

TeamS at VQA-Med 2021: BBN-Orchestra for Long-tailed Medical Visual Question Answering

Sedigheh Eslami^{1,2}, Gerard de Melo² and Christoph Meinel²

¹*D4L data4life gGmbH, Charlottenstraße 109, 14467 Potsdam, Germany*

²*Hasso Plattner Institute, Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany*

Abstract

This work describes our (TeamS) participation in the Medical Domain Visual Question Answering challenge (VQA-Med) at ImageCLEF 2021. We translate the VQA problem to long-tailed multi-class image classification for categorizing abnormalities present in medical images. Our proposed BBN-Orchestra is an ensemble of bilateral-branch networks (BBN) and successfully reduces overfitting to train and validation data in addition to effectively modeling the imbalanced long-tailed image distribution. BBN-Orchestra employs a voting mechanism to assign final predicted classes in the inference phase. Our proposed method achieved a test accuracy of 34.8% and a BLEU score of 39.1%, ranking 3rd in the competition. Our source code is available at <https://github.com/d4l-data4life/BBNOrchestra-for-VQAMed2021>.

Keywords

Medical visual question answering, Long-tailed visual recognition, Ensemble learning, Bilateral neural network

1. Introduction

Digitized medical data brings the potential to develop multi-modal tools such as Visual Question Answering (VQA) systems that can assist patients, clinicians, and radiology trainees in order to expedite patient care. Medical VQA systems are capable of answering questions about a given medical image and thereby aid in assessing and interpreting a radiology image. Despite this enormous potential, the development of medical VQA systems remains in its infancy due to complications arising from the scarcity of available training data, the distribution of this data, as well as the disparity between the natural language question and the medical image modalities. Recent progress has been driven by the VQA-RAD [1] and SLAKE [2] datasets, as well as the ImageCLEF initiative, which has been releasing VQA datasets extracted from PubMed Central articles and hosting challenges for developing task-oriented medical VQA systems. Still, due to the wide range of possible answers and the imbalanced distribution of the training data with respect to such answers, it is non-trivial to train a model to perform well on this task. For many sorts of answers, there are only very few example instances in the training data.


In this work, we investigate the effectiveness of deep learning approaches that overcome these challenges and are able to cope with long-tailed medical VQA data. Our BBN-Orchestra approach recasts the VQA task as a long-tailed multi-class image classification problem and

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

✉ sedigheh.eslami@data4life.care, hpi.de (S. Eslami); gerard.demelo@hpi.de (G. d. Melo); christoph.meinel@hpi.de (C. Meinel)



© 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)

learns an ensemble of bilateral-branch networks (BBNs) to better model imbalanced long-tailed training data and mitigate overfitting.

This paper presents our submissions at VQA-MED 2021 challenge [3]. We developed an ensemble model using Bilateral-Branch Networks (BBN) in order to simultaneously learn effective representative image features and achieve accurate classifiers considering the long-tailed class distribution. We further compare our results with single BBN using different backbone architectures. Our model was the 3rd ranked one in the VQA-Med 2021 Challenge, with 34.8% accuracy and 39.1% BLEU score on the test data.

2. Related work

Medical visual question answering is a challenging problem due to the diversity of the questions, diversity of the image data, as well as the insufficient and imbalanced annotated data distributions. This task has been previously investigated by classifying multi-modal features obtained by fusing encoded questions and images. Vu *et al.*[4] use pre-trained CNN models and Skip-Thought Vectors to encode the image and question, respectively, and combine them via attention mechanisms. Zhan *et al.*[5] enhance the attention-based fusion strategies via a novel question-type-specific conditional reasoning module that further highlights the important segments of the questions. Nguyen *et al.*[6] propose to use publicly available unlabeled image datasets in an unsupervised fashion using meta learning in order to enhance image features and thereby overcome data constraints. The winning team of the VQA-Med 2020 challenge [7] firstly maps similar questions into unified backbones in order to detect the type of the questions in a rule-based fashion. Afterwards, an ensemble multi-task classification network with ResNet, ResNext, VGG and MobileNet backbones is applied for image classification. Kovaleva *et al.*[8] utilize the MIMIC-CXR dataset to create the first publicly available visual dialogue dataset for radiology, which is not only useful for medical VQA, but also draws on the medical history of patients in order to better answer visual-based questions. In the VQA-Med 2021 challenge [3] hosted by ImageCLEF [9], we propose to solve the VQA problem purely by image classification, since the dataset consists entirely of questions sharing a common semantic interpretation, despite being presented in different syntactical forms: “What abnormality is present in the image?”. Inspired by the HCP-MIC team’s work [10], we adopt Bilateral-Branch Networks for classification in the presence of an imbalanced long-tail class distribution.

