Записки научных
семинаров ПОМИ
Том 529, 2023 г.

V. Moskvoretskii, A. Frolov, D. Kuznetsov

IMAD: IMAGE-AUGMENTED MULTI-MODAL DIALOGUE

Abstract. Currently, dialogue systems have achieved high performance in processing text-based communication. However, they have not yet effectively incorporated visual information, which poses a significant challenge. Furthermore, existing models that incorporate images in dialogue generation focus on discussing the image itself. Our proposed approach presents a novel perspective on multi-modal dialogue systems, which interprets the image in the context of the dialogue. By doing so, we aim to expand the capabilities of current dialogue systems and transition them from single modality (text) to multi-modality. However, there is a lack of validated English datasets that contain both images and dialogue contexts for this task. Thus, we propose a two-stage approach to automatically construct a multi-modal dialogue dataset. In the first stage, we utilize text-to-image similarity and sentence similarity to identify which utterances could be replaced with an image. In the second stage, we replace those utterances by selecting a subset of relevant images and filtering them with a visual question answering model. We used this approach, along with additional labeling, to create the IMage Augmented multi-modal Dialogue dataset (IMAD), which can serve as a validated dataset for this task. Furthermore, we propose a baseline model trained on this dataset, which outperforms model trained on the same data without images and BlenderBot.

§1. Introduction

Dialogue systems, also known as conversational agents or chatbots, have become increasingly important in recent years due to their potential to revolutionize human-computer interaction [37].Furthermore, one can see a high activity in this field in the recent year, as the usage of ChatGPT [39] can serve a lot of different goals [10, 20, 35]. In line with ChatGPT, Google has recently announced Bard, based on LaMDA [57], which can serve for the same tasks. Additionally, dialogue systems provide a challenging problem

Key words and phrases: Natural Language Processing and Deep Learning and Machine Learning and IMAD and Dialogue Dataset and Multi-modal Dataset and Dialogue Systems and Multi-modality.

in AI research, as they require a deep understanding of natural language and the ability to generate human-like responses [6]. A good confirmation of this thesis is the abundance of different natural language products, that are widely used, such as DeepPavlov [4].

Contemporary dialogue models such as DialoGPT, BLOOM, and DialogBERT are predominantly text-based [14, 61, 68]. This is reasonable in scenarios where individuals converse solely through textual communication. However, in real-life situations dialogues frequently incorporate images, such as when individuals respond to questions with photographs, provide offers or express emotions [69]. As a result, there is a need for dialogue models that can accommodate multi-modal inputs.

Just as in the dialogue systems, high activity is being spotted in the field of text2image generation [44, 50, 51]. These tools are also helpful in the art sphere [2], healthcare [8], physics [36] and more. However, these models are limited to producing only images and taking text information as input.

These problems are solved with multi-modal deep learning models, that are increasingly important today due to the exponential growth of multimedia data in various domains [3, 13, 19, 54]. Such models have numerous applications in various domains. There is a strong potential of incorporating more modalities to dialogue assistants, such as better emotion recognition [7], visual question answering [29, 62] and operating with a wider range of tasks [1, 17, 46].

Therefore, one could focus on the task of generating an image description in a context of dialogue [23, 69]. This task is more general compared to the response generation [39] or describing a picture [28], because it allows for better response generation with intention knowledge and knowledge about the image sense in a certain dialogue context.

Proposal. That brings us to the task of interpreting an image in the context of dialogue, which would provide a solution for the above-mentioned problem. We present the IMage Augmented multi-modal Dialogue dataset (IMAD) that contains 4864 dialogues where the last utterance was replaced with image. To collect it we utilize multiple sources of dialogue datasets, present a novel approach for image-text dialogue construction, and label a part of it with 3 assessors. We also present baseline models based on the BLIP model [29] that outperform text-only BLIP and BlenderBot 400M [49] on this task. Training data included 4154 samples collected with automated approach and 582 samples labeled with assessors as "Partial Match". Test data included 128 samples, labeled by assessors and authors as "Perfect Match".

Out dataset and code are published at the IMAD Repository[1].

§2. Related Works

Multi-modal Tasks. Multi-modal models involve multiple modalities, such as images [16, 29, 44, 60], video [25, 64, 65], or audio [56, 58]. In the field of text and image modalities there are several popular tasks, including visual question answering [18, 28, 29, 31] and image captioning [22, 26, 59]. We also distinguish as a separate task image-text matching, where the key is to match corresponding images and texts [29, 43, 63]

Multi-modal Embeddings. Modern models [1, 17, 46] focus on connecting multiple modalities in a single model. In contrast, we would like to focus on text with image data. One of the strongest multi-modal model was CLIP [43], which uses image embeddings to align with text embeddings for image-text matching loss [38], so images will correspond to relevant phrases. The same idea was used in BLIP [29] with matching image and text embeddings [31]. This is a key idea to filter pairs of text and image.

Multi-modal Data. Current multi-modal data contains images with captions [5, 18, 21, 34, 40, 52]. BLIP, BLIP2 and Flamingo were trained on these. However, these datasets do not contain dialogue contexts.

Previous research has also encountered a similar challenge, as evidenced by a study on the topic of image-grounded dialogues [69]. The authors of this study utilized dialogues from Chinese social media and crowd-sourcing to develop their model. The authors reported an increase in BLEU [42] scores compared to generated responses that were not conditioned on image data.

