

# TAM at VQA-Med 2021: A Hybrid Model with Feature Extraction and Fusion for Medical Visual Question Answering

Yong Li<sup>1</sup>, Zhenguo Yang<sup>2</sup> and Tianyong Hao<sup>1</sup>(✉)

<sup>1</sup>*School of Computer Science, South China Normal University, China*

<sup>2</sup>*School of Computer Science, Guangdong University of Technology, China*

## Abstract

This paper briefly describes our model for the ImageCLEF Medical Visual Question Answering Task 2021 (ImageCLEF VQA-Med task 2021). Our method is based on a universal VQA framework and consists of image feature extraction module, question feature extraction module and feature fusion module. We employ the modified ResNet-34 as the backbone to construct an image feature extractor, which effectively extracts pixel-level features and enhances the model performance in a deep network. For question feature extraction, we firstly use word embedding to map question tokens to high dimension vectors, and then input them to a long-short-term memory (LSTM) to extract high-level question features. In addition, we leverage Multi-modal Factorized Bilinear Pooling (MFB) with a co-Attention mechanism to fuse these features to predict final answers. Our model achieves the accuracy score of 0.222 and bleu score of 0.255, ranking at the eighth among all participating teams in the VQA-Med task.

## Keywords

VQA, ResNet, LSTM, Co-Attention, MFB

## 1. Introduction

In recent years, the applications of deep learning in Computer Vision (CV) and Natural Language Processing (NLP) have gained remarkable progress. The development of deep learning in single modality facilitates researchers to explore multimodal studies, e.g., image-text retrieval [1], image captioning [2] and Visual Question Answering (VQA) [3]. These techniques have been applied to the domains of finance [4], traffic [5] and medical [6], which are all prosperous. When applying the VQA technique on medical domain, it can fulfill automatic interpretation of radiology images and make clinical decisions, thereby alleviating the shortage of medical resources.

In the ImageCLEF VQA-Med task 2021 [7], given a radiology image with a related question, the class of diseases indicated in the image is needed to be predicted. Compared to other ordinary benchmark datasets such as VQA [8], TDIUC [9] and Visual7W [10], the ImageCLEF VQA-Med task 2021 appears to be more challenging. The images of ordinary datasets contain abundant prior knowledge, such as the object labels including coordinates and category information. However, the images of ImageCLEF VQA-Med task 2021 dataset [11] do not contain object-level


---

*CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*

✉ yonglicutter@gmail.com (Y. Li); yzg@gdut.edu.cn (Z. Yang); haoty@m.scnu.edu.cn (T. Hao)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

labels. Besides, questions of ordinary VQA datasets usually have many entity names. On the contrary, the questions of the ImageCLEF VQA-Med task 2021 have no entity names of diseases. The intuitive comparison of data examples of common VQA tasks and the ImageCLEF VQA-Med task 2021 can be showed in Figure 1. In addition, the amount of data in the two kinds of datasets is dramatically different (Tens of thousands of the former and only 4500 of the latter).

As for the universal VQA benchmark datasets, we can utilize the object-level and pixel-level information of images as well as the entity information of questions for answer prediction. This prior knowledge can greatly enhance the performances of the VQA models. Due to the shortages of ImageCLEF VQA-Med task 2021 dataset, we utilize the modified ResNet-34 [12] as the image feature extraction module. The basic structure of ResNet-34 is convolution neural network (CNN), which can learn the data bias and pixel-level features through a small number of images. Besides, the residual structure can stabilize the information flow during training, which benefits to high-level feature extraction. We utilize long-short-term memory (LSTM) [13] to extract the question features from embedded vectors followed by the word embedding module. After that, Multi-modal Factorized Bilinear pooling (MFB) with co-Attention mechanism [3] is introduced to fuse the image features and question features. Finally, we predict the final answer through doing softmax on the fused features.



Question: What abnormality is seen in the image?  
Answer: pott's disease

(a) Data example of ImageCLEF 2021 VQA-Med dataset.



Question: What is the woman feeding the giraffe?  
Answer: Carrot

(b) Data example of common VQA tasks

**Figure 1:** Comparison of ImageCLEF VQA-Med task 2021 dataset and common VQA datasets.

## 2. Related Work

CNN has been widely used in image feature extraction throughout computer vision. Since LeNet [14] was introduced to extract image features, there have been more and more CNN variants (AlexNet, VGG, GoogleNet) applying in computer vision tasks. With the solutions to gradient disappearing (residual learning and dense learning [15]), deeper CNN can be constructed to promote the model performances on different vision tasks. Recently, many researchers have focused on utilizing transformer [16] as the image features encoder and gained remarkable performances. In transformer, the input images are split into patches and treated as sequences, and then input to the transformer for feature extraction. However, this requires a large number of training data to learn the distribution of the datasets, which is not effective in this task.