3. Approach

By exploring the datasets released in the challenge, we observe that the training data includes two general types of questions:

1. “yes/no” questions, asking about the presence of a medical abnormality in an image,
2. “what” questions, asking about the category of abnormality present in an image.

Although these questions appear in different syntactical forms, their semantics can be categorized into the two mentioned types and can be detected by simple rule-based mechanisms. Furthermore, we notice that the questions in the validation and test sets are only of the aforementioned “what” type. Therefore, we decided to translate the VQA setting in this specific

challenge to a multi-class image classification problem. Denote by $D_T = \{(x_i, q_i, a_i)\}_{i=1}^{n_T}$ the training dataset for a generic VQA model, where n_T is the number of samples in the training set, and x, q, a represent image, question, and answer, respectively. We can enumerate the set of answers and assume $a_i \in \{1, 2, \dots, C\}$, where C is the total number of candidate answers and is typically large. In this task, since all questions solicit the same kind of information, we relax the VQA problem to learn a function f that maps each x_i to a_i and thereby, classify the abnormality for each medical image.

BBN-Orchestra is an ensemble deep learning solution using Bilateral-Branch Networks (BBNs) [11]. As previous work [10, 11, 12] suggests, BBNs achieve effective results when classifying data with long-tail distribution, i.e., when a few classes form most of the data, whereas most classes have very few samples. BBNs consist of three main components:

1. a conventional network for effective representation learning,
2. a re-balancing network for modeling the tail class distribution by reverse sampling,
3. an adaptive cumulative learning component that controls how to shift the attention between the two former components during different epochs and train the classifier by minimizing the training loss.¹

Algorithm 1 BBN Orchestra: Train

Input: $D = \{(x_i, a_i)\}_{i=1}^n, K, backbone_type, criterion, n_epochs$

Output: K BBN models

```

1: procedure ORCHESTRATE( $D, K, backbone\_type, criterion, n\_epochs$ )
2:    $members \leftarrow \emptyset$ 
3:   for  $k$  in  $\{1, \dots, K\}$  do
4:      $train\_data, val\_data \leftarrow \text{random\_split}(D, val\_size = 0.1)$ 
5:      $C \leftarrow \text{num\_classes}(D)$ 
6:      $model \leftarrow \text{initialize\_BBN}(C, backbone\_type)$ 
7:     for  $epoch$  in  $\{1, \dots, n\_epochs\}$  do
8:        $model \leftarrow \text{train\_BBN}(model, train\_data, criterion, epoch)$ 
9:        $val\_acc, val\_loss \leftarrow \text{validate}(model, val\_data, criterion)$ 
10:      if  $val\_acc > best\_result$  then
11:         $best\_model \leftarrow model$ 
12:       $members \leftarrow members \cup \{best\_model\}$ 
  return  $members$ 

```

In BBN-Orchestra, we use ensemble multiple BBNs in order to prevent potential over-fitting with regard to the training and validation sets. To achieve this, the training and validation splits are combined first to form $D = D_T \cup D_V = \{(x_i, a_i)\}_{i=1}^n$, where D_V is the validation set and $n = n_T + n_V$.² We train K different BBN models with diverse backbone networks using different random splits from D . In the inference phase, the voting mechanism selects the most frequently predicted class by the K trained BBNs as the final predicted label for an unseen sample. Training and inference phases of BBN-Orchestra are summarized in Algorithms

¹For further information on BBNs, the reader is referred to the original paper [11].

