Analyzing the Robustness of Vision & Language Models

Alexander Shirnin, Nikita Andreev, Sofia Potapova, and Ekaterina Artemova

***Abstract*—We present an approach to evaluate the robustness of pre-trained vision and language (V&L) models to noise in input data. Given a source image/text, we perturb it using standard computer vision (CV) / natural language processing (NLP) techniques and feed it to a V&L model. To track performance changes, we explore the problem of visual questions answering (VQA). Overall, we utilize 5 image and 9 text perturbation techniques and probe three Transformer-based V&L models followed by a broad analysis of their behavior and a detailed comparison. We discovered several key findings regarding the performance of the models in relation to the impact of various perturbations. These discrepancies in performance can be attributed to differences in their architectures and learning objectives. Last, but not least, we perform an empirical study to assess whether the attention mechanism of V&L Transformers learns to align modalities. We hypothesize, that attention weights for related objects and words, should be on average higher than for random object/word pairs. However, our study shows that, unlike is believed for machine translation models, V&L models do not learn alignment at all or exhibit less evidence to do so. This may support the intuition that V&L Transformers overfit to either of the modalities.**

***Index Terms*—Visual question answering, robustness, black-box interpretation, attention mechanism, spurious correlations.**

I. Introduction

MULTIMODAL machine learning is a novel research field that seeks to design models that can process data from heterogeneous sources. Developing a single interface for vision and language (V&L) has gained wide attention both from academic and industrial communities. During the last few years, multiple V&L models have been proposed, motivated by best practices in computer vision (CV) and natural language processing (NLP). All of these models adopt the pre-training paradigm, which is de facto a standard in modern NLP, and are built upon Transformer architecture [1]. However, more efforts have been put into understanding NLP Transformers' inner workings rather than into the introspection of V&L counterparts.

Recent studies have developed a novel research paradigm, addressed as BERTology [2], which seeks an understanding of how pre-trained language models work. A scope of approaches has been proposed to explore how Transformers encode linguistic properties [3], whether Transformers are capable of generalizing to unseen domains and are robust to noises in input data [4], what the functional roles of the attention heads are [5], and how attention heads align source and target languages in machine translation [6]. In this paper, we built upon BERTology principles and aimed to get a better understanding of how V&L models work.

Since the inception of deep convolutional networks [7], numerous extensions and enhancements have been developed. These networks have become a widespread solution to various computer vision problems due to their efficacy and quality. Resnet [8] and VGGNet [9] are examples of backbones that extract crucial information from images and produce embeddings, which can then be used for a variety of tasks including multimodal models. Recent research has demonstrated the feasibility of using transformer-based computer vision models such as ViT [10], SWIN [11], and their derivatives. Nonetheless, this area is still gaining traction and, despite their strengths, such models have more parameters and require more data for training.

Although significant progress has been made in tasks involving a single modality, it is important to recognize that real-world problems often require the integration of multiple modalities, such as vision and text. Common V&L tasks encompass a wide range of applications, including image captioning [12], VQA [13], visual dialog [14], image-text matching [15], and audiovisual facial analysis [16], [17]. These tasks involve the joint understanding of visual and textual information, facilitating the development of multimodal models that can effectively bridge the gap between images and language. Some of them use open-set vocabulary, which hinders numerical techniques for interpretations. Thus we stick to the tasks with closed-set vocabularies. In particular, the VQA problem is usually formalized as a classification problem and is estimated with accuracy. Since the VQA problem is representative and easily evaluated, we choose it to explore V&L models' robustness. The input to the VQA model is an image and a text question. The task in general is to predict the answer to the question. To formalize the problem as a classification task, the output layer of the model selects one answer from a predefined vocabulary of possible answers.

Every year, new approaches appear and new state-of-the-art results are established, but these models often have a rather strong bias towards the data they work with (e.g. images of a certain kind and resolution, style and vocabulary in a test),

Manuscript received 3 July 2023; revised 8 February 2024; accepted 20 April 2024. Date of publication 9 May 2024; date of current version 28 May 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Yunbin Deng. *(Alexander Shirnin and Nikita Andreev contributed equally to this work.) (Corresponding author: Alexander Shirnin.)*

Alexander Shirnin and Ekaterina Artemova are with the HSE University, 101000 Moskva, Russian Federation (e-mail: ashirnin@hse.ru).

Nikita Andreev is with the CAIT, Applied AI Institute, Moskva 121096, Russian Federation.

Sofia Potapova is with the SmallTalk2Me, Covina 91723 USA.

Digital Object Identifier 10.1109/TASLP.2024.3399061

and if one tries to slightly change the input data (e.g. lower the resolution of an image, change some pixels, rephrase or misspell a question) the model starts to respond incorrectly and its overall quality becomes noticeably worse [18]. This happens to them because they only encounter a certain type of data during pre-training, and in the evaluation phase when data from a shifted distribution is used, their performance decreases. The assessment of VQA performance using accuracy as the evaluation metric enables a quantitative evaluation of the robustness exhibited by V&L models. Robustness refers to the models' ability to maintain stable and accurate predictions when faced with variations in input data mentioned previously. By examining how these changes affect their accuracy on the VQA task, we can draw insights into the underlying factors that contribute to robustness or vulnerability in multimodal learning systems.

Our study aims to investigate the robustness of three V&L models with diverse architectures through the implementation of three practical tests. These tests include black-box attacks on texts and images [19], spurious correlation analyses for texts, and alignment investigations in attention maps. Our approach is comprehensive in terms of the diversity of model types and involves a thorough comparison of the models, including the impact of different architectures on the results. The main contributions of this paper are:

- We introduce three diagnostic tests based on the principles of the black box, to investigate the underlying mechanisms of attention and assess V&L models' robustness.
- We evaluate the proposed diagnostic tests on three V&L pre-trained models with different architecture types.
- We examine cross-modality alignment in diverse model architectures, revealing variations in attention maps and optimal alignment.

The rest of the paper is structured as follows. Section II gives a comprehensive review of the related works. In Section III, we present the experimental setup and a set of reference models used for comparison. Section IV describes, according to the black box technique, how perturbed input affects the performance of the models. In Section V, we discover spurious correlations [20] between artifacts in input data and target predictions. In Section VI, we identify connections among input modalities during the internal robustness evaluation. Limitations and the experimental results are discussed in Sections VII and VIII. Finally, Section IX concludes with future work directions.

II. Related Work

A. Vision-Language Tasks and Models

1) Main Tasks and Challenges: As language models have evolved and become more universal, there have been more and more attempts to combine models from different modalities to solve various multimodal problems. The most popular modalities are textual data as well as images, and the most common tasks for them can be divided into 4 categories [21]:

- *Generation:* Image Captioning (IC), Text-to-Image Generation.
- *Understanding:* Visual Question Answering (VQA) [13], Visual Dialog (VisDial) [14], Visual Reasoning (VR) [22], Visual Entailment (VE) [23]
- *Retrieval:* Image-text Retrieval (IR)/Text-image Retrieval (TR)
- *Grounding:* Phrase Grounding (PG) [24], Reference Expression Comprehension (RE) [25]

Our research focuses on the Visual Question Answering (VQA) task, which requires models to extract information from two input modalities in order to answer natural language questions.

2) Models: The initial attempts to create multimodal models involved combining Recurrent Neural Networks (RNN) [26] and Convolutional Neural Networks (CNN) (more precise, LSTM [27] and VGGNet [9]) and some more complicated fusions like MCB [28]. The introduction of Transformer [1] led to a significant advancement in these models, as transformer-like models were utilized as encoders for textual data [29], [30], [31], [32], [33], [34], [35]. The development of CLIP [36] further enhanced this field, demonstrating the creation of a universal model with extensive pre-training that could be utilized for various multimodal tasks.

