Записки научных
семинаров ПОМИ
Том 529, 2023 г.

A. Alekseev, A. V. Savchenko, E. Tutubalina, E. Myasnikov, S. I. Nikolenko

BLENDING OF PREDICTIONS BOOSTS UNDERSTANDING FOR MULTIMODAL ADVERTISEMENTS

Abstract. The advertising industry employs several content modalities to deliver implied messages: images, videos, text, music, and all of them combined. “Decoding” a message implied by multimodal content often requires both text and visual components. We study the tasks of multimodal symbolism prediction, topic detection, and sentiment type classification. Motivated by the difference in parts of the message conveyed by two modalities in advertisements, we train separate models for images and texts and significantly improve upon current state of the art by blending image- and text-based predictions (with OCR-extracted text), providing a comprehensive experimental validation of our approach.

§1. Introduction

Modern advertising uses visual representations (still images and video) to efficiently convey persuasive messages to potential customers. Much of this persuasion power stems from combining several layers of messaging. Researchers distinguish three layers of messaging in a visual ad: *symbol*, *topic*, and *sentiment* [27,31,37]. The symbol layer shows a symbolic event or an abstract entity that the spectator has to extract a more direct message from; i.e., the WWF campaign ad in Fig. 1 shows trees in the shape of human lungs. This image conveys a symbol of well-being or health, contains the topic of deforestation (more broadly, nature), and has negative sentiment. These three aspects together reinforce the message of preserving the health of the planet and fight against deforestation.

Thus, it is important for both advertisers and consumers to automatically recognize these three components. These aspects are important as features, especially when used to develop better combinations of ads, and for customers it is important to recognize how an ad can affect a user.

Key words and phrases: multimodal and ads understanding and topic detection and sentiment and sentiment classification.

The message is often far from direct, and both modalities are needed: e.g., Fig. 1 has no text explaining that the ad is about deforestation, while sentiment and topic are hard to understand without the slogan “Before it’s too late”. The image on Fig. 2 blends the product (*Audi* symbol) and the “New bad boy on the block” slogan; without the image, there is only a very tenuous connection between the car and the slogan. This intertwining of modalities is very common in advertising, and it presents unique challenges for machine learning.

Hussain et al. [12] present a crowdsourced dataset of such advertisements, including images and videos, and formulate several annotation tasks: topic detection, sentiment type detection, symbolism recognition, strategy analysis, slogan annotation, and Q/A for texts related to the ads’ messages and motivation. We focus on the first three tasks, each of which can be formulated as a classification problem. In *symbolism prediction*, each annotated image has several bounding boxes and textual description of symbols mapped to several categories of symbols. We use multimodal (visual and textual) features to predict specific symbols in a multilabel task setting, extracting text by OCR engines. For symbol prediction, the most important modality is visual: in many cases the symbol is not written. The *topic prediction* task uses the same data, but now both modalities are very important. Finally, *sentiment analysis* is mostly done with extracted texts; for text samples, see Figure 3 and Table 1. In this work, we propose a multimodal blending mechanism that achieves new results in symbolism detection, topic prediction, and sentiment type analysis, significantly exceeding existing state of the art. One key advantage of our blending approach over most multimodal end-to-end models is that it does not need full retraining to change one modality. New data or even a completely new OCR or language model can be plugged in without changing the image-based part and vice versa, which allows for faster experiments, debugging, and deployment. The resulting models are lighter, less complex, and more robust than fully multimodal approaches.

A preliminary version of this work has appeared in [46]. In Section 2 we present related work, Section 3 describes the data and problem setting, Section 4 introduces our approach, Section 5 describes experimental results, Section 6 provides error analysis, and Section 7 concludes the paper.

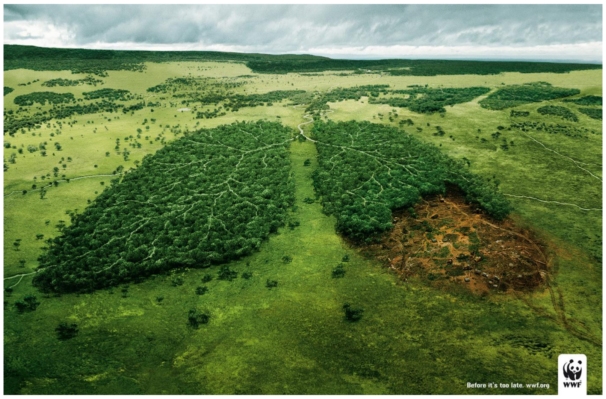


Figure 1. WWF anti-deforestation ad. Topic: environment. Sentiment: alarmed. Symbols: environment.

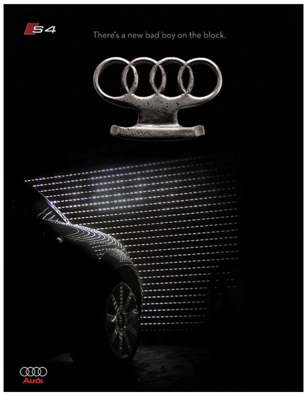


Figure 2. *Audi* ad: a new bad boy on the block. Topic: cars. Sentiment: inspired. Symbols: N/A

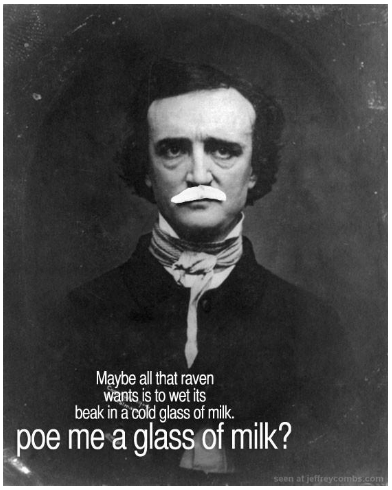


Figure 3. Ad example.

