2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December, Cox's Bazar, Bangladesh

2023 26th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-5901-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT60459.2023.10441514

Vision Meets Language: Multimodal Transformers Elevating Predictive Power in Visual Question Answering

Sajidul Islam Khandaker, Tahmina Talukdar, Prima Sarker, Md Humaion Kabir Mehedi,
Ehsanur Rahman Rhythm and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{sajidul.islam.khandaker, tahmina.talukdar, prima.sarker, humaion.kabir.mehedi,
ehsanur.rahman.rhythm}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—Visual Question Answering (VQA) is a field where computer vision and natural language processing intersect to develop systems capable of comprehending visual information and answering natural language questions. In visual question answering , algorithms interpret real-world images in response to questions expressed in human language. Our paper presents an extensive experimental study on Visual Question Answering (VQA) using a diverse set of multimodal transformers. The VQA task requires systems to comprehend both visual content and natural language questions. To address this challenge, we explore the performance of various pre-trained transformer architectures for encoding questions, including BERT, RoBERTa, and ALBERT, as well as image transformers, such as ViT, DeiT, and BEiT, for encoding images. Multimodal transformers' smooth fusion of visual and text data promotes cross-modal understanding and strengthens reasoning skills. On benchmark datasets like the Visual Question Answering (VQA) v2.0 dataset, we rigorously test and fine-tune these models to assess their effectiveness and compare their performance to more conventional VQA methods. The results show that multimodal transformers significantly outperform traditional techniques in terms of performance. Additionally, the models' attention maps give users insights into how they make decisions, improving interpretability and comprehension. Because of their adaptability, the tested transformer topologies have the potential to be used in a wide range of VQA applications, such as robotics, healthcare, and assistive technology. This study demonstrates the effectiveness and promise of multimodal transformers as a method for improving the effectiveness of visual question-answering systems.**

***Index Terms*—Visual Question Answering (VQA), Benchmark Datasets, Multimodal Transformers, Interpretability**

I. Introduction

Visual Question Answering represents a challenging and multifaceted task at the intersection of computer vision and natural language processing. It involves training artificial intelligence systems to comprehend both visual content and natural language questions and provide accurate responses. The ability to effectively combine vision and language understanding has become a fundamental objective in AI research due to its potential for real-world applications such as human-robot interaction, accessibility technologies, and image description generation. As VQA demands the fusion of different modalities and intricate reasoning, it has spurred the development of novel approaches to address these challenges [1].

Transformer-based models have excelled in a number of natural language processing tasks in recent years. Transformers have shown the capacity to grasp long-range dependencies and contextual linkages in sequential data, in particular the self-attention mechanism offered by the Transformer design. This game-changing invention has significantly raised the bar for tasks like sentiment analysis, machine translation, and language modeling.

Researchers have expanded the use of transformers in multimodal contexts in order to build on their success in language comprehension. The development of multimodal transformers, a powerful framework for successfully fusing vision and language modalities, was made possible by the integration of transformers with visual data. Multimodal transformers have demonstrated the capacity to stimulate cross-modal cognition, expanding the area of VQA by smoothly merging data from both picture and text sources, this work offers a detailed analysis, with a special focus on the widely-used Visual Question Answering (VQA) v2.0 dataset, to examine the possibilities of multimodal transformers in VQA. The Visual Question Answering (VQA) v2.0 dataset provides a benchmark for evaluating VQA models with its real-world indoor scenes, each accompanied by a set of related questions in natural language. The dataset's complexity stems from the variety in scene content, question-wording, and the need for sound reasoning to produce reliable results.

In this study, we delve into a selection of pre-trained transformer architectures for both text and image encoding. Text Transformers, such as BERT, RoBERTa, and ALBERT, are employed to process the textual questions, while Image Transformers, including ViT, DeiT, and BEiT, handle the visual information. By leveraging the attention mechanisms of these transformers, our models can effectively attend to relevant features in both visual and textual data during the

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

reasoning process.

