Visual Question Answering: A Recent State-of-the-art Summary

Giovanni Pica
Master Degree Computer Science
Sapienza University Of Rome
ID: 1816394
pica.1816394@studenti.uniroma1.it

Abstract

Language and vision problems such as image captioning and visual question answering (VQA) have gained popularity recently. In particular, VQA is a task that combines both the techniques of computer vision and natural language processing. It involves challenges in which we have to process the data with visual as well as linguistic processing to find the answer to basic common sense questions related to a given image. This task requires reasoning capabilities on visual features and objects of the image along with that general knowledge to predict the correct answer to the given question. In recent years, the research field of VQA has been expanded. This paper will review and analyze state-of-the-art, datasets, metrics, and evaluation for the VQA task.

1 Task description/Problem statement

Visual Question Answering (VQA) is an inspiring task that includes two major fields of AI namely **Natural Language Processing** and **Computer Vision**. On the first side, computer vision is where we discuss acquiring, processing, and extracting features from the image. On the other side, NLP is where we train our model to understand and process natural language. So that it can take instructions from humans in natural language as well as communicate with them. The main difference between Visual Question Answering (VQA) and traditional computer vision tasks is that in VQA, the model does not have access to the question that needs to be answered until runtime. In order to achieve this goal, a VQA system utilizes algorithms of various types that work together to process an image and a natural language question related to it. The system then generates a natural language answer as its output. Humans are naturally adept at performing this task, except under special circumstances, and artificial intelligence aims to replicate this ability. By combining these different algorithms and approaches, VQA systems aspire to mimic the human ability to comprehend visual information, understand language, and provide meaningful responses to questions about images. The ultimate goal is to develop AI systems that can accurately and effectively perform visual question-answering tasks, thus enhancing human-computer interactions and enabling applications in various domains. In conventional problems, a fixed set of questions is provided for all images, with only the image itself changing as input. However, in the new era of VQA, the questions asked about an image are unknown and vary depending on the specific image. To successfully perform VQA, reasoning capabilities on visual features and objects in the image, as well as general knowledge, are required to predict the correct answer. The task was introduced by Antol et al. [3], who also created the dataset and an improvement of it [14]. The motivation behind VQA arose from the fact that answering open-ended questions requires a wide range of AI capabilities, including fine-grained recognition, object detection, activity recognition, knowledge base reasoning, and common sense reasoning. Examples of such questions include identifying the type of cheese on a pizza, counting the number of bikes in an image, determining if a person is crying, and so on. The aim of this paper is to discuss the state-of-the-art of VQA, datasets, benchmarks, and libraries.

1.1 Examples

VQA is simply based on two inputs, a natural-language question, and an image. It produces as output a natural-language answer that depends on the context given by the question and the image.

Below there is Figure 1 that describes this scenario.

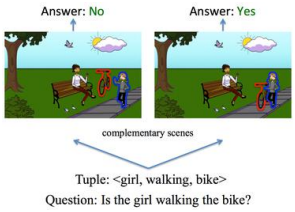

Figure 1: An Example of VQA at a high level.

1.2 Real-world applications

According to Barra et al. [7], there are some interesting areas of applications.

Medical: AI-based medical image understanding, along with related medical question-answering (med-VQA), has become a topic of growing interest among researchers. This field presents new opportunities for assisting medical personnel in making clinical decisions and improving diagnosis through computer-generated "second opinions". The application of AI in this context has the potential to provide valuable support and insights to medical professionals, ultimately benefiting patient care and outcomes. In particular, given a medical image and a clinically relevant question in natural language, the medical VQA system is expected to predict a plausible and convincing answer.

VQA for visually impaired people: Several VQA applications proposed in recent years have focused on assisting blind people as one of their main objectives. The automatic question-answering capabilities of VQA systems enable them to answer everyday questions, which can help visually impaired individuals navigate their lives without visual barriers. By providing real-time answers to their inquiries, VQA technology contributes to enhancing the independence and accessibility of visually impaired individuals, enabling them to overcome obstacles and engage with the world more effectively.

