Kdelab at ImageCLEFmedical 2022: Medical Concept Detection with Image Retrieval and Code Ensemble

Riku Tsuneda[1], Tetsuya Asakawa[1], Kazuki Shimizu[2], Takuyuki Komoda[2] and Masaki Aono[1]

[1]*Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan*
[2]*Toyohashi Heart Center 21-1 Gobutori, Oyama-cho, Toyohashi-shi, Aichi, Japan*

Abstract
ImageCLEFmedical 2022 Concept Detection Task is an example of a challenging research problem in the field of image captioning. The goal of this research is to automatically generate accurate Medical Concept (CUI code) describing a given medical image. We describe three approaches toward the concept detection task:simple image retrieval, CUI code ensemble with retrieval, and modality classification. We submitted 10 runs to the concept detection task, and achieved the F1 score of 0.310 and the Secondary F1 score of 0.412, which ranked 10th among the participating teams. We describe in detail our ways on data analysis and three approaches, and conclude by discussing some ideas for future work.

Keywords
Medical Images, Image Retrieval, Concept Detection, Multi-Label Classification, Concept Unique Identifier

1. Introduction

ImageCLEF is an initiative aimed at advancing the field of image retrieval and improving the evaluation of technologies for annotation, indexing, and retrieval of visual data. ImageCLEF has been held annually since 2003, and since the second edition (2004), tasks such as medical image retrieval have been included. Since the first and the second editions (2003 and 2004), ImageCLEF's Medical Challenge task group has integrated new ones that include medical images, and the Medical Caption task has been in place since 2017. This task consists of two subtasks : concept detection and caption prediction. Although the data used in the most recent version has changed, the goals of this task remain the same. Data has increased significantly over the last year. The goal of the task is to help physicians reduce the burden of manually translating visual medical information (e.g. radiology images) into textual descriptions. This document describes kdelab's participation in the ImageCLEF medical 2022 concept detection task. Our team placed 10th in this task. Our best submission was a system that visually encoded medical images with convolutional neural networks, performed K Nearest Neighbor (KNN) similarity search using Euclidean distance, and ensemble each result.

CLEF 2022 : Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
tsuneda@kde.cs.tut.ac.jp (R. Tsuneda); asakawa@kde.cs.tut.ac.jp (T. Asakawa); shimizu@heartcenter.or.jp (K. Shimizu); komoda@heartcenter.or.jp (T. Komoda); aono@kde.cs.tut.ac.jp (M. Aono)
0000-0002-3063-7489 (R. Tsuneda); 0000-0003-1383-1076 (T. Asakawa); 0000-0002-8345-7094 (M. Aono)

In the following, we first describe related work on Medical Concept Detection task in Section 2, followed by the description of the dataset provided for ImageCLEF medical 2022 Caption Task [1] [2] dataset in Section 3. In Section 4, we describe details of the method we have applied, and then of our experiments we have conducted in Section 5. We finally conclude this paper in Section 6.

2. Related Work

In previous competitions, the best participants used a variety of techniques based primarily on convolutional neural networks, natural language processing, K-Nearest Neighbors, and clustering. In the 2021 concept detection challenge [3], the highest F1 score was 0.505. The winning approach [4] used a convolutional neural network (CNN) [5] as an image encoder, combined with an image retrieval module. This team also took first place in the competition. Second place went to NLIP-Essex-ITESM [6] with an F1 score of 0.469. This team adopted image retrieval methods using various distance calculations. The best distance calculation for this team was cosine similarity.

Looking at past years, the best submissions achieved F1 scores of 0.1108 in 2018 [7], 0.2823 in 2019 [8], 0.3940 in 2020 and 0.505 in 2021.

3. Dataset

For the ImageCLEFmedical 2022 Concept Detection task, organizers have provided us with a training set of 83,275 radiology images with the same number of CUI codes, a validation set of 7,645 radiology images with the same number of CUI codes, and a test set of 7,601 radiology images with the same number of CUI codes. These images are part of ROCO dataset [9]. We are supposed to use these as our datasets. Most of the images in the dataset are non-colored, and they potentially include non-essential logos, arrow symbols, numbers and texts. The image data set included multiple modalities such as CT, MRI, X-ray, ultrasound images, and angiographic images. The task participants have to automatically predict CUI codes based on radiology image data.

