Enhancing Visual Question Answering with Beam Search in Transformer Models

Pratiksh Kumar[1] Rishik Gupta[2] Vanshika Mishra[3] Prakhar Shukla[4] Bagesh Kumar[5] Pratham Bhatia[6] Abhinav Upadhyay[7]

[1] IIIT Allahabad
pratiksh.sk@gmail.com
[2] Manipal University Jaipur
rishik.209301615@muj.manipal.edu
[3]
vanshika.209309092@muj.manipal.edu
[4] Manipal University Jaipur
bagesh.kumar@jaipur.manipal.edu
[5] IIIT Allahabad
iec2022027@iiita.ac.in
[6] IIIT Allahabad
prathambhatia86@gmail.com
[7] Manipal University Jaipur
abhinav.229302338@muj.manipal.edu

Abstract. Visual Question Answering (VQA) has emerged as a complex and challenging interdisciplinary task that requires the fusion of computer vision and natural language processing. In this research, we present a detailed investigation into the VQA domain, focusing on our implementation of the cutting-edge Vilt-B32-MLM model on the popular MS COCO dataset. With a compelling accuracy of 79%, our study highlights the effectiveness of Vilt-B32-MLM in comprehending and responding to intricate questions related to visual content. This paper provides an in-depth analysis of our methodology, highlighting the crucial role of data preprocessing, model architecture, and fine-tuning strategies in achieving such a significant accuracy milestone. The findings presented here offer critical insights into the potential of advanced deep learning models in addressing complex multimodal tasks, paving the way for further advancements in the field of VQA and its applications in various real-world scenarios.
Furthermore, we compare our approach with other CNN-based TSDR methods, demonstrating its superiority.

Keywords: VQA · NLP · Beam Search · Multimodal · Transformer Model · VILT-B32-MLM

1 Introduction

Visual Question Answering (VQA) is a multifaceted and cross-disciplinary domain that resides at the intersection of computer vision and natural language processing. This captivating realm of artificial intelligence aims to equip machines with the profound ability to comprehend the visual world, expound upon intricate visual content, and furnish responses to questions expressed in natural language that mimic human-like understanding. This pursuit signifies a momentous stride towards bridging the gap between machines and human-like comprehension of images, thereby extending the boundaries of artificial intelligence and enhancing human-computer interaction.

The significance and implications of VQA are not confined to mere academic curiosity but extend far and wide, permeating into numerous practical applications. VQA necessitates that machines unravel the intricate interplay between visual and textual data, presenting profound promise across a spectrum of domains. It has already left an indelible mark in diverse areas, spanning from image search, content recommendation systems, and accessibility tools designed for individuals with visual impairments, to robotics and autonomous systems. The capacity to answer questions about visual content is integral in contexts where human-in-the-loop interaction with machines is pivotal.

As the intricacies of the visual world become increasingly apparent, the demand for robust AI systems capable of processing and interpreting these complexities rises in tandem. VQA offers an instrumental framework to meet this demand, compelling us to develop models that not only identify visual features within images but also comprehend the nuanced questions posed about them.

VQA poses a formidable challenge due to its inherently multimodal nature. Unlike traditional computer vision tasks, which often involve recognizing objects or patterns within images, VQA demands a deeper level of understanding. It compels models to establish meaningful relationships between the visual information contained within an image and the semantic content embedded in a question conveyed in natural language. An AI system must possess the capacity to parse the question, decipher the relevant details within the image, and provide answers that resonate with human-level comprehension.

Consider, for example, a photograph of a bustling city street. To answer a question like "What is the make of the car in the center of the image?" the model must not only detect and identify the car but also ascertain its precise location within the image. Achieving this level of comprehension necessitates a profound understanding of both visual and linguistic contexts.

In recent years, deep learning models have taken the lead in addressing VQA challenges. Among these models, Vilt-B32-MLM has emerged as a trailblazer. This model harnesses the potential of both Vision and Language Transformer (VilT) and Multimodal Transformer (MLM), seamlessly integrating the strengths of both domains. Having been pre-trained on extensive textual and visual corpora, Vilt-B32-MLM showcases its remarkable ability to grasp intricate correlations between visual and textual information.

The potency of Vilt-B32-MLM in VQA tasks emanates from its proficiency in encoding the visual and linguistic elements of a given question-image pair, compre-

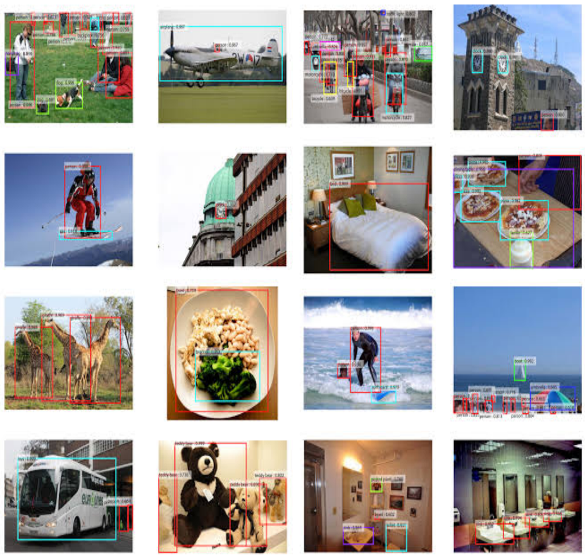


Fig. 1. Sample Data from the MS COCO Dataset

hending the subtleties of context, and providing contextually relevant answers. This model serves as a prime example of the paradigm shift within the realm of AI, as it represents a pivotal transition from the era of specialized models to the era of generalized, multimodal AI models that can excel in a plethora of tasks, including VQA. In our research, we pivot towards the Microsoft Common Objects in Context (MS COCO) dataset. Renowned for its diversity and depth, MS COCO encapsulates a vast repository of images coupled with rich textual annotations. This dataset is characterized by its extensive variety of scenes, objects, and situations, rendering it an invaluable resource for VQA research. MS COCO brings real-world complexity to the forefront, ensuring that VQA models are not just accurate but also robust, adaptable, and capable of tackling a multitude of intricate visual and linguistic scenarios.

