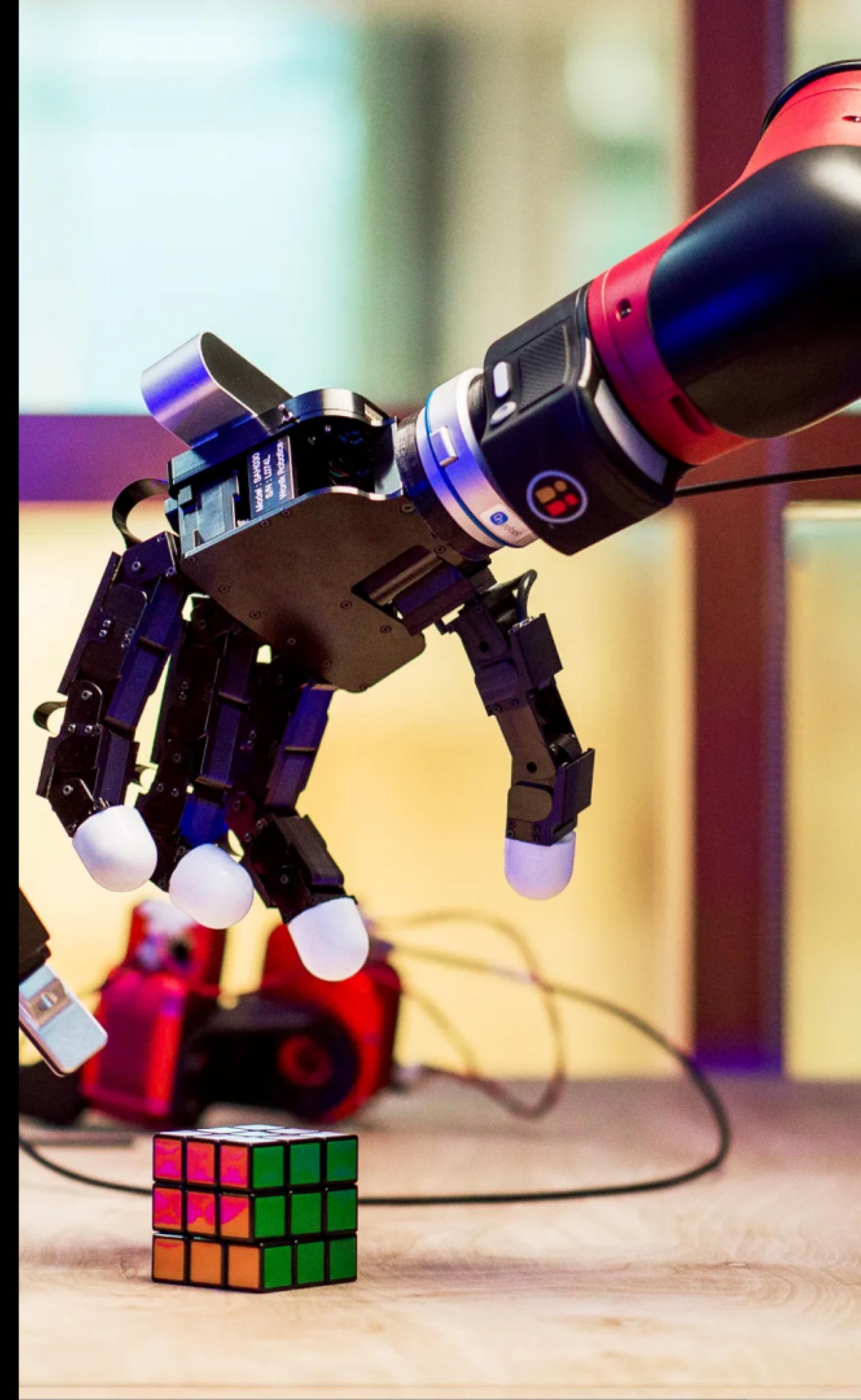


# BINARY VISUAL QUESTION ANSWERING

Ivanov Arseny, 30 nov 2023

#Machine Learning, #Robotics, #VQA, #NLP, #Transformers, #CV

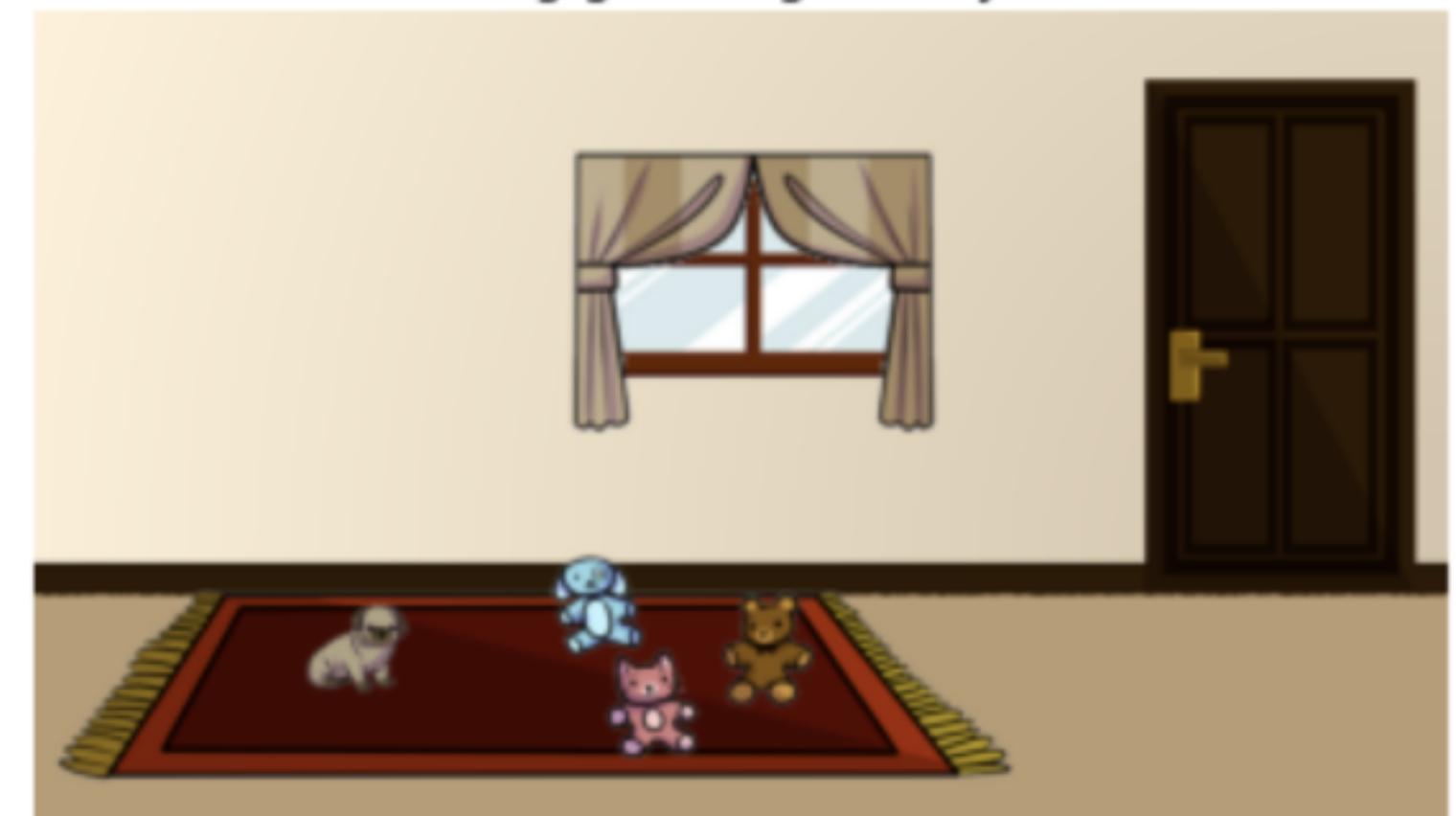
 SBER  
Robotics Lab



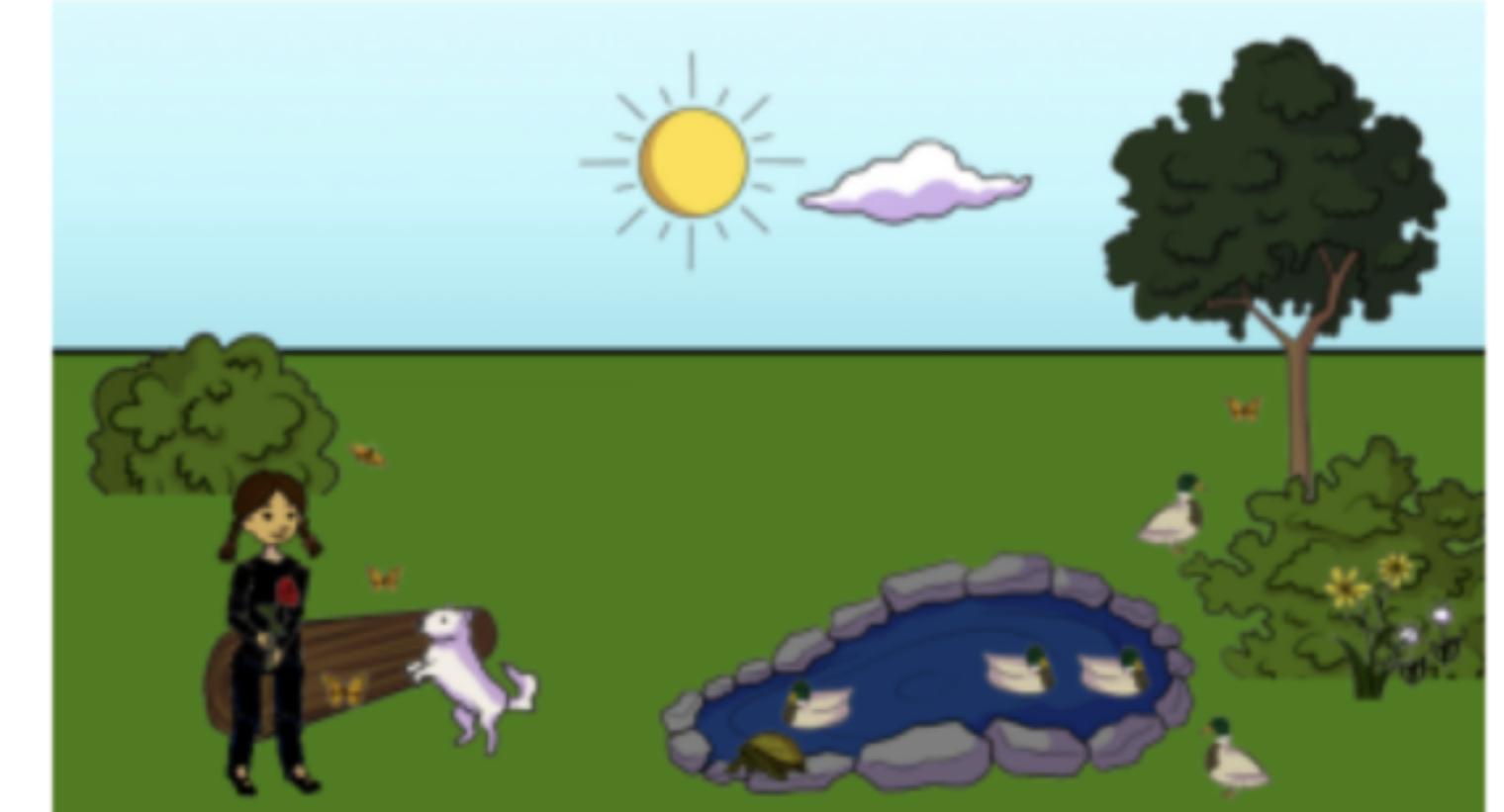
# PROBLEM STATEMENT

- Given an image and a natural language question related to the image, the objective is to produce a natural language binary answer correctly
- Need NLP for two reasons: to understand the question and to generate the answer

Is the dog guarding the toys? - no



Is there an odd number of ducks around the pond? - yes



# RELATED WORK

- Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge
- A Multi -World Approach to Question Answering about Real -World Scenes based on Uncertain Input
- CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning
- VQA: Visual Question Answering
- Stacked Attention Networks for Image Question Answering
- MUTAN: Multimodal Tucker Fusion for Visual Question Answering

**MUTAN: Multimodal Tucker Fusion for Visual Question Answering**

Hedi Ben-younes<sup>1,2\*</sup> Rémi Cadene<sup>1\*</sup> Matthieu Cord<sup>1</sup> Nicolas Thome<sup>3</sup>

<sup>1</sup> Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu, 75005 Paris  
<sup>2</sup> Heuritech, 248 rue du Faubourg Saint-Antoine, 75012 Paris  
<sup>3</sup> Conservatoire National des Arts et Métiers

hedi.ben-younes@lip6.fr, remi.cadene@lip6.fr, matthieu.cord@lip6.fr, nicolas.thome@cnam.fr

**Abstract**

Bilinear models provide an appealing framework for mixing and merging information in Visual Question Answering (VQA) tasks. They help to learn high level associations between question meaning and visual concepts in the image, but they suffer from huge dimensionality issues.

We introduce MUTAN, a multimodal tensor-based Tucker decomposition to efficiently parametrize bilinear interactions between visual and textual representations. Additionally to the Tucker framework, we design a low-rank matrix-based decomposition to explicitly constrain the interaction rank. With MUTAN, we control the complexity of the merging scheme while keeping nice interpretable fusion relations. We show how our MUTAN model generalizes some of the latest VQA architectures, providing state-of-the-art results.

**1. Introduction**

Multimodal representation learning for text and image has been extensively studied in recent years. Currently, the most popular task is certainly Visual Question Answering (VQA) [19, 2]. VQA is a complex multimodal task which aims at answering a question about an image. A specific benchmark has been first proposed [19], and large scale datasets have been recently collected [21, 2, 31] enabling