The top 25 ranking concept names in terms of frequency are summarized in Table 1 and Figure 1. According to our analysis, the minimum number of CUI codes assigned was 1 and the maximum was 50. In addition, an average of 4.7 CUIs were assigned to each image.

For our experiments, we merged the provided training and validation sets and used 10% of the merged data as our validation set, and another 10% of the merged data as our development set in which we evaluated the performance of our models. The remaining 80% served as the training set.

4. Methodology

In this section, we describe the three approaches that were used in our submissions.

Table 1
CUI code frequency in dataset

Rank	CUI	Freq	UMLS defined Name
1	C0040405	28885	X-ray Computed Tomography
2	C1306645	26412	Plain x-ray
3	C0024485	15693	Magnetic Resonance Imaging
4	C0041618	12236	Ultrasoundgraphy
5	C0817096	8030	Chest
6	C0002978	6464	angiogram
7	C0000726	6243	Abdomen
8	C0037303	5175	Bone structure of cranium
9	C0221198	4094	Lesion
10	C0205131	3528	Axial
11	C0030797	3404	Pelvis
12	C0238767	3124	Bilateral
13	C0023216	2753	Lower Extremity
14	C0577559	2497	Mass of body structure
15	C0205129	2243	Sagittal
16	C0205091	1856	Left
17	C0205090	1665	Right
18	C0021102	1564	Implants
19	C0444706	1542	Measured
20	C0009924	1524	Contrast Media
21	C0006660	1412	Physiologic calcification
22	C0205095	1385	Dorsal
23	C0027651	1371	Neoplasms
24	C0023884	1339	Liver
24	C0037949	1339	Vertebral column

4.1. Image Preprocessing

Since most of the images in the dataset are grayscale images, we attempted a pseudo-coloring on the images. For pseudo coloring, use the method of assigning a black and white color map to an RGB color map. We have used the Open-CV [10] JET colormap for the RGB colormap. We show an example of the pseudo-coloring in Figure 2.

4.2. Image Retrieval Approach

Image retrieval methods were one of the major methods in CLEF2021. Last year, AUEB-NLP Gloup [3] and PUC Chile Team [11] adopted this method and achieved top scores. Since the most medical images are grayscale images, retrieval methods may be more effective than deep learning methods. In this section we describe our simple image retrieval and ensemble methods.