Video Surveillance: The adoption of a VQA approach in video surveillance scenarios has the potential to assist operators in improving their understanding of a scene, thereby enabling them to make more informed and timely decisions. By integrating VQA technology into video surveillance systems, operators can receive automated answers to questions related to the observed video footage. This aids in the interpretation of events, identification of suspicious activities, and assessment of potential threats. Ultimately, VQA in video surveillance can enhance the efficiency and effectiveness of operators, leading to fairer and quicker decision-making processes in critical situations.

Education and cultural heritage: VQA is closely related to human perception, as it aims to understand images in a similar manner to humans. For example, there could be an Educational Robot that uses VQA to generate questions and initiate educational dialogues inspired by its surroundings to demonstrate the ability to enhance children's curiosity and desire to explore. VQA can also enhance the desire to explore for adults, particularly in the context of cultural heritage by using VQA to interact with an audio guide when exploring museums and art galleries.

Advertising: When users view an advertisement, they not only perceive the objects within the scene but also consider the accompanying text and the relationships among the objects. Additionally, they interpret all this information within a specific cultural context. For an advertisement to be effective, it needs to be simple enough to be understood by a wide audience while still being interesting and visually appealing. This is where VQA can find a challenging and valuable application in advertising. By utilizing VQA technology, advertisers can gain insights into how users perceive and interpret their advertisements. This information can help in creating more effective and engaging advertisements that resonate with the target audience, resulting in better communication and increased impact.

2 Related work

The first VQA model [3] is based on CNN for extracting the features and LSTM or RNN networks for language processing. After that VQA models had some evolutions like the utilization of the attention to concentrate on particular features of an image. According to Banchnor et al. [4], we have some different approaches for VQA.

Grid Feature models [18]: Through their research, the authors discovered that utilizing grid features can significantly enhance the performance of VQA systems. Grid features were found to be highly efficient, running more than ten times faster than previous methods, while maintaining the same level of accuracy. Significantly, the authors found that grid features, derived from the

same layer of a pre-trained object detector, outperformed region-based counterparts in the context of VQA. This highlights the superiority of grid features for addressing the VQA problem and emphasizes their potential for improving performance in similar tasks.

Differential Networks [39]: The researchers introduced a novel module called Differential Networks (DN), which enables the comparison of pairwise feature items by capturing their differences. By leveraging DN as a tool, they developed a new model called DN-based Fusion (DF) for Visual Question Answering (VQA). In the DF model, the picture and question feature elements are compared using DN, which allows for the extraction of differential representations. These representations capture the variations or discrepancies between the features. The DF model then integrates these differential representations to generate the final response or answer to the given question. By incorporating DN into their fusion model, the researchers aim to enhance the VQA performance by effectively leveraging the differences between feature elements. This approach enables the model to better capture and exploit the complementary information present in the picture and question features, leading to improved accuracy and robustness in answering visual questions.

Multimodal Compact Bilinear Pooling (MCB) [13]: To achieve a joint representation for the VQA task, Fukui et al. introduced a method called Multimodal Compact Bilinear Pooling (MCB). MCB involves approximating the pooling operation by projecting the picture and text representations randomly into a higher-dimensional space. Once in this higher-dimensional space, the vectors are efficiently convolved using the Fast Fourier Transform (FFT) method, specifically through an element-wise product. By leveraging MCB, the researchers were able to anticipate answers for the VQA task by combining information from both the visual and textual modalities. This approach allows for more accurate predictions and improved performance in both the VQA and visual grounding tasks.

Oscar [27]: The Oscar model is a novel approach to Visual-Linguistic Pre-training (VLP). In this approach, training samples are represented as triples, consisting of a word sequence, a set of object tags, and a set of image region features. The researchers were motivated by the fact that modern object detectors can accurately identify important objects in an image, and these objects are often mentioned in the accompanying text. The model goes through a two-step process. First, it undergoes pre-training on a large-scale dataset comprising over 6.5 million pairs of visual and linguistic data. This pre-training step helps the model learn general visual and linguistic representations. After pre-training, the model is fine-tuned and evaluated on seven different tasks that involve both comprehension and generation of visual and linguistic information. By incorporating both visual and linguistic information in its training process, the Oscar model aims to enhance its understanding of the relationships between images and text. This approach enables the model to generate more accurate and comprehensive results across a range of V+L tasks.