The researches of NLP tasks have been greatly promoted by the proposal of transformer. Transformer introduces the full-connected layer (FC) to build the self-attention mechanism, which replaces RNN to learn contextual information of a long-length sentence. Later, Devlin J et al. [17] proposed Bert by stacking encoders of Transformer. With a deeper structure, Bert adds word embedding, segment embedding and position embedding together as input to reach better performances in NLP tasks. In the next years, XLNet [18], GPT-2 [19], GPT-3 [20] were proposed to further promote the performances in NLP tasks. However, both of these models require large amount of training data to fit the distribution of the input text. For the LSTM [13], it can well preserve the contextual features of long-time sequences with a small number of texts. Therefore, we use LSTM as our question feature extractor instead.

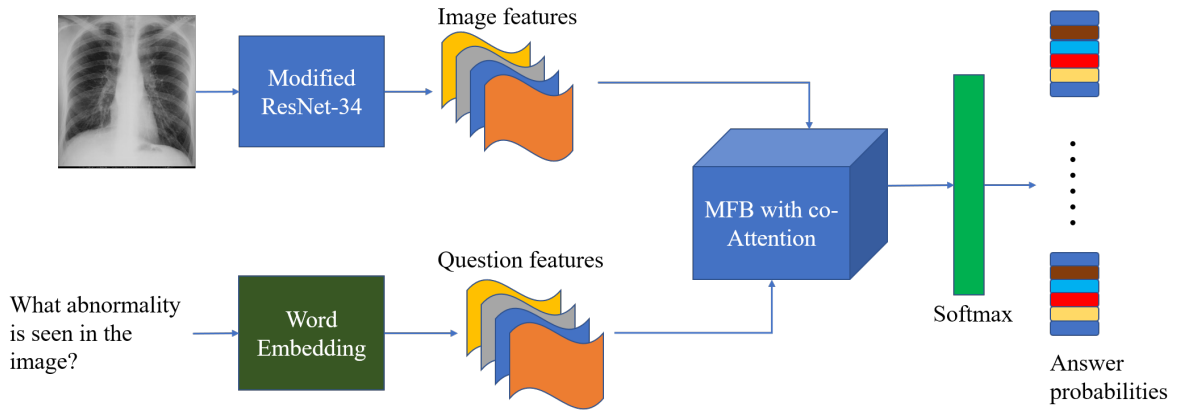
The simple methods of feature fusion in deep neural networks include concatenating the different kinds of feature in channel dimension, making element-wise sum or producting with same feature map sizes. But these might not be expressive enough to fully capture the complex associations between the two different modalities. The Multimodal Compact Bilinear pooling (MCB) [21] projects the images and text representations to a higher dimensional space, and then convolves both vectors by using element-wise product in Fast Fourier Transform (FFT) space. The Multi-modal Factorized Bilinear Pooling (MFB) [22] introduces co-Attention mechanism to jointly learn both image and question attention. The co-Attention mechanism can effectively learn which regions are important for the images related to the questions.

## 3. Method

The overview of the architecture of our proposed model is shown in Figure 2. Our model consists of three components: A image feature extraction module, a question feature extraction module and an attention feature fusion module.

### 3.1. Image Feature Extraction Module

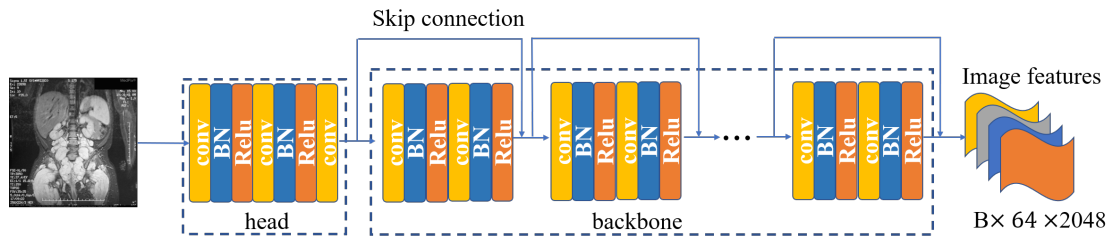
The main component of the original ResNet-34 is convolution neural network (CNN). With the characteristics of translation invariant, translation equivalence, scale invariance and rotation invariance, CNN can learn strong data bias with a small number of data. Besides, the residual structure has a large receptive field and it keeps gradient from vanishing in model training process. Benefitting from these strengthens, this image feature extraction module effectively



**Figure 2:** The overview of the architecture of our proposed model.

learns pixel features and spatial features of the medical images. Although it could be easy to overfit because of the small number of data, we alleviate this situation with a dropout operation. The output of the original ResNet-34 is a 1000-dimension vectors used for 1000 classification, which is not suitable in this task.