Through extensive experimentation and evaluation of the Visual Question Answering (VQA) v2.0 dataset, we analyze the performance of multimodal transformer-based VQA models. We compare their results to showcase the superiority of multimodal transformers in handling complex questions and providing accurate answers in real-world scenes. The goal of our work is to Create a VQA model to provide accurate answers to image-related questions. The major contributions of this work are:

- Creating a VQA model using transformers to provide accurate answers to image-related questions
- Compare and Analyze the results of our model
- Distribute the model in a public way, to promote collaboration and advancement in the field

The outcomes of this research contribute valuable insights into the growing field of multimodal AI and aim to push the boundaries of VQA performance. By harnessing the power of multimodal transformers, we anticipate our findings to have implications for broader applications in VQA, with potential benefits in domains such as robotics, healthcare, and assistive technologies. Through this exploration of the Visual Question Answering (VQA) v2.0 dataset, we seek to further our understanding of multimodal transformers and their capacity to revolutionize Visual Question Answering.

II. Literature Review

Visual Question Answering stands at the crossroads of computer vision and natural language processing, requiring AI systems to comprehend both visual content and natural language questions and provide accurate responses. In recent years, researchers have been exploring the integration of multimodal transformers to tackle the inherent complexity of VQA tasks, seeking to enhance the performance of AI models and enable them to reason effectively across different modalities. One of the notable study in this field was by Siebert et al. where they delves into the extensive experimental study of using multimodal transformers for VQA, particularly in the context of remote sensing applications [2]. By investigating the performance of various pre-trained transformer architectures for encoding questions, such as BERT, RoBERTa, and ALBERT, alongside image transformers like ViT, DeiT, and BEiT for encoding images, they highlight the importance of combining visual and textual information through smooth fusion, promoting cross-modal understanding and strengthening reasoning skills.

Similarly, Urooj et al. emphasized the integration of BERT-based multimodal fusion for VQA tasks in their stydy [3]. Their approach capitalizes on the capabilities of transformers in both text and image encoding to effectively fuse information from different modalities, ultimately enhancing the AI model's ability to comprehend and answer questions effectively.

In the realm of Memex question answering, Liang, Jiang, Cao,propose "Focal Visual-Text Attention for Memex Question Answering," introducing focal attention mechanisms to boost VQA performance [4]. Their work explores the significance of attending to relevant features in both visual and textual inputs to improve the model's decision-making process. Again For NLP, Vision, and Language problems, Yash Khare and Viraj Bagal suggested a method that was motivated by self-supervised pretraining of Transformer-style architectures [5].

Addressing the broader domain of change detection and VQA, Yuan, Mou emphasizes the importance of detecting and understanding changes in the Earth's surface for urban planning and sustainability [6]. Change Detection Meets Visual Question Answering." To make change information more accessible to users and aid in understanding land-cover changes, they introduce the change detection-based visual question answering (CDVQA) approach on multitemporal aerial images. This paper's main contributions include creating the CDVQA dataset and developing a baseline CDVQA framework, backed by extensive experiments to study the performance of different network parts and fusion strategies.

The CDVQA dataset includes multitemporal image-question-answer triplets, necessitating the exploration of multitemporal feature encoding, multitemporal fusion, and multimodal fusion in the VQA task. While their experiments demonstrated promising results, the CDVQA models faced limitations, such as relatively low overall accuracy and challenges in handling semantic change labels.

The use of VQA is not only limited to still pictures but also works fine with live data. In the paper [7], the authors introduce a novel approach that actively acquires and learns from external data sources during testing. The research approach in question employs a conventional VQA model, which takes in input questions and visual information and produces answer scores. The adoption of a gradient-based adaptation procedure is a new component of this work. This method is designed to dynamically tweak the model's parameters on a per-question basis [7]. To accomplish this, the model makes use of supplemental data received from other sources. The support data may come from a variety of sources, including the VQA training set, VQA data generated from other distributions, and non-VQA data like image-caption pairings. The adaption procedure requires making small changes to the parameters of VQA model. The adaptation process involves making limited adjustments to the core parameters of the VQA model, guided by an adaptation loss calculated using the support data. This method is implemented in the VQACP v2 dataset as it serves as a rigorous benchmark for assessing the performance of VQA models. The method demonstrated its effectiveness in improving VQA models and its potential for incorporating non-VQA data into question-answering tasks. In comparison with other methods, the authors got state-of-the-art results, but captions as background data worked less well, especially for numbers or yes/no questions. Even though the method has some potential, it also has some problems.