VL-BERT [34]: The VL-BERT model architecture modifies the original BERT model by incorporating additional components to accommodate image information. It introduces new parts and a distinct type of visual feature embedding in order to integrate visual data into the input feature embeddings. The backbone of VL-BERT consists of a multi-layer bidirectional Transformer encoder [35], similar to BERT, which facilitates the modeling of dependencies between all input elements. VL-BERT combines both visual and linguistic aspects as input. The visual information is described based on regions of interest (ROIs) in the images, while the linguistic input corresponds to subwords derived from input phrases. By incorporating both visual and linguistic modalities, VL-BERT aims to leverage the complementary nature of these sources to improve performance in tasks that involve understanding and processing both images and text. The utilization of a Transformer encoder in VL-BERT enables effective modeling of dependencies between the visual and linguistic components, allowing for a comprehensive understanding of their interactions. This architecture provides a framework for jointly processing visual and textual data, enhancing the model's ability to capture the rich information present in both modalities.

By looking at the benchmarks [1] some new approaches came up that are state-of-the-art. In particular two of the most outstanding are **PaLI** (Pathways Language and Image model) [9] and **BEiT-3** [37].

The PaLI system is capable of generating text by incorporating both visual and textual inputs, enabling it to perform various tasks related to vision, language, and multimodal understanding across multiple languages. To train PaLI, a combination of large pre-trained encoder-decoder language models and Vision Transformers (ViTs) [12] are utilized. This approach takes advantage of the existing capabilities of these models and leverages the substantial training investments already made in them. In order to achieve effective performance, it is crucial to scale both the vision and language components together. While existing Transformers for language are larger compared to their vision counterparts, the researchers train a large ViT model called ViT-e, with a capacity of 4 billion parameters, to evaluate the advantages of employing even larger-capacity vision models. For training PaLI, a diverse range of pre-training tasks is created, encompassing multiple languages. This is accomplished by employing a new image-text training dataset, which consists of a vast collection of 10 billion images and texts spanning over 100 languages. By training PaLI on this extensive multilingual dataset, it gains the ability to perform a wide range of vision and language tasks effectively across various languages.

BEiT-3 is a multimodal model that performs masked "language" modeling on images, texts, and image-text pairs. It is pretrained on large monomodal and multimodal datasets using a shared Multiway Transformer network [5] as its backbone. The Multiway Transformer blocks consist of a shared self-attention module that aligns different modalities and enables deep fusion, along with a set of feed-forward networks to represent each modality. During pretraining, a percentage of text tokens or image patches are randomly masked, and the model is trained to predict the masked tokens. This unified mask-then-predict task [6] allows the model to learn representations and also understand the alignment between different modalities. Text data is tokenized using a SentencePiece tokenizer [24], while image data is tokenized using the tokenizer of BEIT v2 [31] to obtain discrete visual tokens. The masking percentages are set to 15% for monomodal texts and 50% for texts from image-text pairs. For images, a block-wise masking strategy, similar to BEIT [6, 31], is used to mask 40% image patches. This approach helps BEIT-3 learn representations and capture the relationships between different modalities in a unified manner.

3 Datasets and benchmarks

The availability of high-quality datasets is a major challenge in NLP and computer vision, and in the case of Visual Question Answering (VQA) is a particularly challenging domain. It necessitates a large dataset that encompasses a wide range of questions related to real-world images, covering all possible scenarios and situations. Generating such a dataset is complex, involving curated images, annotated questions, and diverse question-image pairs. The expansion of VQA datasets is crucial for advancing the capabilities of VQA models to accurately understand and answer questions about real-world images. In Table 1 there are the datasets and their benchmarks in terms of accuracy.

Dataset	Images	Questions	Benchmark	Accuracy
VQA2.0 test-dev [14]	265016	795048	PaLi [9]	84.3
VQA2.0 test-std [14]	265016	795048	BEiT3 [37]	84.03
VQA [3]	204721	614163	SAAA [22]	64.6
COCO [32]	123287	177684	MCB [13]	66.5
CLEVR [20]	100000	853554	NS-VQA [40]	99.8
Visual7W [42]	47500	2201154	CMN [16]	72.53
GQA [17]	113000	22000000	ProTo [41]	65.140

Table 1: Datasets and Benchmarks

4 Existing tools, libraries, papers with code

Some of the libraries are all GitHub repositories and three of them are selected, that are some of the most used and there are some papers with code.