§2. Related work

The multimodal processing of texts and images reached a great progress nowadays with appearance of CLIP [38] and GPT-4 [32]. A core challenge

Table 1. Sample texts obtained via OCR/captioning.

Tesseract [51]	Maybe all that raven wants is to wet its beak in a'cold glass of milk. poe mea lo EISsTo) aU eg
EAST+ *Tesseract* [21]	Maybe are Gein) eu SH Hostel S to wet its ake cold Me TS 10a milk. eee me glass 0) aa Penal see
PSENet [56]	Elle that eV WM eMey iy is 10 Wed Mey in beak | Ey milk. of (eek glass | ranlll@a a me al eee) glass be at jeffreycombs-com
EasyOCR [15]	Maybe all that raven poe me a glass ofmik? beak in a cold glass ofmik. wants is to wet its seen 3t jefieycomlzesolm
Charnet [60]	MAYBE ALL THAT RAVEN WANTS WET ITS BEAKIN COLD GLASS MILK POE GLASS MILK? SEEN ATJ COM
CloudVision [33]	Maybe all that raven wants is to wet its beak in a cold glass of milk. poe me a glass of milk? seen at jeffreycombs.com

in ads understanding is to join information from two or more modalities since different modalities may have varying predictive power with possibly missing or noisy data [3, 14]. The works [12, 62] provided an annotated dataset of ads, implemented and compared several simple baselines. Several mappings of multimodal data can be mapped into a joint embedding space with triplet losses that bring together features of image segments, textual ad descriptions, and symbol labels [64]; another way to fuse visual and symbolic information used an iterative co-attention mechanism [1]. Extracting opinions from texts or discerning sentiment attitudes between named entities mentioned in texts is a well-established topic, typically explored through machine learning and language models [13,41–43,55]. More information can be extracted from images by fusing image features with bag-of-words features for extracted text for semantic classification and visual question answering [8]. Image captioning can be used in addition to OCR [16], with captions processed with BERT [7]. A visual question answering system has been augmented with extracted texts and a matched *Wikipedia* article [69], while the work [67] notes that image-text alignment is especially difficult for ads and proposes hand-crafted features to resolve this issue. All these works use only still images; video understanding has been considered in [18, 63, 65]. In [30], new ad text is generated from an image; this work also proposes novel approaches for ranking text keyphrases and predicting the tags of appropriate ad images. Multimodal Bitransformers (MMBT) fuse text with images or other information, representing non-textual data as additional tokens in the word sequence [19]. VisualBERT aligns input text and image regions by self-attention [22]. A multimodal factorization model is able to factorize representations into discriminative factors and modality-specific generative factors [54]. In [23], tensor rank regularization is used to learn language, acoustic, and visual representations for multimodal video data. The work [24] uses a recurrent multistage fusion network with cross-modal interactions based on intermediate representations of monologue videos.

Our blending approach differs from most multimodal end-to-end models in the aspect that it does not require retraining "from top to bottom" to update one modality. "Hot swapping" new data/LM/OCR/etc. provides means for faster experiments, debugging, and deployment. The resulting models are less complex, more robust, and often faster than fully multimodal approaches.

§3. Data

The term *symbolism* comes from the classic work by [57], who applied the ideas of de Saussure's semiotics [5] to advertising: the object or content that stands for a symbol is called the "signified" or "concept", and the symbol is the "signifier". Relations between them come from human associations: e.g., in Fig. 1 vegetation symbolizes health. The work [12] presented a list of symbols (concepts, signifiers) and labeled ads with these symbols via crowdsourcing. Annotators first decided if an ad is literal or requires a non-literal interpretation, the latter interpreted as symbolism for simplicity. If most annotators say that an ad is non-literal, it enters the second stage where they are asked to label the signifier and signified, draw a bounding box (marking the signifier), and label it with the symbol it refers to (the signified). There were 221 symbols in total, the most common being "danger", "fun", "nature", "beauty", "death", "sex", "health", and "adventure". Images were also annotated for topics and sentiment, with topics that the images advertise or campaign for coming from a specially developed taxonomy [12]. They also built a taxonomy of sentiments, asking annotators to write free-form topics and sentiments for a small batch of images and videos, similar to "self-reporting" used to measure emotional reactions [45, 47, 61] to ads [37]. Topics and sentiments were

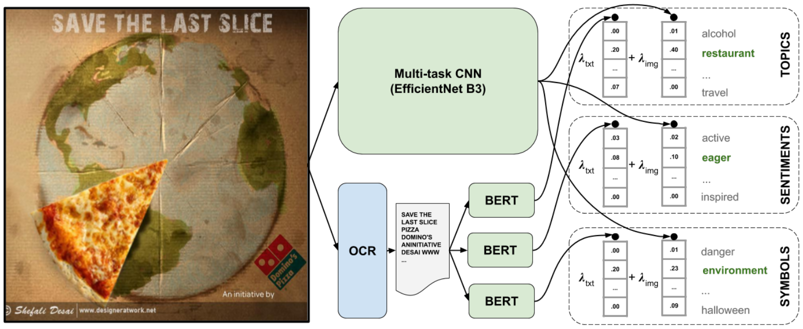


Figure 4. The proposed blending scheme.

clustered into 38 topics and 30 sentiment types, with a representative set of words describing each. In further annotations, an image was labeled with a single topic and one or more sentiments. Note that compared to popular multimodal datasets on visual question answering and commonsense reasoning [2,36,52,66], texts on images in ads are much more varied in font/shape/placement/color etc., which makes it a much harder task for OCR and hence NLP.