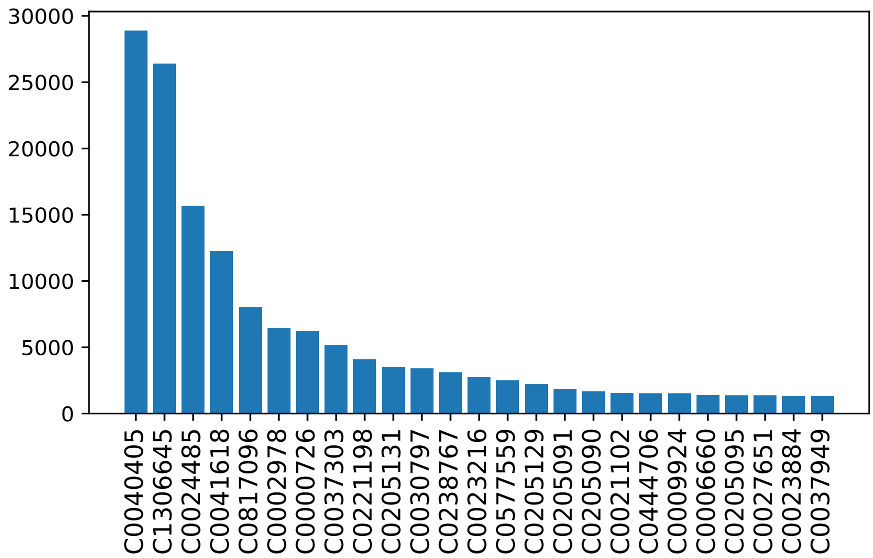

Figure 1: Frequency of CUI in ImageCLEFmedical 2022 Concept Detection Dataset

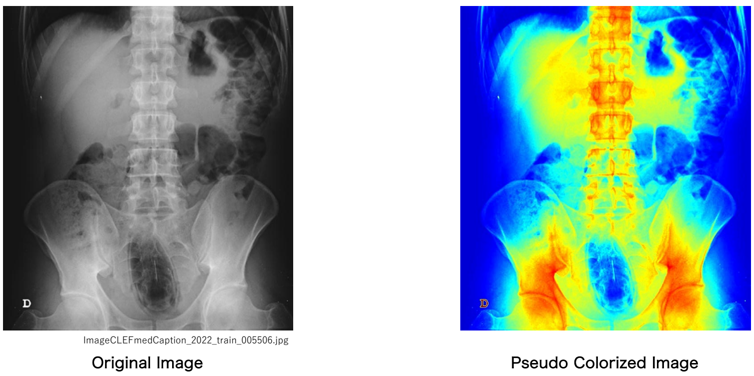

Figure 2: Example of Original Image (Left) and Pseudo Colorization(Right) [CC BY-NC-ND [Peixoto et al. (2015)]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5580006/)

4.2.1. Simple Image Retrieval

We similarly tested the effectiveness of our image retrieval method. We illustrate our image retrieval method in Figure 3.

First, we extracted features from all images using a several feature extractor. Next, we compute approximation based on the features using the Cosine similarity or Euclidean distance. Finally, we assign concepts to test images from the retrieval results.

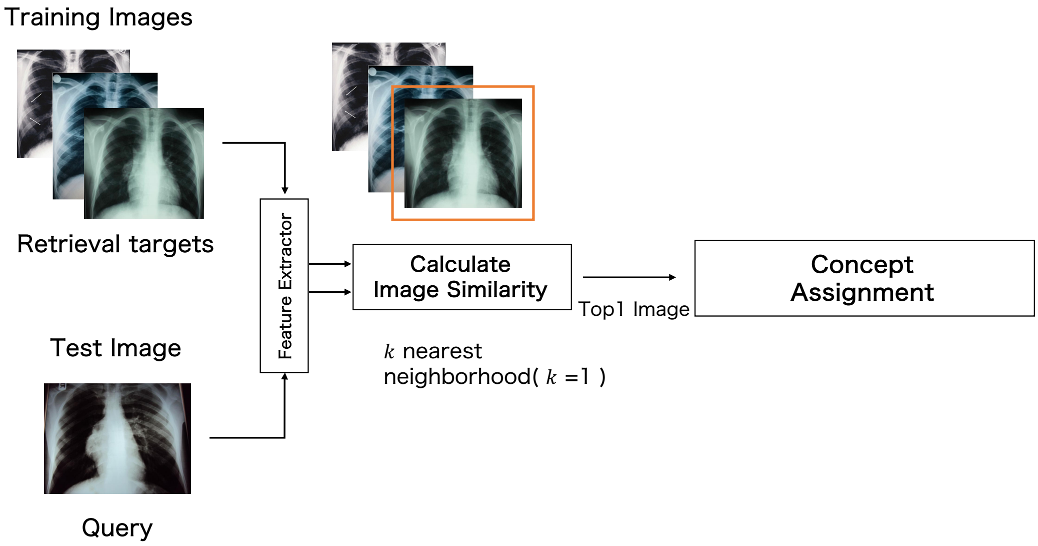

Figure 3: Example of Image Retrieval System, [CC BY-NC [Hekmat et al. (2016)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4835740/), CC BY [Abidi et al. (2015)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4769046/), CC BY [Apaydin et al. (2018)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6202798/), CC BY-NC-ND [Datta et al. (2018)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5925857/)]

4.2.2. Ensemble with Majority Voting

We have attempted an ensemble system with multiple feature extractors. This system is based on a majority voting of the predictions in the five extractors. We have adopted DenseNet-121 [12], DenseNet-201 [12], ResNet-50 [13], ResNet-152 [13], EfficientNet-B0 [14], EfficientNet-B7 [14], Inception-V3 [15], Xception [16], inception ResNet-V2 [17] and Nas Net Large [18] as feature extractor.