There are several primary types of model architectures for VQA. These include single stream models [29], [31], [32], [34], [37], which utilize a single large encoder, and dual stream models [33], [38], [39], which use different encoders for each modality. Additionally, these models may differ in their target task type, either discriminative or generative, though both approaches can effectively solve the VQA task. The task is approached as a classification problem; in the latter, it is tackled as a generation task.

B. Attention Mechanism in V&l Models

Attention mechanisms have arguably become the most indispensable module in vision [40], [41], [42], language [1], [43], [44] and acoustic models [45]. The attention mechanism in transformers is a natural gateway to understanding models, as the heatmaps of attention can be used to highlight salient image regions [46], [47] or text tokens [48], [49], [50], [51].

Multimodal models are more comprehensive for research and understanding. One of the challenges where attention mechanisms could be used is to identify alignment between modalities - cross-modal connections and interactions between elements of multiple modalities. For example, when analyzing the speech and gestures of a human subject, how can we align specific gestures with spoken words? Attention mechanisms have been used as an intermediate (often latent) step enabling better performance and capturing both joint undirected (the connections of different modalities are symmetric in either direction) [31], [52], [53] and cross-modal directed (asymmetric connections) [33], [54], [55], [56] alignments.

Another challenge for multimodal models is understanding how model decisions are formed. While exploring soft and hard attention mechanisms in V&L image captioning generation, [42] showed that attention is a simple yet effective mechanism for a neural network to "focus" on salient features of the input. Given

an input state, attention allows the model to dynamically learn weights to indicate the importance of different parts of the input feature.

Since then, attention maps have been a popular choice for intermediate concepts since they are, to a certain extent, human-interpretable, while retaining differentiability. Several approaches have focused on building interpretable components for unimodal importance through soft [57] and hard [58] attention mechanisms, or designing individual modules that are each parametrized by attention operations [59]. In real-life applications, visualizing attention maps helps users conduct error analysis and debug V&L models [60].

C. Diagnostics of V&l Models

Comparative analysis of V&L models has gained significant research interest due to the abundance of available models and their learning capabilities. To evaluate the visio-linguistic compositional reasoning of V&L models, a novel dataset, Winoground was introduced [61]. Winoground comprises sentence pairs that vary by a single word and two images, each image corresponding to either the correct or an incorrect sentence. The task involves selecting the image that matches the correct sentence, or vice versa. Performance is assessed by the accuracy with which a model can accomplish this task. The authors' findings suggest that state-of-the-art V&L transformers and RNN-based models struggle with compositional reasoning, rarely outperforming random chance. Similarly, there are known limitations on efficient use of language structure for visual referring expression recognition [62] reported, where the difference in performance of ViLBERT [33] and MattNet [63] between diagnostic sets ranges up to 15%.

Nikolaus et al. [64] introduced a benchmark to evaluate the comprehension of predicate-noun dependencies in a controlled setup. The task requires identifying the correct sentence corresponding to an image, where two predicate-noun sentences (a target and a distractor) differ only in either the noun or the predicate. Performance is above chance for some models (e.g., LXMERT [33] and UNITER [65]), while some models such as ViLBERT [33] and CLIP [36] perform at chance level. Hendricks et al. [66] proposed a new benchmark to probe models for their understanding of verbs as well as subjects and objects, finding that verbs were the most challenging while subjects were easier than verbs. Similarly, Zhao et al. [67] introduced a VL-CheckList benchmark, which examines seven V&L models from three aspects: object, attribute, and relation. Their results show that spatial relations are relatively more difficult to learn than action verbs. Meanwhile, Parcalabescu et al. [68] evaluated the abilities of V&L models on various linguistic phenomena, including object recognition, object counting, and identification of action participants and roles. Their study reveals that while V&L models are proficient at recognizing named objects and their presence in images, they face challenges when considering language cues in visual contexts, particularly in establishing interactions.

In video-language domain, Schiappa et al. [69] conducted a thorough robustness analysis on five self-supervised video-language models based on CNN and Transformer architectures: MIL-NCE [70], VideoClip [71], COOT [72], UniVL [73], and FIT [74]. Additionally, they introduced benchmark datasets specifically tailored for perturbations in the context of text-to-video retrieval. Key findings from the study reveal that models exhibit greater robustness when subjected to perturbations in text as opposed to video. Furthermore, pre-trained models demonstrate higher robustness compared to those trained from scratch. Interestingly, the analysis indicates that the models prioritize attention toward scenes and objects over motion and action aspects.

Our research differs from previous works as we employ a diverse set of perturbations separately on both modalities (text and images) for the VQA models without the need for additional model training. This approach allows us to analyze the models' responses to perturbations and gain insights into the relative importance of each modality in their decision-making process. Moreover, we go deeper into the models' architecture and suggest a numerical approach to measure cross-modality alignment in the attention mechanism.

III. Experimental Setup

A. Dataset

We use the validation part of COCO VQA v2.0 dataset [13], which consists of 40504 input images, 214354 input text questions, assigned to each image, and text answers for each pair *(image, question)*.

The images are organized into 12 categories with 90 subcategories. Each image can be assigned to several categories (average 6). Each image is equipped with at least 3 questions and 5.4 questions on average.

Answers to the questions are divided into three categories: "yes/no": 80541 answers, "number": 28134 answers, and "other": 105679 answers. The COCO VQA dataset defines a question type with the first three words of questions.

Table I provides the dataset statistics. The top 10 most frequent nouns are collected via the NLTK Python library [75], and the top 10 most frequent named entities (NE) are extracted with the Spacy toolkit [76], namely, with the `en_core_web_sm` pipeline.

B. Models

We use three state-of-the-art V&L pre-trained models for the experiments are: LXMERT [38], VisualBERT [31], OFA-large [78]. These models follow different architecture designs, namely, single and double Transformer streams. Table II shows the differences between the models.

OFA model is installed from the official GitHub repository.[1] LXMERT and VisualBERT are adopted from the Transformers library [79]. Note, that the published VisualBERT version from the huggingface library, is not fine-tuned on the MS COCO

[1] github.com/OFA-Sys/OFA

TABLE I
COCO VQA VALIDATION SET STATISTICS

Top 10 question prefixes		Top 10 most frequent nouns		Top 10 most frequent NE			Top 10 image categories	
Prefix	Count	Noun	Count	Entity	Type	Count	Category	Count
How many	20462	Color	22327	Two	CARDINAL	1079	accessory	21634
Is the	17265	Man	12931	One	CARDINAL	706	animal	10148
What	15897	People	8447	Night	TIME	342	appliance	9303
What color is the	14061	Picture	8125	Zebras	NORP	276	electronic	8265
What is the	11353	Kind	5954	Daytime	TIME	214	food	7776
Is this	7841	Person	5358	More than one	CARDINAL	207	furniture	7004
Is this a	7492	Photo	4839	First	ORDINAL	197	indoor	6010
What is	6328	Woman	4469	Three	CARDINAL	189	kitchen	5548
What kind of	5840	Type	4387	A sunny day	DATE	188	outdoor	5491
Are the	5264	Room	3509	Winter	DATE	164	person	4496

TABLE II
V&L MODELS INVOLVED IN THE STUDY DIFFER IN TERMS OF ARCHITECTURE, PRE-TRAINING TASKS AND OBJECTIVES

Model	Pretraining Dataset	Pretraining Task	Visual Encoder	Language Encoder	Model Architecture	Transformers Streams	VQA Objective	Number of Parameters	Baseline Performance
LXMERT	COCO+VG +VQA2.0+GQA [77] +VG-QA	MLM+MRM +ITM+IQA	BUTD	Transformer	9 BERTLayers for language, 5 BERTLayers for vision, 5 CrossAttLayers for cross-modality	Double	Classification	218M	72.4
VisualBERT	COCO+VQA2.0	MLM+ITM +Task-specific Pretraining	BUTD	Transformer	12 BERTLayers	Single	Classification	114M	70.8
OFA	VQA2.0+VG-QA +GQA and 17 other datasets for other tasks	VQA and 7 other vision/text tasks	ResNet152	Transformer	12 Encoder layers, 12 Decoder layers	Single	Generation	470M	80.3

dataset and does not reproduce the results of baseline performance.