**Figure 1:** The proposed MUTAN model uses a Tucker decomposition of the image/question correlation tensor, which enables modeling rich and accurate multi-modal interactions. For the same input image, we show the result of the MUTAN fusion process when integrated into an attention mechanism [28]: we can see that the regions with larger attention scores (in red) indicate a very fine understanding of the image and question contents, enabling MUTAN to properly answer the question (see detailed maps in experiments section).

Bilinear models are powerful approaches for the fusion problem in VQA because they encode full second-order interactions. They currently hold state-of-the-art performance [28].

**VQA: Visual Question Answering**

[www.visualqa.org](http://www.visualqa.org)

Aishwarya Agrawal\*, Jiasen Lu\*, Stanislaw Antol\*, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

**Abstract**—We propose the task of *free-form and open-ended* Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ~0.25M images, ~0.78M questions, and ~10M answers ([www.visualqa.org](http://www.visualqa.org)), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance. Our VQA demo is available on CloudCV (<http://cloudcv.org/vqa>).

**1 INTRODUCTION**

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. In particular, research in image and video captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year [16], [9], [12], [38], [26], [24], [53]. Part of this excitement stems from a belief that multi-discipline tasks like image captioning are a step towards solving AI. However, the current state of the art demonstrates that a coarse scene-level understanding of an image paired with word  $n$ -gram statistics suffices to generate reasonable image captions, which suggests image captioning may not be as “AI-complete” as desired.

What makes for a compelling “AI-complete” task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require *multi-modal knowledge* beyond a simple image and question pair, (ii) be *open-ended*, and (iii) be *compositional*.

**Figure 1:** Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense reasoning is required to answer these questions correctly.

**A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input**

Mateusz Malinowski Mario Fritz

Max Planck Institute for Informatics  
Saarbrücken, Germany  
[{mmalinow,mfritz}@mpi-inf.mpg.de](mailto:{mmalinow,mfritz}@mpi-inf.mpg.de)