4.2.3. CUI Code Ensemble

Unlike a simple ensemble, this method combines multiple CUI codes without determining a single predicted image. We illustrate our CUI Code Ensemble method in Figure 4.

First, as in the image retrieval method, each feature extractor is used to estimate the approximate image. Next, the CUI code assigned to the image is obtained. Finally, the CUIs are sorted by frequency of occurrence, and the CUIs up to top l are considered as results. We tried the variable l in two ways :

- `Average Length` : Average length of predicted CUI codes
- `Max Length` : Maximum length of the predicted CUI codes

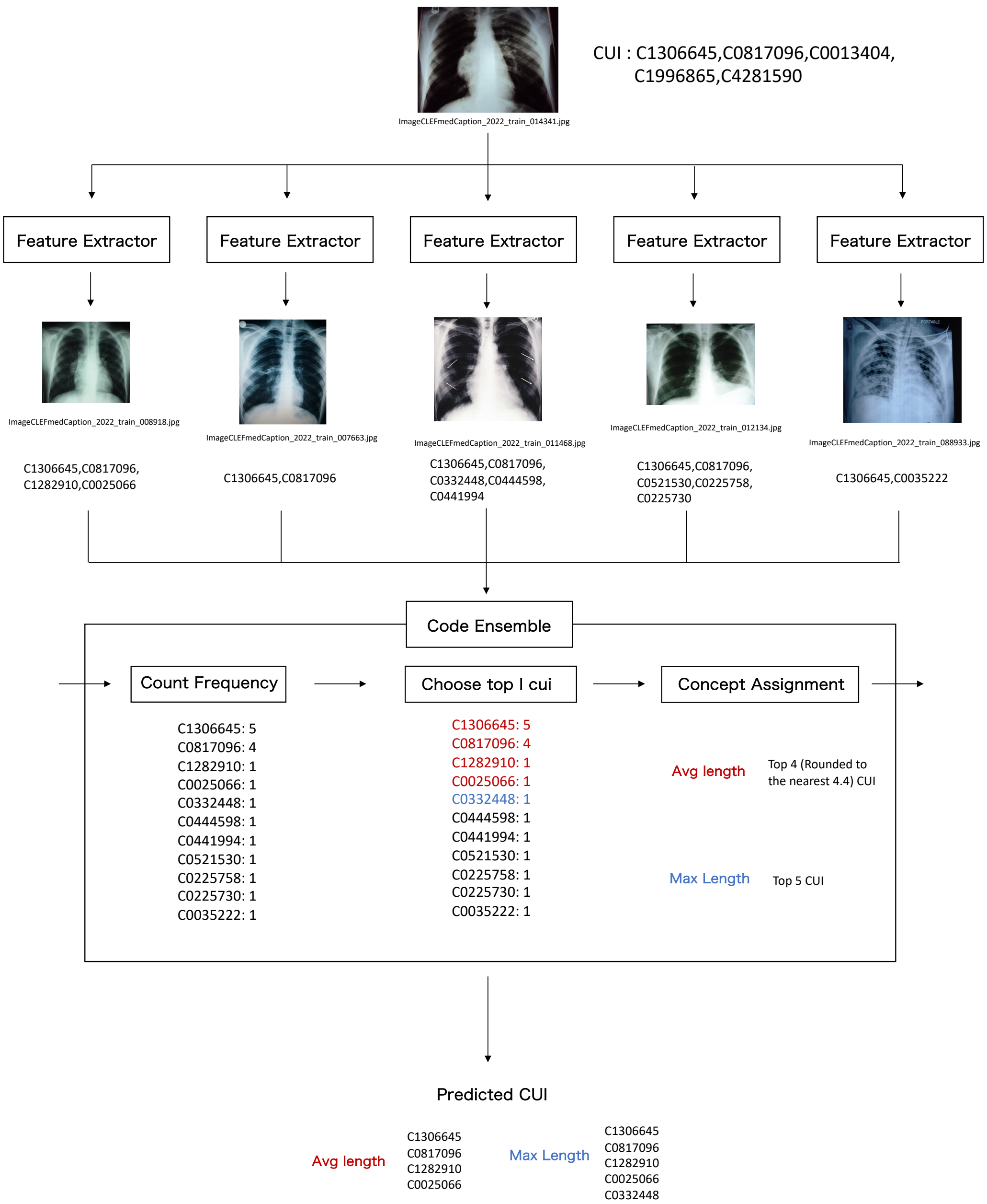

Figure 4: Example of CUI Code Ensemble method, [CC BY-NC [Hekmat et al. (2016)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4835740/), CC BY [Abidi et al. (2015)] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4769046/), CC BY [Apaydin et al. (2018)] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6202798/), CC BY-NC-ND [Datta et al. (2018)] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5925857/), CC BY [Nouri-Majalan et al. (2010)] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2939555/), CC BY [Naz et al. (2020)] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7536294/)]

Table 2
The Classes of Our Modality Classifier

Modality Name	Quantity of Images
CT	28,885
X-Ray	26,412
MRI	15,693
Ultrasoundgraphy	12,236
Angiogram	6,464
Others	1,230

Table 3
The Validation Scores of our Training Modality Classifier on our Development Set

Processing for Imbalanced Data	Accuracy	Precision	Recall	F1 score
None	**0.675**	0.332	**0.416**	**0.367**
Over Sampling	0.493	**0.348**	**0.416**	0.313
Under Sampling	0.179	0.215	0.202	0.085
Class Weighting	0.663	0.326	0.413	0.361

4.3. Modality Classification Approach

This method is a combination of image classification and image retrieval. First, image modality classification is performed, followed by image retrieval among the predicted modalities and approximate image estimation. Approximate image estimation is the same as in section4.2. We tried ResNet-101 as modality classifiers. We trained these classifiers using the dataset Train, Validation. The classifiers are classified into 6 classes : CT, MRI, Plain X-ray, Angiorgam, Ultrasoundgraphy, and others. Class assignment is based on the CUI code that was assigned. The classes and the number of images corresponding to the classes are shown in the Table 2. The training results of the classifiers are