C. Runtime

Calculations are run on GPU TESLA V100 32 GB. LXMERT and VisualBERT take approximately 16 hours to predict answers, while OFA takes approximately 6 hours since it supports batching. Overall there were 42 runs according to the number of perturbations used, 5 image and 9 question perturbations for each of the three models. Models are used for unsupervised scoring only, without fine-tuning or hyperparameters search. We launch all experiments with a fixed random seed. Images and texts are perturbed in advance.

IV. BLACK-BOX ANALYSIS

To test V&L model robustness we follow the black-box paradigm, which simulates real-life cases when a user does not know the model and has no access to training data. We assume to only have access to input data and the ability to collect predictions without having access to the model gradient, loss value, or any internal part of the model. We feed the model with inputs that are modified with controlled perturbations and record which perturbations alter the model's prediction. The perturbation is considered successful if it changes the input slightly, but dramatically changes the output and affects the target metric.

The black-box analysis is widely used to evaluate the robustness of vision and language models towards perturbed inputs [80], [81]. However, there is still room for exploring the effect of black-box perturbations on multimodal models. To this end, we conduct black-box analysis on text and image modalities independently.

A. Images Perturbations

We use standard image perturbations from the Albumentations library [82]. The perturbations are selected as the smallest subset of perturbations yet sufficient for a full study based on related research projects. The main advantage of these perturbations is a real-life relation, as they can easily appear in photos and videos. For example, Random Snow or Sun Flare can easily appear on photos that were taken on a smartphone. Moreover, they are known to keep the image semantics pretty well.

Perturbations, applied to images are:

- Gaussian Blur adds smoothing and blur effect, which makes the image less sharp and objects less visible;
- Grayscale converts color image to grayscale;
- Downscale reduces image resolution, making it fuzzy;
- Sun Flare covers a small fragment of the image with white spot;
- Random Snow adds snow effect to the image;

We evaluate the experimental results with the following metrics for further analysis: D_i for input, Acc and $w2v$ for output of the models:

- D_i is the distance between images, which is computed as the cosine similarity between embedding vectors of source and perturbed images, obtained using a pre-trained ResNet-50 model [8]. It shows how much the image has changed during the perturbation and allows us to compare stronger perturbations with weaker ones (lower D_i means stronger perturbation).

TABLE III
BLACK-BOX ANALYSIS OF IMAGES IN THREE VQA MODELS

Perturbation	LXMERT		VisualBERT		OFA		D_i
	Acc	w2v	Acc	w2v	Acc	w2v	
Gaussian Blur	68.2	0.865	**40.7**	0.935	81.0	0.874	0.993
Downscale	**53.3**	**0.709**	43.8	**0.901**	75.3	0.837	0.985
Grayscale	71.6	0.840	45.2	0.923	**74.1**	0.835	0.990
Sun Flare	66.8	0.845	45.0	0.935	77.8	0.855	0.956
Random Snow	69.9	0.883	44.9	0.912	74.2	**0.832**	0.964
Baseline	73.2	-	45.0	-	82.1	-	

We can see no connection between perturbation strength (D_i) and its effect. Perturbations affect models differently.

- Acc denotes accuracy, the target performance metric for the VQA task.
- $w2v$ stands for the cosine similarity between the model outputs for source and perturbed input images. The output of the VQA task is answer text, which can be presented as a vector using word2vec representations [83]. This additional target performance metric illustrates how semantic similarity changes between answers across perturbations and models, thus in some cases it is more demonstrative than accuracy.

Table III shows the experimental performance of LXMERT, VisualBERT, and OFA on perturbed images. Although we expect lower Acc and $w2v$ for strong perturbations with higher D_i values, there is no evidence of such correlation in either model. On the contrary, the most significant decreases in models' performance are among weak perturbations with lower D_i values: Gaussian Blur for VisualBERT, Downscale for LXMERT, and Grayscale for OFA.

Downscale has the most noticeable effect on all models, while Grayscale has a minor effect. Gaussian Blur significantly affects VisualBERT, but has almost no effect on OFA and a small effect on LXMERT. Also, VisualBERT is robust to strong perturbations (Sun Flare and Random Snow), while both OFA and LXMERT drop in accuracy. Thus, three models are affected by perturbations in three different ways, so that their performance decreases/stays the same on various perturbations. The effect of the perturbation is likely to be connected with the types of augmentations that were used to pre-train each model.

B. Category-Wise Analysis

Gaussian Blur was chosen for categorical analysis as the simplest but effective perturbation. Categories were obtained from dataset labels of image categories. The experiment was run only for LXMERT and OFA since VisualBERT performs poorly and is not of interest for comparative analysis in this case.

The results are summarized in Table IV. Across different categories, there is considerable variation in the performance of the models. Notably, there appears to be no correlation between category-wise performance; higher rates in one category do not necessarily lead to similar rates in another. However, upon closer examination of the models' behavior across categories, two noteworthy cases emerge: "food" and "animals".

TABLE IV
RESULTS BASED ON IMAGE CATEGORY USING Gaussian Blur PERTURBATION

Category	LXMERT			OFA			D_i
	Acc_{before}	Acc_{after}	w2v	Acc_{before}	Acc_{after}	w2v	
accessory	74.9	69.7	0.866	84.0	82.5	0.873	0.993
animal	82.3	76.3	0.885	85.6	84.3	0.881	0.990
appliance	78.2	73.4	0.877	86.3	85.3	0.892	0.995
electronic	75.3	71.2	0.865	83.0	**81.6**	0.866	0.995
food	**62.6**	**58.3**	**0.816**	83.6	82.7	0.872	0.994
furniture	73.4	68.7	0.857	85.1	83.7	0.881	0.994
indoor	71.4	67.2	0.869	83.5	81.8	0.874	0.994
kitchen	70.1	65.5	0.846	84.6	83.5	0.881	0.995
outdoor	74.4	70.8	0.878	83.7	82.4	0.878	0.992
person	74.6	69.2	0.861	83.8	82.2	0.871	0.993
sports	79.6	74.3	0.883	85.9	84.4	0.877	0.993
vehicle	75.2	71.0	0.873	**82.4**	89.3	**0.859**	0.993

There is no impact of the category-wise performance on the overall results.

The "food" category presents a particularly challenging task for the models, resulting in a notable decrease in LXMERT performance of up to 10%. This difficulty can be attributed to the broad scope of the "food" category, encompassing a wide array of diverse meals. Conversely, the "animals" category exhibits a high degree of stability in model performance. This is likely due to the distinct features possessed by animals, which remain recognizable even in transformed images.

C. Questions Perturbations

Text perturbations are divided into symbol-, word-, and sentence-level perturbations. Our selection of text perturbation levels draws upon the methodology designed in NL Augmenter [84]. We use NLPAug[2], NL Augmenter[3] and back-translation from an EasyNMT[4] to craft perturbations.