We use a train/test split similar to [12], with 9856 training and 2464 test images with annotated symbols. We also used 53 clusters of similar symbols for which the train and test sets contain 8394 and 2099 images respectively. Due to heavy class imbalance in all problems, for topic and sentiment recognition we split the dataset at the 80/20 ratio in each class to avoid missing labels or randomly missing training data. Since we need to train multi-task models, training and validation sets for different tasks (topic/sentiment/symbols classification) cannot intersect. As a result, the training set for 39 topic labels (38 topics + "unclear") contains 51460 images, while the test set has 12865 images. We used 24272 training and 6068 test images for 30 categories of sentiments. Images can have multiple labels, so we used the most frequent topic/sentiment as the ground truth, as recommended by [12].

§4. Methods

Image-based methods. We have explored two image-based approaches: training models separately for classifying symbols, topics, and sentiments, and a multitask approach that solves all three tasks simultaneously. We compared several classical feature extraction backbones: ResNet-50 and ResNet-152 [10], MobileNet v1/v2 [6, 11, 44], and EfficientNet B0/B3 [47, 53]. We added a new head with C outputs for each task, where C is the number of categories, learned over 5 epochs with the Adam optimizer [20], freezing backbone weights. Next, the entire model with unfrozen weights is trained over 5 epochs with Adam. Third, the entire model is trained over 3 epochs with stochastic gradient descent, learning rate 0.001. Symbolism prediction is a multi-label task, so binary cross-entropy loss and sigmoid activations were used on the last fully connected layer, while for topics and sentiments we used categorical cross-entropy and softmax activations. For the *multitask approach*, training was implemented by using a multi-head CNN with common feature extraction layers. Every head is trained separately with frozen backbone over 3 epochs, and then the entire network is trained over 10 epochs.

OCR approaches. Text recognition accuracy has proven to be crucial in our approach. We have compared several recent optical character recognition (OCR) methods, including: (1) Tesseract 4 OCR library (Tes 4.00) with an LSTM-based neural network; (2) EAST [68] and Advanced EAST-based two-phase approaches (EastTes): first detect text with EAST and then recognise it with Tesseract; (3) a two-phase approach based on Progressive Scale Expansion Network (PSENet) [56]: first generate a different scale of kernels for each text instance, then gradually expand the minimal scale kernel to the text instance with the complete shape; on the second phase, Tesseract is used to recognize text fragments detected with PSENet; (4) convolutional character networks (*Charnet*) [60], a one-stage CNN that uses characters as basic elements and directly outputs labeled bounding boxes of words and characters; (5) EasyOCR engine [15]; (6) text recognition functionality provided by Google's Cloud Vision API, which was used by the winner of the *Automatic Understanding of Visual Advertisements* challenge for a different task on the same dataset [33]. Note that even if commercial engines perform best, choosing the best openly available engine is still important for practical projects.

Text-based methods. We consider a simple BERT-based [7] model with a classification layer on top of encoded representations. Specifically, we first filter out items without recognized text and lowercase the texts. For each item in the dataset, we combine all recognized text fragments in a consistent order separated by a [SEP] symbol and input the result to the encoder, for which we have compared (i) BERT-base (12 hidden layers, 768 hidden dimensions, 12 attention heads, 110M parameters) [7] and (ii) RoBERTa-base (125M parameters) [26]. We used the *Simple Transformers* library [40] based on [58], fine-tuning the underlying BERT parameters and training all models for 15 epochs with the Adam optimizer, learning rate 4e-5, and other parameters at default values.

Text-based models usually yield results inferior to image-based ones; this is natural because text is not always present, often short, and even the best OCR methods make quite a lot of mistakes. However, combinations of image- and text-based models can yield significant improvements. For comparison, we used text representations from a *Bag-of-Ngrams* baseline, tokenizing extracted texts and preserving only 10,000 most frequent unigrams and bigrams. Importantly for imperfect OCR-extracted text, BERT-based models along with NGram-based baselines help with out-of-distribution text. We have also experimented with multi-output logistic regression trained on SGNS [28,29] and fastText [4] representations, but prediction quality was significantly lower. We have also applied multi-task BERT- and RoBERTa-based approaches, training as above, but they yielded no improvement in prediction quality.

Blending predictions. To avoid overfitting, we use a straightforward ensembling strategy, aggregating per-class probability scores from image-based and OCR+text classifiers. For every ad a, each model in the ensemble yields a vector $f(a) \in [0,1]^{|L|}$, where L is the set of labels (classes). The resulting ensemble outputs 1 if $\lambda f_{\text{img}}(a) + (1-\lambda) f_{\text{txt}}(a) > \theta_0$ for symbolism multi-label prediction, and for topic and sentiment classification outputs argmax $(\lambda f_{\text{img}}(a) + (1-\lambda) f_{\text{txt}}(a))$ (argmax taken over vector components), where $f_*(e) \in [0,1]^{|L|}$ are model predictions, and coefficient λ and threshold θ_0 are tuned parameters.

The final model is illustrated in Fig. 4. For tuning, we have used the predictions of image- and text-based "elementary models", sampling 5 times a fraction of the training set (0.1 and 0.05 for 221 and 53 labels in symbolism prediction, 0.05 for both sentiment and topic prediction). We have tuned the weights on a subset of the training set and not on a special hold-out set since the dataset is relatively small. Then we sample λ from the Dirichlet distribution and evaluate the $F1_{\text{macro}}$ (not $F1_{\text{micro}}$ as an extra measure against overfitting) as implemented in *scikit-learn* [35] on the chosen subsample for every $\theta_0 \in \{0.0, 0.05, ...1.0\}$ and average the 5 (resp. 10, 3) sets of parameters for symbolism prediction (resp. topic and sentiment classification). We have also tested: (i) Multimodal Bitransformers (MMBT) [19], using the original MMBT with ResNet-152 and BERT-base and "updated" MMBT with EfficientNet-B3 backbone instead of ResNet-152; (ii) ConcatBERT with ResNet-152, (iii) VisualBERT [22] with ResNet-152. All models are implemented in the MMF framework [50]. MMBT (orig.) had slightly greater accuracy and F1-score over ConcatBERT and VisualBERT. Specifically, VisualBERT's accuracy is 2.2% (topics) and 3.1% (sentiments) lower than MMBT; for ConcatBERT, F1-score on topic classification is much worse (0.45 vs 0.53 for Charnet OCR).

