PUC Chile team at Concept Detection: K Nearest Neighbors with Perceptual Similarity

Gregory Schuit[1], Vicente Castro[1], Pablo Pino[1], Denis Parra[1] and Hans Lobel[1]

[1]*Pontificia Universidad Católica de Chile, Av. Vicuña Mackena 4860, Macul, 7820244, Chile*

Abstract

This article describes PUC Chile team's participation in the Concept Detection task of ImageCLEFmedical challenge 2021, which resulted in the team earning the fourth place. We made two submissions, the first one based on a naive approach which resulted in a F-1 score of 0.141, and an improved version which leveraged the Perceptual Similarity among images and obtained a final F-1 score of 0.360. We describe in detail our data analysis, our different approaches, and conclude by discussing some ideas for future work.

Keywords

Image Captioning, Concept Detection, Medical Artifitial Intelligence, Deep Learning, Perceptual Similarity, Convolutional Neural Networks

1. Introduction

ImageCLEF [1] is an initiative with the aim of advancing the field of image retrieval (IR) as well as enhancing the evaluation of technologies for annotation, indexing and retrieval of visual data. The initiative takes the form of several challenges, and it is especially 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 those versions, the ImageCLEFmedical challenge group of tasks [2] has integrated new ones involving medical images, with the medical image captioning task taking place since 2017. It consists of two subtasks: concept detection and caption prediction. Although there have been changes in the data used for the newest versions of the challenge, the goal of this task is the same: help physicians reduce the burden of manually translating visual medical information (such as radiology images) into textual descriptions. This year, the concept detection task within the ImageCLEFmedical challenge aims at identifying the presence of relevant biomedical concepts in medical images.

In this document we describe the participation of our team from the HAIVis group [1] within

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
gkschuit@uc.cl (G. Schuit); vvcastro@uc.cl (V. Castro); pdpino@uc.cl (P. Pino); dparra@ing.puc.cl (D. Parra); halobel@ing.puc.cl (H. Lobel)
https://yamadharma.github.io/ (G. Schuit)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]http://haivis.ing.puc.cl/

the artificial intelligence laboratory [2] at PUC Chile (PUC Chile team) in the concept detection task at ImageCLEFmedical 2021 [2]. Our team reached fourth place in the challenge, and our best submission was a combination of deep learning techniques to visually encode the medical images with a VGG convolutional neural network [3], followed by a KNN similarity search using Perceptual Similarity [4] rather than traditional cosine similarity.

The rest of the paper is structured as follows: section 2 briefly describes related work, section 3 our data analysis, and section 4 details of our proposed methods. In section 5 we describe our results, and finally in section 6 we conclude this article.

2. Related work

In previous versions of the competition, the best participants have used a wide variety of techniques, mainly based on Convolutional Neural Networks, Natural Language Processing, K-Nearest Neighbors and Clustering [5]. In the 2020 concept detection challenge, the best F1 score achieved was 0.394 by the AUEB NLP Group. This winning approach consisted of a variation of ChexNet and DenseNet-121 [6] with a feed-forward neural network as the classification head [7]. This team also won the second place in the competition. The second best group was PwC_Healthcare, with a F1 score of 0.3924. They submitted three approaches, using CNN and NLP techniques and clustering, achieving 3rd, 4th and 5th place. Looking at previous years, the best submissions achieved an F1 score of 0.1108 in 2018, 0.2823 in 2019 and 0.3940 in 2020 [5].

3. Data set inspection

The data set provided for this challenge consisted of 2,756 images of varying size for the training set and 500 for the validation set. Since the task is a multi-target classification problem, each sample has one or more labels. Figure 1 shows the distribution of number of labels per image in the training set.

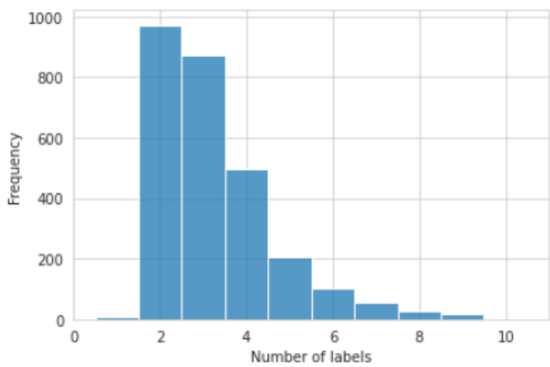

Figure 1: Number of labels per image

[2]http://ialab.ing.puc.cl/

Each label corresponds to a UMLS concept [8]. There are 1,315 different concepts in the training set and 817 in the validation set. Furthermore, only 547 concepts are present in both sets. This is worthy of consideration because it indicates an intrinsic skew in the dataset. We can represent this situation using Figure 2:

Figure 2: Number of different concepts on each data set split.