Perturbations, applied to text questions are:

– *Symbol-level:*

- Keyboard swap swaps 2 random characters in a sentence;
- Insertion symbol randomly adds characters to a sentence, making it longer and slightly less readable;

– *Word-level:*

- U.K. $\rightarrow$ USA converts sentences from British to American English and vice-versa;
- Synonym substitution randomly replaces words with their synonyms;
- Synonym insertion inserts synonyms of randomly selected words;
- Slang (the slangificator from NL augmenter) replaces nouns, adjectives, and adverbs of the original text with their corresponding slang, using the subset of the "Dictionary of English Slang & Colloquialisms";
- Yoda modifies sentences to flip the clauses to sound like Yoda Speaks. For example, "Much to learn, you still have".
- Random Shuffle randomly rearranges words in text;

– *Sentence-level:*

- German (back translation) converts an English sentence to German and back to English

We select realistic perturbations that occur during everyday use. For instance, Random Shuffle and Yoda perturbations might

[2] github.com/makcedward/nlpaug
[3] github.com/GEM-benchmark/NL-Augmenter
[4] github.com/UKPLab/EasyNMT

TABLE V
RESULTS OF BLACK-BOX ANALYSIS ON VQA MODELS

Perturbation	LXMERT		VisualBERT		OFA		D_q	Usage %
	Acc	w2v	Acc	w2v	Acc	w2v		
Keyboard swap	57.8	0.75	37.2	0.71	73.6	0.85	0.82	24.2
Insertion symbol	56.2	0.73	**36.9**	0.70	71.5	0.82	0.83	29.1
UK → USA	**53.4**	0.96	39	0.86	84.6	0.97	0.97	12.7
Synonym substitution	66.1	0.83	38.9	0.78	73.0	0.85	0.86	33.8
Slang	67.5	0.84	42	0.83	77.6	0.89	0.82	11.8
Synonym insertion	65.9	0.85	42.6	0.79	76.3	0.85	0.86	75.8
Yoda	69.5	0.89	43.8	0.83	80.7	0.92	0.95	96.4
Random Shuffle	53.6	**0.68**	37.5	**0.69**	**63.5**	**0.76**	0.9	97.6
German (back translation)	67.9	0.89	41.5	0.85	76.7	0.89	0.93	54.2
Baseline	73.2	-	45	-	82.1	-	-	100

Only questions with $0.75 \leq D_q$ are used. Bold numbers show the worst accuracy and the furthest answers from the previous ones after perturbations. Character-level perturbations are among the most Efficient for VQA models but had the lowest question similarity scores; the U.K. → USA perturbation appeals difficult for LXMERT and visualBERT models, while OFA performs well and had a 2% accuracy increase; random shuffle has a larger Performance hit, but still has high question similarity scores, indicating that sentence transformers currently ignore word order.

imitate a non-English speaker whose native language allows for the rearrangement of words in a sentence. Similarly to image perturbations, we evaluate the experimental results with *Acc*, *w2v* and D_q, i.e. distance between questions. Following related work [85], we define the distance between questions as the cosine distance between source and perturbed question embeddings, obtained from the `all-MiniLM-L6-v2` model from the Sentence Transformers toolkit[5] [86].

For the images an average embedding score was higher than 0.95, thus we left all samples from the validation set for black-box analysis on the images. Whereas for some text perturbations an average D_q was less than 0.7, it was necessary to filter out the questions. Therefore we didn't use questions that were perturbed too much ($D_q < 0.75$) as well as questions that didn't change at all. For example, back translation might return the source sentence or it might be impossible to replace words with slang. Note, that only the remaining questions are used for the calculation of D_q in the table V. In Table V the *Usage %* column shows the percentage of used examples from the entire dataset. As a measure of perturbation success rate, we used the ratio of transformed questions. Slang appears to be the least successful perturbation, as it changes only 11.8% of the data.

Table V shows the results of text perturbations on the models. Judging by the drop in accuracy, character-level perturbations are among the most efficient for models. However, these methods also lead to the lowest similarity values between source and perturbed questions. The U.K. → USA perturbation keeps questions similarity at high scores (0.97) but manages to generate out-of-domain questions, which confuses the LXMERT and VisualBERT models. At the same time, OFA copes well with this perturbation, and its accuracy is increased by around 2%. Word order changes of the Yoda perturbation do not affect the models much, leading to a slight drop in the performance of all models. Another perturbation for changing word orders randomly (Random Shuffle) decreases the performance of models more significantly. However, we expect a larger performance drop, as the shuffled questions have a new meaning, and the previous answers are likely to be incorrect. However, the question similarity remains high (0.9) even though a large sample of 97% of the dataset is used. We might conclude that at the moment sentence model almost ignores the word order and they do not filter shuffled sentences.

[5]sbert.net

D. Results

Overall, the models' behavior is different, that is their performance decreases/stays the same on various perturbations.

The LXMERT performance drops after perturbations by up to 27% along with the $w2v$ metric as the similarity between the predictions (mean of 0.828) for intact and transformed images changes significantly. The performance of VisualBERT stays consistent after image perturbations and drops by about 10% after text perturbations. However, its performance is much lower in comparison to LXMERT. Thus, the little fluctuations in performance can be attributed either to the low quality of the model or to the perturbations, but the predictions do not change much, and hence the $w2v$ values are high (above 0.9) with a mean of 0.921. The performance of OFA also drops after all perturbations, but the effect varies from the LXMERT one. OFA maintains high relative similarity values at the level of 0.847 (mean value across perturbations) and higher and $w2v$ has almost the same changes as accuracy.

Next, we compare the mean δ between the baseline and model accuracy after perturbations. It results in 9.9% for LXMERT, 2.4% for VisualBERT, and 6.8% for OFA for images. For text perturbations, δ was 15.3%, 11.2%, and 8.3%, respectively, for LXMERT, VisualBERT and OFA. This shows that our perturbations are most effective for LXMERT.

We attribute the difference in models' behavior to their architectures and learning objectives. LXMERT processes the input modalities in parallel streams. If one of the streams is affected, the performance of the whole model degrades. VisualBERT performs low from scratch. In VisualBERT the modalities are mixed, so that the model relies less on one modality. This leads to insignificant changes in performance. OFA is pre-trained with image masking, unlike LXMERT and VisualBERT, which were trained on downstream tasks only. Thus, the performance of OFA is more stable and less prone to perturbations in input data.

V. SPURIOUS CORRELATIONS IN QUESTIONS

Language models tend to over-rely on spurious correlations: words, that indicate a certain class for a majority of samples but do not always do so [20], [87]. For example, the sentiment classification model may learn spurious correlations between n-grams "Spielberg" and "New York Subway" as these co-occur frequently with positive and negative classes in the training data [88]. In line with these studies, we explore simplistic text perturbations to reveal the spurious learned by VQA models.