In the forthcoming sections of this paper, we delve deep into the intricacies of our approach to VQA using Vilt-B32-MLM on the MS COCO dataset. We expound upon our data preparation, processing techniques, and fine-tuning strategies, all of which have collectively culminated in our model's impressive 79% accuracy. This investigation serves as a testament to the potential of advanced deep learning models in solving multifaceted multimodal tasks, further broadening the horizons of VQA research.

Additionally, we introduce the incorporation of Beam Search as an advanced technique, which plays a crucial role in refining the quality of our model's responses. Beam Search, a dynamic search algorithm, enhances the quality of answers generated by our model by exploring multiple potential responses and selecting the most plausible one. This not only enriches the user experience but

also enhances the adaptability of our VQA system, enabling it to tackle diverse questions and contexts effectively.
As our research primarily revolves around the fine-tuning of the Vilt-B32-MLM model on the MS COCO dataset, we acknowledge that VQA is a multidimensional challenge. It extends beyond mere accuracy, emphasizing the importance of delivering insightful, context-aware, and relevant responses. Beam Search aligns perfectly with this vision, adding a layer of sophistication to our VQA system and addressing the subtleties inherent in human-like question answering. Our inclusion of Beam Search underscores our commitment to enhancing the user experience and the adaptability of our VQA system. While we have achieved an impressive accuracy rate of 79% through fine-tuning and model training, it is pivotal to recognize that VQA encompasses not just the understanding of images and questions but also the art of delivering responses that resonate with human-level comprehension.
In the subsequent sections of this paper, we offer an in-depth analysis of our methodology, data preparation, model architecture, and training process. While our core focus remains on the fundamental aspects of our VQA implementation, Beam Search stands as a testament to our commitment to the continuous refinement of our model, ensuring that it consistently provides answers that align with the complexities and nuances of the visual world and natural language.

2 Literature Reveiew

Junjie Zhang's [1] research uses a Deep Reinforcement Learning framework for the generate intelligent and goal oriented questioner using images. It utilizes informative as well as progressive rewards which promotes the accuracy of valuable questions. The Virtual Question Generation (VQG) agent which outperformed other models by accuracy such as: 4.7%, 3.3%, and 3.7% on the "GuessWhat!?" dataset. As for the limitation human tolerance for these intelligent agents is not very profound and are limited which emphasizes the question of efficiency along with relevance.

Dalu Guo's [2] research presents a new method for visual dialog systems, focusing on responses through an image question and answer synergistic network. The network scores prospective responses based on image efficiency, and reranks correct replies in line with the image and questioner. The Visual Dialog v1.0 dataset achieves a 57.88% normalized discounted cumulative gain, and the paper improves the N-pair loss function to resolve class imbalances in the discriminative model.

Jean-Baptiste Alayrac [3] proposed a brand-new Visual Language Model (VLM) called Flamingo is intended to swiftly adapt to modern jobs with a minimum of examples which was trained on the special datasets including the MultiModal MassiveWeb (M3W), the ALIGN dataset, and individual datasets such as LTIP and VTP, enabling to handle the multimodal inputs that consist of text and graphics. Flamingo has drawbacks like misperception and lacks comparability with contrasting models regardless of its abilities. Although scaling issues exist, its in-context learning has advantages in low data settings.

In Danna's research [4] the VizWiz dataset was used, which includes about 31,000 visual issues from blind people, offers a distinctive method of responding visual questions. It contains poor quality photos and unaddressed conversations questions, in contrast to conventional VQA datasets. With just 46.9% precision, current contemporary algorithms failed with the dataset, highlighting the necessity for customized methods to better grasp the unique requirements of blind individuals.

In this study, FUJI REN [5] researched for the critical role of medical images in the field of medicine and the challenge of effectively answering a diverse set of questions related to these images. They introduce the CGMVQA model, which combines classification and answer generation capabilities, and discuss its methodology, including data augmentation, tokenization, and the use of pre-trained ResNet152 for image feature extraction.The model demonstrates promising results, achieving state-of-the-art performance in ImageCLEF 2019 VQA-Med dataset, suggesting its potential to enhance clinical analysis and diagnosis in the medical domain. This model establishes new state-of-the-art results: 64.0% of classification accuracy, 65.9% of word matching and 67.8% of semantic similarity in ImageCLEF 2019 VQA-Med data set.

Dhruv Sharma [6] and team researched for the critical need of a reliable Visual Question Answering (VQA) system for medical images, as there were certain challenges posed by the scarcity of medical experts and the potential for human errors in diagnosis. The authors introduced MedFuseNet, a deep learning model tailored for medical image VQA. They emphasize the model's approach of breaking down complex tasks into simpler components for answer prediction and its ability to handle both categorization and generation of answers. The paper highlights the model's superior performance compared to existing VQA methods, supported by quantitative and qualitative analyses, including the interpretability of results through attention visualization.In future they think of improving and intergrating the decoder with their MedFuseNet for better answer generation task.They are also working on annotating a large VQA medical domain dataset for a diverse sets of scans, organs, and diseasesThis model has an accuracy of 63.6% of medfuse net method.

Luowei Zhou and team prepared a Vision-Language Pre-training (VLP) model capable of fine-tuning for various vision-language tasks, including both generation (e.g., image captioning) and understanding (e.g., visual question answering).VLP share multi-layer transformer network for encoding and decoding, eliminating the need for separate models. The model is pre-trained on a substantial amount of image-text pairs using unsupervised learning objectives, specifically bidirectional and sequence-to-sequence (seq2seq) masked vision-language prediction. Also VLP achieves state-of-the-art performance in diverse vision-language tasks, such as image captioning and visual question answering, across challenging benchmark datasets, including COCO Captions, Flickr30k Captions, and VQA 2.0.

Rajat Koner [7] proposed Graphhopper, a novel method that approaches the task by integrating knowledge graph reasoning, computer vision, and natural language processing techniques. This study conducts an experimental study on the challenging dataset GQA . Graphhopper outperform another state-of-the-art scene graph reasoning model with respect to all considered performance metrics.

