SYSU-HCP at VQA-Med 2021: A Data-centric Model with Efficient Training Methodology for Medical Visual Question Answering

Haifan Gong (Co-first author), Ricong Huang (Co-first author), Guanqi Chen and Guanbin Li (Corresponding author)

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

Abstract

This paper describes our contribution to the Visual Question Answering Task in the Medical Domain at ImageCLEF 2021. We propose the method with a core idea that the model design and the training should best suit the feature of the data. Specifically, we design a hierarchical feature extraction structure to capture multi-scale features of medical images. To alleviate the issue of data limitation, we apply the mixup strategy for data augmentation during the training process. Based on the observation that there exist hard samples, we introduce the curriculum learning paradigm to resolve this issue. Last but not least, we apply label smoothing and ensemble training to avoid the model bias on the data. The proposed method achieves 1st place in the competition with 0.382 in accuracy and 0.416 in BLEU. Our code and model are available at https://github.com/Rodger-Huang/SYSU-HCP-at-ImageCLEF-VQA-Med-2021.

Keywords

Medical visual question answering, Classification, Curriculum learning, Mixup, Label smoothing, Ensemble learning

1. Introduction

The task of Visual Question Answering (VQA) is designed to answer the questions by understanding the intrinsic meaning of the corresponding images. By taking advantage of the larger-scale VQA dataset and the ingenious cross-modal feature fusion modules, researchers have made great achievements in the general VQA task. For the sake of promoting the patients' understanding of their disease and supporting the clinical decision, Hasan *et al.* [1] brought the VQA task to the medical domain. To facilitate the lack of the benchmark in the medical VQA, ImageCLEF organizes the 4th edition of the Medical Domain Visual Question Answering Competition named VQA-Med 2021. The examples of VQA-Med 2021 are shown in Figure 1. Since the data from medical VQA is relatively limited comparing to general VQA, we design a data-centric model to collect and make full use of the limited data. Besides, as the questions of VQA-Med 2021 reside in the abnormalities of the medical images, we mainly focus on the

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

gonghf@mail2.sysu.edu.cn (H. Gong); huangrc3@mail2.sysu.edu.cn (R. Huang); chengq26@mail2.sysu.edu.cn (G. Chen); liguanbin@mail.sysu.edu.cn (G. Li)

0000-0002-2749-6830 (H. Gong); 0000-0003-3268-7962 (R. Huang); 0000-0002-1440-3340 (G. Chen); 0000-0002-4499-3863 (G. Li)

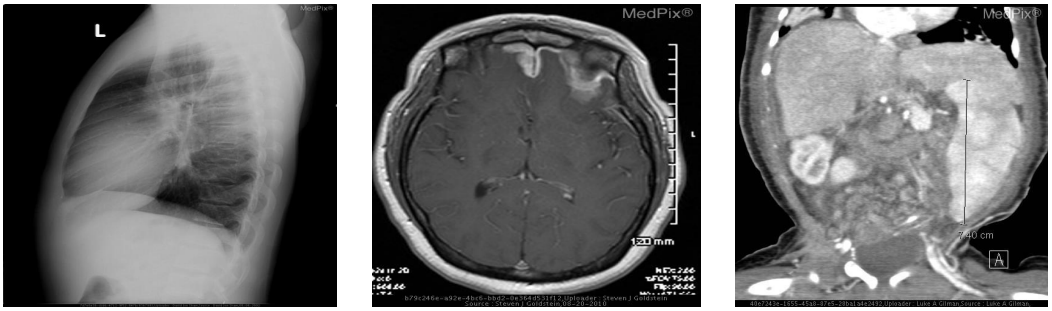

Figure 1: Three examples from the dataset of ImageCLEF 2021 VQA-Med.

design of visual representation to classify the abnormality of the medical images.

To address the issue of data limitation, we expand the dataset with the data in the previous VQA-Med competition (i.e., VQA-Med 2019 [2] and VQA-Med 2020 [3]). To make full use of the limited data, we apply mixup technology to create more samples. For efficient visual feature extraction, we apply label smoothing to stabilize the training progress. Furthermore, we discover the phenomenon that hard samples restrict the performance of the model and use a curriculum learning-based loss function to resolve this issue. Last but not least, to prevent the model from the infliction of intrinsic model bias, we propose to ensemble different types of models in exchange for higher model accuracy(e.g., VGG [4], ResNet [5]).