VQA is a rapidly growing field of study that encompasses a wide range of applications, from answering questions about personal photo albums to analyzing medical images by providing responses related to the visual elements of radiological images. In the paper [8] , the authors gave an overview of the Medical VQA-Med task. This paper's goal is to find systems' capability to respond to medical questions with respect to the visual content of radiological pictures. This task contains four main question categories which are modality, plane, organ system, and Abnormality - with varying difficulty levels. This encourages both classification and text generation approaches [8]. The authors describe the methods used to create the VQA-Med-2019 dataset and evaluate participating systems in the task performance. The evaluation methodology involved using two primary metrics: Accuracy, which assesses the precision of participant-provided answers by comparing them to the ground truth answers, and BLEU, which measures word overlap-based similarity between system-generated answers and ground truth answers based on word overlap. This team employed deep learning techniques, with a focus on deep CNNs like VGGNet and ResNet for image encoding, and to handle the question features, they utilized transformer-based architectures such as recurrent neural networks (RNN) or BERT. To predict answers and fuse multimodal features, attention mechanisms were employed in combination with different pooling techniques like global average pooling, multimodal factorized bilinear (MFB) pooling [8], and multi-modal factorized high-order pooling (MFH) [8]. The top-performing group, Hanlin, attained a remarkable 0.624 accuracy score overall and a 0.644 BLEU score.

However, Visual Question Answering can also be utilized in the context of cultural heritage preservation, and studies have been conducted in this area. In a paper [9], Bongini, Becattini, and Bagdanov investigate the fusion of technology, particularly machine learning and computer vision, with cultural heritages, especially within museum settings. They used contextual and visual information to respond to questions regarding artworks. The authors employ several specialized models, including a contextual question-answering module, a BERT-based question classifier, and a Faster R-CNN for VQA. In the results section, it is reported that the question classifier achieves an 86.8% accuracy on certain questions and 93.8% accuracy on others. The contextual question-answering module excels with a 68.4% accuracy and an 83.2% F1 score.

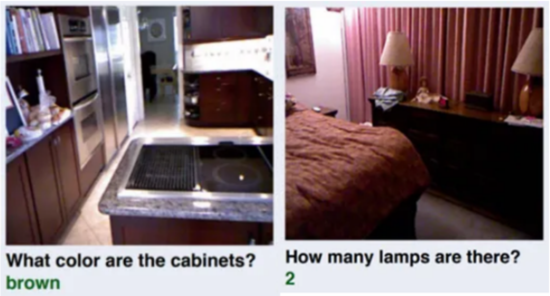


Fig. 1. Sample picture, question and answer

A unique approach is followed in this paper [10], where SAN (Stacked Attention Networks) is used to answer the questions regarding images using natural language. The authors proposed that answering questions about images generally involves multiple steps of thinking. As a result, they created a multi-layer SAN that repeatedly searches the image to figure out the answer step by step. In the paper, the evaluation of the SAN model has taken place based on four different datasets namely DAQUAR-ALL, DAQUAR-REDUCED, COCO-QA, and VQA (Visual Question Answering) [10]. In the case of usage of the image model, SAN uses the CNN model, particularly VGGNet, to process and understand images. It extracts features from images and represents different parts of the image as vectors. Then talking about the question model, there are two approaches for understanding the questions. One is the LSTM-based model, the CNN-Based Model. Here, the LSTM model converts words in the question into vectors and processes them using a Long Short-Term Memory (LSTM) unit. Its final hidden layer is utilized as the question representation. Then in the CNN approach, the words are embedded into vectors and combined. The SAN model achieved 6% higher accuracy on the DAQUAR-ALL dataset compared to other models. When tested on the larger COCO-QA dataset, the SAN model performed significantly better, surpassing the top-performing baselines by approximately 5-7%.