shown in the following Table 3, which shows the results of our Modality Classification Approach using development set. As can be seen from the results, we failed to produce a highly accurate modality classifier. In fact, the best modality classifier in the Table 3 was used to search for CUI codes, resulting in a best F1 score of 0.034. This method was ineffective and we did not submit it.

5. Submission and Results

We performed 10 submissions using Simple Image Retrieval, CUI Code Ensemble and pre-processing described in the previous section. Since the official evaluation metric for concept detection is F1 score, we evaluated models using this metric in the development set to determine which models to submit (each participant was allowed a maximum of 10 submissions). Table 4, Table 5 shows the scores for the development set, and Table 6 shows the final scores for our model on the unknown test caption.

First, we describe our results and findings in the development set. In simple image retrieval

Table 4
The scores of our Image Retrieval systems on our pseudo colorized development set

ID	Approach	Calculation	F1 score
exc01	DenseNet121	Cosine Similarity	0.269
exc02	EfficientNetB0	Cosine Similarity	0.266
exc03	EfficientNetB7	Cosine Similarity	0.256
exc04	DenseNet-201	Cosine Similarity	0.271
exc05	ResNet-50	Cosine Similarity	0.261
exc06	ResNet-152	Cosine Similarity	0.259
exc07	Xception	Cosine Similarity	0.253
exc08	inceptionResNetV2	Cosine Similarity	0.241
exc09	NasNetLarge	Cosine Similarity	0.232
exc10	inceptionV3	Cosine Similarity	0.251
exc11	Ensemble1 (exc01, exc02, exc04, exc05, exc06)	Cosine Similarity	**0.286**
exc12	Ensemble2 (exc01, exc02, exc03, exc04, exc07)	Cosine Similarity	0.284
exc13	Ensemble1 (exc01, exc02, exc04, exc05, exc06)	Euclidean Distance	0.281
exc14	Ensemble2 (exc01, exc02, exc03, exc04, exc07)	Euclidean Distance	0.276
exc15	CUI code Ensemble (average length)	Cosine Similarity	0.283
exc16	CUI code Ensemble (max length)	Cosine Similarity	0.281

methods, accuracy was found to improve when using ensembles with simple majority voting. Ensemble 1 has a higher BLEU score than Ensemble2. Comparing Cosine similarity and Euclidean distance, Cosine Similarity provides better retrieval accuracy. Second, we describe our results and findings in the test set. We submitted to AIcrowd the systems that scored highly in each of the approaches in our development set. The highest scoring submission was simple image retrieval system using Euclidean distance.