We test the VQA model's sensitivity to word removal. Given a question, we loop over all words and remove the words one by one. At each step, we feed the perturbed question to the VQA model. We record if the model's output differs from the output of the source question, the current word, its position, and its morpho-syntactic features. Towards this, we apply the Stanza part-of-speech (POS) tagger and dependency parser to label the

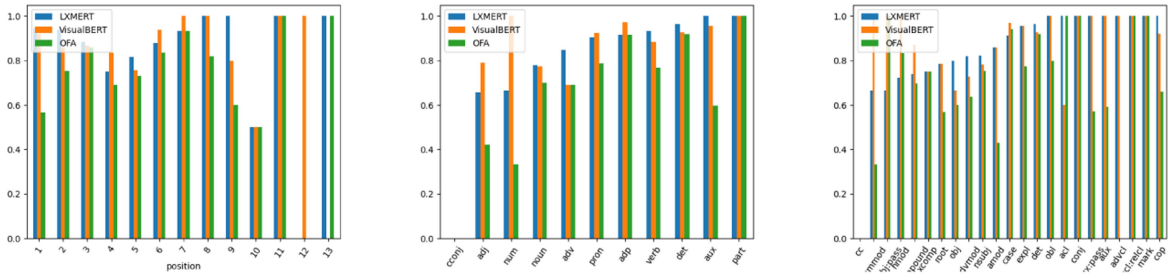


Fig. 1. Analysis of spurious correlations. The decrease in accuracy according to the position (**left**), part-of-speech tags (**middle**), and dependency tags (**right**) of removed words. The performance is the most affected when the words positioned towards the end of the sentence are removed. Models generally perform well when functional words are removed. The performance of all models decreases the most when a coordinating conjunction (cc) is deleted from the questions. The experiment showed that OFA is the most affected model, while VisualBERT is the most robust.

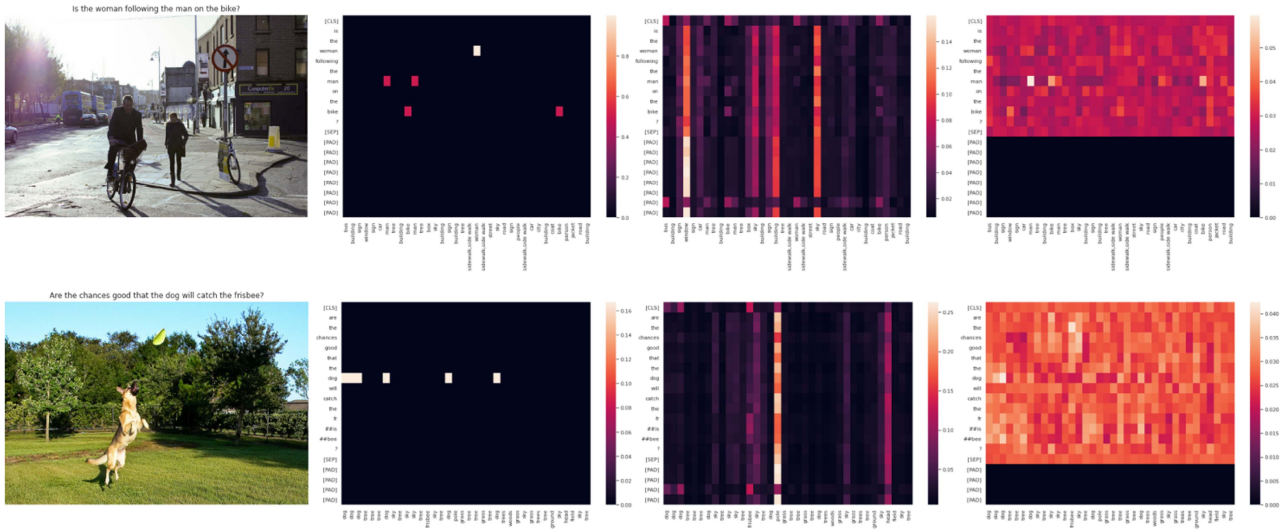


Fig. 2. Example of image (**left**), golden matrix (**middle-left**), LXMERT attention (**middle-right**) and VisualBERT attention (**right**). Here we can see the difference in attention maps: LXMERT has more attention to key objects on `[CLS]` and `[PAD]` tokens whose weights differ from other tokens, while VisualBERT has many uniform weights and more attention on the key object's token.

words with their (i) Universal POS tags[6] and (ii) tags of Universal Dependency Relation[7] to their dependents. For the sake of time, we randomly select 500 sentences, to which all models produce correct answers.

Removing words from input questions generally decreases VQA performance by up to 20%. Fig. 1 shows the decrease with respect to the position, the part of the speech tag, and the dependency tag of the removed word. The left-hand side plot shows, that the performance is the most affected when the words positioned towards the end of the sentence are removed. In this experiment OFA is the most affected, losing up to 40% performance when the first word is removed from the question. VisualBERT seems to be the most robust model, maintaining performance above 80% in most cases.

We analyze the impact of removed word dependency tags on models. LXMERT and VisualBERT are affected rather similarly. For about half of the cases, the accuracy is not affected and stays at about 1, in other cases for both models it went from 1 to 0.6 on average, depending on the word's dependency tag. However, OFA's accuracy in most cases is between 0.3 to 0.8 and less than 0.6 in half of cases. Finally, the performance of all models decreases the most when a coordinating conjunction (cc) is deleted from the questions. The middle part of Fig. 1 shows the changes in the performance concerning the part-of-speech tag of the removed word. The models cope well, when functional

TABLE VI
ALIGNMENT SCORES STATISTICS

		LXMERT			VisualBERT		
		top	middle	bottom	top	middle	bottom
TP + TN	mean	0.268	0.146	0.268	0.198	0.195	0.197
	std	0.149	0.080	0.149	0.139	0.137	0.143
FP + FN	mean	0.346	0.182	0.346	0.176	0.173	0.176
	std	0.250	0.114	0.250	0.193	0.184	0.195
all question	mean	0.284	0.153	0.284	0.187	0.184	0.187
	std	0.176	0.089	0.176	0.168	0.162	0.171

LXMERT shows higher alignment scores when compared to visualBERT. Higher metrics for the top row mean that the [CLS] token has important information about key objects in the image

[6]universaldependencies.org/u/pos/
[7]universaldependencies.org/u/dep

words, such as a participle (part), auxiliary verb (aux), or determiner (det) are removed. The only exception is a coordinating conjunction.

VI. Cross-Modality Alignment

Multiple recent works study the role of the attention mechanism in Transformer-based models. The attention mechanism in machine translation models learns a shallow alignment between source and target languages [44]. The attention heads in pre-trained language models exhibit several well-defined patterns such as attending to functional tokens (SEP) or next or previous tokens [5]. Following this line of research, we hypothesize that the attention of V&L models should learn to align semantically close units, such as visual objects and corresponding words. We seek to conduct an experimental evaluation of this hypothesis.

The attention matrix of LXMERT and VisualBERT models is responsible for matching modalities. We seek to explore whether the attention matrix learns to align visual objects (e.g. the image patch of a cat) with the semantically close parts of questions (words "cat", "kitten", "pet", etc.). We compute the model's attention map by averaging the attention weights of different head. The x-axis of the attention map corresponds to visual objects, and the y-axis — to question tokens. The names of visual objects are obtained from the built-in Faster R-CNN [89]. We would expect the (i, j) weight to be high (or close to unity) for aligned visual objects and tokens, and lower otherwise because. This is similar to a human's attention when answering a question using an image. For this purpose, we craft a `golden` matrix with ideal elements (with attention only to the target objects) that aims to emulate aligned vision objects and words.

For each question-image pair, we design a golden alignment matrix in the following way. The rows of a matrix stand for the question tokens, while the columns stand for the visual objects. The (i, j) element is set to 1 if the cosine similarity score between the token embeddings and the object name is greater than the predefined threshold, which we set manually to take words with high semantic similarity into account. Otherwise, it is 0. The shape of the question-image pair attention map and a golden matrix coincide allowing for element-wise comparison.

Next, we select the first non-zero vector from the golden matrix (and name it `golden` vector, and it has `golden` index in the matrix), as we suppose that it stands for the main object in the image and we expect main attention on it, (again, like the human focus). We compare this vector to 3 rows of the original attention map: `top`, `middle`, and `bottom`. Finally, we compute the alignment score using Algorithm 1.