**Abstract**

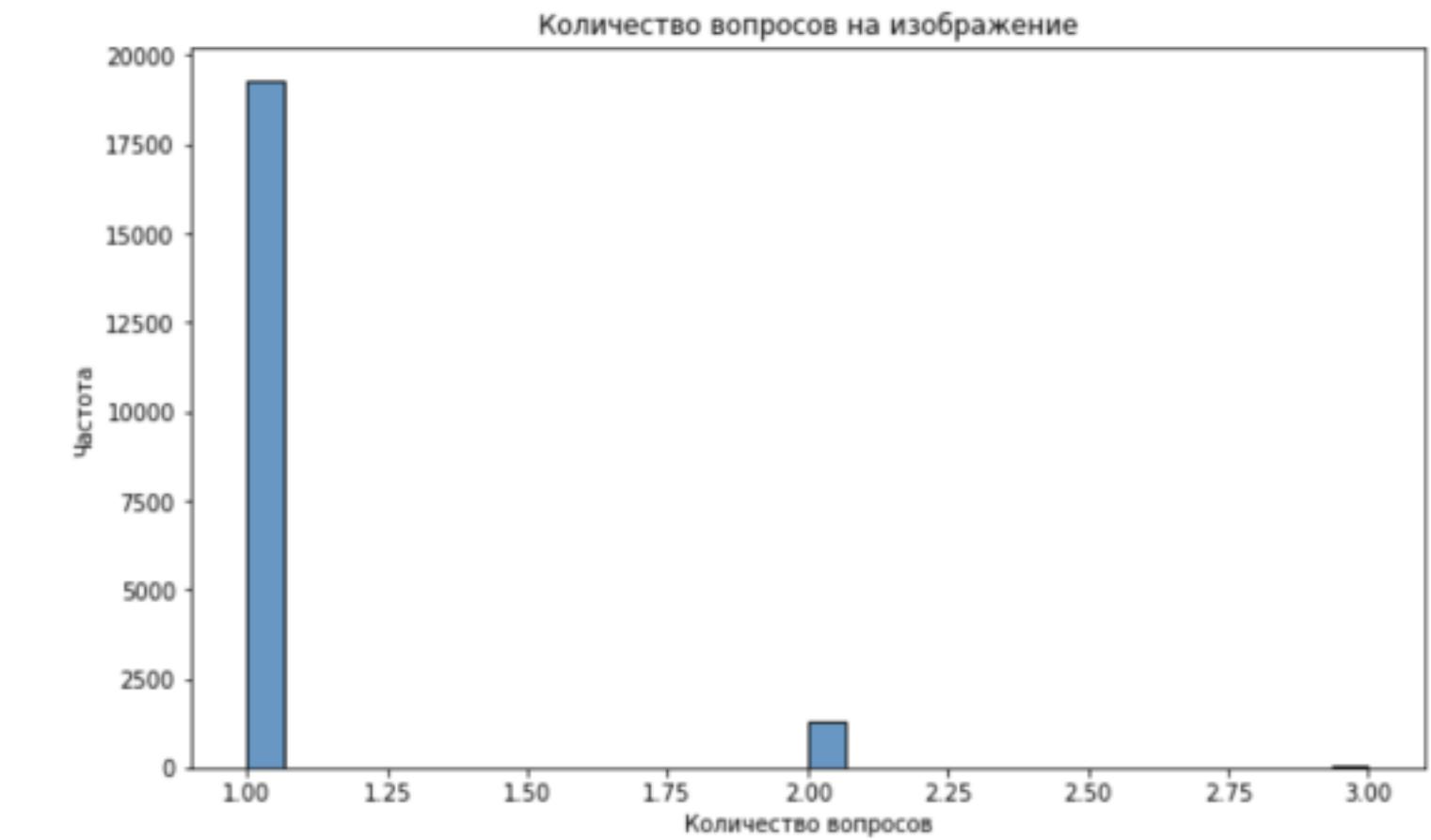
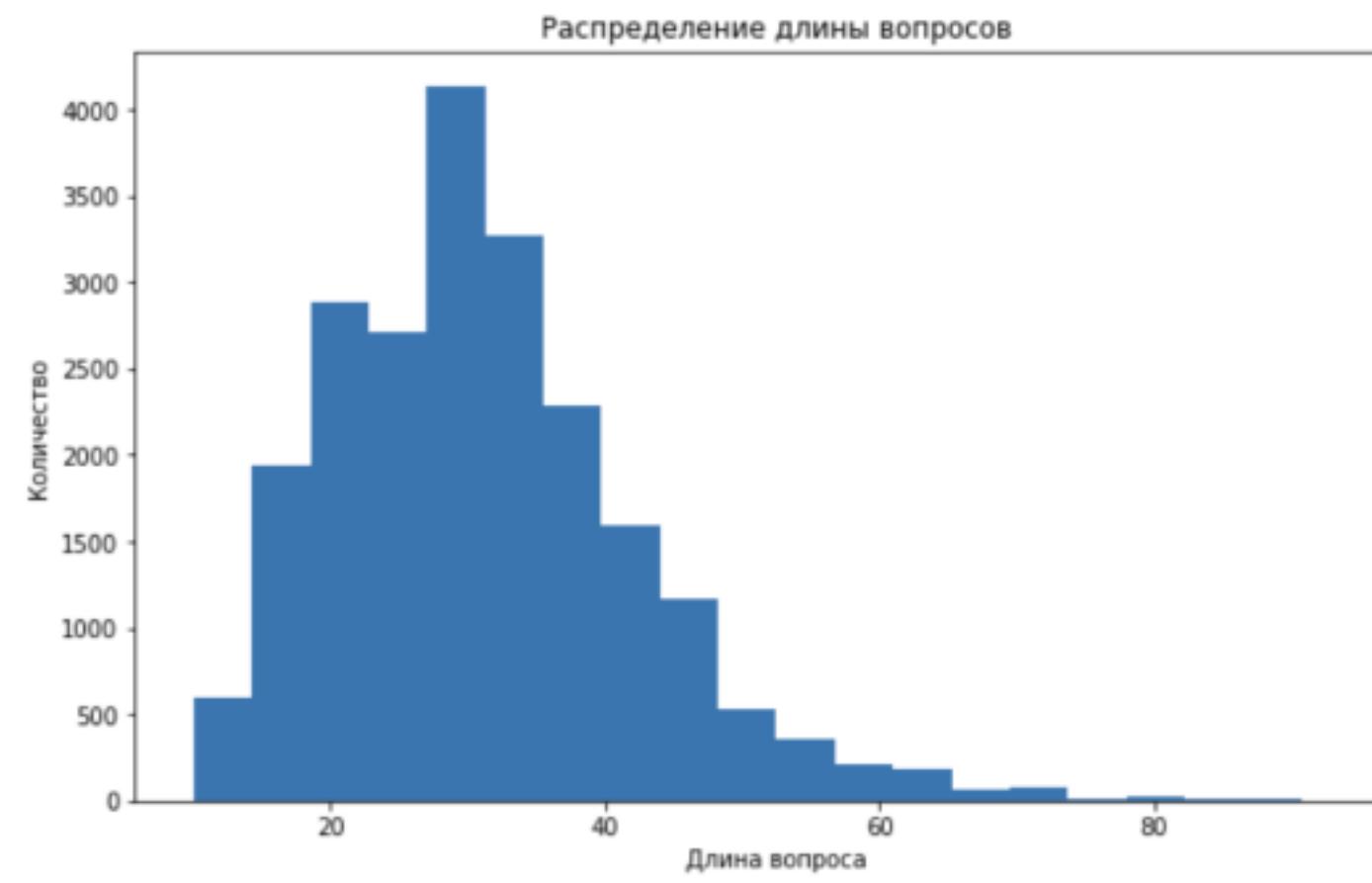
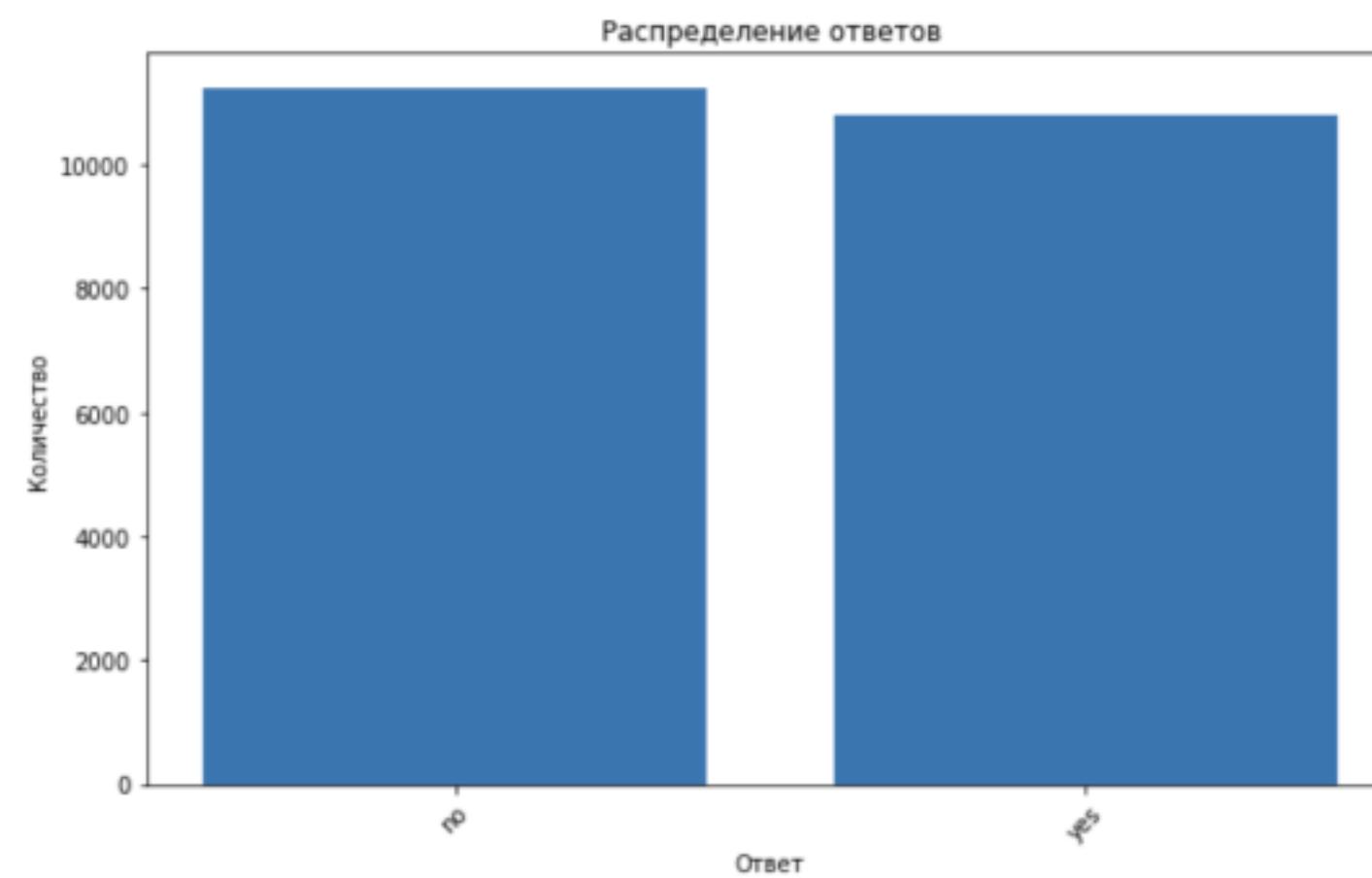
We propose a method for automatically answering questions about images by bringing together recent advances from natural language processing and computer vision. We combine discrete reasoning with uncertain predictions by a multi-world approach that represents uncertainty about the perceived world in a bayesian framework. Our approach can handle human questions of high complexity about realistic scenes and replies with range of answer like counts, object classes, instances and lists of items. The system is directly trained from question-answer pairs. We establish a first benchmark for this task that can be seen as a modern attempt at a visual turing test.