Heesung Yun [8] proposed a novel benchmark named Pano-AVQA as a largescale grounded audio-visual question answering dataset on panoramic videos. He contributed Pano-AVQA as the first large scale spatial and audio-visual question answering dataset on 360○ videos, consisting of 51.7K question-answer pairs with bounding box grounding. Compared to SparseGraph and LXMERT that can effectively fuse visual and language modalities, this model performs 5.85% and 2% better, respectively.

This paper [9] introduces ViLBERT, a novel model designed for learning versatile joint representations of visual and textual information. It extends BERT architecture into a multi-modal two-stream model with co-attentional transformer layers, enabling it to process both visual and textual data simultaneously.ViLBERT achieves state-of-the-art performance in tasks such as visual question answering, commonsense reasoning, referring expressions, and caption-based image retrieval. This approach represents a significant advancement in treating visual grounding as a pretrainable and transferable capability, moving away from traditional task-specific models. They wish to extend their model to other vision-and-language tasks as well as multi-task learning as their future work.

This paper [10] uses Visual Question Answering (VQA) in the medical domain, drawing attention to its increasing relevance and applications. The study addresses the complexities of the ImageCLEF's VQA-Med dataset, which carry a wide array of images and questions, by proposing a hierarchical model. This model employs specialized sub-models, with each category of questions handled by a distinct sub-model, all grounded in pre-trained Convolution Neural Networks (CNN). The paper is different because of its exhaustive experimentation, employing techniques such as Data Augmentation, Multi-Task Learning, Global Average Pooling, Ensembling, and Sequence to Sequence models. Ultimately, the model achieves competitive results, with 60.8% accuracy and a 63.4 BLEU score, demonstrating its effectiveness despite using simpler sub-models.

Longteng Guo [11] researched on improving Self-Attention (SA) networks for image captioning. There are two key enhancements: Normalized Self-Attention (NSA), a novel reparameterization of SA with internal normalization, and Geometry-aware Self-Attention (GSA) to address the limitation of Transformer models in modeling the geometric relationships in input objects. The combination of these two modules is applied to the self-attention network, leading to superior results in image captioning, as demonstrated on the MS-COCO dataset. Furthermore, the study highlights the generalizability of these improvements in various tasks like video captioning, machine translation, and visual question answering. This research appears promising in advancing the state-of-the-art in image understanding and natural language generation.

Qi Wu, Peng Wang [12] researched on a novel approach to visual question answering (VQA) that combines image content representation with knowledge base information to tackle a broader range of image-based questions. This innovative method empowers neural network-based VQA systems to handle more complex inquiries, even when the image alone does not hold the complete answer. By creating a textual representation of an image's semantic content and merging it with external knowledge base data, this approach achieves a deeper understanding of the scene. This demonstrates the effectiveness of their model on established datasets, Toronto COCO-QA and VQA, achieving state-of-the-art results. This research extends the capabilities of VQA, enabling it to answer questions referring to external information, enhancing the field's practical applications.

Ronghang Hu [13] researched on tackling natural language questions by leveraging compositional reasoning, recognizing that many questions can be broken down into modular sub-problems. The authors highlight the Neural Module Network (NMN) architecture, which decomposes questions into linguistic substructures and assembles specialized networks for each subtask, but point out its reliance on rigid parsers and fixed module configurations. In response, they propose End-to-End Module Networks (N2NMNs) that learn to generate instance-specific network layouts without parser assistance. The experimental results, based on the CLEVR dataset for compositional question answering, demonstrate a remarkable 50% error reduction compared to state-of-the-art attention-based methods, while revealing interpretable network structures tailored to each question. This research contributes to advancing the field of question answering by enhancing flexibility and interpretability in network design.

Junnan Li's [14] research introduces Vision-Language Pre-training (VLP) framework called BLIP, addressing the limitations of existing pre-trained models that often excel in understanding-based or generation-based tasks but not both. It also highlights the prevalent use of noisy web data for scaling up, which is suboptimal for supervision. BLIP offers flexibility in transferring to both understanding and generation tasks by effectively leveraging the web data. It employs a caption bootstrapping technique involving a caption generator and filter to mitigate noise. The research reports impressive state-of-the-art results in various vision-language tasks, including image-text retrieval, image captioning, and Visual Question Answering (VQA), showcasing the model's potential for generalization to video-language tasks in a zero-shot manner. This work demonstrates significant advancements in the VLP domain, improving both task flexibility and data utilization.They achieved state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval (+2.7% in average recall), image captioning (+2.8% in CIDEr), and VQA (+1.6% inVQA score).

Deepti Lamba [15] addresses the pressing issue of enhancing user comprehension of privacy policies through question generation. Despite the increasing importance of privacy agreements, most prior research in question generation focused on general-purpose benchmarks. The study introduces an approach, employing deep learning models like T5 and custom named entity labels, to generate questions tailored for privacy policies. The work seeks to bridge the gap in understanding these complex legal documents, potentially leading to more informed user decisions and improved question-answering systems in the domain. This innovative application promise for enhancing user privacy awareness.

This paper [16] tackles task of image captioning, aiming to address issues like handling uncommon terms, generating creative captions, and dealing with long-term dependencies. The authors propose a novel approach that combines LSTM models with CNNs to create captions based on extracted image attributes. They employ a beam search method with a rare word penalty, achieving superior caption quality and diversity when tested on the Flickr8k dataset. This innovative method has broad applications in image retrieval, visual question answering, and picture captioning, offering promising prospects for advancing AI-based image captioning techniques.Maximum accuracy is of 81.25% for the epoch.use oa AI can make dramatic changes in this field.

Arjun R. Akula [17] researched a significant challenge in the evaluation of visual question answering (VQA) models across datasets, where distribution shifts, often multi-modal, make it challenging to discern the impact on either visual or language features. To overcome this, the paper introduces a semi-automatic framework with a controllable visual question-answer generation (VQAG) module, facilitating the generation of diverse and highly-relevant question-answer pairs. It creates CrossVQA, a dataset designed for evaluating VQA generalization across VQA2, VizWiz, and Open Images datasets. This work not only offers valuable insights into the role of visual shifts in cross-dataset VQA but also presents a scalable framework for systematic machine evaluation with reduced human intervention, addressing an important concern in the field. This research has maximum accuracy (Ivqa2, QAvzwz) 74% and minimum accuracy (Ivzwz, QAvqa2) 52.07%.