§5. Results and discussion

Symbolism prediction. The best purely image-based model, which has been using EfficientNet-B3 for feature extraction, is able to obtain an F1-score of 0.1912, while the previous state of the art for this task was 0.1579 [12]. Thus, even at this level we have already significantly exceeded state of the art by using better convolutional backbones and tuned training schedules. Fig. 5 shows how the average F1-score for all classes depends on the threshold t_0, which is used to select classes for which the CNN prediction exceeds this threshold. The best result is obtained for $t_0 = 0.1$. Next, we repeated the training procedure of EfficientNet-B3 for 53 clusters of symbols and obtained F1-score 0.2774, which is also slightly better than the previously best F1-score of 0.2684 for this problem [12].

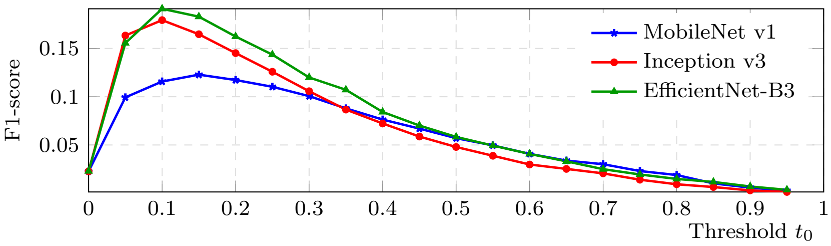


Figure 5. F1-score for image-based symbol recognition (221 categories).

Topic/sentiment recognition. Table 2 shows test set accuracies for topic and sentiment classification. Here the baseline results of [12] were obtained with ResNet-152. Our best models are 2% and 6% more accurate than the baseline in topic and sentiment recognition respectively. Note

Table 2. Image-based topic/sentiment classification.

CNN	Topics	Sentiments
Baseline [12]	60.34	27.92
Curriculum learning [34]	—	27.96
ResNet-50	53.90	34.34
Resnet-152	52.67	27.58
Resnet-152 V2	52.12	27.64
MobileNet v1	50.56	33.50
MobileNet v2	54.76	**34.58**
EfficientNet-B0	60.06	34.03
EfficientNet-B3	**62.62**	34.12
Our multitask model	**62.99**	**36.27**

that the absolute numbers are low in sentiment recognition in part due to the fact that most classes are underrepresented, and their recognition accuracy is virtually zero.

Multi-task model. We used the multi-task learning procedure shown above with the best architecture from our previous experiments, namely EfficientNet-B3. The multitask model showed the following results on test sets for the three tasks: accuracy 0.6299 for topics, accuracy 0.3627 for sentiment types, and precision 0.5508 and recall 0.0826 for symbols. It is clear that multitask learning has led to an improvement for all tasks; we expect that further improvements may be possible with more research investigating other backbone networks and developing a better joint loss function for the three tasks. Individual networks have also led to quality improvement compared to [12], with the most significant improvement achieved for sentiment classification.

Blending. Tables 3 and 4 present our results on texts extracted by Charnet and Cloud Vision for topic, sentiment, and multilabel symbol classification in both 221 and 53 label settings. They show $\mathrm{F1}_{\mathrm{micro}}$ and $\mathrm{F1}_{\mathrm{macro}}$ scores; “text-based” rows show results for OCR-extracted text from ads with nonempty extracted text; “Blend (backoff)”, results for the entire test set with image-based prediction when text is empty; “MMBT (ResNet)”, multimodal bitransformers with ResNet-152 for images; “MMBT (EffNet)”, same with EfficientNet-B3.

Unsurprisingly, the best performance is shown by Transformer-based models BERT and RoBERTa, which achieve accuracy higher than 0.75 and 0.3675 for topic and sentiment classification respectively with both

Table 3. Symbol classification results.

OCR	Model	Text-based (w/texts)		Blend (backoff)	
		F1_{micro}	F1_{macro}	F1_{micro}	F1_{macro}
Multilabel symbol classification, 221 labels.					
Image-based results: F1_{micro} 0.1928, F1_{macro} 0.1025.					
Charnet	Bag-of-NGrams	0.1684	0.0967	0.2156	0.1146
	BERT	0.1354	0.0203	0.1964	0.0998
	RoBERTa	0.1441	0.0253	0.1974	0.0995
	MMBT (orig.)	–	–	0.0962	0.0671
	MMBT (upd.)	–	–	0.1078	0.0757
Google Cloud Vision	Bag-of-NGrams	**0.1830**	**0.1060**	**0.2249**	**0.1175**
	BERT	0.1520	0.0252	0.1968	0.1014
	RoBERTa	0.1580	0.0263	0.2017	0.1004
	MMBT (orig.)	–	–	0.1202	0.0825
	MMBT (upd.)	–	–	0.1099	0.0812
Multilabel symbol classification, 53 labels.					
Image-based results: F1_{micro} 0.2796, F1_{macro} 0.2182.					
Charnet	Bag-of-NGrams	0.2434	0.1914	0.3025	0.2345
	BERT	0.2618	0.2080	0.3264	0.2528
	RoBERTa	0.2769	**0.2254**	0.3315	0.2606
	MMBT (orig.)	–	–	0.2424	0.2069
	MMBT (upd.)	–	–	0.2781	0.2358
Google Cloud Vision	Bag-of-NGrams	0.2446	0.2015	0.3041	0.2344
	BERT	0.2624	0.2033	0.2966	0.2213
	RoBERTa	**0.2896**	0.2150	**0.3137**	0.2324
	MMBT (orig.)	–	–	0.2611	0.2291
	MMBT (upd.)	–	–	0.2951	**0.2619**