**1 Introduction**

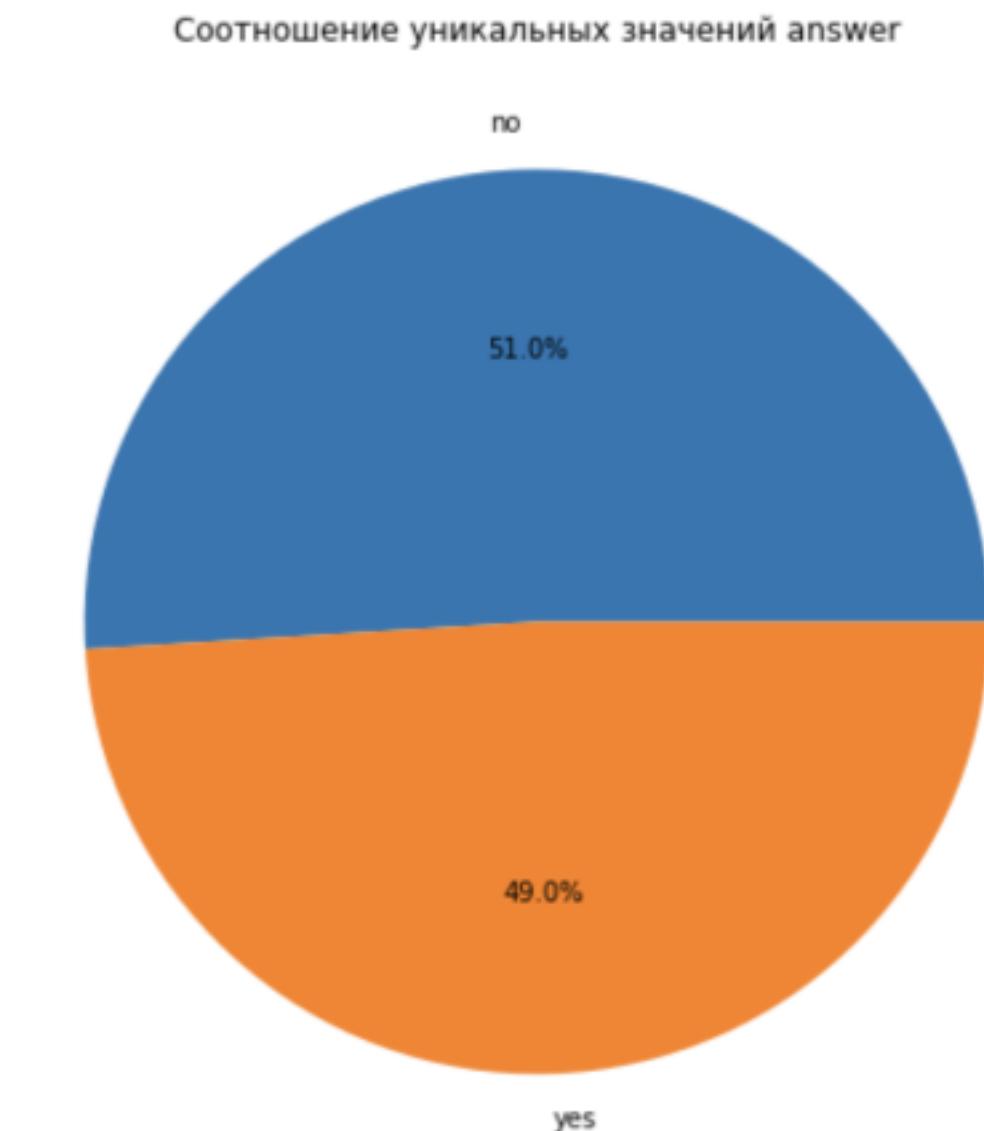
As vision techniques like segmentation and object recognition begin to mature, there has been an increasing interest in broadening the scope of research to full scene understanding. But what is meant by “understanding” of a scene and how do we measure the degree of “understanding”? Most often “understanding” refers to a correct labeling of pixels, regions or bounding boxes in terms of semantic annotations. All predictions made by such methods inevitably come with uncertainties attached due to limitations in features or data or even inherent ambiguity of the visual input.

Finally strong progress has been made on the language side, where methods have been proposed

# DATASET



- Dataset for Visual Question Answering on abstract images (Abstract VQA 2017 V002).
- It contains 22055 question-answer pairs, based on images from the Abstract Scenes. That means about 1-2 pairs per image on average.
- The answers are a binary “yes” or “no”. Data sample is almost balanced: label of answer “no” - 51%, “yes” - 49%.
- Besides we can see how many characters there are in the question, that is, how detailed the question is, the sample has a normal distribution and basically it is about 30 characters per question.



# PRE-PROCESSING

Initially, we should do preprocessing with data:

- All text questions and answers are given in lowercase and for text we use lemmatization and stemming operations.
- Moreover, we have multi-label classification problem (as multiple answers are possibly valid). Moreover, rather than just creating a one-hot encoded vector, one creates a soft encoding, based on the number of times a certain answer appeared in the annotations.
- For instance, in the example, the answer "YES" seems to be selected way more often than "MAYBE". Hence, we want the model to give more emphasis on "yes" then on "maybe". We can achieve this by giving a score of 1.0 to labels which are counted at least 3 times, and a score < 1.0 for labels that are counted less.
- Bifurcated data into train, test and validation: 80:10:10, that used in each experiment with models so that our metrics are comparable.

```
[340]: annotations[6]
```

```
[340]: {'question_type': 'is there',
'multiple_choice_answer': 'yes',
'answers': [{'answer': 'no', 'answer_confidence': 'yes', 'answer_id': 1},
{'answer': 'yes', 'answer_confidence': 'maybe', 'answer_id': 2},
{'answer': 'yeah', 'answer_confidence': 'yes', 'answer_id': 3},
{'answer': 'no', 'answer_confidence': 'yes', 'answer_id': 4},
{'answer': 'yes', 'answer_confidence': 'yes', 'answer_id': 5},
{'answer': 'yes', 'answer_confidence': 'maybe', 'answer_id': 6},
{'answer': 'no', 'answer_confidence': 'yes', 'answer_id': 7},
{'answer': 'yes', 'answer_confidence': 'maybe', 'answer_id': 8},
{'answer': 'yes', 'answer_confidence': 'yes', 'answer_id': 9},
{'answer': 'no', 'answer_confidence': 'maybe', 'answer_id': 10}],
'image_id': 15675,
'answer_type': 'yes/no',
'question_id': 156750}
```

```
[209]: # lemmatize and lowercase with questions and answers
dataset[['question', 'answer']].apply(lambda x: x.str.lower(), axis=0)
```

	question	answer
0	is the boy having wine?	yes
1	is the boy having wine?	no
2	is it night time?	no

```
[356]: dataset_test.sample(5)
```

	image_id	question	question_id	answer	image_path	labels	scores
880	900109972	Is the baby happy?	900109972	no	/home/jovyan/ars/vqa-research/VQA/data/scene_i...	[9]	[1.0]
134	900195452	Is this man looking at the carpet?	900195452	no	/home/jovyan/ars/vqa-research/VQA/data/scene_i...	[9]	[1.0]
3150	9759	Is the baby boy playing with a doll?	97591	yes	/home/jovyan/ars/vqa-research/VQA/data/scene_i...	[9, 3]	[1.0, 1.0]
3106	900008111	Is the old lady getting drunk with the baby?	900008111	no	/home/jovyan/ars/vqa-research/VQA/data/scene_i...	[9]	[1.0]
3202	900153271	Is he falling in the pond?	900153271	no	/home/jovyan/ars/vqa-research/VQA/data/scene_i...	[9]	[1.0]