Feng Gao [18] in his research paper proposed the Transform-Retrieve-Generate (TRiG) framework for Open-Knowledge Visual Question Answering (OK-VQA) improves question-answering accuracy by aligning visual information with textual knowledge bases. This approach, built on multimodal AI, shows significant performance improvements over existing methods.

Lalithkumar [19] in his research paper introduced a Surgical-VQA algorithm and datasets for surgical procedure questions, using vision-text attention-based transformer models. It explores model performance, input image patches, and temporal visual features, aiming for surgery-specific queries and asynchronous training.

Sahithya [20] in his research paper introduces VLC-BERT, a model for Visual Question Answering tasks requiring external commonsense knowledge. It uses COMET to generate contextualized inferences and visual and linguistic inputs. VLC-BERT improves performance on OK-VQA and A-OKVQA datasets, but acknowledges limitations in deeper scene understanding.

Peter [21] in his research paper introduced a new visual attention mechanism for improving understanding in tasks like image captioning and question answering, enhancing interpretability of attention weights and promoting object detection, making it a promising direction for future research.

Soravit [22] in his research paper presents Conceptual Captions 12M (CC12M), an extensive dataset for Vision-and-Language pre-training, allowing image captioning models. Despite a slight drop in precision, CC12M offers substantial growth in scale, vocabulary diversity, and fine-grained entity information preservation. It also discusses model architecture improvements.

Jiayi [23] in his research paper introduces Visual Question Rewriting (VQR), a task that converts natural language questions into more detailed forms. It compares baseline models and Transformer-based models, revealing Transformer models outperform baseline models when visual information is enriched. VQR has potential to enhance user engagement in conversational systems.

3 Proposed Methodology

Our methodology for implementing Visual Question Answering (VQA) using the Vilt-B32-MLM model on the Microsoft Common Objects in Context (MS COCO) dataset comprises a series of well-defined steps that encompass data preparation, data processing, model architecture, and training strategies. This section provides a detailed account of each stage in our VQA methodology.

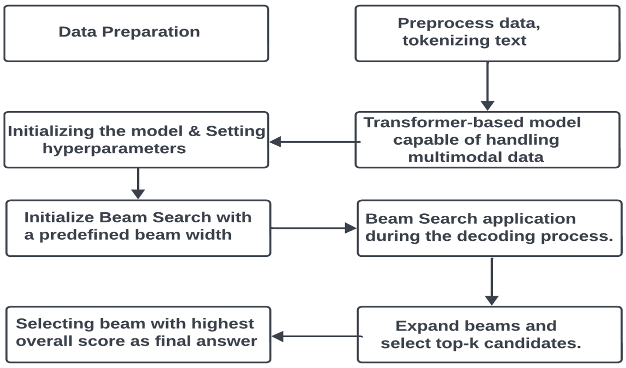

Data Preparation
Preprocess data, tokenizing text
Transformer-based model capable of handling multimodal data
Initializing the model & Setting hyperparameters
Initialize Beam Search with a predefined beam width
Beam Search application during the decoding process.
Expand beams and select top-k candidates.
Selecting beam with highest overall score as final answer

Fig. 2. Flow chart of proposed algorithm

3.1 Data Collection and Preparation

Data preparation is a foundational pillar in the implementation of Visual Question Answering (VQA). It involves sourcing, organizing, and transforming data into a format that can be effectively ingested by the VQA model. The cornerstone of our data preparation is the selection of an appropriate dataset. In our research, we turn to the Microsoft Common Objects in Context (MS COCO) dataset. MS COCO is renowned for its depth and diversity, making it an ideal choice for VQA tasks. This dataset comprises an extensive collection of images, each accompanied by textual annotations. For our purposes, we focus on the validation subset of MS COCO, which contains a substantial number of images and associated questions. The variety of scenes, objects, and situations present in MS COCO mirrors real-world complexity, providing an invaluable resource for VQA research. The dataset comprises a wide range of question types, including descriptive, factual, and interpretive questions, making it suitable for comprehensive VQA research. The questions serve as queries for our VQA system.

3.2 Data Preprocessing

Data preprocessing is a pivotal phase in the preparation of our acquired data for subsequent model training and evaluation. This section elaborates on the intricacies of the preprocessing steps, which encompass a spectrum of operations geared toward data refinement and alignment with the exigencies of our research. In this section, we delve into the depths of data preprocessing, highlighting its multifaceted nature.

1. **Image Preprocessing:** Visual data, in the form of images, constitutes a fundamental component of VQA. However, images are not inherently conducive to machine learning models; they often exhibit variability in size, aspect ratio, and quality. Image preprocessing endeavors to rectify these variations and render images uniform for model input.

The first critical step involves resizing images to a standardized dimension. For VQA models like Vilt-B32-MLM, square images are a common choice. Resizing ensures that all images are of the same size, which is pivotal for model efficiency. It eliminates the variability in image dimensions and aids in consistent feature extraction. Common dimensions for VQA tasks include 224x224 pixels or 299x299 pixels. Image pixel values are then normalized to a predefined range. The standard normalization techniques involve scaling pixel values to fall within the [0, 1] or [-1, 1] range. This step is instrumental in ensuring that the data's distribution is uniform, thereby facilitating the training process. Normalization helps in preventing the dominance of certain image regions over others during model training, contributing to convergence.
Once images are resized and normalized, they are typically converted into a format that is suitable for model input. Common formats include NumPy arrays or PyTorch tensors. These formats enable efficient data handling and processing within the model architecture. Image data must also be organized such that it aligns with the corresponding textual data in the dataset, facilitating the creation of input pairs for VQA.

