ImageCLEFmedical Caption Task, Concept Detection, Finding Duplicates, SDVA-UCSD Approach

Amilcare Gentili

San Diego VA HCS, 3350 La Jolla Village Drive, San Diego, CA 92161, USA
UCSD, 9300 Campus Point Drive #7756, La Jolla, CA 92037-7756, USA

Abstract

The ImageCLEFmedical Caption task, concept detection goal, was to detect relevant concepts in a large corpus of medical images. For this task, a subset of the extended Radiology Objects in COntext (ROCO) dataset was provided. We utilized 2 approaches to classify images: hashing similarities and convolutional neural networks. Hashing similarity was able to detect duplicate images in the dataset, but did poorly in classifying images. Neural networks did better in classifying images, but our model had problems with low frequency concepts.

Keywords 1

Convolutional neural network, hashing, caption detection

1 Introduction

ImageCLEF [1] is part of the Conference and Labs of the Evaluation Forum (CLEF) since 2003 and ImageCLEF has included medical tasks every year since 2004. Since 2017 it has a Caption Task [2]. Interpreting and classifying medical images such as radiology images is a time-consuming task that involves highly trained experts. Manual labeling images to input in machine learning pipeline is often a slow and expensive process. There is a considerable need for methods that can automatically label medical images. This year concept detection task was to detect relevant concepts in a large corpus of medical image. The images were labeled with concepts generated using a reduced subset of the UMLS 2020 AB release.

2 Methods

2.1 Data

The dataset provided for this task was a subset of the extended Radiology Objects in COntext (ROCO) dataset [1], without imaging modality information. The dataset originates from biomedical articles of the PMC OpenAccess subset. It included 83,275 radiology images in the training set, 7,645 radiology images in validation set and 7,601 radiology images in the test set. The images in the training and validation dataset were accompanied by UMLS concepts extracted from the original image captions.

2.2 Equipment

We used a workstation with 2 Nvidia 24 GB (Titan RTX and Quadra 6000) video cards, 128 GB RAM, and a 2 TB solid state drive.

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: agentili@ucsd.edu
ORCID 0000-0002-5623-7512)


2.3 Hashing

The ImageHash hashing library [4] written in Python was used for hashing images, of the available hashing algorithms provided by the ImageHash library, the following 4 were utilized:

- average hashing (aHash)
- perception hashing (pHash)
- difference hashing (dHash)
- wavelet hashing (wHash)

Results of each hashing algorithm were combined and similarities among all image pairs were calculated [5].

2.4 Neural Network

Fastai library [6] was used for multilabel classification. Each model was trained for 15 epochs, using an image size of 384, batch size of 16, gradient accumulation of four, and learning rate in the range of 0.1 and 0.001. Pretrained Resnet [7] and Densenet [8] architectures were tested and based on results on the validation data set, best models were chosen and ensembled for final submission. The outputs of the best models were the average of the probabilities that each concept was a label for the image. The label was considered present if the average probability was greater than 0.5.

3 Results

3.1 Exploratory Data Analysis

8374 different concepts were used to label the 83,275 images in the training dataset. The most common concept, CUIS CU0040405 was used 25989 times, while the list commonly used concept CUIS C0004760 was used only twice. The top 25 concept account for one third of the labels Figure 1.