Another study [23] proposes constructing a dataset utilizing image-text matching via the Visual Semantic Reasoning Network (VSRN) [32] with images sourced from the MS COCO [34] and Flicker 30k [66] datasets, just as it was done before [24].

§3. Dialogue Datasets

To produce clean data, it is important to have dialogues created with humans. They could be collected either from crowd-sourcing or written by humans, such as in English books. Therefore, we have chosen the dataset

[1]`https://github.com/VityaVitalich/IMAD`

sources listed below to make our data valid and diverse in terms of dialogue content: DailyDialog, Persona Chat, MuTual, DREAM, Common-sense dialogues, and Empathetic Dialogues [9, 33, 45, 55, 67, 70]. Detailed reasons are provided in Appendix B.

With these sources, we have collected 451,611 pairs of utterances with context and images. This dataset contains all the features and model predictions described below.

The filtered version of the dataset is named IMage-Augmented Multi-modal Dialogue Dataset (IMAD) and is used later for modeling. It is constructed with a combination of the Text-Image Replacing approach and Human Annotated part. Basic statistics are shown in Table 3.

§4. Text-Image Replacing

Previously, the filtering approach with a filter from VSRN did not show good results [23]. Therefore, to obtain a more precise and appropriate dataset we utilized a two-step process to determine the feasibility of replacing an utterance with an image. The first step involved predicting whether the utterance could be substituted with an image. The second step focused on matching a better picture utilizing VQA from BLIP.

4.1. Find Replaceable Utterances. The first step is to find utterances that could be replaced with an image. To accomplish this task, we have labeled a small subset, created features (we will name them scores), and built a random forest model [41].

4.1.1. *Human Annotation.* For the initial step, we labeled 1000 random samples from the DailyDialogue [33], denoting them as $U = \{u_1, \ldots, u_{1000}\}$, that are utterances with contexts. To label them, we were using a heuristic formulated as "*This phrase potentially can be described with a picture*".

4.1.2. *Replacing Features.* The matching process involves pairing each utterance with an image from *Unsplash*. This is achieved by maximizing the cosine similarity between the embeddings of the utterance and the image extracted from CLIP [43].

To optimize the matching process, we conducted experiments using various text and image features that were deemed important for predicting if the utterance is replaceable. The most promising results were obtained when each utterance was accompanied by the following features: Image

Score, Maximum Entity Score, Sentence Similarity, BLEU Score, Threshold. A detailed description is provided in Appendix A.

4.1.3. *ML Labeling.* For classification we employed the random forest algorithm that demonstrated the best precision [41]. This behavior is likely attributed to the high variance in the data, as evidenced by the standard deviation.

Multiple tests were conducted with stratified K-fold cross-validation with 3 folds and 40 repeats. Precision was deemed to be the key metric, given the importance of minimizing errors, as even the exclusion of valid utterances is preferable to making errors.

The resulting model metrics are shown in Table 6 and feature importances are shown in Fig. 2.

4.2. Text-Image Matching. The criterion from the first step, which considers only the text and not the image, serves as a primary filter for the text-image task. However, it is insufficient on its own, as evidenced by a relatively low correlation in previous works [23]. The image dataset limitations impose constraints on the ability to form a pair of utterance and image, even if the utterance is replaceable. To overcome this, it is necessary to introduce a step to match utterances with better images.

4.2.1. *ML Labeling.* In order to improve the quality of the images, we utilized the BLIP VQA [30]. To select proper images we use confidence of the model output, which is defined as sum of the log probabilities of each token in the utterance. Confidence$(out, u) = \sum_{i=1}^{length\, token(u)} out_{i,token(u)_i}$.

The process of selecting a better image for an utterance was carried out using the following steps.

(1) **Create scoring for all images**. For each utterance we have cosine similarity with every image in dataset

$$\Big\{\text{cosine}(\text{emb}_{CLIP}(u), \text{emb}_{CLIP}(i)) \quad \forall i\Big\} =: ISS_u.$$

(2) **Create set of N images**. From all that set we take set of N images, which are the top-N for cosine similarity

$$\Big\{i \mid \text{cosine}(\text{emb}_{CLIP}(u), \text{emb}_{CLIP}(i)) \geq ISS_{u_{(N)}}\Big\} =: TopImg_{u,N}.$$

(3) **Query the VQA**. The model was queried with the text input "Which phrase can describe this image?" for each image in the aforementioned set.

(4) **Calculate confidences**. For each image in aforementioned set we calculate confidence $\Big\{\text{Confidence}(\text{out}, \text{u}) \;\Big|\; out = \text{VQA}(quest, i)$ $\forall i \in TopImg_{u,N}\Big\} =: ConfSet_{u,N}$.

(5) **Select the most confident**. Then we select image with the highest confidence score $\text{argmax}\{ConfSet_{u,N}\} =: img_N$.

We conducted a test of our methodology by labeling image-text matching in the context of dialogue pairs that were previously identified as replaceable. The labeling process involved 3 classes: "*Image matches*", "*Image does not match*" and "*Unknown*" in cases where determination was difficult. Our results, as presented in Table 2, confirm that our initial assumptions were correct. Specifically, our findings indicate that, at best, only half of the pairs with replaceable utterances were found to have matching images.