We sample 100 random image-question pairs from the COCO VQA validation part. These 100 image-question pairs can be categorized into: (i) True Positive + True Negative (those, to which the model produces correct answer), (ii) False Positive + False Negative (those, to which the model fails to produce correct answers), (iii) all question (combination of (i) and (ii)). Table VI shows the results by category in an aggregated way

Table VI shows that the top and bottom (`[CLS]` and `[PAD]` are the special rows) rows for LXMERT are equivalent and have much higher scores than those of a regular row. The most important weights of target objects are encoded in the special rows. The VisualBERT model does not perform well and its attention matrix is quite sparse. Thus low alignment scores and high standard deviation values can be attributed to attention map scarcity.

Algorithm 1: Alignment Score Calculation.

Input: `att_row` is an attention row which score we want to obtain, `golden` is a golden row, n is the length of the vector
Output: `res` is the alignment score (float)

```
function alignment_scoreatt_row, golden, n
  res := 0
  for i=1,...,n do
    if golden[i] != 0 then
      res += att_row[i]
    end if
  end for
  return res
end funtion
```

Overall, we can not confirm that there is an obvious alignment between visual objects and question tokens. One of the possible directions of the future work would be to evaluate in a head-wise fashion.

Lastly, to address the lack of alignment, the next-generation V&L models, such as UNITER [29], have been trained to enforce visual object-to-word alignment using additional pre-training tasks, resulting in a significant performance boost due to the learned inter-modality alignment.

VII. Limitations

a) The choice of the downstream task: We employ Visual Question Answering (VQA) as a downstream task to assess the robustness of the Vision and Language (V&L) model. VQA stands out as one of the most extensively researched tasks, offering multiple advantages. It serves both as a well-studied benchmark and as a task designed as a cloze-ended prediction. Unlike VQA, tasks like image captioning and text-to-image generation yield free-form outputs, making them challenging to evaluate formally with performance metrics.

b) Black-box setup: We focus mostly on black-box analysis, which requires prior domain knowledge of model vulnerabilities to be effective. This approach entails studying the behavior and responses of a model without access to its inner workings or parameters, relying solely on input-output interactions. Other commonly used approaches to evaluate model robustness, such as adversarial attacks, search for counterfactuals, or probing tasks, assume access to models' representation, prediction probabilities, and gradients, and are out of the scope of black-box setting.

c) The choice of perturbations: Our choice of perturbations is inherently constrained: we opt for a small number of perturbations inspired by real-life use cases, but it is not computationally feasible to account for an extensive set of perturbations. Each inference run takes around 30–40 hours on our GPU.

d) Ungrammatical questions: The perturbed question may become ungrammatical. Removing one word from the question may change the meaning of questions and lead to incorrect answers as well. As a result, we may rely on ill-formed questions. It might require additional human evaluation to judge how grammatical and natural perturbed questions are. However, this is out of the scope of the current study.

e) Distance between texts: There is no universally agreed-upon method for measuring distances between images or texts. Although we use one of the best practices for measuring the distance between questions (Sentence Transformers), it has its limitations. Random shuffling of words does not change the distance significantly, despite potentially affecting the meaning.

VIII. Discussion

a) Difference between modalities: As we generate out-of-domain data to test the models, we assume that the models would not consistently provide accurate responses to perturbed questions. Simultaneously, the decrease in quality is more pronounced in response to text perturbations compared to image perturbations. We assume that it is because perturbed images do not lead to a considerable failure of object recognition, thereby leading to only a minor reduction in performance. Furthermore, manual exploration of the perturbed data indicates that changes made to the text are of greater significance than those made to the images.

b) Spurious correlations: Analysis of spurious correlations shows that there is no overfitting on surface forms or words of particular parts of speech. The exception is a coordinating conjunction. However, a reduction in the quality of results stemming from the removal of coordinating conjunctions may be reasonable, as their absence could lead to a change in the intended meaning of the question and consequently, the appropriate answer. In other cases, there is some decrease in accuracy, but there is no clear correlation between it and the word being removed. We hypothesize that the utilization of two modalities, namely images and text, in the models accounts for the absence of overfitting in the textual modality.

Our studies on cross-modality alignment show that with different architectures attention maps can vary as well as optimal alignment. Moreover, different models connect modalities in different ways (depending on encoders, architectures, training procedures, etc.). Therefore, the common structure is not possible in this case. However, in general, some alignment still can be found in special cases.

c) Cross-modality alignment: Our research involves choosing two subsets of perturbations to compute performance metrics concerning how effectively models perform real-world tasks. Although users typically fine-tune or modify a model's parameters to fit specific distributions when employing it with real data, our work remains valuable because it sheds light on whether such models require extra customization before being deployed on real-world datasets or can function optimally out of the box.

d) Model reliability: We strongly believe that the quality of data plays a crucial role in V&L models' applicability within practical contexts. Unfortunately, several factors contribute to diminished data quality in real-world circumstances. Consequently, we consider it necessary to examine and tailor V&L models not only using standard benchmarks but also with artificial data customized for real-life scenarios and vulnerabilities specific to target domains.

e) Model interpretability: To build trust in the deployment of V&L models, there is a need for increased transparency and explainability. Researchers and developers should strive to make these models interpretable, allowing users to understand the decision-making processes and potential limitations. The *golden matrix* approach proposed in the paper or similar approaches could be used to measure the relevance of modalities.

f) Adversarial defenses: Our findings underscore the importance of developing effective adversarial defense strategies, particularly in the current landscape dominated by LLMs and widespread API-level deployments. As the models are susceptible to various perturbations, investing in defense mechanisms becomes imperative to enhance the resilience of V&L models against potential attacks.

IX. Conclusion and Future Work

In this work, we evaluated the robustness of recent multimodal Vision and Language (V&L) transformers to various basic yet effective text and image perturbations, applied to the task of Visual Questions Answering (VQA). We conducted a comparative analysis of VQA models with various architecture types to determine any potential vulnerabilities. The decrease in performance was more pronounced in response to text perturbations compared to image perturbations. Furthermore, we delved deeper into the models' inner workings and explored the alignment in attention maps. To this end, we proposed an algorithm for measuring modalities alignment. Additionally, we investigated spurious correlations in textual data. Our findings indicate that there is no evidence of significant overfitting regarding words belonging to various parts of speech, except for coordinating conjunctions.

In our future work, we plan to further investigate model robustness on other V&L generation and retrieval tasks. Such tasks might not have a tractable performance metric. Thus, we are going to adopt suitable metrics to evaluate the robustness of the model in such cases. Another direction of our research is to explore whether V&L models are prone to ethical biases. Since V&L models are integrated into various applications and systems, it is critical to ensure that they do not discriminate against certain societal groups or expose ethical issues.

Finally, we plan to compare the results of our diagnostic tests with other existing methods, focusing on how V&L models perform under adversarial attacks, handle counterfactuals, and analyze the performance of models' hidden representations in probing tasks. This would entail developing novel attacking methods and probing suites that incorporate both modalities or adopting efforts from the Computer Vision and Natural Language Processing communities. These comparisons will inform the future enhancement of V&L models and aid machine learning practitioners in deploying safer and more trustworthy models.