2. Related Work

A retrospective analysis is conducted in this section. We first literately review the research of general VQA, then we conclude the methods of VQA in the medical domain.

2.1. General VQA

The prevailing VQA framework in the general domain is mainly composed of four components: a visual encoder, a linguistic encoder, a cross-modal feature fusion module, and a classifier. The visual encoders are usually based on deep CNNs such as ResNet [5], VGG [4], Faster-RCNN [6], etc. To extract the linguistic feature, researchers apply the Transformer or RNN based models (e.g., Bert [7], LSTM [8]). The cross-modal feature fusion modules are dominant in current VQA systems. To capture the relationship between images and languages, Fukui *et al.* [9] and Kim *et al.* [10] apply the compact bilinear pooling methods. In the meanwhile, Yang *et al.* [11], Cao *et al.* [12], and Anderson *et al.* [13] investigate to design the model that focus on the question-related region of the image. MLP-liked classifier is usually used to select the final answer.

2.2. Medical VQA

Different from general VQA, medical VQA usually suffers from limited data and the distinction between common-sense knowledge and medical domain-specific knowledge. Thus, we first discuss the current methods to alleviate the data limitation in medical VQA. After that, we summarize the previous methods for the medical VQA.

Data limitation in medical VQA. Data limitation is an unavoidable topic in the field of medical image analysis, especially in the domain of medical VQA. To address this issue, Nguyen *et al.* [14] combine the meta-learning and the denoising auto-encoder to make use of large-scale unlabeled data. Nevertheless, they neglect the compatibility between the visual concept and the questions. Gong *et al.* [15] propose a novel multi-task pre-train framework, the image encoders of which are mandatory to not only learn the linguistic compatibility feature but also the vision concept by performing the original task (i.e., classification & segmentation) on the external dataset. Still, the easiest way to overcome the data limitation is to collect more data as was done by Chen *et al.* [16]. In this work, we not only collect the data from the previous VQA-Med competition but also apply the mixup strategy for data augmentation.

Previous methods on VQA-Med challenges. In the 2018 VQA-Med challenge [1], the top three groups applied the analogous pipeline as the VQA in the general domain. Specifically, they apply CNNs (i.e., ResNet-152 [5], VGG [4], Inception-ResNet-v2 [17]) for visual feature extraction, LSTM [8] or Bi-LSTM for language modeling, and attention based feature fusion modules (i.e., MFH [18], SAM [11], BAN [10]).

In the 2019 VQA-Med challenge [2], the leading three groups [19, 20, 21] used the Inception-Resnet-v2 [17] or a tailor-designed VGG [4] to extract image feature. Bert [7] is used to capture the semantic of the questions, and MFH [18] is applied for feature fusion. The tailor-designed VGG proposed by Yan *et al.* [19] replaced the conventional global average pooling layer of VGG with the hierarchical average pooling layer, which could efficiently capture the multi-scale feature. Besides, it is worth noting that the top 3 teams apply the question classification method to figure out the category of the questions.

In the 2020 VQA-Med challenge [3], two [22, 23] of the top three teams abandon the conventional VQA framework. Only Bumjun *et al.* [24] applies the conventional VQA framework, which uses VGG backbone for visual feature extraction, BioBert [25] for question encoding, and MFH [18] for feature fusion. The other two teams chose to direct classify the image with the deep neural networks. The winner of VQA-Med-2020 [22] designed a question skeleton-based approach to take full use of the linguistic feature, and integrate multi-scale and multi-architecture models to achieve the best result.

3. Datasets

In VQA-Med task [26] of ImageCLEF 2021 [27], the original dataset includes a training set of 4500 radiology images with 4500 question-answer (QA) pairs, a validation set of 500 radiology images with 500 QA pairs, and a test set of 500 radiology images with 500 questions. These questions focus on the abnormalities of medical images. Figure 1 shows three examples in the dataset.