4.2.2. *Human Annotation.* With the above method we obtained a subset of 4154 samples. In order to enlarge the dataset, samples with lower RF scores were selected and labeled by three expert assessors, resulting in the addition of 4644 more samples to the dataset. The labeled dataset was organized into 4 distinct categories: "*Perfect Match*", "*Partial Match*", "*Undefined*" and "*No Match*". A detailed description and labeling instructions are provided in Appendix D.

The inter-rater reliability of their annotations were evaluated using the Fleiss kappa statistic. The refined version of the IMage Augmented multi-modal Dialogue dataset (IMAD) is constructed from samples obtained with our initial approach and samples from dataset labeled with assessors, that had the label "Perfect Match" or "Partial Match". Basic statistics are shown in Table 3, numbers of samples from different sources and Fleiss kappa are shown in Table 1.

§5. Multi-modal Dialogue Language Model

To validate our approach we train a model on the proposed dataset using both image and text data and compare it to text-only models. We choose the task of reconstructing the substituted utterance since the visual signal is clearly beneficial in this setting.

Table 1. Accuracy for each data source for each label in assessor's validation dataset, where the ground truth labeling was done with authors. Fleiss kappa across assessors for each data source. Number of samples in the resulting dataset from each data source.

Dataset	1 class	2 class	3 class	4 class	Fleiss Kappa	# of samples
PersonaChat	0.29	0.43	0.82	1.0	0.83	2483
DailyDialog	-	0.50	-	-	0.80	899
EmpatheticDialogues	-	-	-	-	0.76	754
Commonsense-Dialogues	-	-	1.0	-	0.83	220
MuTual	-	-	-	-	0.88	333
DREAM	-	-	-	-	0.81	175
Mean across all sources	0.29	0.46	0.9	1.0	0.82	810.67

Table 2. Number of image-text matches for different N in VQA image-text matching approach.

N	Image Matches	No Match	Unknown
1	37	51	8
5	46	46	4
10	48	44	4
15	47	45	4
50	42	51	3

Table 3. Basic statistics of the dataset.

Total Dialogues	4864
Average Speaker Turns Per Context	5.1
Average Number of Tokens Per Context	56.4
Average Number of Tokens Per Replaced Utterance	14.5
Size of Context Vocabulary	12375
Size of Replaced Utterances Vocabulary	7962

We choose a pre-trained BLIP [29] model for experiments as it is one of the best open source models utilising both visual and text modalities and has a convenient interface in the LAVIS library [27]. The model consists of a visual transformer [12] for image encoding and a BERT [11] with a language modeling head for text decoding, both initialised from a BLIP checkpoint. Training details shown in Appendix C.

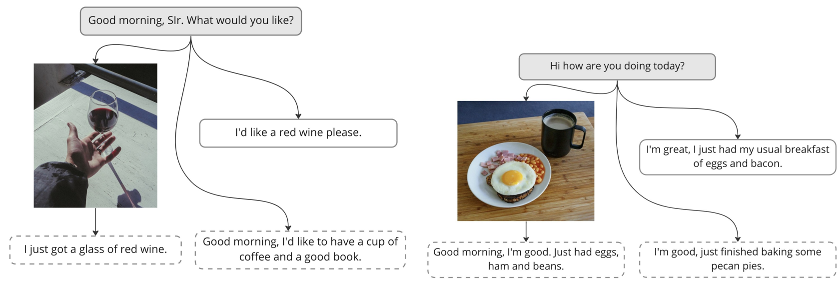

Figure 1. Examples of finetuned models' generation: grey blobs represent the context, white blobs represent ground truth utterances, and dashed blobs represent model generation outputs with or without using visual input.

To validate the resulting dataset we finetuned two BLIP models from a pre-trained checkpoint. These models consist of ViT-B/16 image encoder and BERT text decoder with 12 layers, 12 attention heads, hidden size of 768, intermediate size of 3072, and GeLU activations [15]. The total number of parameters is 224M.

5.1. Evaluation. For evaluation we use 128 samples annotated by both us and the assessors. For each trained model we choose the best checkpoint using validation metrics and use it to compute metrics on the test split. We compare BLIP finetuned on the proposed dataset to zero-shot performance of a distilled BlenderBot 400M [49] and BLIP finetuned in text-only setting. We choose BLEU [42] with n-grams lengths from 1 to 4 as quality metrics and also report perplexity for the models we trained. During evaluation we use beam search sampling with 3 beams. We also divide the test split of the dataset into parts corresponding to the source dialog corpuses and report the metrics for each of them in Table 5 and for the whole set in Table 4.

5.2. Generation examples. Figure 1 show some examples of model generation results. We find that the model finetuned using visual data uses that information in its answers as opposed to the one finetuned using constant visual inputs. As we have only one ground truth label for one sample sometimes model outputs do not exactly align with them, but usually make sense and come close.

Table 4. Test split metrics for both finetuned models and BlenderBot 400M.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Perplexity
image+text	**23.73** $\pm$ 1.25	**14.37** $\pm$ 1.06	**9.19** $\pm$ 0.84	**6.33** $\pm$ 0.77	**44.19** $\pm$ 1.00
text-only	10.63 $\pm$ 0.60	5.76 $\pm$ 0.35	4.01 $\pm$ 0.28	3.23 $\pm$ 0.24	90.21 $\pm$ 1.04
BlenderBot	10.93	4.75	2.62	1.61	-

Table 5. Metrics on test split by source for both finetuned models and BlenderBot 400M. We report metrics for top-3 most frequent sources.