In the training set, the most frequent concept is C0040398 (Tomography and Emission-Computed), with 1,159 images. This corresponds to 42% of the images in the training set. The ten most frequent concepts can be seen in Figure 3.

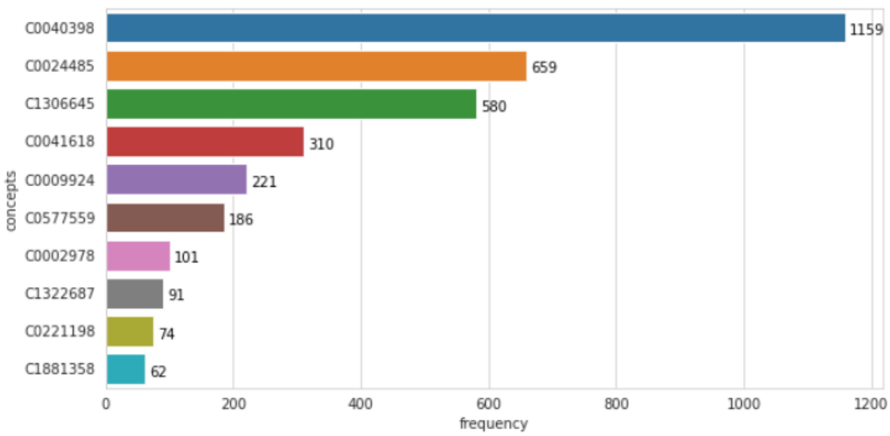

Figure 3: Top 10 most frequent concepts in the training set

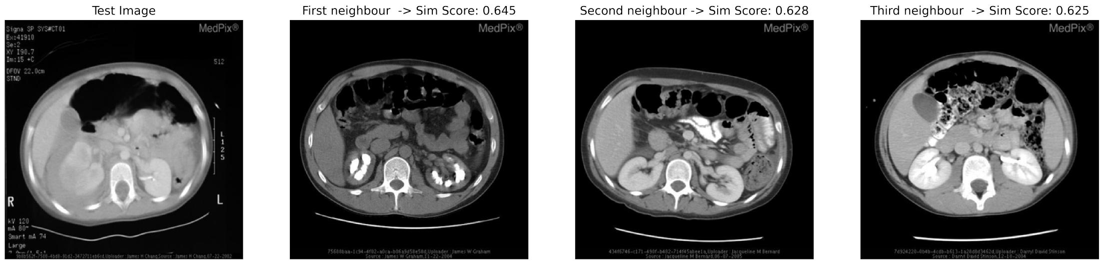

Figure 4: 3-Nearest neighbours for test image with LPIPS

4. Methods

4.1. Naive approach

The data set analysis motivated us to try a naive method as a benchmark that consisted in classifying all inputs as the most frequent concept in the training set, C0040398 (Tomography and Emission-Computed). When submitted, this approach gave us an F1 score of 0.141 over the testing set.

4.2. Perceptual Similarity

We tried using a KNN algorithm over the images to detect concepts. This is, given a test image, we find the K closest training examples according to a specific distance metric, which is detailed further in the article. Then, the concepts assigned to the test image are the concepts that are present among all selected (K) neighbours. Figure 4 shows an example using K=3 neighbours.

The metric used to calculate the distance between different images was Learned Perceptual Image Path Similarity (LPIPS) [4][3], a learned metric based on the similarity between deep features from several layers of a neural network, typically a VGG model [3]. This distance tries to capture the similarity between two images according to human perception. The use of this metric as a good estimator for this dataset arose in the caption prediction task, where a similar approach meant a significant performance gain.

4.3. Multi-Label classification

Another method used at an early stage of the study was the ResNet-18 [9] model pre-trained on the ImageNet dataset, from PyTorch[4]. The model was fine-tuned with the ImageCLEFmedical 2021 data set. However, this approach strongly overfitted to the training set. Over the validation set, the model almost always predicted the class to be the most frequent concept in the training set, C0040398 (Tomography and Emission-Computed), so the results achieved on the testing set was 0.141 F1 score, same as the naive approach. This is probably due to the imbalanced nature of the dataset, and that no regularization or drop-out layers were used.

[3]The implementation is available for PyTorch @ https://github.com/richzhang/PerceptualSimilarity
[4]https://pytorch.org/hub/pytorch_vision_resnet

5. Results

Using Perceptual Similarity as the distance metric for the KNN algorithm leads to images representing similar concepts being grouped together in a neighbourhood. Figure 5 shows the performance in F1 score of the KNN approach for different number of neighbours. The best performance of the KNN classifier results when only choosing the closest neighbour.