Since the previous dataset [2, 3] in the VQA-Med competition is allowed to use, we leverage the

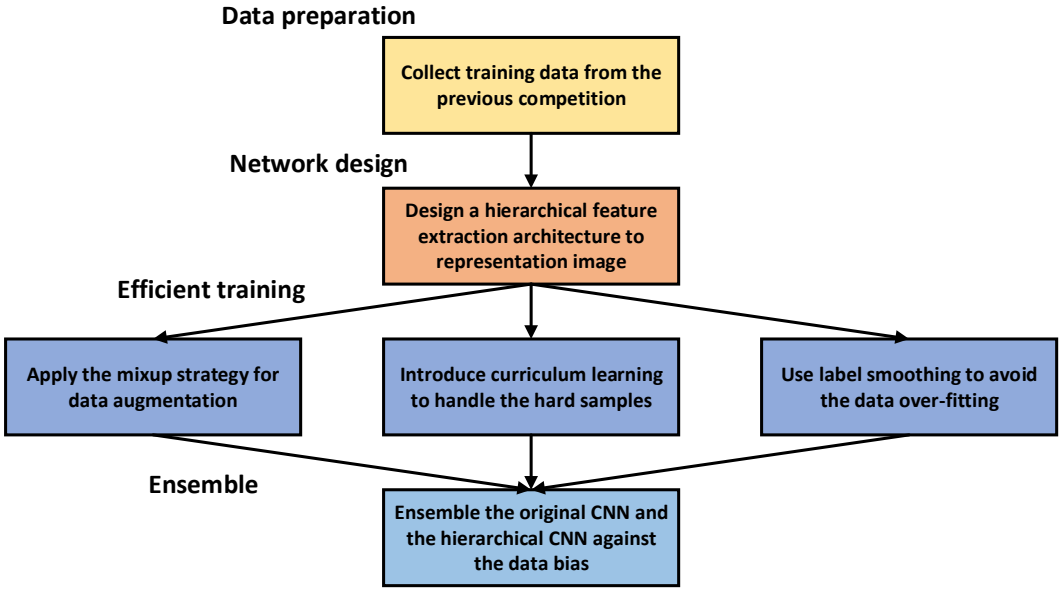

Figure 2: Overview of the proposed data-centric medical VQA framework. Followed by the instruction of the organizers, we collect the data from the previous competition. Then we design a hierarchical structure to better represent the feature of the image. After that, we adopt three efficient training strategies according to the characteristic of the data. Finally, we ensemble the conventional CNN and the hierarchical CNN towards eliminating data bias.

abnormality subset from the VQA-Med 2019, the test set of VQA-Med 2020, and the validation set of VQA-Med 2021 to extend the VQA-Med 2021 training set. The final training set is composed of 6183 VQA pairs.

4. Methodology

As this competition focus on the questions about abnormalities, we discard the conventional VQA framework and regard VQA as an image classification task. To make full use of the data, we design a data-centric model which is shown in Figure 2. This framework mainly consists of four parts: data preparation, network design, data-centric training methodologies, and model ensemble. The data preparation is illustrated in Section 3. Other parts are detailed below.

4.1. Network Architecture

Inspired by Yan *et al.* [19] and Aisha *et al.* [23] that multi-scale features contain more abundant information of medical images, we design a hierarchical feature extraction architecture to capture the multi-scale features of medical images. Different from the conventional high-level semantic feature representation architecture with fully-connected layers (Fig. 3 (a)), our