Compared to the original ResNet-34, we remove the global average pooling layer (GAP) and full-connected layer (FC), as showed in Figure 3. Before the images are fed to this module, each image is resized to  $128 \times 128$  with the INTER\_AREA algorithm. In order to fit the input size of the feature fusion module, we reshape the shape of output image feature maps from  $B \times 512 \times 16 \times 16$  to a new shape of  $B \times 64 \times 2048$ , where B represents the batch size in training. We take this reshaped feature maps as the extracted image features.



**Figure 3:** The structure of modified ResNet-34.

### 3.2. Text Feature Extraction Module

In this task, we introduce the word embedding and LSTM to extract the question features. Given the questions of the raw data, the preprocessing of them includes two steps: tokenizing the words of questions and fixing the length of sentences to 12. After that, the tokenized

sequences are sent to the word embedding module and generate the embedded word vectors in the dimension of  $B \times 12 \times 600$  (Note the tokens are encoded using GLOVE word embeddings). These embedded vectors are fed into the LSTM module to acquire high-level question features. The input layer dim and hidden layer dim are set to 600 and 1024 respectively. The number of LSTM unit in feature extraction module is set to 1. In this task, we use the whole sequence output features instead of the last token output features to guarantee the information integrity of sentence structure.

While inputting the word vectors to the LSTM, the forget gate of LSTM determines whether the information flows from previous moment can pass through to the next moment with a sigmoid function, which prompts LSTM to keep useful information and filter useless ones. The input gate of LSTM determines which information needs to be updated at current moment, and the output gate outputs updated information to next moment or as final output. With the gate units, LSTM has a strong ability in storing state information, which benefits to catch contextual correlation in long-time sequences.

### 3.3. Attentional Feature Fusion Module

After extracting the image features and question features, we feed them to the Multi-modal Factorized Bilinear Pooling (MFB) with the co-Attention mechanism together to obtain the fused features.

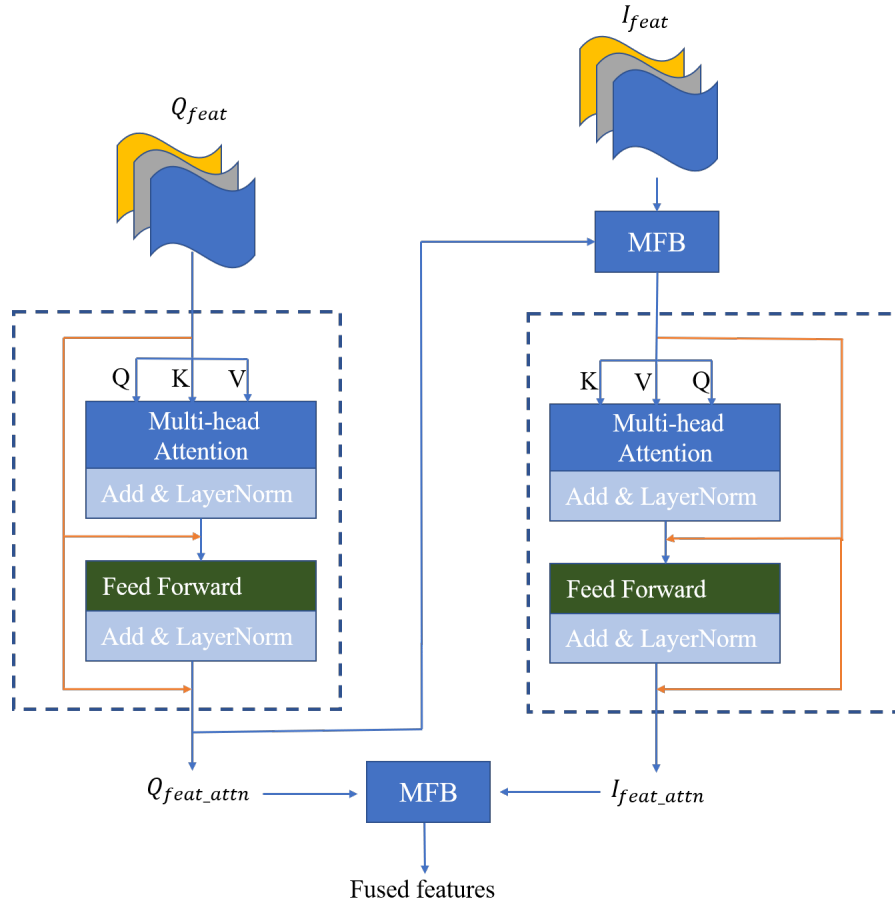
The MFB keeps the robust expressive capacity when compacts the features from different modalities by using the matrix factorization tricks. The co-Attention mechanism consists of a self-attention mechanism (SA) and a guided-attention (GA). Given two modalities features  $X$  and  $Y$ ,  $X$  first makes the self-attention operation to generate attention features, denoted as  $X_{attn}$ . Furthermore, the  $X_{attn}$  is used to guide the attention learning to obtain attention features of  $Y$ , denoted as  $Y_{attn}$ . This operation enhances the connection of two modalities in the learning process. By introducing the co-Attention mechanism to the MFB, the joint feature representation learning can be more accurate and effective.