Huggingface's Transformers [38] (https://github.com/huggingface/transformers) is an open-source library that aims to make recent advancements in Transformer architectures accessible to the broader machine learning community. The library provides a unified API and includes well-designed, state-of-the-art Transformer architectures. It also offers a curated collection of pre-trained models created by the community and available for public use. The primary objectives of Transformers are to be extensible for researchers, user-friendly for practitioners, and efficient and reliable for industrial applications. In this library there is the paper with code ViLT [23].

Lavis from SalesForce [25] (https://github.com/salesforce/lavis) is a Python deep learning library for LAnguage-and-VISion intelligence research and applications. This library aims to provide engineers and researchers with a one-stop solution to rapidly develop models for their specific multimodal scenarios, and benchmark them across standard and customized datasets. In this library there is the implementation of BLIP-2 [26].

UniLM by Microsoft [11] (https://github.com/microsoft/unilm) that stands for Unified pre-trained Language Model (UniLM) is a versatile model that can be fine-tuned for both natural language understanding and generation tasks. It is pre-trained using three types of language modeling tasks: unidirectional, bidirectional, and sequence-to-sequence prediction. The model achieves its unified modeling capability by employing a shared Transformer network and utilizing self-attention masks to control the context in which predictions are made. In this library there are pre-trained models and papers with code like BEiT-3 [37] and VL-BEiT [6].

5 State-of-the-art evaluation

To evaluate VQA models the metrics used are the **classic VQA** [3] based on the classification aspect of answers, and others that consider the phrasal structure.

The classic one is based on:

$$Acc = min(\frac{\#htpta}{3}, 1)$$

#htpta stands for humans that provided that answer.

BLEU [30], is designed for machine translation in the context of open-ended answers in the Visual Question Answering (VQA) task by evaluating the fluency, adequacy, and length of machine-translated candidate answers compared to human-annotated reference answers. It focuses on evaluating the similarity in terms of n-gram matches (typically up to 4 grams) and also considers the brevity penalty to encourage translations of appropriate length.

ROUGE-L [28], is a machine translation evaluation system that consists of four distinct metrics. Each metric has its complexities in interpreting the adequacy of a translation. The system is commonly used to assess the quality of machine-generated summaries by comparing them to reference summaries. The four metrics focus on measuring the recall of n-gram overlap, including unigram, bigram, trigram, and higher-order n-grams, between the candidate translation and the reference translation.

CIDEr [36], is a metric used to assess the similarity between machine-translated descriptions and human-annotated ones. The authors of CIDEr have defined three principles that should be considered when evaluating automatic translations in terms of consensus, and these principles are incorporated into the calculation process. The first one is Frequency-based Consensus, the measure of consensus should take into account the frequency with which candidate n-grams (word or phrase sequences) appear in their corresponding reference translations for the same image. The second one is the Absence Penalty, the measure of consensus should penalize candidate n-grams that are not present in the reference translations for the same image. The third one is Inverse N-gram Frequency Weighting, the measure of consensus should assign weights to n-grams inversely proportional to their frequency across all images.

Hit@K [2], is a performance evaluation measure that provides information on how well a solution or system is performing. It assesses how well a solution performs in terms of ranking and retrieving correct test triples from a Knowledge Graph [10]. It measures the rate at which the correct test triples are ranked within the top K triples, where K represents the desired cutoff or number of triples to consider.