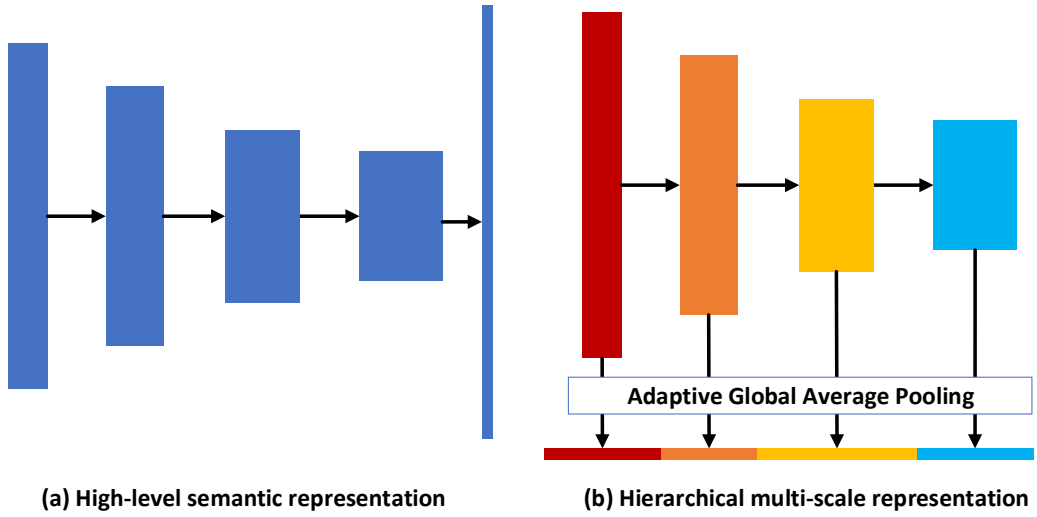

Figure 3: The conventional high-level semantic representation architecture and the proposed hierarchical multi-scale architecture.

proposed architecture replaces the fully-connected layers with hierarchical adaptive global average pooling layers (Fig. 3 (b)). Compared with a similarity work [19] that uses global average pooling to construct feature vector, the proposed method with adaptive global average pooling is more flexible to receive arbitrary input size of the image. This hierarchically adaptive global average pooling (HAGAP) structure is applied to ResNet-50 [5], ResNeSt-50 [28], VGG-16 and VGG-19 [4] to extract image feature.

4.2. Data-centric efficient training

In this part, we introduce three efficient strategies to better utilize the training data according to their characteristic.

Mixup. To alleviate the issue of data limitation in medical image representation learning, we adopt a simple yet effective data augmentation method called Mixup [29]. Given two samples (x_i, y_i) and (x_j, y_j), we create a new image $\hat{x}$ with label $\hat{y}$ by linear interpolation with the following operation:

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda) x_j \\ \hat{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{1}$$

where $\lambda \in [0, 1]$ is a random value drawn from the $Beta(\alpha, \alpha)$ distribution with the hyper-parameter $\alpha = 0.2$. It is worth noting that we only use the newly created images during the training process.

Curriculum learning. Based on the observation that in the training set of this competition, one disease could occur in various images modalities (e.g., CT, MRI.). Some modalities are of

numerous samples while others are of few. Thus, the imaging modality of the diseases that occur infrequently is hard to learn. Furthermore, the training set is unavoidable to contain noise. To resolve these issues, we introduce the idea of curriculum learning [30] into the training process. To simplify this process, we apply the *SuperLoss* [31], which automatically down-weights the hard samples with a larger loss.

Label smoothing. The label smoothing methodology is first proposed to train the Inception-V2 [32] network. It works by adjusting the probability of the target label by:

$$p_i = \begin{cases} 1-\varepsilon & \text{if } i = y \\ \varepsilon/(K-1) & \text{otherwise} \end{cases} \tag{2}$$

where ε is a small constant, K is the number of classes, and p_i denotes the possibility of category i. As label smoothing groups the representations of the examples from the same class into tight clusters, the model could achieve a better generalization ability.

4.3. Model Ensemble

As the model unavoidably contains bias, we apply multi-architecture ensemble to further improve the model performance. Comparing to the winner of VQA-Med 2020 [22] that takes more than 30 models to an ensemble, our best submission only contains 8 models by taking the advantage of the HAGAP structure. Specifically, the 8 models are ResNet-50, ResNeSt-50, VGG-16, VGG-19, ResNet-50-HAGAP, ResNeSt-50-HAGAP, VGG-16-HAGAP and VGG-19-HAGAP. The input size of ResNet-based networks is 256×256 while the input size of VGG based networks is 224×224. All backbones are initialized with the ImageNet pre-trained weight. Since the data is limited, we do not set the validation set during the training process and select the model of a fixed epoch for evaluation.