	Persona-Chat		DailyDialog		EmpatheticDialogues	
	BLEU-1	BLEU-4	BLEU-1	BLEU-4	BLEU-1	BLEU-4
image+text	**22.81**	**5.52**	**26.51**	**6.27**	**27.64**	**9.23**
text-only	9.73	2.68	13.81	3.79	8.98	0.00
BlenderBot	13.29	2.83	12.17	0.00	10.24	1.62

§6. Results and Discussion

Dataset. We have presented a dataset (IMAD) with a considerable amount of multi-modal dialogues, sourced from validated text-only datasets. The creation of our dataset is automated with a two-step process, involving the filtering of the most relevant utterances and the selection of the most suitable image. Additionally, we developed a methodology for labeling the data, which proved to be easy to understand and valid. This is demonstrated by the high Fleiss kappa score, which measures the consistency between 3 assessors, as shown in Table 1. The basic statistics of our dataset are presented in Table 3, and statistics per dialogue are quite similar to those of DailyDialogue, which is a well validated dataset.

Baseline Model. In this study, we propose a model for the task of generating an utterance that is replaced with an image, using the IMAD. It is based on the BLIP architecture and achieves a relatively high BLEU score compared to other models that use only text information (as shown in Table 4). This result demonstrates the validity of our dataset, which is consistent across different sources of data (as shown in Table 5). We have overcame the issue of noisy and irrelevant pairs of utterances and images by incorporating a filtering stage and creating the IMAD, resulting in a model that potentially could outperform a previous approach due to cleaner data [69].

Further Work. These promising findings are expected to significantly contribute to the advancement of research in the field of multi-modal dialogue models. The methodology presented in this paper could aid researchers in the creation of more accurate filters for dialogue data, which could lead to improvements in the quality and efficiency of collecting multi-modal dialogues. We were limited with resources and yet tested model accuracy with a lot of repeats on a subset, that could have led to distribution bias across datasets. As well, low recall means we do not include a lot of valid samples, which reduces our total dataset size. Therefore, the approach presented in this work could facilitate the development of more effective multi-modal models. There is also potential to further improve the size of the dataset through labeling or model upgrades, as mentioned above.

Moreover, it is essential to conduct extensive research to compare multi-modal approaches for solving this task. This involves comparing models using cross-attention and concatenation of embeddings to that task, as well as conducting experiments with different language models and visual encoders. Such an investigation can lead to the development of more effective and accurate multi-modal models.

§7. Conclusion

In conclusion, our work presents a new task of interpreting images in the context of dialogue and proposes a novel approach to construct a multi-modal dialogue dataset to tackle this challenge. We utilize a two-stage process that involves identifying utterances that can be replaced with images and selecting relevant images using visual question answering models. Through this process, we have created the IMage Augmented multi-modal Dialogue dataset (IMAD), which is validated and labeled, providing a valuable resource for further research in this area. Additionally, we have proposed a baseline model trained on IMAD, which outperformed existing models that do not incorporate images. Our work demonstrates the potential of incorporating visual information in dialogue systems and highlights the need for more research in this area. Future work can explore the use of more advanced techniques for identifying relevant images and developing more sophisticated models that can effectively incorporate visual information into dialogue systems.

Appendix A. Implementation details

Formally pairs of utterance and images are made in the following way. Given a set of images I, we created pairs

$$\forall u \quad \exists\,(u, img) : img = \mathrm{argmax}_{i \in I}\{\mathrm{cosine}(\mathrm{emb}_{CLIP}(u), \mathrm{emb}_{CLIP}(i))\}$$

, where $\mathrm{emb}_{CLIP}(u)$ is taking embedding from CLIP for utterance and $\mathrm{emb}_{CLIP}(i)$ is taking embedding from CLIP for images and $\mathrm{cosine}(a, b) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$ is calculating cosine similarity between vectors in a usual sense.

Main features are described in depth below:

Image Score. For utterance $\mathrm{IS}(u)$ is the maximum cosine similarity between utterance and all images embeddings extracted from CLIP.

$\max\limits_{i \in I}\{\mathrm{cosine}(\mathrm{emb}_{CLIP}(u), \mathrm{emb}_{CLIP}(i))\}$

Maximum Entity Score. We follow the idea, that most of entities in the NER datasets are nouns [48]. First, each utterance undergoes a noun extraction process and has corresponding noun set $ENT_u = \{noun \,|\, \forall noun \in u\}$. Second, for each entity in the set Image Score is calculated $\mathrm{IS}(entity)$. That forms set of Image Scores of utterance nouns which we call Entity Scores for utterance $ES_u = \{\mathrm{IS}(entity) \,|\, \forall entity \in ENT_u\}$. Finally, we take maximum of Entity Scores $\mathrm{MES}(u) = \max ES_u$.