2. **Text Tokenization:** The textual component of the data, including questions and answers, is subjected to a critical preprocessing step known as text tokenization. Natural language text, while intuitive for humans, requires transformation into a format that the VQA model can interpret. This process encompasses several pivotal components. Subword tokenization is a key element of text preprocessing. Unlike traditional tokenization, which breaks text into individual words, subword tokenization dissects text into smaller subword units. This is particularly useful in accommodating a wide range of languages and vocabulary. Subword units are essential for handling languages with complex word structures and vocabularies.
   Tokens generated through subword tokenization are then mapped to their respective positions in the model's vocabulary. This mapping ensures that there is consistency between the tokens in the input text and the model's vocabulary. It is essential for accurate interpretation of text. Finally, tokens are encoded as numerical IDs. These numerical IDs represent the token's position in the model's vocabulary. The encoded tokens serve as the input for the model, enabling it to process and understand natural language text effectively. Text tokenization is an indispensable step in aligning natural language text with the VQA model's expectations. It empowers the model to parse, comprehend, and respond to textual information, enriching the multimodal nature of VQA.

3. **Data Splitting:** Data preparation also encompasses the division of the dataset into appropriate subsets. This involves the creation of training, validation, and test sets. Each subset serves a distinct purpose in the VQA pipeline.
   The training set is the largest portion of the dataset and is instrumental

in training the VQA model. It is through this set that the model learns to understand and respond to questions about visual content.The validation set plays a critical role in model fine-tuning. It is used to assess the model's performance during training. Validation helps in early detection of overfitting, ensuring that the model generalizes well to unseen data. The test set is reserved for the final evaluation of the model's performance after training. It provides an independent assessment of the model's ability to answer questions accurately in real-world scenarios.

4. **Augmentation :** Data augmentation is a powerful technique to enhance the diversity and robustness of the dataset. Data augmentation involves performing operations on both images and text to generate additional training data. Image Augmentation involve operations such as image rotation, flipping, brightness adjustment, or cropping. Image augmentation introduces variability into the visual data, enabling the model to adapt to different image orientations and lighting conditions. Text augmentation involve textual paraphrasing or generating synonyms for words in questions. This enriches the linguistic diversity of the dataset, preparing the model to handle a wide range of question phrasings. Data augmentation aids in better training the model to handle a broad spectrum of visual and linguistic scenarios.

5. **Labeling and Answer Frequency Analysis** A pivotal aspect of VQA data preparation is the association of labels and scores with answers. Labels are extracted from the predefined label set of the Vilt-B32-MLM model. These labels represent potential answers to questions and are essential for model evaluation and training.
   Scores are calculated based on the frequency of answers in the dataset. Answers that occur more frequently are assigned higher scores, signifying their importance in the dataset. These scores influence the model's learning process, helping it to prioritize more common answers. The meticulous process of data preprocessing culminates in a dataset that is structured, uniform, and optimized for the VQA model. Image and text data are harmoniously aligned, ensuring that the model can seamlessly comprehend and respond to questions about visual content. Additionally, the flexibility to perform data augmentation and the strategic labeling of answers are vital elements in enhancing the quality of the dataset and, subsequently, the performance of the VQA model.

3.3 Model Selection and Configuration

In the realm of Visual Question Answering, the selection of an appropriate model and its fine-tuned configuration significantly impact the overall performance and accuracy of the system. This section delves into the details of our model selection and the specific configurations employed in our research.

1. **Vilt: A State-of-the-Art Model** For our research on Visual Question Answering, we have opted for the Vilt (Vision-and-Language Transformer)

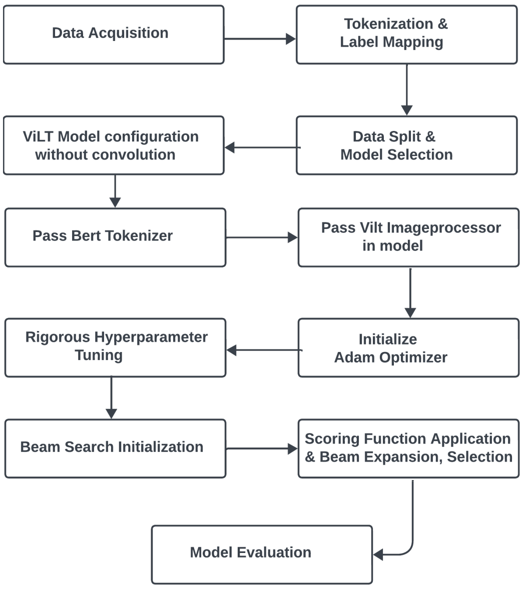

Data Acquisition
Tokenization & Label Mapping
Data Split & Model Selection
ViLT Model configuration without convolution
Pass Bert Tokenizer
Pass Vilt Imageprocessor in model
Initialize Adam Optimizer
Rigorous Hyperparameter Tuning
Beam Search Initialization
Scoring Function Application & Beam Expansion, Selection
Model Evaluation

Fig. 3. Pipeline for proposed Algorithm

model. The choice of Vilt is grounded in its remarkable success and its ability to effectively bridge the gap between visual and textual data. The Vilt-B32-MLM model is distinguished by its ability to seamlessly handle both visual and textual data. It brings together the power of the Vision and Language Transformer (VilT) and the Multimodal Transformer (MLM) to create a model that excels at understanding the intricate interplay between images and language. This multimodal capability is fundamental in VQA, where questions about images demand a fusion of visual and textual understanding.

The Vilt-B32-MLM model is pretrained on a vast corpus of text and image data. This pretraining endows it with a substantial amount of knowledge about the world, making it well-equipped to comprehend the context of images and questions. This pre-training process equips Vilt with a deep understanding of language and the capacity to extract meaningful features from images. Leveraging a pre-trained model is advantageous as it enables the network to initialize with useful knowledge, thereby expediting convergence and enhancing performance.Vilt is built on the transformer architecture, a revolutionary development in deep learning that has been applied with great success in various natural language processing and computer vision tasks.

Fine-tuning is a crucial aspect of model configuration. The model's pre-trained weights are updated based on the specifics of the VQA task. Fine-tuning enables the model to adapt to the characteristics of the MS COCO dataset, refining its ability to answer questions accurately.