5. Experiments

5.1. Implementation details

As for training data, we leverage the data in Section 3. The models for our best submission are trained with the combination of mixup loss, SuperLoss, and Label smoothing loss. We used the SGD optimizer with momentum set to 0.9. The initial learning rate is set to 1e-3, and the weight decay is 5e-4. All models are trained for 60 epochs, and we select the model for inference on a fixed epoch (e.g., 50).

5.2. Evaluation

The VQA-Med competition applies accuracy and BLEU[33] as the evaluation metrics. Accuracy is calculated as the number of correct predicted answers among all answers. BLEU measures the similarity between the predicted answers and ground truth answers. As shown in Fig.4, we achieved an accuracy of 0.382 and a BLEU score of 0.416 in the VQA-Med-2021 test set, which won the first place in this competition.

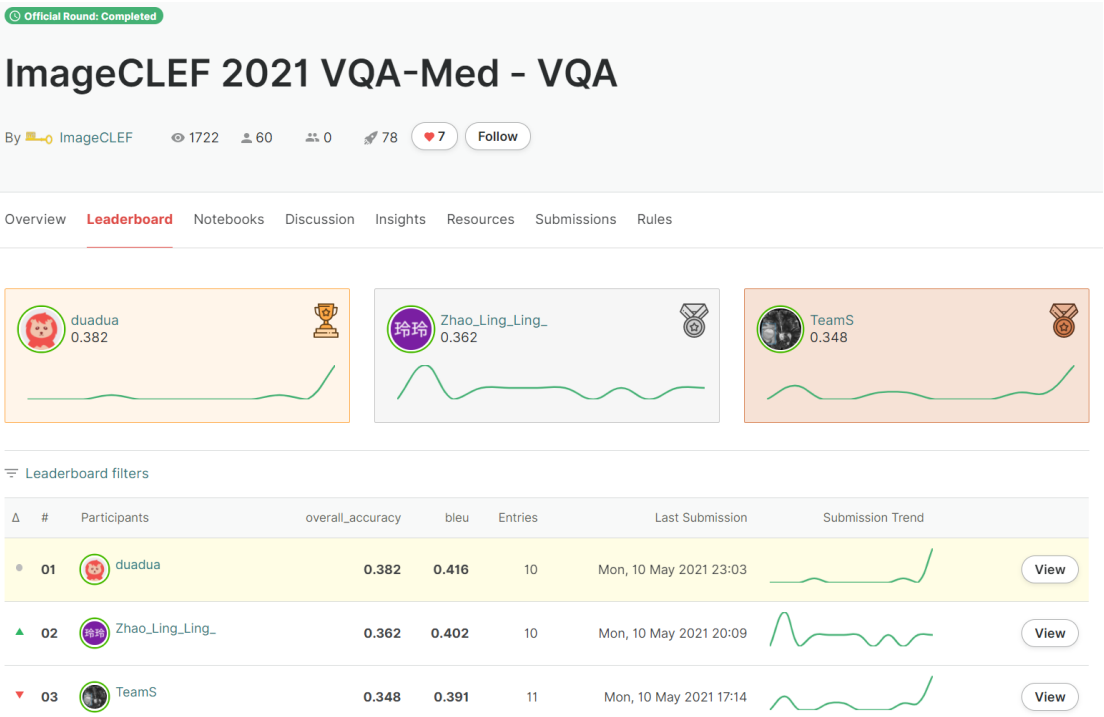

Figure 4: The leaderboard of the ImageCLEF VQA-Med 2021 challenge.

Table 1
Ablation study on the VQA-Med-2021 validation set.

Method	Accuracy	Improvement
VGG16	66.6%	-
+Mixup strategy	68.4%	+1.8%
+Label smoothing	68.8%	+0.4%
+HAGAP and Curriculum Learning	69.2%	+0.4%

5.3. Ablation study

To demonstrate the effectiveness of our proposed model, we conduct ablation study with the VGG-16 network, which is shown in Table 1. Specifically, the input image size is 224×224. The training set contains 5664 images while the validation set contains 500 images. The original VGG-16 network is set as the baseline, which achieves an accuracy of 66.6%. We utilize the mixup strategy for data augmentation, which surpasses the baseline by 1.8%. Furthermore, we adopt label smoothing to avoid the over-fitting of the model, which improves the accuracy to 68.8%. After that, we apply hierarchical architecture to represent the feature of the medical image, and introduce the curriculum learning paradigm into the framework, which brings an accuracy gain of 0.4%. With the efforts mentioned above, we achieve 69.2% accuracy on the

VQA-Med-2021 validation set.