Sentence Similarity. It is obtained from comparing image caption and initial utterance. First, for each corresponding image captions are generated $\mathrm{caption}(img)$. We do this with VIT-GPT2 model [22]. Then we find similarity between utterance and generated caption with cosine distance between their embeddings from SentenceBert $\mathrm{SS}(u, img) = \mathrm{cosine}(\mathrm{emb}_{SB}(u)$, $\mathrm{emb}_{SB}(\mathrm{caption}(img))$ [47] .

BLEU Score. BLEU [42] metric for only unigrams between generated caption and utterance $\mathrm{BLEU}_1(u, \mathrm{caption}(img))$.

Threshold. Binary feature that shows if utterances features listed above are greater than founded thresholds. We found thresholds for Image Score t_{IS}, Sentence Similarity t_{SS} and Maximum Entity Score t_{MES} by maximizing precision on labeled subset U via grid-search on triplets $(t_{IS}, t_{SS}, t_{MES})$. To reduce the computations, each threshold was chosen as k-th statistic in set of train values sorted in descending order. Therefore we were grid searching through k-th statistic for each parameter with step equals 10. Formally, threshold is $THR_{u,img} = \mathbb{1}\{\mathrm{SS}(u, img) >= t_{SS}\}\mathbb{1}\{\mathrm{MES}(u) >= t_{MES}\}\mathbb{1}\{\mathrm{IS}(u) >= t_{IS}\}$

The hyperparameters were determined via grid search that maximized the precision score. The resulting model consisted of 500 estimators with class weights of 5 to 1 for not-replaceable and replaceable, respectively. The model used the Gini criterion, a maximum depth of 2, and the square root of the number of features in each estimator.

Thresholds features were tested on labeled dataset and brute forced on the 10 step grid. We report on listed below thresholds, that results in precision = 0.921171 and recall = 0.068.

- Image Score = 0.33265801843083337
- Sentence Similarity = 0.12116438820166667
- Maximum Entity Score = 0.3103302687291667

There is a list of pictorialization features that were tested, but resulted in worse metrics.

- Embedding representations from SentenceBERT and subsets of embeddings representation
- Image-text matching loss from BLIP
- Answers from VQA model to the question "does statement *utterance* describe picture well?" transitioned to binary feauture for model output "yes" and "no"
- Answers from VQA model to the question "Can the utterance *utterance* be described by the picture?" transitioned to binary feauture for model output "yes" and "no"
- Cardinality of intersection between nouns from utterance and objects that VQA answers to the question "what objects are in the picture?"
- Confidence of model on "The picture shows *utterance*". It was calculated as sum of logarithmic probabilities of tokenized phrase without taking in account pattern phrase "The picture shows".
- Text feauteres: Number of parts of speech in utterance, punctuation in utterance, utterance length, number of words, SMOG Index, LIX Index, Flesch–Kincaid readability tests, Coleman–Liau index. Lexical diversity metrics: TTR, RTTR, CTTR, HTTR, STTR, MTTR, DTTR, MATTR, MSTTR, MTLD, MAMTLD, HD-D [53]

Appendix B. Dataset Reasoning

DailyDialog [33]. Daily Dialog is a popular source of human-written dialogues, contains 76k utterances with context. Moreover, dialogues stick to the certain topic or object and they end after reasonable turn. These features are essential for building context related models.

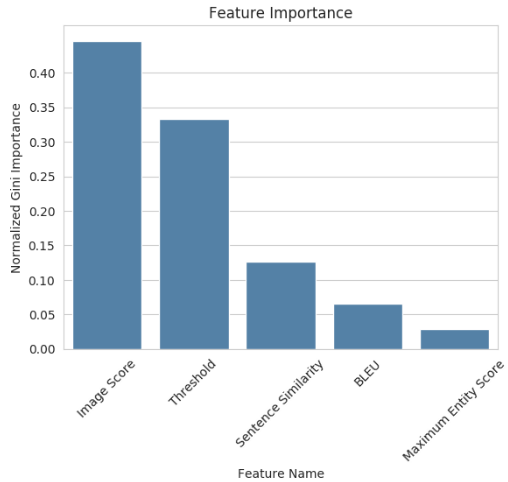

Figure 2. Feature Importance in the Random Forest model for predicting if an utterance could be replaced with an image. Image Score is the maximum cosine similarity between utterance and images embeddings from CLIP. Threshold is binary feature, indicating if Image Score, Sentence Similarity and Maximum Entity Score is bigger than empirically founded values. Sentence Similarity is cosine distance similarity between utterance embedding and caption embedding, that were generated from corresponding image with VIT-GPT2, where embeddings come from SentenceBert. BLEU is calculated between utterance and generated caption with taking only unigrams into account. Maximum Entity Score is maximum out of Image Scores for each noun in utterance.

Mutual dataset [9]. This dataset was crawled from English students books, writtend with expert linguists and contains dialogues with reasonings. Contains 18k utterance with context.

Common Sense Dialogues [70]. This dataset focuses on real-life common sense dialogues. It contains 43k utterances with context from human-written dialogues, collected from assessors on Amazon Mechanical Turk (MTurk).

Table 6. Metrics for best (in terms of Precision) Random Forest model for predicting if utterance is replaceable.

Metric	Mean	Standart Deviation	Median
Precision	0.975000	0.156125	1.000000
Recall	0.030990	0.006374	0.031250
F1	0.060042	0.012072	0.060606

Table 7. Comparison for different ML Algorithms in terms of Precision and Recall.