²Since the training and validation sets are independent and have no intersection.

1 and 2, respectively.

Algorithm 2 BBN Orchestra: Inference

Input: $D' = \{(x'_i)\}_{i=1}^{n'}$, *members*
Output: $\{(x'_i, \hat{a}_i)\}_{i=1}^{n'}$

- 1: **procedure** PREDICT(D' , *members*)
- 2: $data_size \leftarrow \text{len}(D')$
- 3: $predictions \leftarrow \text{dict}()$
- 4: $\hat{a} \leftarrow \text{dict}()$
- 5: **for** *model* **in** *members* **do**
- 6: $predicted_labels \leftarrow \text{predict}(model, D')$
- 7: **for** d **in** D' **do**
- 8: $predictions[d].append(predicted_labels[d])$
- 9: **for** d **in** D' **do**
- 10: $\hat{a}[d] \leftarrow \text{most_frequent}(predictions[d])$

return (D' , \hat{a})

4. Experiments

4.1. Dataset

We conduct our experiments using the datasets released in the VQA-Med 2021 challenge. The train, validation, and test datasets respectively include 4 500, 500, and 500 images as well as question–answer pairs. This means that for each image, there exists one question–answer pair among all sets. Additionally, we exploited the train, validation, and test datasets from the VQA-Med 2019 challenge [13] to increase the amount of data available for training. Since both validation and test sets only include “what” questions asking about the type of abnormality in the image, we omit the “yes/no” questions via the simple rule of checking whether the answer is “yes” or “no” and retain only “what” questions. The final training set includes a total of 5 435³ training samples with 330 distinct answers.

4.2. Experimental Setup

4.2.1. Data setup and augmentation

In order to develop ensemble models, we combine the 5 435 train and 500 validation samples. In each iteration, 10% of the combined data is randomly selected to serve as a validation set and the rest is used for training. Similar to the original BBN experiments [11], we perform random resized cropping with size 224 and random horizontal flipping with probability 0.5 for data augmentation.

³Including VQA-Med 2019. Before mixing with the validation set from VQA-Med 2021.

Table 1

Accuracy scores on validation and test sets

	BBN ResNet 34	BBN ResNeSt 50	BBN Orchestra 1	BBN Orchestra 2	BBN Orchestra 3
Validation	59.8%	61.3%	54.4%	57.7%	55.9%
Test	29.9%	30.4%	32.2%	34.8%	32.8%

4.2.2. BBN-Orchestra setup

We evaluated three ensemble models:

1. Orchestra 1: $K = 4$ with ResNet34 [14] backbone⁴ in BBNs,
2. Orchestra 2: $K = 4$ with ResNeSt50 [15] backbone in BBNs,
3. Orchestra 3: $K = 8$ with 4 BBNs with ResNet34 and 4 BBNs with ResNeSt50 backbone.

All backbones are trained end-to-end from scratch with the medical VQA data described above. The adaptive parameter in BBN that controls the attention between learning universal features and the long-tail class distribution is directly proportional to the epoch number [11]. Hence, we aim to set the maximum number of epochs relatively high, namely 450, in order to give BBN a better chance to model the tail distribution.⁵ For all BBNs, we use cross-entropy loss. Stochastic gradient descent optimization is used with momentum 0.9 and a weight decay of 0.0004. The initial learning rate is set to 0.1 decaying at the 150th, 250th and 300th epochs by a factor of 0.1. All pipeline implementations are based on the PyTorch framework [16].

4.3. Results and Insights

The experimental results of our submissions are given in Tables 1 and 2. We provide the accuracy and BLEU evaluation scores, as they are the official evaluation metrics in the VQA-Med 2021 challenge. The reported values for the validation case of BBN-Orchestras are the averaged performance on different random validation splits over K models. In contrast, for single BBNs, the scores are computed using the original validation set released by the challenge.