APPENDIX

TABLE VII
TEXT TRANSFORMATIONS, QUESTIONS ARE SAMPLED FROM THE VALIDATION SET

Transformation	Original → Transformed
Keyboard swap	Where is he looking? Where is he lokkiHg?
Insertion symbol	Where is he looking? Where is he lxookinsg
UK → USA	What is the color of the building in the background? What is the colour of the building in the background?
Synonym substitution	Is the building large? Is the construction big?
Slang	What are the people in the background doing? What are the peeps in the background doing?
Synonym insertion	How many giraffes can been seen? How many giraffes giraffe can been seen visit?
Yoda	Where is he looking? He looking, where is?
Random Shuffle	What are the people in the background doing? Doing the what people are in the background?
German (back translation)	What website copyrighted the picture? Which website has the image protected by copyright?

A. Alignment Rows

As for selected rows for the LXMERT model we use `[CLS]`, *golden* index, and penultimate `[PAD]` rows for comparison. For VisualBERT we use `[CLS]`, `golden index`, and `[SEP]` rows.

B. Alignment Score Example

The alignment score for each row is defined as the sum of the elements x_j of the row, which index has a non-zero element in the golden vector $\texttt{golden vector}[i] \neq 0$, or in the matrix form $\texttt{att\_row}[\texttt{golden} > 0]).sum()$. For example, if the golden row is [0, 1/2, 0, 1/2, 0] and the attention map row is [0, 1/3, 1/3, 1/3, 0], the resulting score is 2/3.

C. Golden Matrix Example

Suppose we have such normalized attention matrix for question *Where is the beer?:*

	beer	*table*	*bottle*	*floor*
[CLS]	0.5	0.08	0.02	0.4
where	0	0.33	0.33	0.34
is	0.4	0.3	0.15	0.15
the	0.6	0.2	0.2	0
beer	0.7	0.05	0.2	0.05
?	0.8	0.2	0	0
[SEP]	0.25	0.25	0.25	0.25
[PAD]	0.5	0.08	0.02	0.4
[PAD]	0.3	0.3	0.3	0.1

We compute tokens and objects embeddings (simple word embeddings) cosine similarity (for each token-object pair) and get *golden* matrix (again, we normalize each row):

	beer	*table*	*bottle*	*floor*
[CLS]	0	0	0	0
where	0	0	0	0
is	0	0	0	0
the	0	0	0	0
beer	0.5	0	0.5	0
?	0	0	0	0
[SEP]	0	0	0	0
[PAD]	0	0	0	0
[PAD]	0	0	0	0

Then, we take first non-zero row in our *golden* matrix (for the beer token) [0.5, 0, 0.5, 0] (we call it *golden row*) and the same row in the original matrix + rows for [CLS] and penultimate [PAD]:

	beer	*table*	*bottle*	*floor*
[CLS]	0.5	0.08	0.02	0.4
beer	0.7	0.05	0.2	0.05
[PAD]	0.5	0.08	0.02	0.4

After, for each non-zero position in *golden row* (first and third in our case) we sum values in the above matrix, so:

- for `[CLS]` we have 0.5 + 0.02 = 0.52
- for *beer* we have 0.7 + 0.2 = 0.9
- for `[PAD]` we have 0.5 + 0.02 = 0.52

And now we have our alignment metric scores 0.9, 0.52, 0.52.

For VisualBERT we adapt our algorithm as it has a different attention architecture, so we use columns instead of rows, `[SEP]` token instead of `[PAD]` and we compute only cross-modal (token to object) scores (token to token scores are not taken into account).

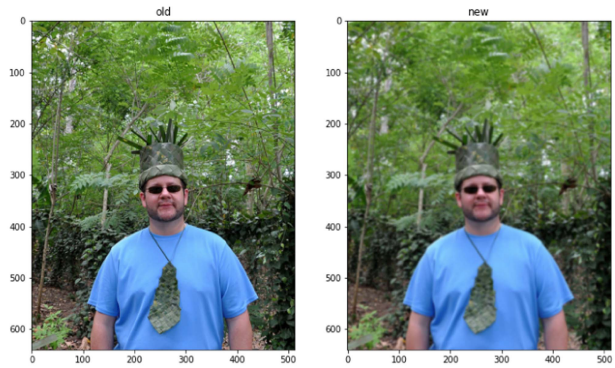


Fig. 3. Gaussian Blur before **(left)** and after **(right)**.

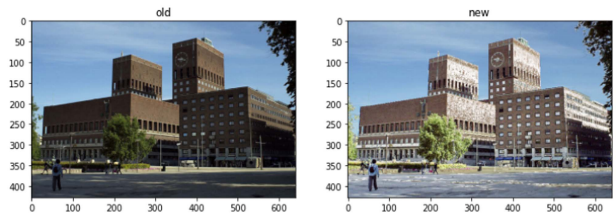


Fig. 4. Random Snow: before **(left)** and after **(right)**.

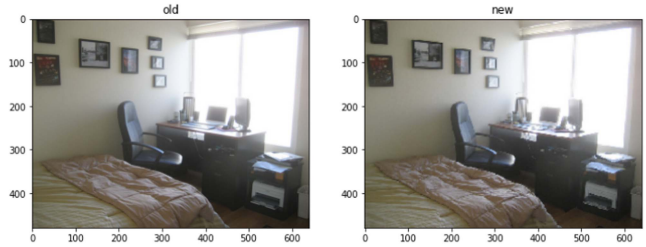


Fig. 5. Downscale before **(left)** and after **(right)**.

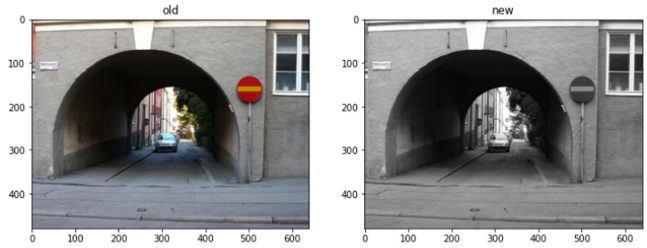


Fig. 6. Grayscale before **(left)** and after **(right)**.

Fig. 7. Sun Flare before **(left)** and after **(right)**.

ACKNOWLEDGMENT

This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University). We acknowledge the computational resources of HPC facilities at the HSE University.

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[2] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54

[3] A. Conneau and D. Kiela, "SentEval: An evaluation toolkit for universal sentence representations," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, pp. 1699–1704. [Online]. Available: https://aclanthology.org/L18-1269

[4] B. Wang et al., "Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–13.

[5] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of BERT's attention," in *Proc. ACL Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2019, pp. 276–286. [Online]. Available: https://aclanthology.org/W19-4828

[6] Y. Chen, Y. Liu, G. Chen, X. Jiang, and Q. Liu, "Accurate word alignment induction from neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 566–576. [Online]. Available: https://aclanthology.org/2020.emnlp-main.42

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–21.

[11] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[12] A. S. Kharchevnikova and A. V. Savchenko, "Visual preferences prediction for a photo gallery based on image captioning methods," *Comput. Opt.*, vol. 44, no. 4, pp. 618–626, 2020.

[13] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.

[14] A. Das et al., "Visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 326–335.

[15] A. F. Biten, A. Mafla, L. Gómez, and D. Karatzas, "Is an image worth five sentences? A new look into semantics for image-text matching," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1391–1400.

[16] A. V. Savchenko, "MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2022, pp. 45–59.

[17] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct.–Dec. 2022.

[18] K. Mahmood, R. Mahmood, and M. V. Dijk, "On the robustness of vision transformers to adversarial examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7838–7847.

[19] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2484–2493.

[20] L. Tu, G. Lalwani, S. Gella, and H. He, "An empirical study on robustness to spurious correlations using pre-trained language models," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 621–633, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.40

[21] F. Li et al., "Vision-language intelligence: Tasks, representation learning, and large models," 2022, *arXiv:2203.01922*.