6. Discussion

The VQA-Med is challenging task due to the limited data. As this work mainly focused on distinguish the abnormality between the medical images, we focus on designing the training scheduler and the feature extract module to make better use of the limited data. It is worth noting that as the training set, the validation set, and test set may not obey the same distribution, the Table 1 is of limited value. In other words, Label smoothing, HAGAP, and curriculum learning may be effective in the test set, but it not brings significant improvement on the validation set. For the same distribution inconsistent issue, we directly classify the images rather than use the long-tailed based methods [16]. Though we achieves 1 st place at this competition, our score is not high and there is still a long way to go to achieve applicable medical VQA.

For the future works in the medical VQA, we may digger deeper into better feature representation of image or words with the help of large amount unlabeled data. Besides, generating the answer word by word rather than regard the answer as a label is more valuable research topic.

7. Conclusion

In this paper, we describe our participation at the ImageCLEF 2021 VQA-Med challenge. Considering most of the questions are about abnormality, we abandon the conventional complex cross-modal fusion methodologies. With the firm brief that **the characteristics of the data should be fully considered in the construction of the model**, we design a data-centric model with efficient training strategies. Our proposed method achieves the best results among all participating groups with an accuracy of 0.382 and a BLEU score of 0.416.

References

[1] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, M. P. Lungren, Overview of imageclef 2018 medical domain visual question answering task, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

[2] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.

[3] A. Ben Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, H. Müller, Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[4] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning

Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[6] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[9] 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).
[10] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: Advances in Neural Information Processing Systems, 2018, pp. 1564–1574.
[11] Z. Yang, X. He, J. Gao, L. Deng, 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.
[12] Q. Cao, X. Liang, B. Li, G. Li, L. Lin, Visual question reasoning on general dependency tree, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.
[14] 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.
[15] H. Gong, G. Chen, S. Liu, Y. Yu, G. Li, Cross-modal self-attention with multi-task pre-training for medical visual question answering, in: ACM International Conference on Multimedia Retrieval (ICMR), 2021.
[16] G. Chen, H. Gong, G. Li, HCP-MIC at vqa-med 2020: Effective visual representation for medical visual question answering, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, 2020.
[17] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: S. P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 4278–4284.
[18] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp.

1839–1848.
[19] X. Yan, L. Li, C. Xie, J. Xiao, L. Gu, Zhejiang university at imageclef 2019 visual question answering in the medical domain, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.
[20] Y. Zhou, X. Kang, F. Ren, TUA1 at imageclef 2019 vqa-med: a classification and generation model based on transfer learning, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.
[21] M. H. Vu, R. Sznitman, T. Nyholm, T. Löfstedt, Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.
[22] Z. Liao, Q. Wu, C. Shen, A. van den Hengel, J. Verjans, AIML at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.
[23] A. Al-Sadi, H. Al-Theiabat, M. Al-Ayyoub, The inception team at vqa-med 2020: Pretrained VGG with data augmentation for medical VQA and VQG, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.
[24] B. Jung, L. Gu, T. Harada, bumjun_jung at vqa-med 2020: VQA model based on feature extraction and multi-modal feature fusion, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.
[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinform. 36 (2020) 1234–1240.
[26] 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.
[27] B. Ionescu, H. Müller, R. Péteri, A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, V. Kovalev, S. Kozlovski, V. Liauchuk, Y. Dicente, O. Pelka, A. G. S. 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, LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.
[28] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A. J. Smola, Resnest: Split-attention networks, CoRR abs/2004.08955 (2020).

arXiv:2004.08955.

[29] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).

[30] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on MachineLearning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009, volume 382 of *ACM International Conference Proceeding Series*, ACM, 2009, pp. 41–48.

[31] T. Castells, P. Weinzaepfel, J. Revaud, Superloss: A generic loss for robust curriculum learning, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2818–2826.

[33] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL, 2002, pp. 311–318.