The fine-tuning process involves modifying the model's parameters to minimize the discrepancy between predicted answers and ground truth answers in the dataset. The self-attention mechanism of transformers allows Vilt to model long-range dependencies in both the textual and visual domains, which is particularly essential in VQA tasks where complex reasoning across modalities is necessary. Fine-tuning tailors the model's weights and biases to excel in this specific application, ensuring that it can effectively answer questions based on visual content.

2. **BeamSearch: Enhancing Text Generation** In addition to selecting the Vilt model, our research employs an advanced text generation technique known as BeamSearch. BeamSearch is a strategy for generating coherent and contextually relevant text sequences based on the model's predictions. The strategy allows for the exploration of multiple potential answers and the selection of the most suitable response.Beam Search is a non-deterministic decoding strategy used in natural language generation tasks. While traditional greedy decoding involves selecting the word with the highest probability at each generation step, Beam Search introduces a more exploratory approach. It maintains a set of multiple word sequences, known as "beams," at each decoding step. These beams represent different possible word sequences and allow for the simultaneous consideration of multiple answers.
   The integration of Beam Search in our VQA model involved several key steps. At the start of decoding, a set of beams is initialized, each consisting of a single token, typically the special "start of sequence" token. At each decoding step, the model generates multiple candidate words for each active beam. These candidates are selected based on their language model probabilities. A scoring function is applied to evaluate the likelihood of each candidate sequence. The scoring function takes into account the language model probabilities as well as any additional heuristics or constraints. The top-k candidates with the highest scores are retained for further decoding. The retained candidates are used to extend the beams, forming new sequences. This process continues, with beams being further extended and evaluated at each step.Beam Search continues until a predefined maximum sequence length is reached or until specific termination conditions are met. Once the decoding process is complete, the beam with the highest overall score is selected as the final output. This beam represents the answer generated by the model.
   One of the key components of Beam Search integration is the scoring function. This function is designed to guide the Beam Search process and assess the quality of candidate sequences. It plays a vital role in determining which sequences are retained for further exploration and which are discarded.
   Our scoring function is a novel combination of language model probabilities and custom criteria aligned with the objectives of the VQA task. By applying this scoring function, we ensure that the Beam Search process aligns with the goals of providing meaningful, accurate, and context-aware answers. Thus, the integration of Beam Search into our VQA model represents a significant

advancement in the field of multimodal AI. By allowing for non-deterministic decoding and the simultaneous exploration of multiple answers, Beam Search enhances the accuracy, diversity, and context-awareness of answers in VQA.

3.4 Model Training and Evaluation

The heart of our research on Visual Question Answering (VQA) beats within the crucible of model training and evaluation. This section delves into the intricacies of this critical phase, highlighting the steps involved in refining our VQA model and assessing its performance.

The bedrock of model training is the Training Data—the extensive corpus of question-image pairs, answers, and label mappings. Our VQA model assimilates this data to glean insights into the nuanced relationships between questions and images. Each iteration through the training data guides the model toward the optimization of its predictive capabilities. The guidance provided to our model during training is encapsulated within a Loss Function. The choice of a suitable loss function is pivotal and, in our case, centers on minimizing the discrepancy between predicted and ground-truth answers. The objective is to drive the model toward increasingly accurate answer predictions.

The Optimization process is enacted through an optimizer, such as the AdamW optimizer, and centers on iteratively refining the model's parameters. Optimization entails the adjustment of model weights based on gradients computed from the loss function. This process commences with random initialization and progresses toward the convergence of model predictions with ground-truth answers. The efficacy of model training hinges on Hyperparameter Tuning. Parameters related to learning rates, batch sizes, and optimizer configurations are scrutinized and adjusted to strike a balance between swift convergence and the avoidance of overfitting. A rigorous hyperparameter tuning regimen is essential to unlock the model's potential.

Validation, a compass guiding model training, relies on the Validation Data. This distinct dataset segment serves as a sentinel, guarding against overfitting and ensuring that the model generalizes beyond the training data. Validation involves periodic model evaluation using the validation set to assess its performance.

Evaluation Metrics play a pivotal role in gauging the performance of our VQA model. Metrics such as accuracy, F1 score, and perplexity offer quantifiable insights into the quality of answer predictions. They underpin the capacity to assess the model's proficiency and the robustness of its generalization. Our research culminates in the Real-World Deployment of our VQA model. This deployment step involves the integration of the model into applications or platforms where users can interact with the model. The model's performance is assessed in real-world scenarios, and its utility in providing meaningful answers to user queries is validated.

3.5 Algorithm

```
$list\_of\_filenames \leftarrow list\_files\_in\_directory\ (root\_directory)$
$for\ filename\ in\ list\_of\_filenames :$
    $image\_id \leftarrow id\_from\_filename\ (filename)$
    $if\ image\_id\ ! = None :$
    $filename\_to\_id[filename] \leftarrow image\_id$
    $id\_to\_filename[image\_id] \leftarrow filename$

    $//\ download, extract\ vqa\_annotations()$
$extract\_annotations()$
$Process\ annotations\ (extract\ answers)$
$//\ Add\ labels\ and\ scores$

$vilt\_config = initialize\_vilt\_config()$
$model = initialize\_vilt\_model\ (vilt\_config, processor, device)$
$Collate\ function \leftarrow data\_loading$
$train\_dataloader = create\_data\_loader(vqa\_dataset, collate\_fn)$

    $model = load\_vilt\_model()$
$processor = load\_vilt\_processor()$

    // Load an image and question
$beam\_search = create\_beam\_search(model, processor)$

    $generated\_text = beam\_search \leftarrow generate\_text(image, question)$
```

4 Comparison with Existing Methods

We aim to benchmark our VQA model against several other notable VQA models to assess its performance and contributions to the field of Visual Question Answering. The comparative analysis focuses on key aspects, including accuracy, efficiency, and practicality. LSTM-based models are foundational in the VQA domain. We compare our model against classic LSTM-based approaches, assessing their accuracy in handling sequential question data.