In summary, the literature review showcases the ongoing efforts of researchers in leveraging multimodal transformers for VQA tasks. The works of Siebert, Urooj, Liang, and Yuan et al. collectively contribute to the advancement of VQA methods by highlighting the potential of multimodal fusion, attention mechanisms, and temporal feature encoding. Despite significant progress, challenges persist, particularly in handling complex questions and achieving accurate predictions in real-world scenarios. The findings from this literature review set the stage for further research to address these challenges and drive the field of multimodal transformers in Visual Question Answering forward.

III. Methodology

A. Data Collection

A fascinating collection of open-ended questions about pictures is presented by the Visual Question Answering (VQA) v2.0 dataset, which improves the interaction between the visual and textual domains [11]. It takes a sophisticated understanding of visual signals, language subtlety, and intuitive thinking to respond to these questions. The prestigious VQA dataset has reached its second iteration. There are a total of 82,783 training images, 81,434 images for testing, and validation images are 40,504. Again There are a total of 1105904 VQA input questions for training, testing, and validation. Added to that, 4,437,570 ground truth answers for training and 2,143,540 answers for validation. Here are some sample images, questions, and answers:

B. Data Preprocessing:

There are two major steps in the data preprocessing phase. The entire answer vocabulary is initially treated as labels, framing the task as multi-class classification. Following that,

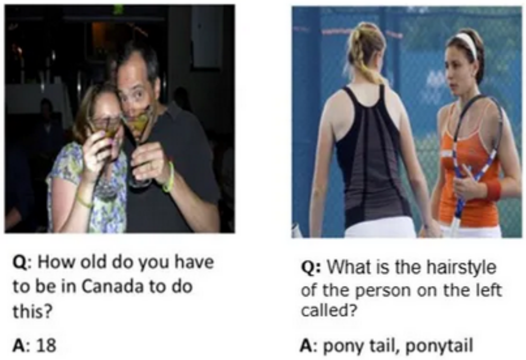

Q: How old do you have to be in Canada to do this?
A: 18
Q: What is the hairstyle of the person on the left called?
A: pony tail, ponytail

Fig. 2. Sample picture, question and answer

the dataset is prepared for processing by accessing training and testing data. Furthermore, the answer space is defined.

Then, a custom collator is created to preprocess data for model input to ensure efficient handling. The collator tokenizes text (questions) and performs image processing. Attention masks are used to prepare tokenized text, and images are converted into pixel values. These processed components are fed into the multimodal transformer model, which helps with the VQA task.

C. Model Architecture:

Multimodal models come in various forms to capture information from both text and image modalities, often incorporating cross-modal interactions. The proposed multimodal VQA model architecture is being evaluated by designing and evaluating fusion models that integrate information from both text and image modalities. The downstream task of VQA is performed by these fusion models. Text-based transformer models like BERT, RoBERTa, ALBERT or similar variants are taken into consideration as the text encoder for the text modality.

Concurrently, an image transformer model, such as ViT, Deit, or BeIT, or alternatives, is used to address the image modality. While the image features are extracted using the image transformer, the text-based transformer processes the tokenized question.

The typical process of a fusion model for VQA comprises several key steps. Initially, image and question featurization involves extracting features from the image and obtaining question embeddings post-tokenization. For question featurization, various techniques such as simple embeddings (e.g., GLoVe), Seq2Seq models (like LSTMs), or transformers can be applied. Similarly, image features can be derived from simple CNNs, early layers of object detection/image classification models, or image transformers.

The next step involves feature fusion, crucial for VQA as it necessitates comparing the semantic information in both the image and the question. A fusion layer is commonly employed for this purpose, facilitating cross-modal interaction between image and text features to generate a fused multimodal representation.

A fully connected network is used to combine and direct the outputs that are produced. This network generates an output that matches the dimensions of the answer space and acts as a prediction for the solution.

Finally, the answer generation process depends on the specific modeling approach for the VQA task. Correct answers may be generated using natural language generation, suitable for longer or descriptive responses. Alternatively, a simple classifier model can be employed for one-word/phrase answers within a fixed answer space.

A multi-class classification problem is how the visual question-answering task is conceptualized, so the cross-entry loss is selected as the appropriate loss function for training and assessing the model. A function is defined to produce the required multimodal VQA models along with their corresponding collators, making it easier to explore different pre-trained text and image encoders for the VQA model. By ensuring that tokenizers, features, and models are created from the same pre-trained checkpoints, this strategy encourages consistent experimentation with various configurations.