Cloud Vision and Charnet, while end-to-end neural multimodal approaches MMBT and VisualBERT shows performance inferior to Bag-of-Ngrams. However, replacing the original ResNet-152 backbone with EfficientNet-B3 we have improved over state of the art for topics but not for sentiment types. The results should be compared to the baseline values of 0.6034 for topic classification in [12] and 0.6923 in [9] (best known result) and 0.2792 for sentiment classification. Symbolism prediction models in prior works scored no higher than 0.1579 for 221 labels and 0.2684 for 53 labels in terms of the F1_{micro}-score [12]. Interestingly, Bag-of-Ngrams-based logistic regression is almost on par with BERT/RoBERTa for sentiment classification, and results of open source *Charnet* are only a little worse than Google Cloud Vision OCR. This correlates with one of our basic assumptions: we do not see a reason here to unite latent spaces for textual

Table 4. Classification results.

OCR	Model	Text-based (w/texts)		Blend (backoff)	
		Acc.	$\mathrm{F1}_{\mathrm{macro}}$	Acc.	$\mathrm{F1}_{\mathrm{macro}}$
Topic classification. Image-based: accuracy 0.6299, $\mathrm{F1}_{\mathrm{macro}}$ 0.3800.					
Charnet	Bag-of-Ngrams	0.6340	0.4502	0.7213	0.4816
	BERT	0.6985	0.5515	0.7536	**0.5793**
	RoBERTa	0.6933	0.5473	0.7545	0.5722
	VisualBERT	–	–	0.676	0.514
	VisualBERT (COCO)	–	–	0.677	0.518
	ConcatBERT	–	–	0.736	0.454
	MMBT (ResNet)	–	–	0.6821	0.5357
	MMBT (EffNet)	–	–	0.7534	0.5441
Google Cloud Vision	Bag-of-Ngrams	0.6391	0.4531	0.7227	0.4840
	BERT	**0.7149**	**0.5573**	0.7599	0.5736
	RoBERTa	0.7109	0.5492	0.7557	0.5623
	VisualBERT	–	–	0.667	0.497
	VisualBERT (COCO)	–	–	0.655	0.524
	ConcatBERT	–	–	0.728	0.450
	MMBT (ResNet)	–	–	0.7031	0.5396
	MMBT (EffNet)	–	–	**0.7686**	0.5700
Sentiment classification. Image-based: accuracy 0.3627, $\mathrm{F1}_{\mathrm{macro}}$ 0.1041.					
Charnet	Bag-of-Ngrams	0.2705	0.0905	0.3675	0.1062
	BERT	0.2497	0.1000	0.3675	0.1093
	RoBERTa	**0.2774**	**0.1211**	0.3717	0.1093
	VisualBERT	–	–	0.340	0.108
	VisualBERT (COCO)	–	–	0.344	0.103
	ConcatBERT	–	–	0.371	0.096
	MMBT (ResNet)	–	–	0.2836	0.1049
	MMBT (EffNet)	–	–	0.3053	0.1141
Google Cloud Vision	Bag-of-Ngrams	0.2641	0.0859	0.3676	0.1061
	BERT	0.2595	0.1023	**0.3731**	0.1117
	RoBERTa	0.2750	0.1165	0.3697	0.1072
	VisualBERT	–	–	0.336	0.109
	VisualBERT (COCO)	–	–	0.338	0.112
	ConcatBERT	–	–	0.371	0.088
	MMBT (ResNet)	–	–	0.3152	0.0925
	MMBT (EffNet)	–	–	0.3224	**0.1219**

and visual embeddings, as many multimodal methods do when text and image are assumed to represent the same concepts. We train models for image and text classification independently and create diverse ensembles.

Our blending approach uses a very small number of parameters to join the two modalities, which makes it important to analyze the per-class performance on each of them and on the final resulting ensemble. For this purpose, we have used Google Cloud Vision OCR texts and BERT-based and EfficientNetB3 predictions for text- and image-based predictions respectively.

Results for topic and sentiment type classification are shown in Tables 5 and 6, respectively. We show per-class recall and support (total number of data points with the corresponding label in the test set) for EfficientNet-B3 multi-task model image-based predictions and BERT model (fine-tuned on Google Cloud Vision OCR texts) for text-based predictions. When no text was extracted, we assume that the text-based model made a mistake. For many labels, text-based prediction is superior in both tasks, but for several large classes it is not (see the support column "#"), and hence the overall accuracy is higher. The tables show that the $\mathrm{F1}_{\mathrm{macro}}$ score often increases for blended models, probably due to the complementary nature of the multimodal ensemble: when one classifier fails, another modality may enable a correct prediction.

§6. Discussion and Error Analysis

We have analyzed the outputs of three topic classification models: text-based, image-based, and their blend (Table 7). We identify four major error types. First, ads contain text fragments in non-standard font and small size, so input to the text-based model is noisy and limited (Ex. 1). Second, ads may contain the company name instead of product details (Ex. 2 and 4); this confirms that additional information about the company can help, which we leave for future work. Third, the difference between some topics is vague; e.g., Ex. 5 shows that topics *soda* and *alcohol* are both related to *drinks*, which may indicate the need for a better data annotation scheme. Finally, advertisers use highly abstract visual metaphors and symbols that require additional knowledge (recall also Figs. 1 and 2). In Ex. 3, the deer head is shown as a reward for hunting sports, so text-based models predict the topic correctly due to words such as *hunter* and *archery*, while the image-based model sees the deer head as a sign for animal rights advertising. Ex. 5 shows that pretrained OCR models that support only Latin characters and English words may work incorrectly when other languages are present: a Russian disclaimer was partially recognized by Charnet with English chars. We believe that such scenarios should be handled by multilingual OCR models and adding more variability to the dataset, which may allow for better fine-tuning of language models in future work. Visual

Table 5. Per-class topic prediction performance; darker color corresponds to higher scores.