In this task, we input the image features and question features (denoted as  $I_{feat}$  and  $Q_{feat}$ ) extracted from the modified ResNet-34 and LSTM to the MFB with the co-Attention mechanism, as showed in Figure 4. We firstly made a self-attention operation on  $Q_{feat}$  to obtain question attention features  $Q_{feat\_attn}$ , then we used the MFB to fuse the  $Q_{feat\_attn}$  and  $I_{feat}$  to guide the image features attention learning to generate image attention features  $I_{feat\_attn}$ . After that, we used the MFB to fuse the  $I_{feat\_attn}$  and  $Q_{feat\_attn}$  with vector multiplication and projected the fused features to a linear dimension. At last, we employed a softmax function on the fused features to predict the probability of each answer.

## 4. Experiments

### 4.1. Data Description

This dataset of ImageCLEF VQA-Med task 2021 contains a training set of 4000 image-question pairs, a validation set of 500 pairs and a test set of 500 pairs. The triplet data of images, questions and answers are one-to-one associated. The training set of ImageCLEF VQA-Med task 2021 is



**Figure 4:** The co-Attention mechanism in our model.

totally the same as the training set of ImageCLEF VQA-Med task 2020. The organizers replaced the samples of image, question and answer in validation set and test set in ImageCLEF VQA-Med task 2021. Therefore, we merged the validation set of ImageCLEF VQA-Med task 2020 with the training set of ImageCLEF VQA-Med task 2021 in order to extend the training set to 4500 paired image-question-answer data.

By analyzing the dataset, we found that the data can be divided into two types: open-ended, i.e., the answers are “yes/no” or the questions start with “Is/Does/...”, or close-ended, i.e., the questions do not have limited structures and could have multiple correct answers. Through statistics, there are 88 close-ended paired data and the rest of them are open-ended in the training set. However, there is no close-ended data in the validation set and test set. The total classes of answers are 332 in the training set and 236 (subset of training set) in the validation set. The statistical result is summarized in Table 1.

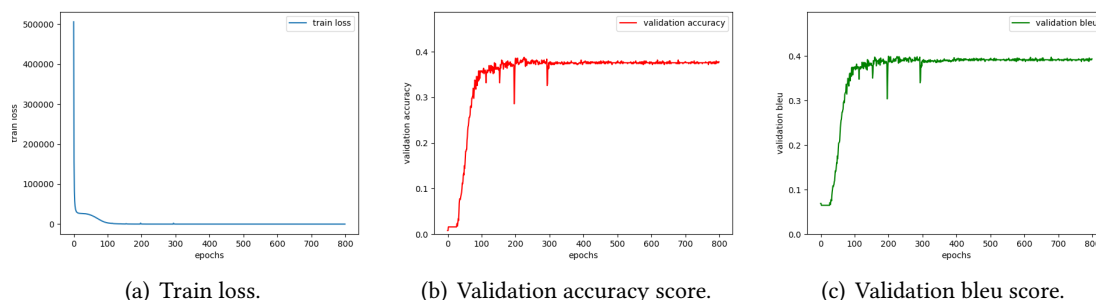
**Table 1**

Statistics of the ImageCLEF VQA-Med task 2021 dataset.