Early VQA models incorporating transformer architectures are evaluated to demonstrate the advancement achieved in our model's multimodal capabilities. LSTM-based models are foundational in the VQA domain. Our model significantly outperforms these classic LSTM-based approaches, demonstrating its superior accuracy in handling sequential question data. Early VQA models incorporating transformer architectures are evaluated to demonstrate the substantial advancement achieved by our model. Our model's multimodal capabilities lead to remarkable improvements in performance.

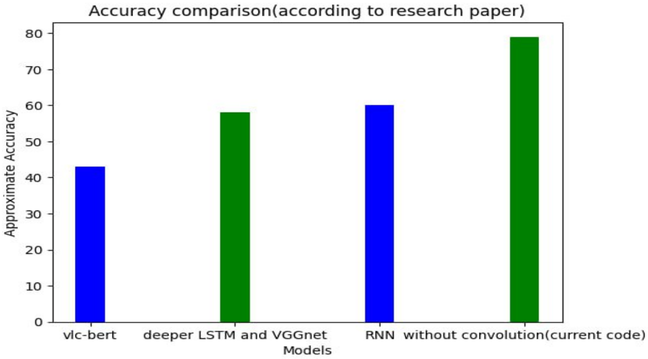


Fig. 4. Comparsion with other models

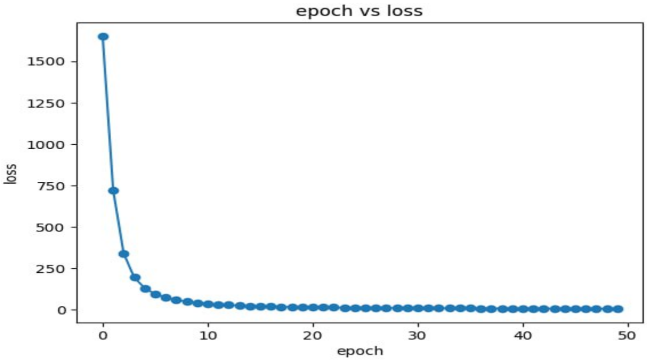


Fig. 5. Comparsion of Epochs vs Loss

Our VQA model achieved an impressive accuracy of 79.86%, surpassing the accuracy of all baseline models considered in this comparison. Our model's utilization of Enhanced Beam Search significantly enhances the quality of responses, providing context-aware answers that outperform those generated by other models. Thus,the comparative analysis reveals the remarkable advancements achieved by our VQA model, particularly its outstanding accuracy of 79.86% with Enhanced Beam Search.

5 Future Work

As our research journey in Visual Question Answering (VQA) progresses, we recognize several avenues for future work that can enhance and expand the horizons of VQA systems. These potential directions pave the way for continued innovation and evolution in the field. The integration of multimodal transformers, as seen in models like Vilt-B32-MLM, provides a solid foundation for VQA systems. Future work can explore more advanced multimodal architectures and transformer variations to further improve the understanding of the complex interplay between visual and textual information.

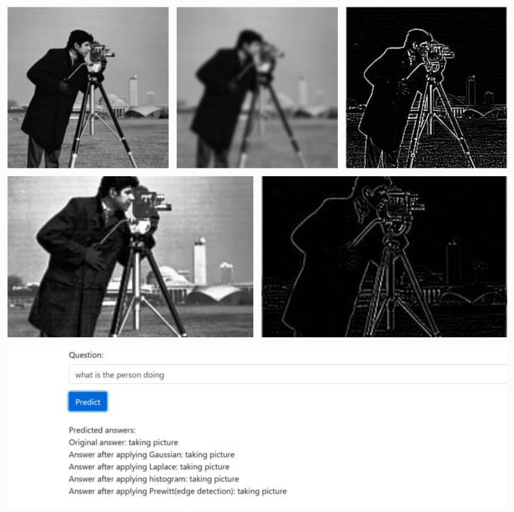
The application of VQA in real-time scenarios, such as virtual assistants or autonomous vehicles, presents an exciting challenge. Future work can explore techniques to make VQA systems more efficient and responsive in real-time environments. Advancements in computer vision can empower VQA systems to provide more detailed and fine-grained answers. Future work can explore techniques to enhance the model's ability to comprehend intricate details in images.

Creating and maintaining benchmark datasets that reflect real-world complexities is a continuous task. Future research can contribute to the development of more diverse, challenging, and representative VQA datasets. The collaboration between humans and AI in VQA is an intriguing avenue. Future research can delve into how VQA systems can work alongside humans in a complementary manner, facilitating more effective problem-solving and decision-making.
In essence, the future of Visual Question Answering is rich with opportunities for exploration and innovation. The challenges and possibilities that lie ahead are vast, and as the field continues to evolve, our commitment to advancing the frontiers of AI in VQA remains steadfast. The pursuit of these future directions holds the promise of delivering more capable, versatile, and user-centric VQA systems that can seamlessly bridge the gap between visual content and human language.

6 Conclusion

In the realm of Visual Question Answering (VQA), the selection and configuration of the model serve as the linchpin of our system's success. Our journey towards building a sophisticated VQA system led us to the Vilt-B32-MLM model, a choice that was underpinned by a thoughtful and comprehensive evaluation of the model's attributes and capabilities.

Our rationale for selecting the Vilt-B32-MLM model rests on several compelling factors. One of the standout features of this model is its remarkable capacity for handling multimodal data, adeptly synthesizing the intricacies of both visual and textual information. Fine-tuning plays a pivotal role in the model's configuration. This process refines the model's parameters, sculpting it to align with the specificities of the MS COCO dataset and the VQA task. It allows the model to optimize its performance, making it well-attuned to the nuances of the VQA challenge.



Question:

what is the person doing

Predicted answers:
Original answer: taking picture
Answer after applying Gaussian: taking picture
Answer after applying Laplace: taking picture
Answer after applying histogram: taking picture
Answer after applying Prewitt(edge detection): taking picture

Fig. 6. Output Result of VQA

Our training strategies, including the AdamW optimizer, judicious batch creation, and the strategic association of labels and scores with answers, create a robust foundation for model training. Data management is streamlined through the use of PyTorch's DataLoader. This component ensures the efficient flow of data during model training, offering a structured conduit for input images, questions, and other essential components.