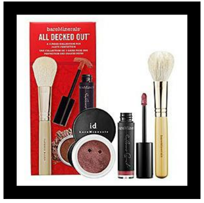
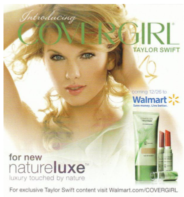
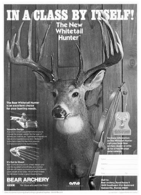
Class	#	Recall (%)		
		Image	Text	Blend
alcohol	485	66.80	73.61	71.34
animal right	107	51.40	67.29	54.21
baby	16	37.50	43.75	50.00
beauty	1124	80.96	77.49	83.36
cars	1267	85.40	83.82	87.21
charities	15	0.00	33.33	6.67
chips	316	53.16	66.46	58.86
chocolate	681	72.83	73.72	76.65
cleaning	36	25.00	38.89	30.56
clothing	1585	79.43	72.05	82.33
coffee	130	59.23	73.08	61.54
dom. violence	37	37.84	54.05	40.54
education	44	4.55	43.18	6.82
electronics	806	71.46	72.46	74.69
environment	73	34.25	42.47	42.47
financial	323	44.27	71.83	50.77
gambling	7	0.00	28.57	0.00
game	115	28.70	51.30	30.43
healthcare	204	23.53	60.29	30.88
home appliance	117	43.59	56.41	43.59
home improv.	53	9.43	33.96	9.43
human right	45	11.11	53.33	22.22
media	256	19.92	42.97	26.95
other service	240	9.58	30.83	15.42
petfood	6	0.00	16.67	0.00
phone tv internet	167	23.35	71.86	31.74
political	15	13.33	20.00	13.33
restaurant	772	74.74	75.00	77.07
safety	86	37.21	66.28	43.02
seasoning	135	62.22	68.15	64.44
security	16	0.00	25.00	0.00
self esteem	39	15.38	51.28	25.64
shopping	330	52.12	59.70	53.64
smoking alc. abuse	98	61.22	62.24	67.35
soda	717	69.04	76.29	73.36
software	81	4.94	28.40	3.70
sports	464	44.61	51.51	49.57
travel	403	49.88	67.00	57.82
unclear	91	4.40	7.69	3.30

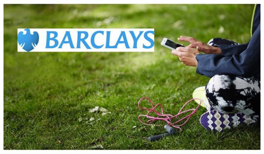
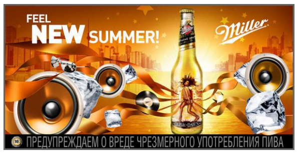
Table 6. Per-class sentiment prediction performance.

Class	#	Recall (%)		
		Image	Text	Blend
active	605	33.06	26.94	35.54
afraid	31	0.00	6.45	0.00
alarmed	74	1.35	4.05	0.00
alert	529	18.71	16.64	19.85
amazed	103	0.00	1.94	0.97
amused	205	0.49	9.76	2.93
angry	23	0.00	0.00	0.00
calm	135	0.00	4.44	0.00
cheerful	233	1.29	10.30	2.15
confident	148	0.00	4.73	0.00
conscious	240	2.50	7.50	2.92
creative	984	49.80	30.28	50.41
disturbed	27	0.00	0.00	0.00
eager	856	76.52	43.93	78.62
educated	141	0.71	4.96	1.42
emotional	33	0.00	0.00	0.00
empathetic	15	0.00	0.00	0.00
fashionable	634	76.50	47.00	76.97
feminine	176	12.50	18.75	14.77
grateful	3	0.00	0.00	0.00
inspired	117	0.85	8.55	1.71
jealous	0	–	–	–
loving	6	0.00	0.00	0.00
manly	46	0.00	4.35	0.00
persuaded	35	0.00	2.86	0.00
pessimistic	0	–	–	–
proud	1	0.00	0.00	0.00
sad	1	0.00	0.00	0.00
thrifty	32	40.62	12.50	40.62
youthful	26	11.54	0.00	11.54

and textual components are both necessary for successful predictions: text and image in advertisements are complementary in that sense.

Table 7. Five sample results of three models; the text-based model is based on RoBERTa and Charnet for OCR.

	(1)	(2)	(3)
			
OCR (Charnet)	ALL DECKED OUT	IRL TAYLOR SWIFT COMING 12/26 WAIMART SAVE MONEY LIVE BETTER. FOR NEW NATURE ELUXE LUXURY TOUCHED NATURE FOR EXCLUSIVE TAYLOR SWIFT CONTENT VISIT WAIMART.COM /COVERGIRL	CLASS WHITETAIL THE NEW HUNTER THE CHOKE HUNTER BEAR ARCHERY
Text	restaurant	clothing	sports
Image	beauty	beauty	animal_rights
Blend	beauty	beauty	animal_rights
Gr. truth	beauty	beauty	sports

	(4)	(5)
		
OCR (Charnet)	ARC LAYS	FEEL NEW MILER SUMMER! ANN YPE3MEPHORO
Text	cleaning	soda
Image	electronics	alcohol
Blend	electronics	soda
Gr. truth	financial	alcohol

§7. Conclusion

In this work, we have presented a unified blended system that combines several state of the art models, e.g., BERT for text embedding and MobileNet and EfficientNet for image embedding. This led to significant improvements in symbolism detection (by 3% over baseline quality of 16% F_1) and sentiment detection (by 7% over 28% F_1), all of this using only well-known models for individual components. Interestingly, topic detection/classification was improved by a mere 2% from the baseline quality of 60% F_1; this could be due to an already high quality achieved by the baselines, but future research may reveal other interesting reasons for this effect. Another important point is that we train our system in the multi-task fashion, so the image embedding is shared among all three tasks, while the text embeddings are specific for each of the three tasks. This discrepancy also leaves an open question: why has this led to better results than other approaches? One possible answer is that the language used to detect sentiments, topics, and symbols is different, but further research is needed to test this answer. Another interesting observation is that Charnet, a non-commercial OCR system, shows performance comparable to Google Cloud Vision on the tasks in question. We also note that our models have been able to significantly outperform an end-to-end neural pipeline from [19], which also suggests that even better results may appear in this direction. In closing, we hope that our work will foster more research on multimodal problems and draw attention to advertisement understanding in general.