6 Conclusions

VQA has had a huge progress when compared to previous years, there is still space for improvement. It is an area of great opportunity and can be used in medicine, robotics, security, welfare, etc. This is possible through the use of Deep Learning, which in general shows a lot of advances for recognizing and evaluating images and texts together. The current standards for evaluating Visual Question Answering (VQA) approaches are considered inadequate in determining if an approach has successfully solved the VQA task. To address this, larger and more diverse datasets will be needed. Although there has been an increase in the quantity and variety of VQA datasets available, algo-

rithms still require sufficient data for training and evaluation purposes. According to Banchnor et al. [4], it has been observed that increasing the dataset size, even if it is imbalanced or skewed, can lead to significant improvements in accuracy. However, simply increasing the dataset size alone does not guarantee a useful benchmark because human-generated questions often contain biases [21]. To enhance the learning capabilities of VQA models, datasets with less skewness are required. Data balancing techniques can be employed to mitigate the impact of strong language biases in VQA datasets. By reducing the prevalence of certain language biases, models are less likely to rely on those biases to produce answers [14]. This helps in creating a more fair and balanced evaluation environment for VQA models. In this work, there are studies about the state-of-the-art, datasets and benchmarks, and libraries that can be used to learn this task.

Next-generation level VQA models can focus on different techniques such as deeper models, segmentation masks, question sentiments, object-level details, etc. To expand the scope of VQA could be introduced some future directions that lack in the literature, or are only preprints, or they are not the main focus:

- **Generative VQA**: the focus is on generating novel and imaginative content in response to questions about visual input. This goes beyond providing factual answers and aims to foster creativity and imagination. It involves generating detailed and contextually relevant captions, diverse and imaginative stories, and even visual representations to support the generated content. The goal is to push the boundaries of VQA by encouraging more creative and expressive responses.

- **RL VQA**: in the context of Visual Question Answering (VQA) involves using Reinforcement Learning techniques [29] to train VQA models that can interact with an environment, receive feedback based on their actions, and learn to generate optimal answers to visual questions. In RL-based VQA, the VQA model takes on the role of an agent that interacts with the environment. The environment provides visual input (images or videos) and questions, and the agent generates answers. The agent's performance is evaluated based on the quality and accuracy of its answers, and it receives rewards or penalties based on the correctness of its responses.

- **Explainable VQA**: is focused on making VQA models more explainable [15]. Researchers aim to develop models that can provide explanations or visualizations to justify their answers. This enhances interpretability, builds trust, and helps users understand the reasoning behind the model's decisions.

- **Robustness to Adversarial Inputs**: aims to develop models that are more robust and resilient to adversarial attacks [8]. Adversarial attacks involve introducing imperceptible modifications to input images or questions, which can mislead VQA models. By enhancing the models' resilience to such adversarial inputs, researchers seek to improve their reliability and accuracy in answering questions. In this, there are some studies about adversarial VQA [33, 19] but not recent robust VQA models.

References

[1] PapersWithCode Benchmark Results. https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-test-dev. Accessed: 2023-06-23. 3

[2] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8825–8845, dec 2022. 5

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 4, 5

[4] Mrinal Banchhor and Pradeep Singh. A survey on visual question answering. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–5, 2021. 2, 6

[5] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, 2022. 4

[6] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vl-beit: Generative vision-language pretraining, 2022. 4, 5

[7] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. Visual question answering: Which investigated applications? *Pattern Recognition Letters*, 151:325–331, 2021. 2

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee*

symposium on security and privacy (sp), pages 39–57. Ieee, 2017. 6

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. 3, 4

[10] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. Zero-shot visual question answering using knowledge graph, 2021. 5

[11] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation, 2019. 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas, November 2016. Association for Computational Linguistics. 3, 4

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 4, 6

[15] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019. 6

[16] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks, 2016. 4

[17] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 4

[18] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10267–10276, 2020. 2

[19] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018. 6

[20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. 4

[21] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, oct 2017. 6

[22] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering, 2017. 4

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. 4

[24] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018. 4

[25] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 5

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 5

[27] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *CoRR*, abs/2004.06165, 2020. 3

[28] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004. 5

[29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 6

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. 5

[31] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers, 2022. 4

[32] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering, 2015. 4

[33] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. 2021. 6

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. *CoRR*, abs/1908.08530, 2019. 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3

[36] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 5

[37] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 3, 4, 5

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. 4

[39] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Ruifan Li. Differential networks for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8997–9004, 2019. 3

[40] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding, 2019. 4

[41] Zelin Zhao, Karan Samel, Binghong Chen, and Le Song. Proto: Program-guided transformer for program-guided tasks, 2021. 4

[42] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016. 4