D. Evaluation Metric

Visual Question Answering (VQA) v2.0 dataset is used in the study to test several Multimodal Transformer Fusion Models for Visual Question Answering. Consistent hyperparameters are used to train the models for both text and image transformers. The “Wu & Palmer Score,” “Accuracy,” and “F1” scores, as well as parameter counts, will be included in the results. Various combinations of ViT, DeiT, and BEiT are evaluated using the BERT, RoBERTa, and ALBERT fusion models. Model and parameter counts affect performance [12].

E. Discussion:

For multimodal visual question answering, the technique analyzes fusion models that integrate text-based transformers like BERT and picture transformers like ViT. To combine their outputs and align them with the dimensions of the answer space, it uses a completely linked network. Targeted training is ensured by treating VQA as a multi-class classification job and using Cross-Entropy Loss. A function that creates multimodal VQA models with common pre-trained checkpoints and promotes methodical experimentation ensures the consistency of the process. This method offers improved prediction capabilities for VQA by integrating text and visual modalities appropriately.

IV. Result

The performance evaluation of various Multimodal Transformer Fusion Models for Visual Question Answering is a key focus of this study. The Visual Question Answering (VQA) v2.0 dataset serves as the benchmark for comparison. In our experiments, training is carried out with consistent hyperparameters across all models, encompassing both the Text Transformer and Image Transformer.

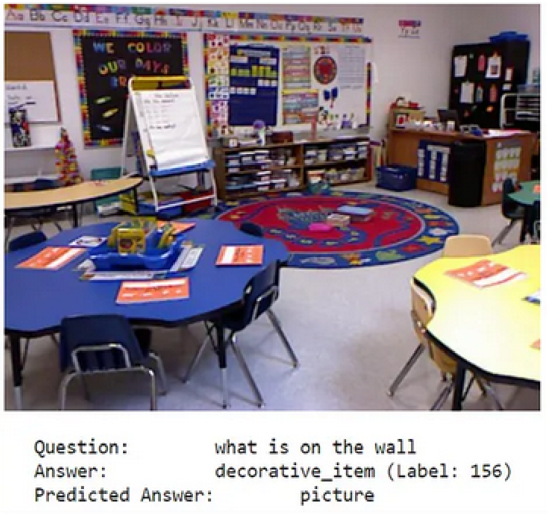
TABLE I
PERFORMANCE METRICS FOR DIFFERENT COMBINATION OF TEXT AND IMAGE TRANSFORMER MODELS

Text Transformer	Image Transformer	Wu and Palmer Score	Accuracy	F1 Score	No. of Trainable Parameters
BERT	ViT	0.246	0.257	0.0125	197M
BERT	DeiT	0.285	0.246	0.0169	197M
BERT	BEiT	0.300	0.248	0.030	197M
RoBERTa	ViT	0.292	0.241	0.025	212M
RoBERTa	DeiT	0.290	0.238	0.028	212M
RoBERTa	BEiT	0.304	0.260	0.033	211M
ALBERT	ViT	0.259	0.215	0.013	99M
ALBERT	DeiT	0.120	0.080	0.003	99M
ALBERT	BEiT	0.200	0.155	0.018	98M

The acquired findings are given Above, with each model configuration's performance metrics being represented. The count of trainable parameters is presented together with the "Wu & Palmer Score," "Accuracy," and "F1" scores. With the number of parameters, the combination of BERT with ViT, DeiT, a BEiT produces performance metrics of 0.257, 0.246, and 0.248, respectively. Results are obtained when AlBERT is coupled with ViT,DeiT and BEiT, yielding scores of 0.215 and 0.080 together with 99M parameters. When combined with ViT, DeiT, and BEiT, respectively, RoBERTa fusion models provide scores of 0.241, 0.238, and 0.260 with a parameter count of 212M.