One of important future research directions is application of our approach to understanding of video advertisements [12]. In such case, it will be possible to obtain text not only by using OCR, but with the automatic speech recognition techniques [18, 39, 49]. Another important future study is related to usage of modern visual transformers instead of convolutional networks [17, 25]. Finally, it is important to implement the proposed approach for mobile applications [48] to measure emotional reactions and analyze understanding of symbolism by a concrete user [34, 37, 59].

Acknowledgments

The work has been supported by the Russian Science Foundation grant # 23-11-00358.

References

1. K. Ahuja, K. Sikka, A. Roy, A. Divakaran, *Understanding visual ads by aligning symbols and objects using co-attention*, arXiv preprint arXiv:1807.01448 (2018).
2. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C Lawrence Zitnick, and Devi Parikh, *VQA: Visual question answering*, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2425–2433.
3. T. Baltrušaitis, C. Ahuja, L.-P. Morency, *Multimodal machine learning: A survey and taxonomy.* — IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, No. 2 (2018), 423–443.

4. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, *Enriching word vectors with subword information.* — Transactions of the Association for Computational Linguistics **5** (2017), 135–146.
5. F. de Saussure, *Course in general linguistics*, Duckworth, London, 1983, (trans. Roy Harris).
6. P. Demochkina, A. V. Savchenko, *MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices*, Proceedings of Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V, Springer, 2021, pp. 266–274.
7. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018.
8. A. U. Dey, Su. K. Ghosh, E. Valveny, *Don't only feel read: Using scene text to understand advertisements*, arXiv preprint arXiv:1806.08279 (2018).
9. A. U. Dey, S. K. Ghosh, E. Valveny, G. Harit, *Beyond visual semantics: Exploring the role of scene text in image understanding*, 2019.
10. K. He, X. Zhang, S. Ren, J. Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
11. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, arXiv preprint arXiv:1704.04861 (2017).
12. Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, Ad. Kovashka, *Automatic understanding of image and video advertisements.* — Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1705–1715.
13. V. V. Ivanov, E. V. Tutubalina, N. R. Mingazov, I. S. Alimova, *Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars*, Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2015, pp. 22–33.
14. S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, A. Jabbar, *A review on methods and applications in multimodal deep learning.* — ACM Transactions on Multimedia Computing, Communications and Applications **19**, No. 2s (2023), 1–41.
15. JaidedAI, *EasyOCR: Ready-to-use ocr with 70+ languages supported including chinese, japanese, korean and thai.*, `https://github.com/JaidedAI/EasyOCR`, 2020.
16. K. Kalra, B. Kurma, S. V. Sreelatha, M. Patwardhan, S. Karande, *Understanding advertisements with BERT*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7542–7547.
17. A. Karpov, I. Makarov, *Exploring efficiency of vision transformers for self-supervised monocular depth estimation*, Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2022, pp. 711–719.
18. Ya. I. Khokhlova, A. V. Savchenko, *About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems*, Optical Memory and Neural Networks **23** (2014), 34–42.
19. D. Kiela, S. Bhooshan, H. Firooz, D. Testuggine, *Supervised multimodal bitransformers for classifying images and text*, arXiv preprint arXiv:1909.02950 (2019).

20. D. P. Kingma, J. Ba, *Adam: A method for stochastic optimization*, International Conference on Learning Representations (ICLR), 2015.
21. L. Kopeykina, A. V. Savchenko, *Automatic privacy detection in scanned document images based on deep neural networks*, Proceedings of the International Russian Automation Conference (RusAutoCon), IEEE, 2019, pp. 1–6.
22. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, *VisualBERT: A simple and performant baseline for vision and language*, arXiv preprint arXiv:1908.03557 (2019).
23. P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, L.-P. Morency, *Learning representations from imperfect time series data via tensor rank regularization*, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1569–1576.
24. P. P. Liang, Z. Liu, AmirAli Bagher Zadeh, L.-P. Morency, *Multimodal language analysis with recurrent multistage fusion*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 150–161.
25. Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y, Zhang, Z. Shi, J. Fan, Z, He, *A survey of visual transformers*, IEEE Transactions on Neural Networks and Learning Systems (2023).
26. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, *RoBERTa: A robustly optimized BERT pretraining approach*, 2019.
27. D. McDuff, R. El Kaliouby, J. F. Cohn, R. W. Picard, *Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads*, IEEE Transactions on Affective Computing **6** (2014), no. 3, 223–235.
28. T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781 (2013).
29. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, *Distributed representations of words and phrases and their compositionality*, Advances in neural information processing systems, 2013, pp. 3111–3119.
30. S. Mishra, M. Verma, Y. Zhou, K. Thadani, W. Wang, *Learning to create better ads: Generation and ranking approaches for ad creative refinement*, (2020), 2653–2660.
31. L. C. Olson, C. A. Finnegan, D. S. Hope, *Visual rhetoric: A reader in communication and american culture*, Sage, 2008.
32. OpenAI, *GPT-4 technical report*, arXiv preprint arXiv:2303.08774 (2023).
33. M. Otani, Y. Iwazaki, K. Yamaguchi, *Unreasonable effectiveness of OCR in visual advertisement understanding*, 2018.
34. R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, A. K. Roy-Chowdhury, *Contemplating visual emotions: Understanding and overcoming dataset bias*, Proceedings of European Conference on Computer Vision (ECCV) (Cham) (Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, eds.), Springer International Publishing, 2018, pp. 594–612.
35. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python.* — J Machine Learning Research **12** (2011), 2825–2830.