Algorithm	Precision	Recall
Random Forest	0.975000	0.030990
Kernel SVM	0.958333	0.037500
Gradient Boosting	0.816667	0.026302
KNN	0.500188	0.035677

Empathetic Dialogues [45]. This dataset was also collected via MTurk. The main goal is to focus on emotional and personal dialogues. It contains 76k utterances with context.

Dream Dataset [55]. This dataset contains 14k utterances with context from dialogues from English students books. The goal is to make wide range of dialogues, that would include both common knowledge and reasoning.

Persona Chat [67]. We have utilized Persona Chat because it was collected using assessors and contains 223k utterances with context from small real life dialogues. It was well validated and contains dialogues with personas.

Appendix C. Training Details

To evaluate the contribution of visual information into solving the task we finetune two models in different settings: using both image and context as input and replacing image feature vector with constant input. We expect the model which has access to visual modality to perform better, which would mean that images in the composed dataset are actually relevant and useful for utterance prediction.

To train our models we utilise a prefix LM learning paradigm as opposed to usual next token prediction used to finetune BLIP on captioning tasks.

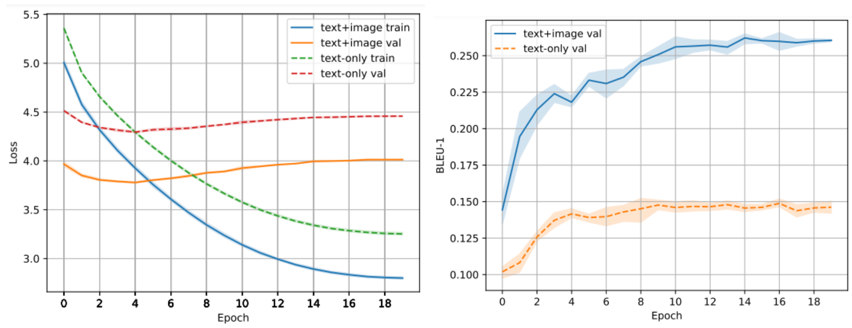

Figure 3. Training metrics: left figure shows train and validation loss for both finetuned BLIP models during training, right figure shows BLEU-1 metrics on validation during training.

We use the image and the last utterance before the image as inputs for the model and aim to predict the substituted utterance. We let the decoder use a square attention mask for the context tokens and a triangular attention mask for the utterance tokens we want to predict.

For both models we freeze the image encoder part during training for efficiency and to reduce compute cost.

Both models are trained on 4336 samples from the composed dataset, leaving 400 samples for validation, for 20 epochs with batch size of 10 samples, learning rate of 1e-5, and cosine learning rate scheduler with linear warmup (Figure 3). For each setting we train 5 models initializing from different seeds for better quality estimation.

Appendix D. Labeling methodology

The detailed descriptions for classes are as follows.

(1) **Perfect Match**. This label is assigned when utterance has only 1 sense, and it is fully transferred with an image. If image could not transfer fully the sense of utterance and it could be done only with context knowledge, then also this label is assigned. There should be no factual mistakes, image should not be specific to cultural differences. The heuristic rule was to question yourself "Could i possibly came to this phrase, knowing image and context?".

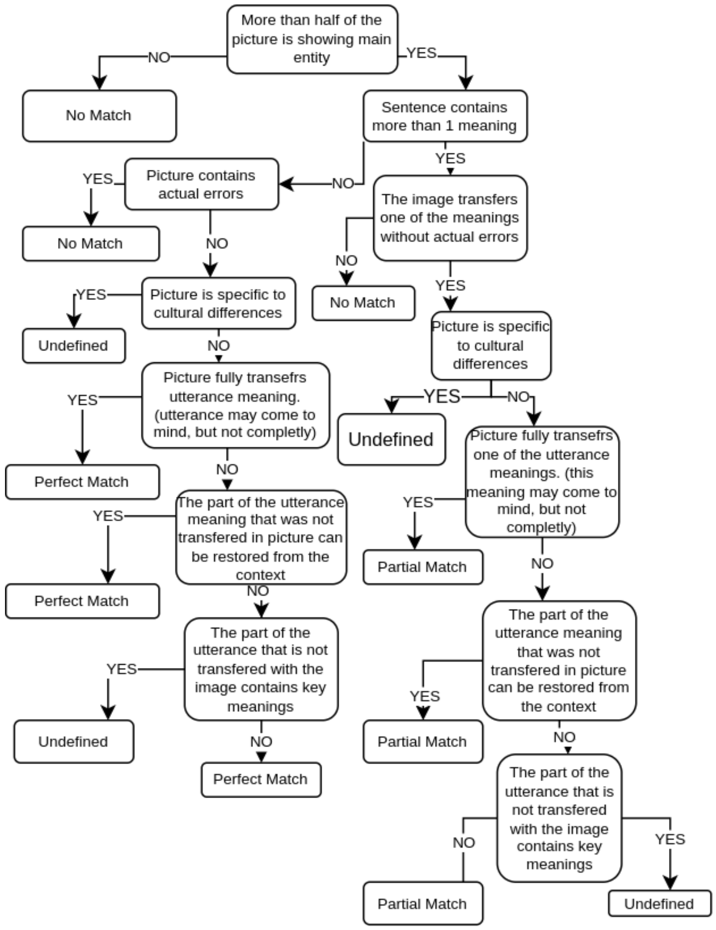

Figure 4. Labeling Methodology.

