

PUC Chile team at VQA-Med 2021: approaching VQA as a classification task via fine-tuning a pretrained CNN

Ricardo Schilling¹, Pablo Messina¹, Denis Parra¹ and Hans Löbel¹

¹*Pontificia Universidad Católica, Chile*

Abstract

This paper describes the submission of the IALab group of the Pontifical Catholic University of Chile to the Medical Domain Visual Question Answering (VQA-Med) task. Our participation was rather simple: we approached the problem as image classification. We took a DenseNet121 with its weights pre-trained in ImageNet and fine-tuned it with the VQA-Med 2020 dataset labels to predict the answer. Different answers were treated as different classes, and the questions were disregarded for simplicity since essentially they all ask for abnormalities. With this very simple approach we ranked 7th among 11 teams, with a test set accuracy of 0.236.

1. Introduction

ImageCLEF [1] is an initiative with the aim of advancing the field of image retrieval (IR) as well as enhancing the evaluation in various fields of IR. The initiative takes the form of several challenges, and it is specially aware of the changes in the IR field in recent years, which have brought about tasks requiring the use of different types of data such as text, images and other features moving towards multi-modality. ImageCLEF has been running annually since 2003, and since the second version (2004) there are medical images involved in some tasks, such as medical image retrieval. Since then, new tasks involving medical images have been integrated into the ImageCLEFmedical challenge group of tasks [2], and that is how the task of medical visual question-answering (VQA) has been taking place since 2018. The goal of this task is as follows: given a medical image accompanied with a clinically relevant question, participating systems are tasked with answering the question based on the visual image content [3]. This task could help patients get a second opinion by an automated system, helping them better understand their conditions and supporting clinical discisions [2].


In this document we describe the participation of our team from HAIVis group ¹ within the artificial intelligence laboratory ² at PUC Chile (PUC Chile team) in the VQA-Med task at MedicalImageCLEF 2021 [2]. Our team earned the 7th place out of 11 in the challenge, and

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

✉ reschilling@uc.cl (R. Schilling); pamessina@uc.cl (P. Messina); dparra@ing.puc.cl (D. Parra); halobel@ing.puc.cl (H. Löbel)



© 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)

¹<http://haivis.ing.puc.cl/>

²<http://ialab.ing.puc.cl/>

our best submission was as simple as a DenseNet121 for visual input encoding to a classifier of answers, which did not take into account the questions and obtained a test accuracy of 0.236.

The rest of the paper is structured as follows: Section 1 presents relevant related work, Section 2 describes our data analysis, in section 3 we provide a description of our method, and section 4 describes our results. Lastly, in section 5 we conclude our article.

2. Related work

We addressed the task of medical VQA by modeling it as a classification task. Previously, there have been other attempts at medical image classification, such as the work of Kumar et al [4] that used an ensemble of fine tuned models to classify medical images, using the ImageCLEF 2016 medical image public dataset and achieving over 82% accuracy on its predictions.

Also the work of Li [5] which solved a similar problem but worked on lung images exclusively using the ILD dataset, reporting a significant improvement to regular image descriptors.

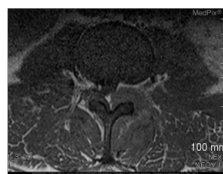
Similar works, such as Ren [6] also use a classification approach for visual question answering, in this case for the COCO-QA dataset, using a CNN together with an LSTM which predicts a single word.

Other similar approaches include the work of Yan, Xin, et al [7], using a VGG network with BERT[8] encodings for the 2019 Imageclef Med task.

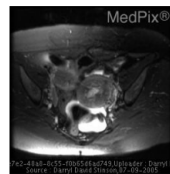
Unlike these approaches, we attempted to see how a regular CNN would work in a more general Visual Question Answering problem, where the answers are previously known and we rank them.

3. Dataset

Our dataset consisted of triples of {image, question, answer}, where the question asks about the possible existence of abnormalities in the image, and the answer provides a list of relevant abnormalities plus some possible additional details.



Q: what is most alarming about this mri?
A: spine epidural abscess



Q: what is most alarming about this mri?
A: bicornuate uterus

Figure 1: Example of dataset triples

Because in this task every question asks about the abnormalities present in the image, we ignored the questions and treated the problem as classification.

In the 2020 dataset[9], there are a total of 3940 medical images with associated answers. After some basic preprocessing, such as simplifying whitespaces, removing punctuation and so on,

we obtained a total of 325 distinct labels.