36. B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, S. Lazebnik, *Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models*, Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.
37. K. Poels, S. Dewitte, *How to capture the heart? reviewing 20 years of emotion measurement in advertising.* — J. Advertising Research **46**, No. 1 (2006), 18–37.
38. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S.Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., *Learning transferable visual models from natural language supervision*, Proceedings of International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763.
39. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, *Robust speech recognition via large-scale weak supervision*, Proceedings of International Conference on Machine Learning (ICML), PMLR, 2023, pp. 28492–28518.
40. T. Rajapakse, *Simple transformers*, 2020.
41. N. Rusnachenko, N. Loukachevitch, E. Tutubalina, *Distant supervision for sentiment attitude extraction*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (Varna, Bulgaria) (Ruslan Mitkov and Galia Angelova, eds.), INCOMA Ltd., September 2019, pp. 1022–1030.
42. A. Sakhovskiy, Z. Miftahutdinov, E. Tutubalina, *Kfu nlp team at smm4h 2021 tasks: Cross-lingual and cross-modal bert-based models for adverse drug effects*, Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021, pp. 39–43.
43. A. Sakhovskiy, E. Tutubalina, *Multimodal model with text and drug embeddings for adverse drug reaction classification.* — J. Biomedical Informatics **135** (2022), 104182.
44. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.
45. A. Savchenko, *Facial expression recognition with adaptive frame rate based on multiple testing correction*, International Conference on Machine Learning, PMLR, 2023, pp. 30119–30129.
46. A. Savchenko, A. Alekseev, S. Kwon, E. Tutubalina, E. Myasnikov, S. Nikolenko, *Ad lingua: Text classification improves symbolism prediction in image advertisements*, Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1886–1892.
47. A. V. Savchenko, *MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data*, Proceedings of European Conference on Computer Vision Workshops (ECCVW), Springer, 2022, pp. 45–59.
48. A. V. Savchenko, K. V. Demochkin, I. S. Grechikhin, *Preference prediction based on a photo gallery analysis with scene recognition and object detection.* — Pattern Recognition **121** (2022), 108248.
49. V. V. Savchenko, A. V. Savchenko, *Criterion of significance level for selection of order of spectral estimation of entropy maximum.* — Radioelectronics and Communications Systems **62**, No. 5 (2019), 223–231.

50. A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, D. Parikh, *MMF: A multimodal framework for vision and language research*, `https://github.com/facebookresearch/mmf`, 2020.
51. R. Smith, *An overview of the Tesseract OCR engine*, Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), vol. 2, IEEE, 2007, pp. 629–633.
52. A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, Y. Artzi, *A corpus for reasoning about natural language grounded in photographs*, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6418–6428.
53. M. Tan, Q. V. Le, *EfficientNet: Rethinking model scaling for convolutional neural networks*, (2019), 6105–6114.
54. Y.-H. H.Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, *Learning factorized multimodal representations*, International Conference on Learning Representations (ICLR), 2018.
55. E. Tutubalina, S. Nikolenko, *Inferring sentiment-based priors in topic models*, Mexican International Conference on Artificial Intelligence, Springer, 2015, pp. 92–104.
56. W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, *Shape robust text detection with progressive scale expansion network*, 2019.
57. J. Williamson, *Decoding advertisement*, 1978.
58. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R'emi Louf, M. Funtowicz, J. Brew, *Huggingface's transformers: State-of-the-art natural language processing*, ArXiv **abs/1910.03771** (2019).
59. L. Xiao, X. Li, Y. Zhang, *Exploring the factors influencing consumer engagement behavior regarding short-form video advertising: A big data perspective.* — J. Retailing and Consumer Services **70** (2023), 103170.
60. L. Xing, Z. Tian, W. Huang, M. R, Scott, *Convolutional character networks*, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9126–9136.
61. J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, H. Huang, *EmoSet: A large-scale visual emotion dataset with rich attributes*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20383–20394.
62. K. Ye, N. Honarvar Nazari, J. Hahn, Z. Hussain, M. Zhang, and A. Kovashka, *Interpreting the rhetoric of visual advertisements*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019), 1–1.
63. K. Ye, K. Buettner, A. Kovashka, *Story understanding in video advertisements*, arXiv preprint arXiv:1807.11122 (2018).
64. K. Ye, A. Kovashka, *Advise: Symbolism and external knowledge for decoding advertisements*, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 837–855.
65. A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, *Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

66. R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, *From recognition to cognition: Visual commonsense reasoning*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6720–6731.
67. M. Zhang, R. Hwa, A. Kovashka, *Equal but not the same: Understanding the implicit relationship between persuasive images and text*, arXiv preprint arXiv:1807.08205 (2018).
68. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, *EAST: an efficient and accurate scene text detector*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5551–5560.
69. Y. Zhou, S. Mishra, M. Verma, N. Bhamidipati, W. Wang, *Recommending themes for ad creative design via visual-linguistic representations*, Proceedings of The Web Conference (WWW), 2020, pp. 2521–2527.

Поступило 12 октября 2023 г.

Steklov Institute of Mathematics
at St. Petersburg, Russia
E-mail: anton.m.alexeyev@gmail.com

Sber AI Lab, Russia

Sber AI, Russia;
Kazan Federal University, Russia

Samara National Research University, Russia

Steklov Institute of Mathematics
at St. Petersburg, Russia
E-mail: snikolenko@gmail.com