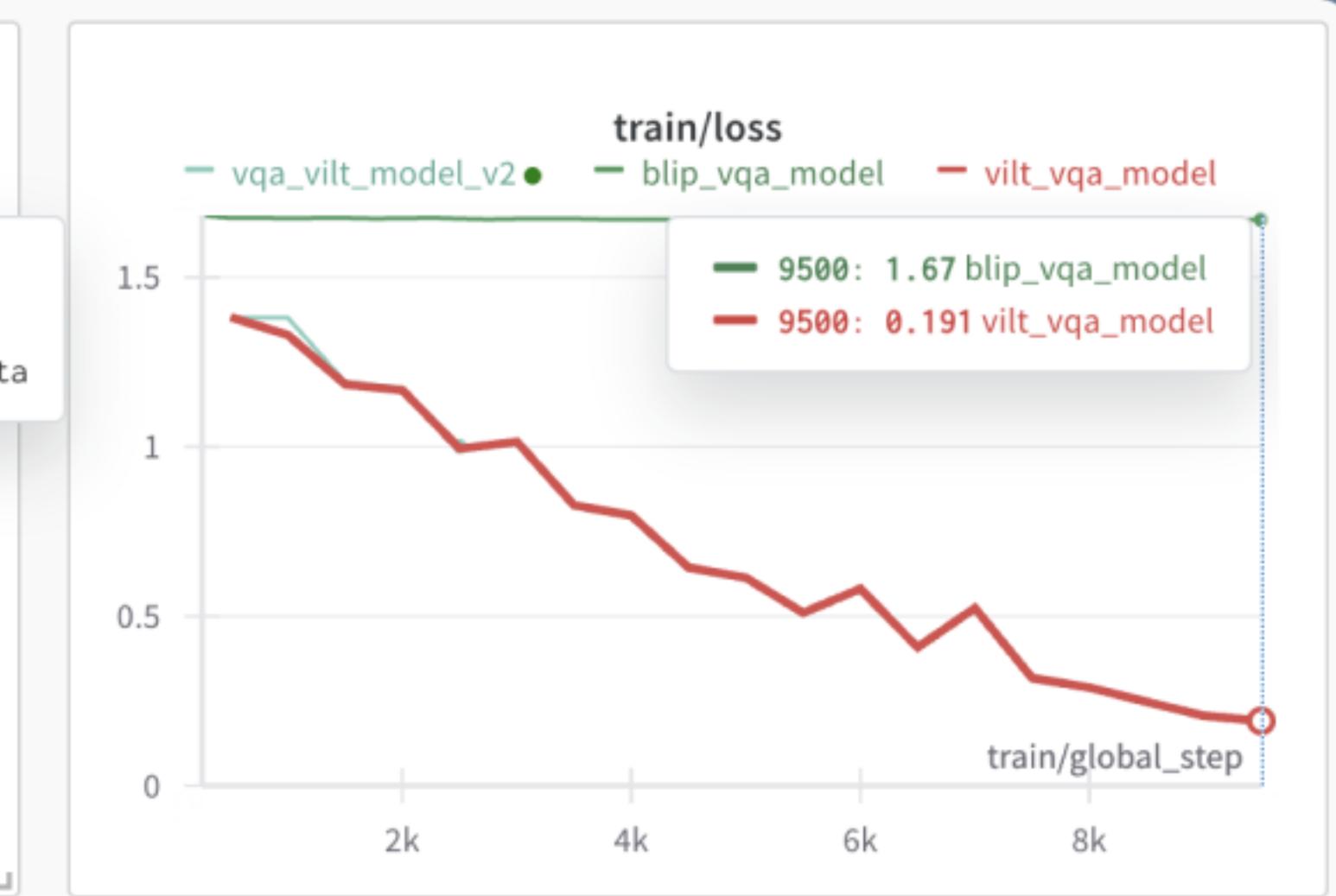
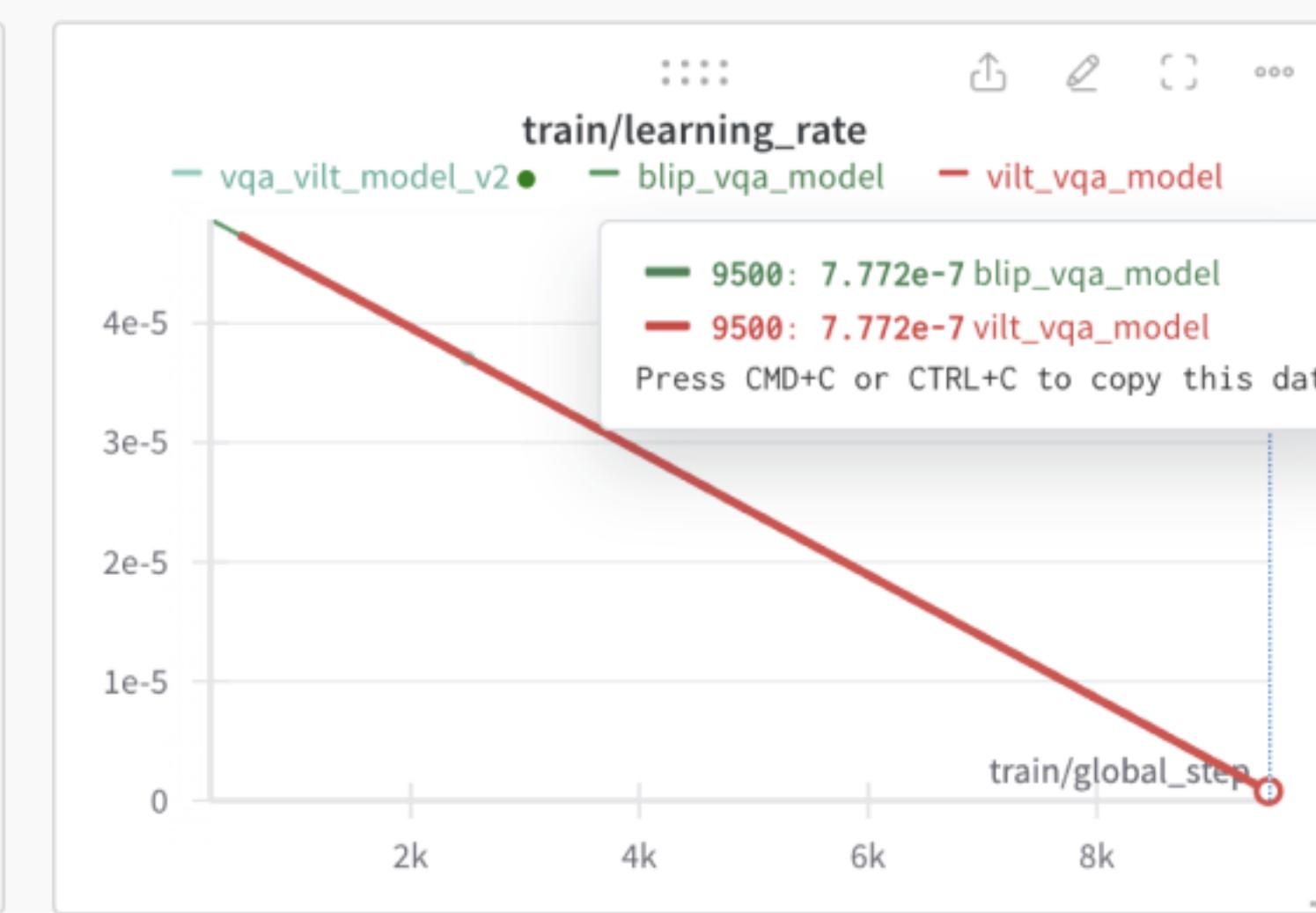
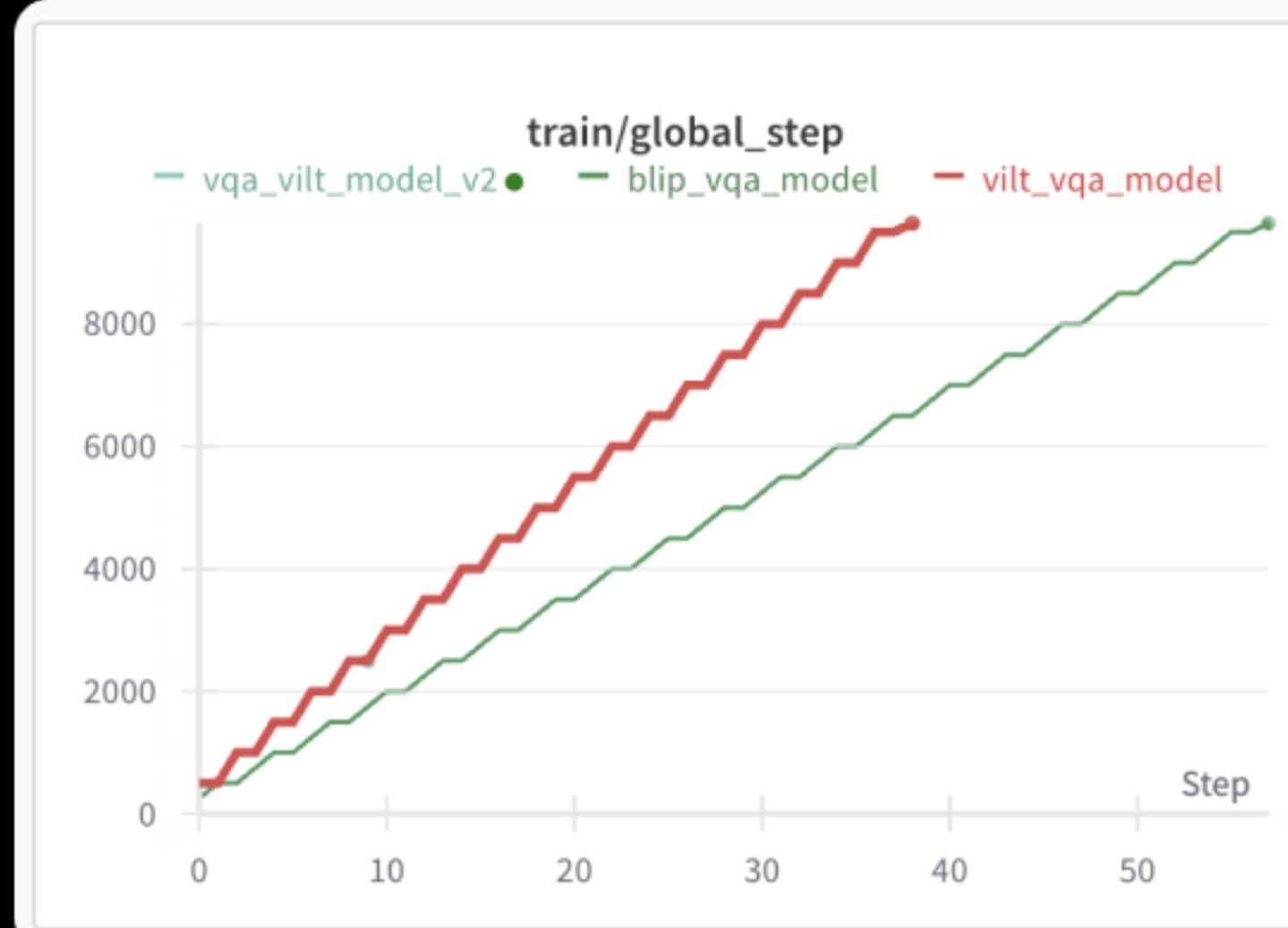
# EVALUATION METRIC