The results show that all three orchestrated BBN models achieve better test accuracy in comparison to single BBN networks. The best performance on the test set is achieved by BBN-Orchestra 2. Comparing BBN-Orchestra 2 with a single BBN-ResNeSt 50 and BBN-Orchestra 1 with a single BBN-ResNet 34, we observe that the increase in test accuracy occurs while the validation accuracy decreases. This means that the orchestrated models were partially able to mitigate the potential overfitting with respect to the validation set. Furthermore, using the ResNeSt backbone performs better in comparison to ResNet 34. The reason for this is two-fold: 1. fewer layers in ResNet 34 leads to underfitting, 2. as shown in an empirical analysis [15], the Split-Attention mechanism of the ResNeSt architecture improves the performance of residual

⁴Both conventional learning and rebalancing branches.

⁵In early epochs, BBN focuses on learning the universal features and in later epochs, it learns to model the tail class distribution.

Table 2

BLEU scores on validation and test sets

	BBN ResNet 34	BBN ResNeSt 50	BBN Orchestra 1	BBN Orchestra 2	BBN Orchestra 3
Validation	63.1%	64.6%	57.3%	61.9%	59.3%
Test	33.2%	33.8%	35.8%	39.1%	36.6%

networks, e.g., the mean average precision of Cascade-RCNN is improved by 3% when using ResNeSt 50 instead of ResNet 50. With the same reasoning, Orchestra 3 achieves better results in comparison to Orchestra 1, since it benefits from ResNeSt, but cannot outperform Orchestra 2, since it also uses ResNet 34 models that underfit the data.

5. Conclusion

This work describes the submissions of our team (TeamS) to the VQA-Med challenge at ImageCLEF 2021. Considering the simplicity of the questions in the challenge datasets, we mapped the medical VQA problem to a multi-class image classification problem and mainly utilized Bilateral-Branch Networks to effectively address the resulting long-tailed abnormality classification task. In order to prevent potential overfitting, we developed BBN-Orchestra, an ensemble version of BBN. Our best submission exploited BBN-Orchestra with ResNeSt 50 backbone, which achieved 34.8% accuracy on the test data and ranked 3rd in the competition.

Acknowledgement

We would like to thank Matthias Steinbrecher for his helpful comments and discussions.

References

- [1] J. J. Lau, S. Gayen, A. B. Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, *Scientific data* 5 (2018) 1–10.
- [2] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, *arXiv preprint arXiv:2102.09542* (2021).
- [3] 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*.
- [4] M. H. Vu, T. Löfstedt, T. Nyholm, R. Sznitman, A question-centric model for visual question answering in medical imaging, *IEEE transactions on medical imaging* 39 (2020) 2856–2868.
- [5] L.-M. Zhan, B. Liu, L. Fan, J. Chen, X.-M. Wu, Medical visual question answering via conditional reasoning, in: *Proceedings of the 28th ACM International Conference on Multimedia, 2020*, pp. 2345–2354.

- [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.
- [7] 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, *CLEF*, 2020.
- [8] O. Kovaleva, C. Shivade, S. Kashyap, K. Kanjaria, J. Wu, D. Ballah, A. Coy, A. Karagyris, Y. Guo, D. B. Beymer, et al., Towards visual dialog for radiology, in: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 2020, pp. 60–69.
- [9] 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. 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.
- [10] G. Chen, H. Gong, G. Li, HCP-MIC at VQA-Med 2020: Effective visual representation for medical visual question answering, *CLEF*, 2020.
- [11] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.
- [12] Y. Liang, T. Qian, Recommending accurate and diverse items using bilateral branch network, *arXiv preprint arXiv:2101.00781* (2021).
- [13] A. B. 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: *CLEF (Working Notes)*, 2019.
- [14] L. Lei, H. Zhu, Y. Gong, Q. Cheng, A deep residual networks classification algorithm of fetal heart CT images, in: *2018 IEEE international conference on imaging systems and techniques (IST)*, IEEE, 2018, pp. 1–4.
- [15] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, *arXiv preprint arXiv:2004.08955* (2020).
- [16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems* 32, Curran Associates, Inc., 2019, pp. 8024–8035.