Dataset	Training set	Validation set	Test set
Open-ended	4412	500	500
Close-ended	88	0	0
Classes of Answers	332	236	Unknown

## 4.2. Optimization

Our model was optimized with BCE loss function on RTX2080 GPU devices by training 800 epochs. We trained the network from the scratch without using any pretrained model weights. We employed Adam with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  as the optimizer, and used dynamic learning rate to update the weights. In terms of convergence, we visualized the optimizations of the objective of our proposed model in Figure 5, from which we could observe that the training loss was decreasing, the accuracy score and bleu score were increasing. We utilized the model weight that generated the highest validation accuracy score as the final model weight. We submitted the result generated by this trained model on the test set and achieved the accuracy score of 0.222 and bleu score of 0.255. The top 10 result of this competition is shown in Table 2.

**Figure 5:** Optimization of our model

## 4.3. Discussion

An intuitive observation of Figure 5 (b) and (c) is that the performance of our method on validation set is much better than on test set. The reason could be summarized as follows: The sampling method of the test set was different from that of the training and validation set. Besides, in order to fit the input size of the image feature extraction model, the input images were resized to  $128 \times 128$  roughly, which might lose some spatial information and introduce the noises. In addition, the questions of the dataset provided little entity information of the related classes.

**Table 2**

Official results of ImageCLEF VQA-Med task 2021.

Participants	Accuracy	bleu
duadua	0.382	0.416
Zhao_Ling_Ling	0.362	0.402
TeamS	0.348	0.391
jeanbenoit_delbrouck	0.348	0.384
riven	0.332	0.361
Zhao_Shi	0.316	0.352
IALab_PUC	0.236	0.276
<b>TAM (ours)</b>	<b>0.222</b>	<b>0.255</b>
sliencec	0.220	0.235
sheerin	0.196	0.227

## 5. Conclusion

This paper describes the model designed in the ImageCLEF VQA-Med task 2021 competition. We proposed a modified ResNet-34 + LSTM + MFB with a co-Attention mechanism to predict final answers of VQA-Med. In image features extraction process, CNN was employed to learn the image bias and dropout function was introduced to suppress the over-fitting situation. For question feature extraction module, LSTM was utilized to extract the question features that did not rely on a large number of data. The two modalities features were fused by the MFB with co-Attention mechanism to generate the fused features for answer prediction. The best result of our model is 0.222 in accuracy score and 0.255 in bleu score.

## References

- [1] Q. Zhang, Z. Lei, Z. Zhang, S. Z. Li, Context-aware attention network for image-text retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3536–3545. doi:10.1109/CVPR42600.2020.00359.
- [2] Y. Feng, L. Ma, W. Liu, J. Luo, Unsupervised image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4125–4134. doi:10.1109/CVPR.2019.00425.
- [3] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290. doi:10.1109/CVPR.2019.00644.
- [4] G. Liu, Y. Mao, Q. Sun, H. Huang, W. Gao, X. Li, J. Shen, R. Li, X. Wang, Multi-scale two-way deep neural network for stock trend prediction (2020) 4555–4561. doi:10.24963/ijcai.2020/621.
- [5] A. Prakash, K. Chitta, A. Geiger, Multi-modal fusion transformer for end-to-end autonomous driving, arXiv preprint arXiv:2104.09224 (2021).
- [6] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, Overcoming data limitation in medical visual question answering, in: International Conference on Medical



Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 522–530. doi:10.1007/978-3-030-32251-9\_57.

- [7] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, V. Kovalev, S. Kozlovski, V. Liauchuk, Y. Dicente, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid, A. Popescu, J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021)*, LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.
- [8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433. doi:10.1109/ICCV.2015.279.
- [9] K. Kafle, C. Kanan, An analysis of visual question answering algorithms, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1965–1973. doi:10.1109/ICCV.2017.217.
- [10] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004. doi:10.1109/CVPR.2016.540.
- [11] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, H. Müller, Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain, in: *CLEF 2021 Working Notes, CEUR Workshop Proceedings*, CEUR-WS.org, Bucharest, Romania, 2021.
- [12] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: *European conference on computer vision*, Springer, 2016, pp. 630–645. doi:10.1007/978-3-319-46493-0\_38.
- [13] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.
- [14] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: *Proceedings of the IEEE*, volume 86, Ieee, 1998, pp. 2278–2324. doi:10.1109/5.726791.
- [15] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. doi:10.1109/CVPR.2017.243.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *arXiv preprint arXiv:1706.03762* (2017).
- [17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, *arXiv preprint arXiv:1906.08237* (2019).
- [19] K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, T. Nilsson, Gpt2: Empirical slant delay model for radio space geodetic techniques, *Geophysical research letters* 40 (2013) 1069–

1073. doi:10.1002/grl.50288.

- [20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).
- [21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, arXiv preprint arXiv:1606.01847 (2016).
- [22] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1821–1830.