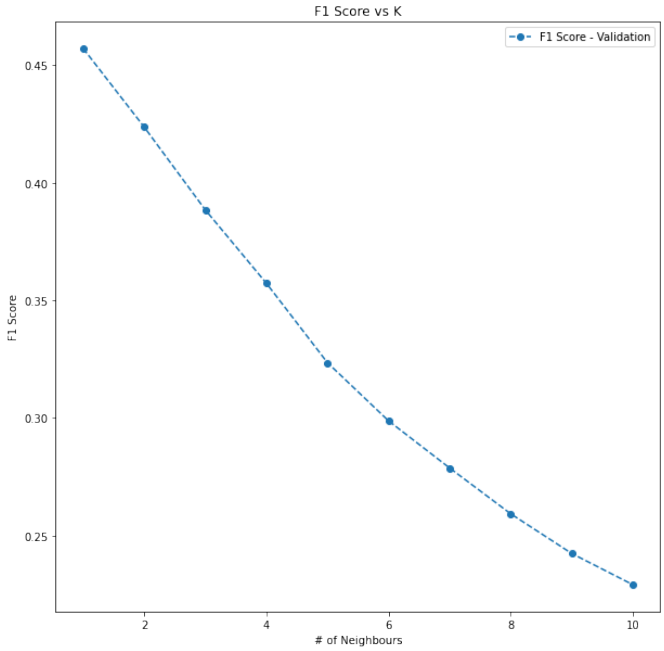

Figure 5: Change in F1 Score with respect to K neighbours

Figure 6 shows the neighborhood of an image in the validation set. In general, experimental results show that LPIPS similarity is able to capture significant information such as image type and the general shape of the observed object, which are important attributes when selecting concepts.

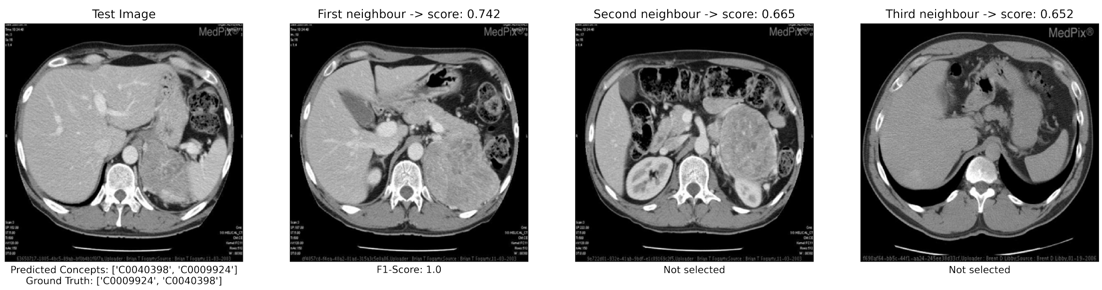

Figure 6: Sample neighborhood of a test image and F1 Score

5.1. Runs

Two submissions were made for the task, the first one using a naïve approach in order to get insights of the results with a baseline, and a second approach based on K-NN using the perceptual similarity implemented in LPIPS, as shown in Table 1. We improve the score more than twice by K-NN employing LPIPS.

Table 1
Results of the two submission made by the PUC Chile team to the concept detection task.

	Method	Score
Submission 1	Naive Approach - most popular concept	0.141
Submission 2	KNN with LPIPS as similarity metric	0.360

6. Conclusion

This article describes the participation of the PUC Chile team for the concept detection task in ImageCLEFmedical challenge 2021. Our best submission included a VGG deep neural network for visual encoding integrated with a KNN image search using Perceptual Similarity [4] to select concepts. This metric, as the original article states, provides a robust similarity approach, although more expensive to calculate than traditional approaches such as cosine-based similarity.

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, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, 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] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.

[3] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[5] O. Pelka, C. M. Friedrich, A. Garcıa Seco de Herrera, H. Müller, Overview of the imageclefmed 2020 concept detection task: Medical image understanding, in: CLEF2020 Working Notes, CEUR Workshop Proceedings. CEUR-WS. org, Thessaloniki, 2020.
[6] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.
[7] B. Karatzas, J. Pavlopoulos, V. Kougia, I. Androutsopoulos, AUEB NLP group at imageclefmed caption 2020, in: CLEF 2020 Working Notes, Thessaloniki, Greece, September 22-25, 2020, 2020.
[8] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.
[9] 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 (CVPR), 2016.