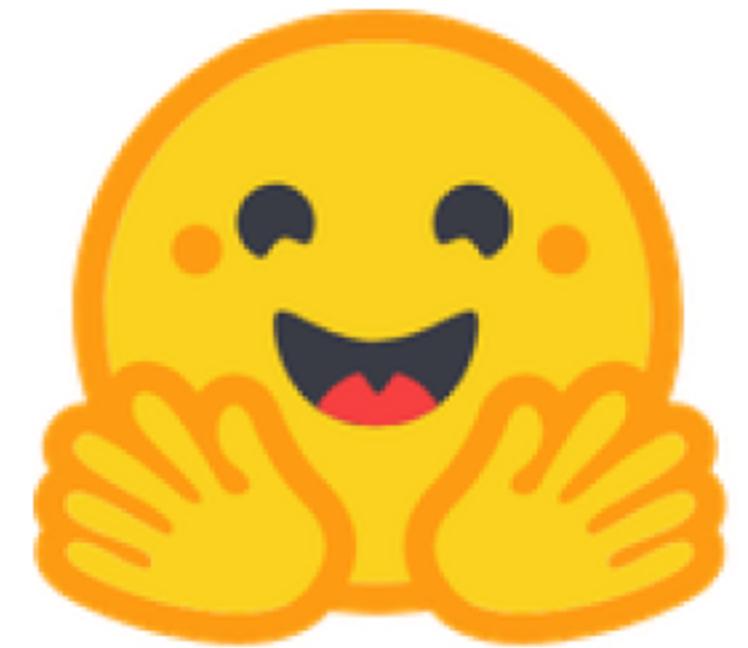
- ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation):  
ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum
- Accuracy:  $\text{correct\_predictions} / \text{total\_predictions}$
- F1\_Score

```
In [245...]: rouge_score = calc_metrics(finetunned_model, processor, test_df_dataloader, rouge_metric, device)
print(f'ROUGE: {rouge_score}')

100% |██████████| 276/276 [03:39<00:00, 1.26it/s]
ROUGE: {'rouge1': 0.8191296464188577, 'rouge2': 0.0, 'rougeL': 0.8186763372620127, 'rougeLsum': 0.8182230281051677}
```