[22] J. Johnson, B. Hariharan, L. V. D. Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2901–2910.

[23] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," 2019, *arXiv:1901.06706*.

[24] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2641–2649.

[25] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process*., 2014, pp. 787–798. [Online]. Available: https://aclanthology.org/D14-1086

[26] M. I. Jordan, "Serial Order: A Parallel Distributed Processing Approach," in *Advances in Psychology*, vol. 121. Amsterdam, The Netherlands: Elsevier, 1997, pp. 471–495.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.

[29] Y.-C. Chen et al., "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.

[30] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," 2019, *arXiv:1909.02950*.

[31] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "What Does BERT with Vision Look At?," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 5265–5275, doi: 10.18653/v1/2020.acl-main.469.

[32] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11336–11344.

[33] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[34] W. Su et al., "VL-BERT: Pre-training of generic visual-linguistic representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[35] A. Savchenko, A. Alekseev, S. Kwon, E. Tutubalina, E. Myasnikov, and S. Nikolenko, "Ad lingua: Text classification improves symbolism prediction in image advertisements," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1886–1892.

[36] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[37] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6616–6628.

[38] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5100–5111.

[39] F. Yu et al., "ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3208–3216.

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[41] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6964–6974.

[42] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[43] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[44] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[45] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[46] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 782–791.

[47] E. Aflalo et al., "VL-InterpreT: An interactive visualization tool for interpreting vision-language transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21406–21415.

[48] B. Hoover, H. Strobelt, and S. Gehrmann, "exBERT: A. visual analysis tool to explore learned representations in transformer models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2020, pp. 187–196.

[49] V. Lal et al., "InterpreT: An interactive visualization tool for interpreting transformers," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations*, 2021, pp. 135–142. [Online]. Available: https://aclanthology.org/2021.eacl-demos.17

[50] R. Li, W. Xiao, L. Wang, H. Jang, and G. Carenini, "T3-Vis: A visual analytic framework for training and fine-tuning transformers in NLP," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, Dominican Republic, Nov. 2021, pp. 220–230, doi: 10.18653/v1/2021.emnlp-demo.26.

[51] I. Tenney et al., "The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstration*, 2020, pp. 107–118.

[52] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-attention architectures for task-oriented language grounding," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2819–2826.

[53] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7463–7472.

[54] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 570–585, 2021.

[55] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4334–4340.

[56] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[57] D. H. Park et al., "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8779–8788.

[58] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[59] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2016, pp. 1545–1554.

[60] P. P. Liang et al., "MultiViz: Towards visualizing and understanding multimodal models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–49.

[61] T. Thrush et al., "Winoground: Probing vision and language models for visio-linguistic compositionality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5238–5248.

[62] A. Akula, S. Gella, Y. Al-Onaizan, S.-c. Zhu, and S. Reddy, "Words aren't enough, their order matters: On the robustness of grounding visual referring expressions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6555–6565.

[63] L. Yu et al., "MAttNet: Modular attention network for referring expression comprehension," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1307–1315.

[64] M. Nikolaus, E. Salin, S. Ayache, A. Fourtassi, and B. Favre, "Do vision-and-language transformers learn grounded predicate-noun dependencies?," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 1538–1555.

[65] T. Brown Askell et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[66] L. A. Hendricks and A. Nematzadeh, "Probing image-language transformers for verb understanding," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP* 2021, pp. 3635–3644.

[67] T. Zhao et al., "VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations," 2022, *arXiv:2207.00221*.

[68] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 8253–8280.

[69] M. Chantry, S. Vyas, H. Palangi, Y. Rawat, and V. Vineet, "Robustness analysis of video-language models against visual and language perturbations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 34405–34420.

[70] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9879–9889.

[71] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6787–6800.

[72] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "COOT: Cooperative hierarchical transformer for video-text representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22605–22618.

[73] H. Luo et al., "UniVL: A unified video and language pre-training model for multimodal understanding and generation," 2020, *arXiv:2002.06353*.

[74] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1728–1738.

[75] S. Bird, "NLTK: The natural language toolkit," in *Proc. COLING/ACL Interactive Presentation Sessions*, 2006, pp. 69–72.

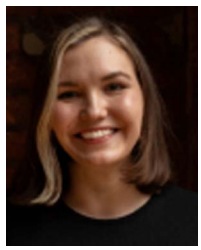
[76] M. Honnibal, I. Montani, S. V. Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://spacy.io

[77] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6693–6702.

[78] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 23318–23340. [Online]. Available: https://proceedings.mlr.press/v162/wang22al.html
[79] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.
[80] T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu, "Black-box tuning for language-model-as-a-service," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 20841–20855.
[81] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black-box deep neural networks for computer vision: A survey," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 4, pp. 966–989, 2021.
[82] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020, Art. no. 125.
[83] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[84] K. Dhole et al., "NL-augmenter: A framework for task-sensitive natural language augmentation," *Northern Eur. J. Lang. Technol.*, vol. 9, no. 1, pp. 1–41, 2023.
[85] N. T. McDermott, J. Yang, and C. Mao, "Robustifying language models with test-time adaptation," in *Proc. Workshop Pitfalls Limited Data Computation Trustworthy ML*, 2023, pp. 1009–1016.
[86] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3982–3992.
[87] J. Eisenstein, "Informativeness and invariance: Two perspectives on spurious correlations in natural language," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 4326–4331. [Online]. Available: https://aclanthology.org/2022.naacl-main.321
[88] T. Wang, R. Sridhar, D. Yang, and X. Wang, "Identifying and mitigating spurious correlations for improving robustness in NLP models," in *Proc. Findings Assoc. Comput. Linguistics: NAACL*, 2022, pp. 1719–1729. [Online]. Available: https://aclanthology.org/2022.findings-naacl.130
[89] S. Ren, K. He, R. Girshick, and J. Sun, "F. R- CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

Alexander Shirnin is currently working toward the postgraduation in data science with the Laboratory for Models and Methods of Computational Pragmatics, Higher School of Economics. He is currently a Scientist with the Laboratory for Models and Methods of Computational Pragmatics, Higher School of Economics. He is also distinguished as a kaggle.com competitions expert. He has a deep passion for developing state-of-the-art solutions in diverse machine-learning domains, maximizing performance metrics, and pushing the boundaries of innovation.

Nikita Andreev received the undergraduation degree with the Higher School of Economics. He is currently working toward the master's degree with CAIT and Applied AI Institute. He has a strong background in computer science and artificial intelligence. During his time there, he developed a passion for multimodal transformers and pursued several research projects in this field. His research focuses on generative models, particularly optimal transport and its applications in image translation.



Sofia Potapova received the master's degree from Yandex Data School, Russia. She is currently the Head of Data Science with SmallTalk2Me, leveraging her expertise in computer science earned through the master's degree. With a background as a Teaching Assistant of Deep Learning and Computer Vision courses with the Higher School of Economics and Yandex Data School. She has transitioned to the industry, where she is currently focused on developing deep learning algorithms. Her experience extends to her work with Yandex, where she improved perception algorithms for self-driving cars.

Ekaterina Artemova received the graduation degree from the Higher School of Economics, Moscow, Russia, and the Ph.D. degree in computer science from the Russian Academy of Science, Moscow. She is currently a Research Scientist with Toloka.AI. She completed her postdoctoral studies with LMU, Munich, Germany. She actively authored or coauthored in top computer linguistics venues, focusing on language model evaluation, low-resource language technologies, and natural language understanding.