(2) **Partial Match**. This label is assigned when utterance has 2 or more distinct senses and image transfers one of them fully. It also should not be specific to culture or contain mistakes. In fact rules are the same as for Perfect Match, but applied to one of the senses in the utterances contains 2 or more senses.

(3) **Undefined**. This label is assigned when image is specific to cultural differences or when image can not transfer one of the senses of the utterance and the context could not help to recover untransferred sense.
(4) **No Match**. This label is assigned when image contains factual errors about utterance or when none of the entities from utterance present in the image.

A decision tree was designed as the primary instruction for the labeling process, which aimed to assist the assessors in assigning appropriate labels to the samples. The decision tree was presented in Figure 4 and consisted of closed questions at each node, or terminal nodes containing the desired label. The assessors were instructed to follow the decision tree, starting from the root node and answering the questions until the label for each sample was reached. This strategy yielded high inter-rater reliability among the 3 assessors, indicating its efficacy in achieving consistent labeling outcomes.

References

1. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, *Flamingo: a visual language model for few-shot learning*, 2022.
2. N. Anantrasirichai, D. Bull, *Artificial intelligence in the creative industries: a review*, Artificial Intelligence Review **55** (2021), no. 1, 589–656.
3. T. Baltrušaitis, C. Ahuja, L.-P. Morency, *Multimodal machine learning: A survey and taxonomy*, 2017.
4. M. S. Burtsev, A. V. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, et al., *Deeppavlov: Open-source library for dialogue systems.*, ACL (4), 2018, pp. 122–127.
5. S. Changpinyo, P. Sharma, N. Ding, R. Soricut, *Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts*, 2021.
6. H. Chen, X. Liu, D. Yin, J. Tang, *A survey on dialogue systems.* — ACM SIGKDD Explorations Newsletter **19**, No. 2 (2017), 25–35.
7. J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, X. Zhang, Z. Huang, M. Liu, T. Yang, *HEU emotion: a large-scale database for multimodal emotion recognition in the wild.* — Neural Computing and Applications **33**, No. 14 (2021), 8669–8685.
8. Y. Chen, Xu-Hua Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, Q. Guan, *Generative adversarial networks in medical image augmentation: a review.* — Computers in Biology and Medicine (2022), 105382.
9. L. Cui, Y. Wu, S. Liu, Y. Zhang, M. Zhou, *MuTual: A dataset for multi-turn dialogue reasoning.* — Proceedings of the 58th Annual Meeting of the Association

for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 1406–1416.
10. D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, P. Ebel, *The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems*, 2021.
11. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018.
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2020.
13. J. Gao, P. Li, Z. Chen, J. Zhang, *A Survey on Deep Learning for Multimodal Data Fusion.* — Neural Computation **32**, No. 5 (2020), 829–864.
14. X. Gu, K. M. Yoo, J.-W. Ha, *Dialogbert: Discourse-aware response generation via learning to recover and rank utterances*, 2020.
15. D. Hendrycks, K. Gimpel, *Gaussian error linear units (gelus)*, 2016.
16. X. Hu, Zhe Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, L. Wang, *Scaling up vision-language pre-training for image captioning*, 2021.
17. S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, F. Wei, *Language is not all you need: Aligning perception with language models*, 2023.
18. C. Jia, Y. Yang, Ye Xia, Yi-Ting Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, T. Duerig, *Scaling up visual and vision-language representation learning with noisy text supervision*, (2021).
19. G. Joshi, R. Walambe, K. Kotecha, *A review on explainability in multimodal deep neural nets.* — IEEE Access **9** (2021), 59800–59821.
20. A. B. Kocaballi, *Conversational ai-powered design: Chatgpt as designer, user, and product*, 2023.
21. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, Li-Jia Li, D. A. Shamma, M. S. Bernstein, F.-F. Li, *Visual genome: Connecting language and vision using crowdsourced dense image annotations*, 2016.
22. A. Kumar, *The illustrated image captioning using transformers*, ankur3107.github.io (2022).
23. N. Lee, S. Shin, J. Choo, H.-J. Choi, S.-H. Myaeng, *Constructing multi-modal dialogue dataset by replacing text with semantically relevant images*, 2021.
24. Y.-J. Lee, B. Ko, H.-G. Kim, H.-J. Choi, *Dialogcc: Large-scale multi-modal dialogue dataset*, 2022.
25. J. Lei, T. L. Berg, M. Bansal, *Revealing single frame bias for video-and-language learning*, 2022.
26. C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, L. Si, *mplug: Effective and efficient vision-language learning by cross-modal skip-connections*, 2022.
27. D. Li, J. Li, H. Le, G. Wang, S. Savarese, S. C. H. Hoi, *Lavis: A library for language-vision intelligence*, 2022.
28. J. Li, D. Li, S. Savarese, S. Hoi, *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*, 2023.