Figure 1

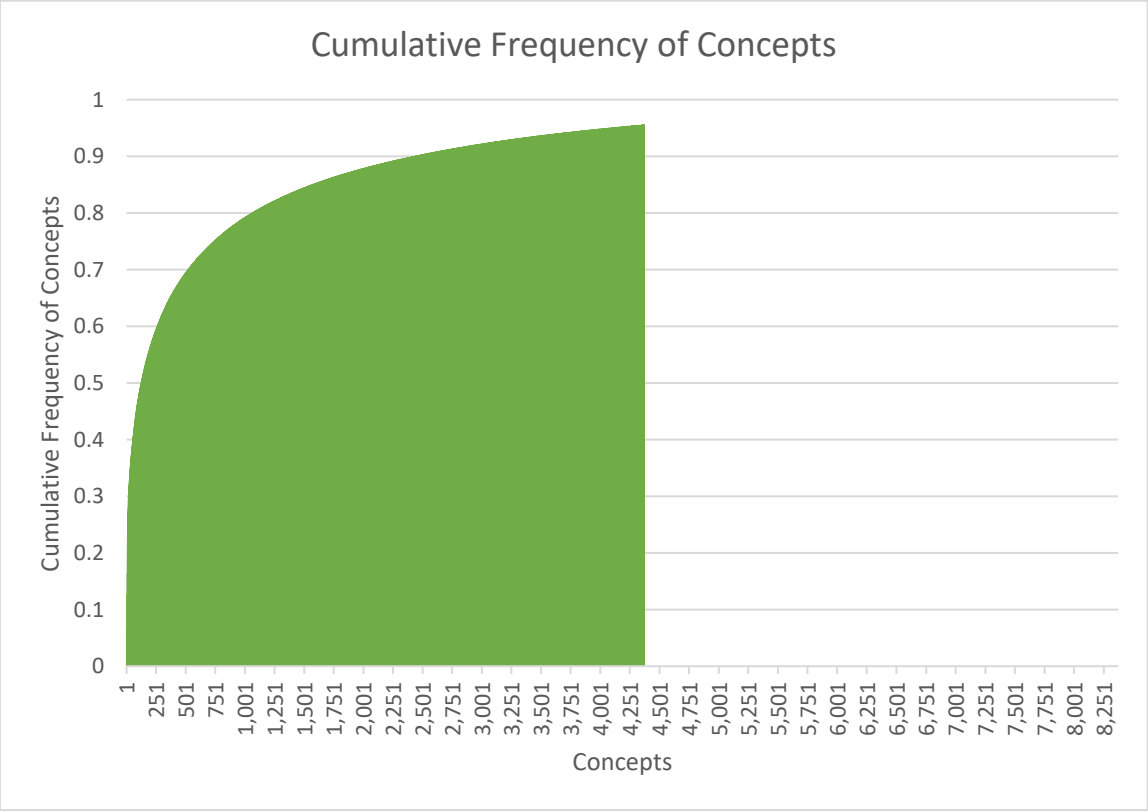

Figure 1. Cumulative frequency of concepts.

The number of concepts assigned to each image varied from 1 to 50, with most images being assigned 3 concepts.

Figure 2

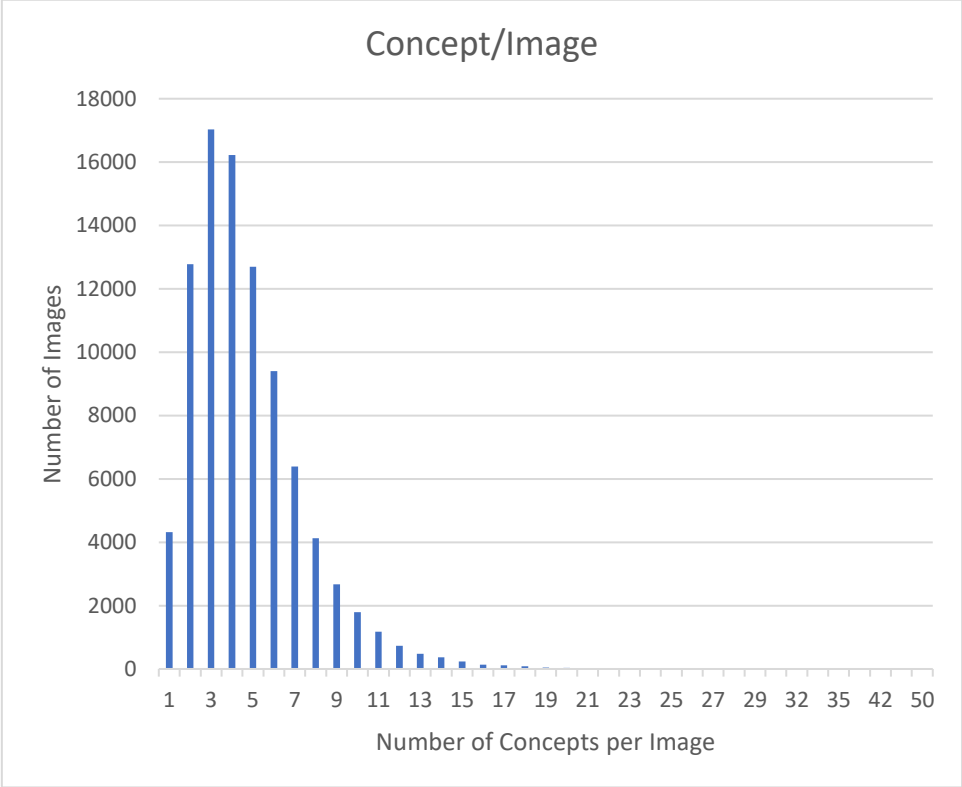

Figure 2: Histogram of number of concepts assigned to each image in the validation and training set combined.