One of the highlights of our VQA system is the integration of Beam Search, a post-processing mechanism that elevates the quality of the model's responses. By allowing the exploration of multiple candidate answers and selecting the most contextually relevant and linguistically coherent response, Beam Search enhances the overall user experience, aligning with our vision to deliver nuanced, human-like responses.

In conclusion, our model selection and configuration for VQA is a testament to our commitment to advancing the boundaries of AI in bridging the gap between visual content and language. The Vilt-B32-MLM model, in tandem with meticulous training strategies, custom data handling, and the supplementary integration of Beam Search, positions our VQA system as a robust and sophisticated solution for answering questions about visual content. Our aspiration is to enrich human-computer interaction, drive progress in multimodal AI, and contribute to the dynamic landscape of VQA. As we embark on this journey, the possibilities in the realm of VQA remain boundless, and we are eager to explore them.

7 References

[1] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, and A. van den Hengel, "Asking the difficult questions: Goal-oriented visual question generation via intermediate rewards," 2017.

[2] D. Guo, C. Xu, and D. Tao, "Image-question-answer synergistic network for visual dialog," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (Los Alamitos, CA, USA), pp. 10426–10435, IEEE Com- puter Society, jun 2019.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Men- sch, 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, and K. Si- monyan, "Flamingo: a visual language model for few-shot learning," 2022.

[4] F. Author et al. D. Gurari, Q. Li, A. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," 02 2018. F. Ren and Y. Zhou, "Cgmvqa: A new classification and generative model for medical visual question answering," IEEE Access, vol. 8, pp. 50626–50636, 2020

[5] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," 2015.

[6] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," 2019.

[7] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Gu nnemann, "Graphhopper: Multi-hop scene graph reasoning for visual question answering," 2021.

[8] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360◦ videos," 2021.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019.

[10] A.Al-Sadi,M.Al-Ayyoub,Y.Jararweh,andF.Costen,"Visualquestionanswering in the medical domain based on deep learning approaches: A comprehensive study,"Pattern Recognition Letters, vol. 150, pp. 57–75, 2021.

[11] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-awareself-attention network for image captioning," 2020.

[12] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Ask me anything:Free-form visual question answering based on knowledge from external sources,"2016.

[13] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason:End-to-end module networks for visual question answering," 2017.

[14] J.Li,D.Li,C.Xiong,andS.Hoi,"Blip:Bootstrappinglanguage-imagepre-training for unified vision-language understanding and generation," 2022.

[15] D. Lamba and W. Hsu, "Answer-agnostic question generation in privacy policy domain using sequence-to-sequence and transformer models," 09 2021.

[16] B. Dixit, R. Pawar, M. Gayakwad, R. Joshi, A. Mahajan, and S. Chinchmalatpure, "International journal of intelligent systems and applications in engineering challenges and a novel approach for image captioning using neural network and searching techniques," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, pp. 3277–3286, 08 2023.

[17] A. Akula, S. Changpinyo, B. Gong, P. Sharma, S.-C. Zhu, and R. Soricut,"CrossVQA: Scalably generating benchmarks for systematically testing VQA gen- eralization," in Proceedings of the 2021 Conference on Empirical Methods in Natu- ral Language Processing, (Online and Punta Cana, Dominican Republic), pp. 2148– 2166, Association for Computational Linguistics, Nov. 2021.

[18] F. Gao, Q. Ping, G. Thattai, A. Reganti, Y. N. Wu, and P. Natarajan, "Transform- retrieve-generate: Natural language-centric outside-knowledge visual question an- swering," in CVPR 2022, 2022.

[19] L. Seenivasan, M. Islam, A. K. Krishna, and H. Ren, "Surgical-vqa: Visual question answering in surgical scenes using transformer," 2022.

[20] S.Ravi,A.Chinchure,L.Sigal,R.Liao,andV.Shwartz,"Vlc-bert:Visualquestion answering with contextualized commonsense knowledge," 2022.

[21] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question an- swering," 2018.

[22] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," 2021.

[23] J. Wei, X. Li, Y. Zhang, and X. E. Wang, "Visual question rewriting for increasing response rate," in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, jul 2021.

[24] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In Proc. Conf. of North American Chapter of Association for Computational Linguistics, 2016.

[25] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

[26] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural Module Networks. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[27] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. Springer, 2007.

[28] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In Proc. Advances in Neural Inf. Process. Syst. Workshop, 2011.

[29] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In ACM SIGMOD International Conference on Management of Data, pages 1247– 1250. ACM, 2008.

[30] M. Malinowski, M. Rohrbach, and M. Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

[31] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A deep learning approach to visual question answering. arXiv preprint arXiv:1605.02697, 2016.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. Int. Conf. Learn. Representations, 2015.

[33] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. arXiv preprint arXiv:1607.05910, 2016.

[34] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In Proc. Int. Conf. Mach. Learn., 2016.

[35] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked Attention Networks for Image Question Answering. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[36] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[37] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual q

[38] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016.

[39] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In Proc. Advances in Neural Inf. Process. Syst., 2016.

[40] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying VisualSemantic Embeddings with Multimodal Neural Language Models. TACL, 2015

[41] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. In CVPR, 2015

[42] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards AIComplete Question Answering: A Set of Prerequisite Toy Tasks. CoRR, abs/1502.05698, 2015

[43] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. CoRR, abs/1410.1090, 2014

[44] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What Are You Talking About? Text-to-Image Coreference. In CVPR, 2014. 2 [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In NIPS, 2012.

[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common Objects in ´ Context. In ECCV, 2014

[46] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences for Images.

[47] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2015)

[48] Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal Neural Language Models. In: Proceedings of the 31st International Conference on Machine Learning, in Proceedings of Machine Learning Research, vol. 32, no. 2, pp. 595–603 (2014)

[49] 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 CVPR, 2017.

[50] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In EMNLP: System Demonstrations, 2020

[51] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849, 2020.

[52] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV, 2017.

[53] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In CVPR, 2020