I use WandB for my experiments





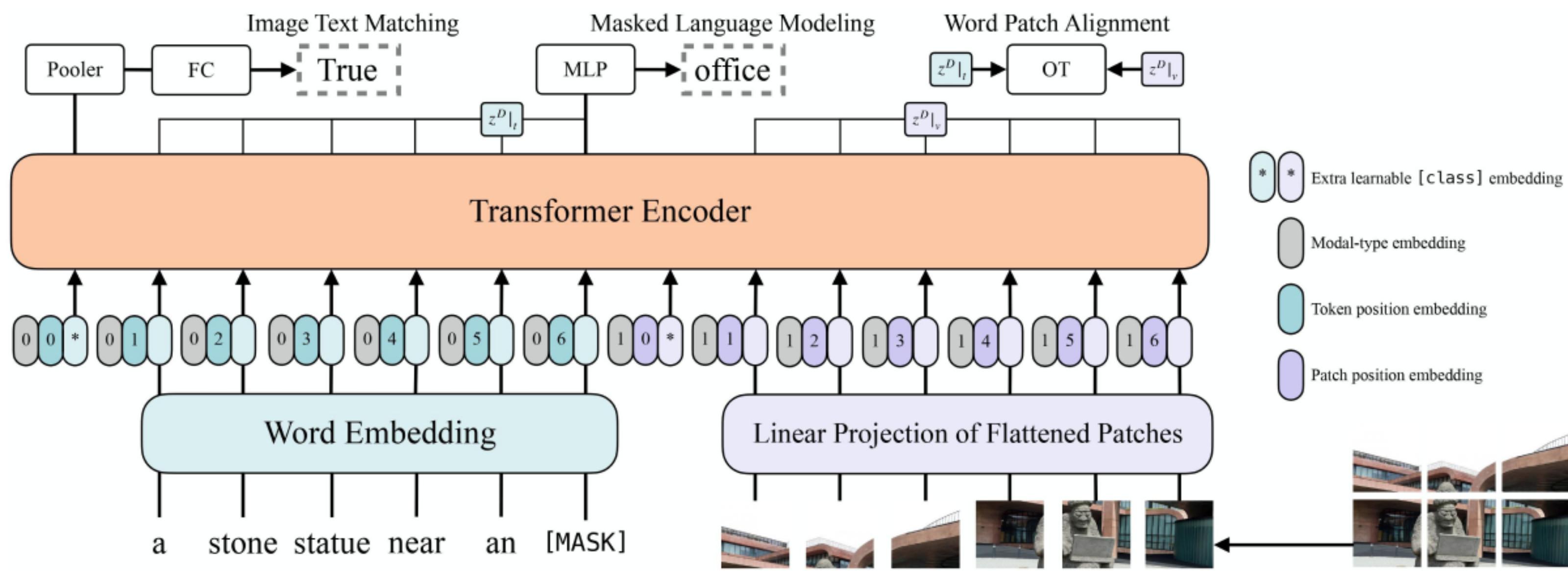
# Hugging Face

**WORKFLOW:  
ARCHITECTURES AND TECHNIQUES**

# ViLT: VISION-AND-LANGUAGE TRANSFORMER WITHOUT CONVOLUTION OR REGION SUPERVISION

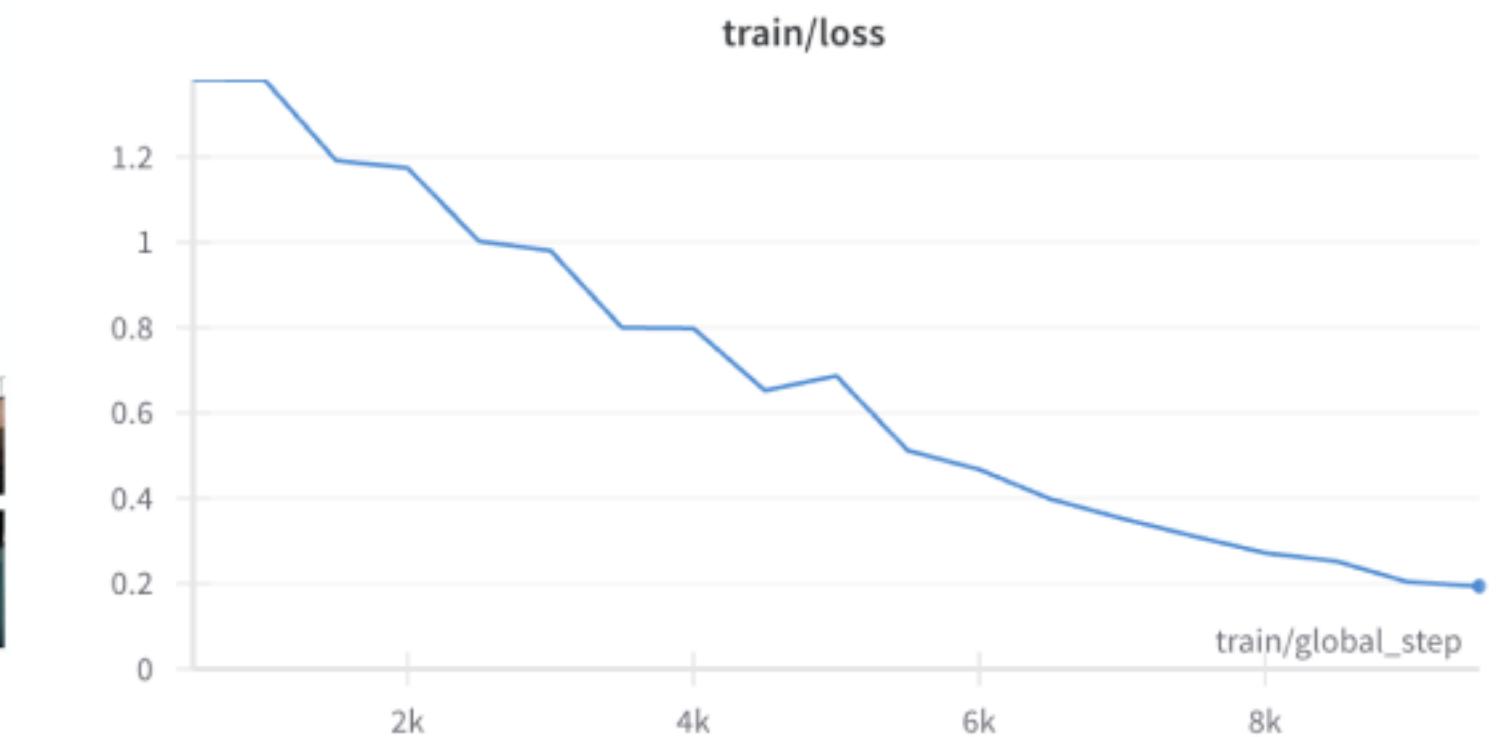


- ViLT incorporates text embeddings into a Vision Transformer (ViT), allowing it to have a minimal design for Vision-and-Language Pre-training (VLP).
- Load processor, config and model from HF “dandelin/vilt-b32-finetuned-vqa”. Initialized PyTorch dataset for our pair of samples image-question.
- Use Trainer to fine-tune last classification head in ViLT.
- For inference model we use forward() with logits.



```
print(predict(finetunned_model, processor, device, cat_image, 'is there dog here?'))
print(predict(finetunned_model, processor, device, cat_image, 'is there cat face here?'))
```

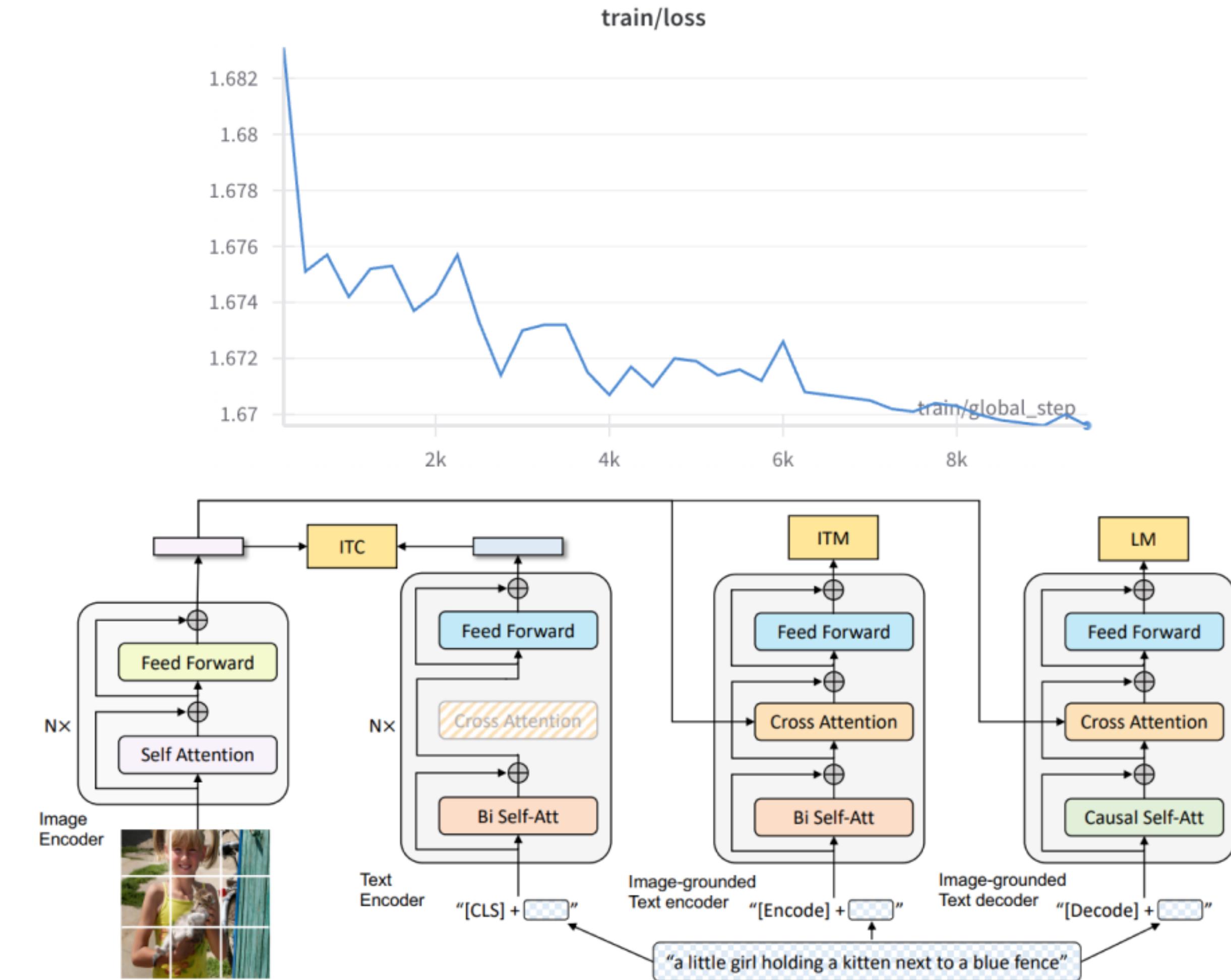
no  
yes



# BLIP: BOOTSTRAPPING LANGUAGE-IMAGE PRE-TRAINING FOR UNIFIED VISION-LANGUAGE UNDERSTANDING AND GENERATION



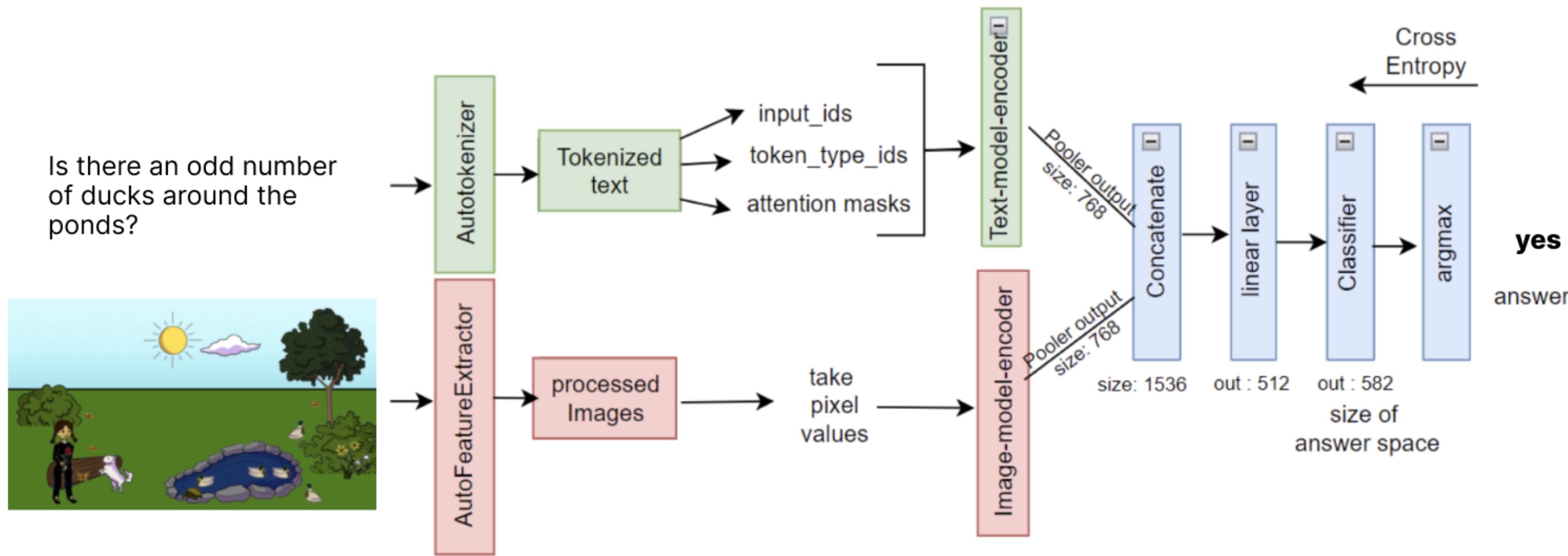
- BLIP is a model that is able to perform various multi-modal tasks including:
  - **Visual Question Answering**
  - Image-Text retrieval (Image-text matching)
  - Image Captioning
- Load BlipProcessor, BlipForQuestionAnswering, BlipConfig from HF “Salesforce/blip-vqa-base”. Initialized PyTorch dataset for our pair of samples image-question.
- Use Trainer to fine-tune last classification head in BLIP.
- For inference model we use forward() with logits.