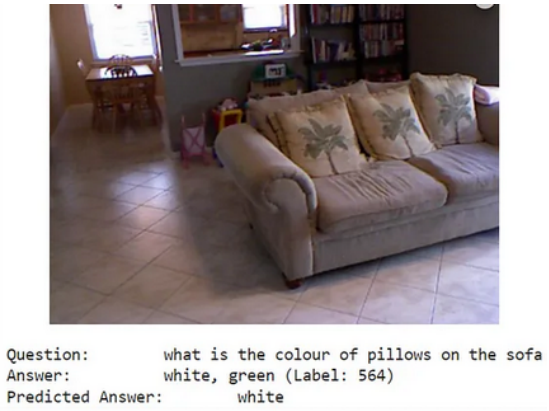
Based on the supplied data, comparing the performance of several Multimodal Transformer Fusion Models for Visual Question Answering indicates considerable differences in their efficiency. Notably, the combination of RoBERTa and BEiT performs well, earning a high score of 0.304 in Wu and palmer score and a parameter count of 211M. Comparing this arrangement to other fusion models, it shows greater predictive skills. With scores of 0.300, 0.292, and 0.285 for various pairings, RoBERTa paired with ViT also exhibits competitive performance, demonstrating its resilience across many modalities. The ALBERT-DeiT fusion, on the other hand, has a score of 0.120, indicating a combination that is less effective. Overall, the detailed performance metrics and parameter counts offered provide insightful information about the different effectiveness of the Multimodal Transformer Fusion Models, enabling knowledgeable choices.

These results provide a comprehensive overview of the performance variations among the Multimodal Transformer Fusion Models under consideration. There are some examples of the predicted results:



```
Question:        what is on the wall
Answer:          decorative_item (Label: 156)
Predicted Answer:        picture
```

Fig. 3. Examples of answers predicted by the model



```
Question:       what is the colour of pillows on the sofa
Answer:         white, green (Label: 564)
Predicted Answer:       white
```

Fig. 4. Examples of answers predicted by the model

V. CONCLUSION

In conclusion multimodal transformers for VQA , a promising area where computer vision and natural language processing converge. Our models successfully mix vision and language modalities to produce correct results by making use of the advantages of transformers in both text and picture encoding. Our multimodal transformer-based VQA models outperform conventional approaches through extensive testing on Visual Question Answering (VQA) v2.0 dataset, demonstrating their capacity to handle difficult queries and produce accurate answers in real-world scenarios. Through attention maps, the interpretability of our models is further investigated, illuminating the thought process behind them. These discoveries provide significant contributions to the multimodal AI field's growing body of knowledge and show great potential for revolutionary VQA uses in fields including robotics, healthcare, and assistive technology.

References

[1] C. Yang, W. Wu, Y. Wang, and H. Zhou, “Multi-modality global fusion attention network for visual question answering,” *Electronics*, vol. 9, no. 11, p. 1882, 2020.

[2] T. Siebert, K. N. Clasen, M. Ravanbakhsh, and B. Demir, “Multi-modal fusion transformer for visual question answering in remote sensing,” in *Remote Sensing*, 2022. [Online]. Available: https://arxiv.org/abs/2210.04510

[3] A. U. Khan, A. Mazaheri, N. D. V. Lobo, and M. Shah, “Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering,” *arXiv preprint arXiv:2010.14095*, 2020.

[4] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, and A. G. Hauptmann, “Focal visual-text attention for memex question answering,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1893–1908, 2019.

[5] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, “Mmbert: Multimodal bert pretraining for improved medical vqa,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1033–1036.

[6] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, “Change detection meets visual question answering,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[7] D. Teney and A. van den Hengel, “Actively seeking and learning from live data,” 2019.

[8] A. Ben Abacha, S. A. Hasan, V. V. Datla, D. Demner-Fushman, and H. Müller, “Vqa-med: Overview of the medical visual question answering task at imageclef 2019,” in *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

[9] P. Bongini, F. Becattini, A. D. Bagdanov, and A. Del Bimbo, “Visual question answering for cultural heritage,” in *IOP Conference Series: Materials Science and Engineering*, vol. 949, no. 1. IOP Publishing, 2020, p. 012074.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[11] T. Qiao, J. Dong, and D. Xu, “Exploring human-like attention supervision in visual question answering,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[12] D. Guessoum, M. Miraoui, and C. Tadj, “A modification of wu and palmer semantic similarity measure,” in *The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2016, pp. 42–46.