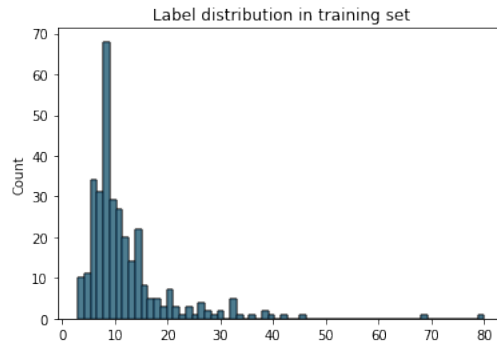


Figure 2: Distribution of labels in the training set

As Figure 3 shows, the amount of images per label varied, with the least frequent label appearing 5 times, and the most frequent label up to 80 times.

We also tried including data from the 2019 dataset[10], however this ended up adding over a thousand new labels, with a total of over 1400 distinct labels, while only doubling our amount of images.

4. Methods

We tried out two distinct approaches, and applied them to both the normal and extended datasets.

The best results were achieved with our first approach, where we finetuned a densenet-121[11] pre-trained on imagenet, to predict the image label.

We first treated each of our distinct labels as a different class and tasked the network with predicting the corresponding label, which we then mapped into the corresponding label. To achieve this, we changed the last layer of the pretrained model with three dense layers, size 1024, 1024 and a final classification layer with 325 neurons.

For this approach, the network performed slightly better using only the smaller dataset. We finetuned the network for around 100 epochs achieving around 30% accuracy and similar BLEU in the validation set. We used simple techniques such as early stopping, l2 normalization and dropout to avoid unnecessary overfitting of the networks.

We also tried using perceptual similarity[12] in an attempt to classify the image accordingly, however this approach failed to generalize properly and didn't manage to beat our initial approach.

This approach consisted in using a simple k-Neighbors Classifier to try to predict the corresponding image label, using an image vector obtained directly from the lpi module.

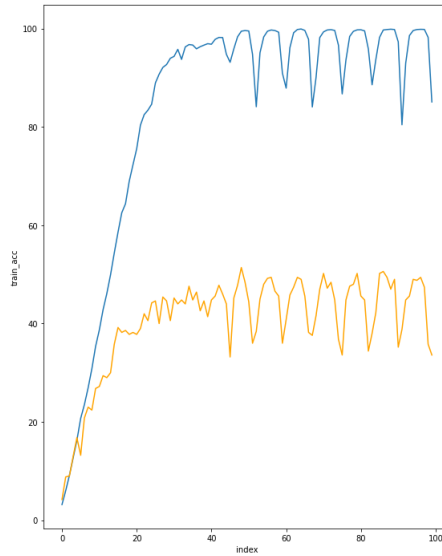


Figure 3: Early results of the evolution of model accuracy for the extended dataset.

5. Results

With our simple approach, we were able to achieve 30% accuracy and a similar BLEU score on the 2021 validation set[2], however we realized that adding the data from 2019 only decreased both metrics to around 20%.

The model’s performance on the test set ended up being slightly lower at a 23% accuracy and 0.276 BLEU score, which was expected and shows the model’s inability to generalize well.

As for the perceptual similarity model, unfortunately the results were very bad, not even making over 5% accuracy. This however, was expected as it was only an initial approach and this technique proved useful for some similar tasks.

6. Conclusion

Although our approach did not attain high accuracy, we believe it is simple enough and could serve as the base for a more complex model which includes actual text generation for better generalization. In the future, we plan to also include the question as input along the image to output the generated answer.

We strongly believe that this task can be distilled into a simpler classification or captioning problem, however we failed to make a model that could accurately determine the exact description of the input image. One of the reasons we believe this is the case is that our model does not currently generate new text, but could serve as the base of a bigger, more complex model.

Acknowledgments

This work was partially funded by ANID - Millennium Science Initiative Program - Code ICN17_002 and by ANID, FONDECYT grant 1191791.

References

- [1] 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. Stefan, 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.
- [2] 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.
- [3] 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: *CLEF (Working Notes)*, 2018.
- [4] A. Kumar, J. Kim, D. Lyndon, M. Fulham, D. Feng, An ensemble of fine-tuned convolutional neural networks for medical image classification, *IEEE Journal of Biomedical and Health Informatics* 21 (2017) 31–40. doi:10.1109/JBHI.2016.2635663.
- [5] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, M. Chen, Medical image classification with convolutional neural network, in: *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, 2014, pp. 844–848. doi:10.1109/ICARCV.2014.7064414.
- [6] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, 2015. arXiv:1505.02074.
- [7] X. Yan, L. Li, C. Xie, J. Xiao, L. Gu, Zhejiang university at imageclef 2019 visual question answering in the medical domain., in: *CLEF (Working Notes)*, 2019.
- [8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
- [9] 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: *CLEF 2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org*, Thessaloniki, Greece, 2020.
- [10] 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: *CLEF2019 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org* <<http://ceur-ws.org>>, Lugano, Switzerland, 2019.

- [11] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, 2018. [arXiv:1608.06993](#).
- [12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, 2018. [arXiv:1801.03924](#).