29. J. Li, D. Li, C. Xiong, S. Hoi, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 2022.
30. J. Li, D. Li, C. Xiong, S. Hoi, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, ICML, 2022.
31. J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, S. Hoi, *Align before fuse: Vision and language representation learning with momentum distillation*, 2021.
32. K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, *Visual semantic reasoning for image-text matching*, ICCV, 2019.
33. Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, *Dailydialog: A manually labelled multi-turn dialogue dataset*, (2017).
34. T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, *Microsoft coco: Common objects in context*, 2014.
35. B. D. Lund, T. Wang, *Chatting about chatgpt: how may ai and gpt impact academia and libraries?*, Library Hi Tech News (2023).
36. O. M. Manyar, J. Cheng, R. Levine, V. Krishnan, J. Barbič, S. K. Gupta, *Physics informed synthetic image generation for deep learning-based detection of wrinkles and folds.* — J. Computing and Information Science in Engineering **23**, No. 3 (2023), 030903.
37. Q. Motger, X. Franch, J. Marco, *Software-based dialogue systems: Survey, taxonomy, and challenges.* — ACM Comput. Surv. **55**, No. 5 (2022).
38. A. van den Oord, Y. Li, O. Vinyals, *Representation learning with contrastive predictive coding*, 2018.
39. OpenAI, *ChatGPT: Optimizing language models for dialogue.*, 2022.
40. V. Ordonez, G. Kulkarni, T. Berg, *Im2text: Describing images using 1 million captioned photographs*, Advances in Neural Information Processing Systems (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, eds.), vol. 24, Curran Associates, Inc., 2011.
41. 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.
42. M. Post, *A call for clarity in reporting BLEU scores*, Proceedings of the Third Conference on Machine Translation: Research Papers (Belgium, Brussels), Association for Computational Linguistics, October 2018, pp. 186–191.
43. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, *Learning transferable visual models from natural language supervision*, 2021.
44. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, *Zero-shot text-to-image generation*, 2021.
45. H. Rashkin, E. M. Smith, M. Li, Y.-Lan Boureau, *Towards empathetic open-domain conversation models: a new benchmark and dataset*, 2018.
46. S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barthmaron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, N. de Freitas,

A generalist agent, Transactions on Machine Learning Research (2022), Featured Certification.

47. N. Reimers, I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*. — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 11 2019.
48. A. J. C. Rodriguez, D. C. Castro, S. H. Garcia, *Noun-based attention mechanism for fine-grained named entity recognition*. — Expert Systems with Applications **193** (2022), 116406.
49. S. Roller, E. Dinan, N. Goyal, Da Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y-Lan Boureau, J. Weston, *Recipes for building an open-domain chatbot*, 2020.
50. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2021.
51. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G.Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, *Photorealistic text-to-image diffusion models with deep language understanding*, 2022.
52. C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, *Laion-400m: Open dataset of clip-filtered 400 million image-text pairs*, 2021.
53. S. Shkarin, *Ruts, a library for statistics extraction from texts in russian*, Moscow, 2023.
54. J. Summaira, Xi Li, A. M. Shoib, S. Li, J. Abdul, *Recent advances and trends in multimodal deep learning: A review*, 2021.
55. K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, C. Cardie, *Dream: A challenge dataset and models for dialogue-based reading comprehension*, 2019.
56. Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu, *Naturalspeech: End-to-end text to speech synthesis with human-level quality*, 2022.
57. R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Yu Du, YaGuang Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, Ch.-Ch. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, Q. Le, *Lamda: Language models for dialog applications*, 2022.
58. R. Valle, K. Shih, R. Prenger, B. Catanzaro, *Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis*, 2020.
59. P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, *Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework*, 2022.

60. Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, Y. Cao, *Simvlm: Simple visual language model pretraining with weak supervision*, 2021.
61. BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, and et. Al Yvon, *Bloom: A 176b-parameter open-access multilingual language model*, 2022.
62. Y. Xi, Y. Zhang, S. Ding, S. Wan, *Visual question answering model based on visual relationship detection*, Signal Processing: Image Communication **80** (2020), 115648.
63. X. Xu, C. Wu, S. Rosenman, V. Lal, W. Che, N. Duan, *Bridgetower: Building bridges between encoders in vision-language representation learning*, 2022.
64. S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, J. Yu, *Videococa: Video-text modeling with zero-shot transfer from contrastive captioners*, 2022.
65. A. Yang, A. Miech, J. Sivic, I. Laptev, C. Schmid, *Zero-shot video question answering via frozen bidirectional language models*, 2022.
66. P. Young, A. Lai, M. Hodosh, J. Hockenmaier, *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*. — Transactions of the Association for Computational Linguistics **2** (2014), 67–78.
67. S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, *Personalizing dialogue agents: I have a dog, do you have pets too?*, (2018).
68. Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, JJ (Jingjing) Liu, B. Dolan, *Dialogpt: Large-scale generative pre-training for conversational response generation*, arXiv:1911.00536, November 2019.
69. Y. Zheng, G. Chen, X. Liu, J. Sun, *Mmchat: Multi-modal chat dataset on social media*, 2021.
70. P. Zhou, K. Gopalakrishnan, B. Hedayatnia, S. Kim, J. Pujara, X. Ren, Y. Liu, D. Hakkani-Tur, *Commonsense-focused dialogues for response generation: An empirical study*, 2021.

Поступило 6 сентября 2023 г.

DeepPavlov.ai
E-mail: vvmoskvoretskiy@yandex.ru

DeepPavlov.ai
E-mail: antonylovanto@gmail.com

DeepPavlov.ai
E-mail: kuznetsov.den.p@gmail.com