3.2 Hashing

Using hashing similarities [5] greater than 96%, 275 duplicate images were present in the training dataset, 16 duplicate images were present in the validation dataset. Of the 191 images duplicated in the train or validation dataset, 222 had different concepts or captions assigned. 196 images of the test set were present in either the train or validation set. We made several attempted to use hashing similarities to label the validation images. We assign concepts to the validation images using variable thresholds of similarities between 0.5 and 0.9 and using concepts in common in the images above the similarity threshold. We also used concept from the 1, 3 or 5 most similar images in the training set, when 3 or 5 most similar images were used, we used 2 strategies to combine the results, we assigned the concept to validation images either if it was present in any image or if it was present in the majority of the images. Despite these multiple strategies, the F1 on the validation set were all below 0.16 and these attempts were not submitted to the competition.

3.3 Neural Network

Resnet 18, Resnet 34, Resnet 50, Resnet 101, Resnet 152, Resnet 201, Densenet 121, Densenet 161, Densenet 169 were used, based on validation results, predictions from Resnet 101, Densenet 161 and Densenet 169 were ensembled. For the final submission, the concepts of the validation or training dataset with a similarity greater than 0.96 were used as labels of the test images, instead of the concepts derived by the ensemble of the neural networks. Our single submission achieved F1 Score of 0.307932

and Secondary F1 score of 0.552432. Review of the submission classification demonstrated that only top 10 concepts were recognized by our neural network.

4 Discussion/Conclusion

Using hashing similarities, we were able to find duplicate images in the dataset but did not help in assign correct concepts to not identical images based on hash similarities. The dataset was very unbalanced, with top 25 concepts accounting for half of the labels. As we did not use any correction for the unbalanced data, our neural networks only learned how to recognize the most common concepts.

5 Perspectives for Future Work

Hashing similarities can be used to recognize duplicate images, but not images with similar concepts. Neural networks, without correction for imbalanced data, did well in classifying common concepts but did poorly on rare concepts. Strategies to correct for imbalanced data, such as enriching dataset with images with rare concepts or weighting more images with rare concepts during training will be necessary to improve concept detection. Other options include using a different neural network architecture, such as Siamese neural network.

6 Acknowledgements

The author was supported in part by the Office of the Assistant Secretary of Defense for Health Affairs through the Accelerating Innovation in Military Medicine Program under Award No. W81XWH-20-1-0693.

7 References

[1] Bogdan Ionescu, Henning Müller, Renaud Péteri, Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Serge Kozlovski, Yashin Dicente Cid, Vassili Kovalev, Liviu-Daniel Ștefan, Mihai Gabriel Constantin, Mihai Dogariu, Adrian Popescu, Jérôme Deshayes-Chossart, Hugo Schindler, Jon Chamberlain, Antonio Campello, Adrian Clark, Overview of the ImageCLEF 2022: MultimediaRetrieval in Medical, Social Media and Nature Applications, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), Springer Lecture Notes in Computer Science LNCS, Bologna, Italy, September 5-8, 2022.

[2] Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Henning Müller and Christoph M. Friedrich. Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings (CEUR-WS.org), Bologna, Italy, September 5-8, 2022.

[3] O. Pelka, S. Koitka, J. Rückert, F. Nensa und C. M. Friedrich “Radiology Objects in COntext (ROCO): A Multimodal Image Dataset“, Proceedings of the MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS 2018), Granada, Spain, September 16, 2018, Lecture Notes in Computer Science (LNCS) Volume 11043, Page 180-189, DOI: 10.1007/978-3-030-01364-6_20, Springer Verlag, 2018.

[4] Johannes Buchner, ImageHash, 2022. URL: https://github.com/JohannesBuchner/imagehash.

[5] Appian. Let's find out duplicate images with imagehash. 2022 URL: https://www.kaggle.com/code/appian/let-s-find-out-duplicate-images-with-imagehash/notebook.

[6] Jeremy Howard, Sylvain Gugger. fastai: A Layered API for Deep Learning, CoRR 2020, https://arxiv.org/abs/2002.04688.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385 https://doi.org/10.48550/arXiv.1512.03385.
[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger. Densely Connected Convolutional Networks. arXiv:1608.06993. https://doi.org/10.48550/arXiv.1608.06993.