# ROBERTA + ViT



- RoBERTa: A Robustly Optimized BERT Pretraining Approach, ViT: Vision Transformer
- Let's try to process each modality separately. To obtain textual vector representations, I use bert-like encoder model (RoBERTa), and for images embedding I use visual transformers (ViT).
- After that we had embeddings, we can concatenate this and apply linear layers for binary classification (yes or no).
- In the rep you can see created RobertaVitFusionNetwork() and process of training



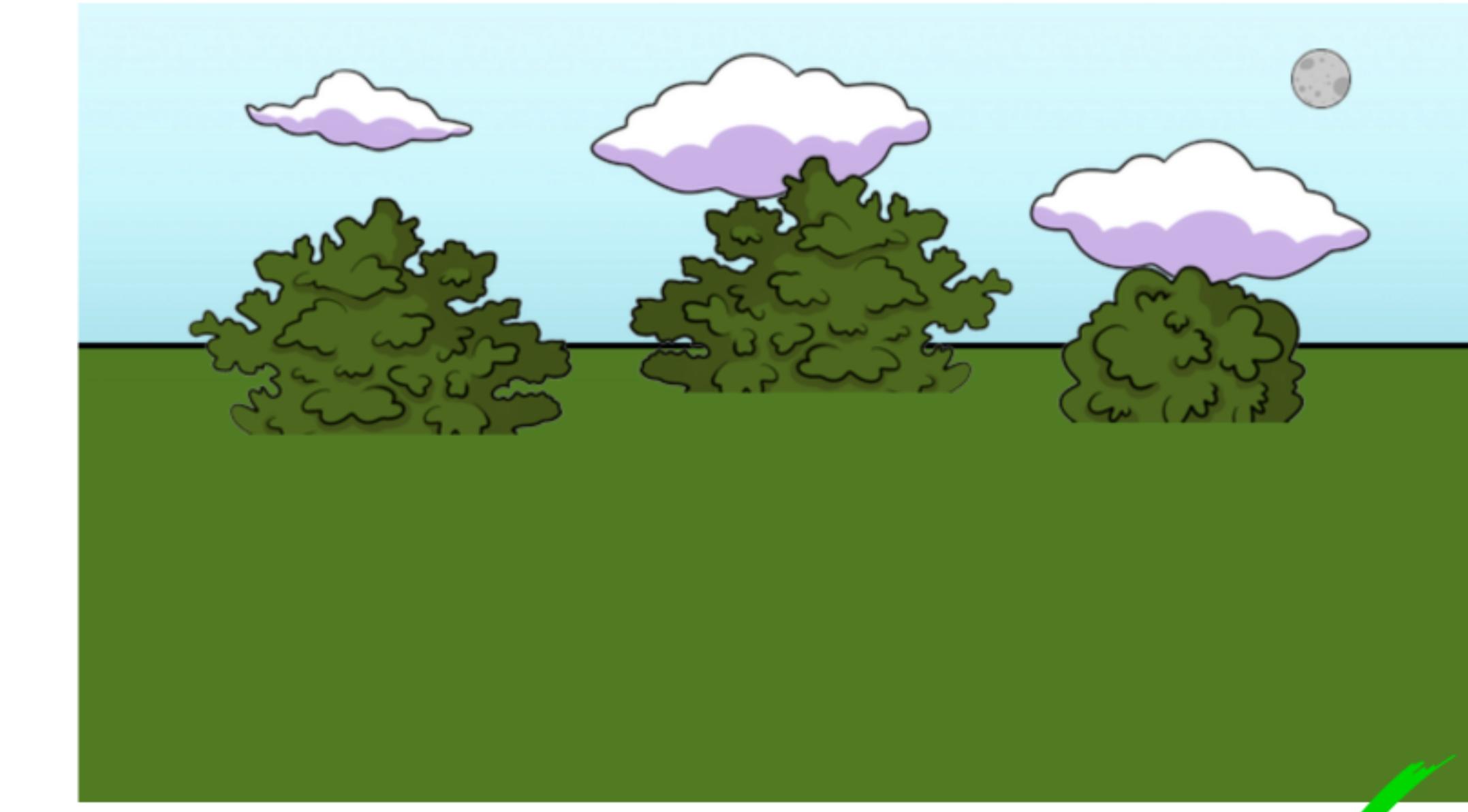
# RoBERTA + DeiT



- RoBERTa: A Robustly Optimized BERT Pretraining Approach, DeiT: Training data-efficient image transformers & distillation through attention
- To obtain textual vector representations, I use bert-like encoder model (RoBERTa), and for images embedding I use visual transformers (DeiT).
- After that we had embeddings, same like in the previous exp, embeddings concatenated and apply linear layers for binary classification (yes or no).



Question: Are the man and woman dancing?  
Answer: yes (Label: 1)  
Predicted Answer: no



Question: Is the sun up?  
Answer: no (Label: 0)  
Predicted Answer: no

# RESULT AND ANALYSIS

	Rouge1	Rouge2	RougeL	RougeLSum	
<b>ViLT</b>	<b>0.819129</b>	0.818371	<b>0.818676</b>	0.818223	
BLIP	0.728153	0.725714	0.727038	0.725792	
RoBERTa + ViT	0.538641	0.536012	0.537826	0.537285	
RoBERTa + DeiT	0.512494	-	0.513892	0.512836	

# WE SHOULD TRY TO FINE-TUNE LLAVA MODEL!

Visual input example, Extreme Ironing:



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User  
LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[Start a new conversation, and clear the history]

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User  
GPT-4 [32]

What is unusual about this image?

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

User  
BLIP-2

Can you explain this meme in detail?

a man is sitting on the back of a yellow cab

User  
OpenFlamingo

Can you explain this meme in detail?

The man is drying his clothes on the hood of his car.

# LLaVA: Large Language and Vision Assistant

## Visual Instruction Tuning

NeurIPS 2023 (Oral)

Haotian Liu\*, Chunyuan Li\*, Qingyang Wu, Yong Jae Lee

► University of Wisconsin-Madison ► Microsoft Research ► Columbia University

\*Equal Contribution

arXiv arXiv (LLaVA-1.5) Code Demo Dataset Model

## Science QA: New SoTA with the synergy of LLaVA with GPT-4



# CONCLUSIONS

- Data analysis and sample preprocessing were performed.
- Modules of training and inference of models have been produced.
- A large number of works of the VQA problem have been studied. Several models were trained to test the results: VILT and BLIP from hugging face, as well as RoBERTa + VIT, RoBERTa + DEiT, which were processed in own way and the output was a network with two model handlers.
- VILT shows the best quality according to the rouge metric, which means that the proportion of correctly guessed values is much higher than that of other models.
- As a research scale, it is worth trying to train the LAVA model. On sbs mlu tests, this model shows very good results.

# DEMO



```
print(predict(finetunned_model, processor, device, image_abstract, 'Is the person sitting?'))  
print(predict(finetunned_model, processor, device, image_abstract, 'Is the person reading the book?'))  
print(predict(finetunned_model, processor, device, image_abstract, 'Is there window here?'))  
print(predict(finetunned_model, processor, device, image_abstract, 'Is it night here?'))  
print(predict(finetunned_model, processor, device, image_abstract, 'Is it daytime here?'))  
print(predict(finetunned_model, processor, device, image_abstract, 'Is the carpet color red here?'))
```

yes  
no  
yes  
no  
yes  
yes



★GITHUB  
REPOSITORY  
WITH  
PROJECT