015.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural inf. processing systems.*

Kuznetsova, P., Chen, J., & Choi, Y. (2013). Understanding and quantifying creativity in lexical composition. In *Proceedings of EMNLP 2013.*

Lievers, F. S., & Winter, B. (2017). Sensory language across lexical categories. *Lingua*.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Res. Methods*, *41*(2).

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behav. Res. Methods*, *45*(2).

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, *8*(4).

Özbal, G., Pighin, D., & Strapparava, C. (2013). Brainsup: Brainstorming support for creative sentence generation. In *Proceedings of ACL 2013.*

Ozbal, G., Strapparava, C., Tekiroğlu, S. S., & Pighin, D. (2016). Learning to identify metaphors from a corpus of proverbs. In *Proceedings of EMNLP 2016.*

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014.*

Percy, L. (1982). Psycholinguistic guidelines for advertising copy. *ACR North American Advances*.

Pricken, M. (2008). *Creative advertising ideas and techniques in the world's best campaigns*. Thames&Hudson.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. in neural inf. processing systems.*

Sartori, A., Yan, Y., Ozbal, G., Salah, A., Salah, A., & Sebe, N. (2015). Looking at Mondrian's victory boogie-woogie: What do I feel? In *Proceedings of IJCAI 2015.*

Schroeder, J. E., & Zwick, D. (2004). Mirrors of masculinity: Representation and identity in advertising images. *Consumption Markets & Culture*, *7*(1), 21–52.

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features.

Smith, R. E., MacKenzie, S. B., Yang, X., Buchholz, L. M., & Darley, W. K. (2007). Modeling the determinants and effects of creativity in advertising. *Marketing science*, *26*(6).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the CVPR 2016.*

Tekiroğlu, S. S., Özbal, G., & Strapparava, C. (2014). Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of EMNLP 2014.* Doha, Qatar.

Winter, B. (2016). *The sensory structure of the english lexicon*. Unpublished doctoral dissertation, University of California.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems.*

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250.*