Finally, from organizer's evaluation, we have achieved a F1 score of 0.310 and a secondary F1 score of 0.412 in the ImageCLEF2022medical Concept Detection task, placing us 10th.

6. Conclusion

We have described our systems with which we submitted to the ImageCLEFmedical 2022 Concept Detection task. In our system, we have done our own data pre-processing, and have attempted to automatically generate concepts with image retrieval.

The results demonstrate that some of experiment have improved the concept detection accuracy of the image retrieval. Pseudo colorization and code ensemble approach turns out to be ineffective in this task.

Acknowledgments

A part of this research was carried out with the support of the Grant for Toyohashi Heart Center Smart Hospital Joint Research Course and the Grant-in-Aid for Scientific Research (C) (issue numbers 22K12149 and 22K12040)

Table 5
The scores of our Image Retrieval systems on our development set

ID	Approach	Calculation	F1 score
ex01	DenseNet121	Cosine Similarity	0.277
ex02	EfficientNetB0	Cosine Similarity	0.276
ex03	EfficientNetB7	Cosine Similarity	0.261
ex04	DenseNet201	Cosine Similarity	0.280
ex05	ResNet-50	Cosine Similarity	0.273
ex06	ResNet-152	Cosine Similarity	0.272
ex07	Xception	Cosine Similarity	0.261
ex08	InceptionResNetV2	Cosine Similarity	0.253
ex09	NasNet Large	Cosine Similarity	0.226
ex10	Inception-V3	Cosine Similarity	0.264
ex11	Ensemble1 (ex01, ex02, ex04, ex05, ex06)	Cosine Similarity	0.311
ex12	Ensemble2 (ex01, ex02, ex03, ex04, ex07)	Cosine Similarity	0.308
ex13	Ensemble1 (ex01, ex02, ex04, ex05, ex06)	Euclidean Distance	**0.312**
ex14	Ensemble2 (ex01, ex02, ex03, ex04, ex07)	Euclidean Distance	0.290
ex15	CUI code Ensemble (average length)	Cosine Similarity	0.296
ex16	CUI code Ensemble (max length)	Cosine Similarity	0.295

Table 6
The scores of all of systems on our submission

Approach	Image Preprocessing	F-1	Secondary F1	Run ID
Ensemble1 Retrieval with Cosine Similarity	Pseudo Colorization	0.309	0.409	181907
Ensemble2 Retrieval with Cosine Similarity	Pseudo Colorization	0.290	0.396	182181
Ensemble1 Retrieval with Cosine Similarity	None	0.309	0.409	181905
Ensemble2 Retrieval with Cosine Similarity	None	0.296	0.409	181906
DenseNet201 Retrieval with Cosine Similarity	None	**0.310**	0.408	182182
DenseNet121 Retrieval with Cosine Similarity	None	0.309	0.409	182183
Code Ensemble (Average Length)	None	0.292	0.406	182235
Code Ensemble (Max Length)	None	0.239	0.292	182237
Ensemble1 Retrieval with Euclidean Distance	None	**0.310**	**0.412**	182346

References

[1] B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia retrieval 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), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.

[2] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-

Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – caption prediction and concept detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[3] 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: 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.

[4] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: [19], 2021, pp. 1184–1200. URL: http://ceur-ws.org/Vol-2936/#paper-96.

[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[6] J. Jacutprakart, F. P. Andrade, R. Cuan, A. A. Compean, G. Papanastasiou, A. G. S. de Herrera, Nlip-essex-itesm at imageclefcaption 2021 task : Deep learning-based information retrieval and multi-label classification towards improving medical image understanding, in: [19], 2021, pp. 1264–1274. URL: http://ceur-ws.org/Vol-2936/#paper-103.

[7] A. García Seco de Herrera, C. Eickhoff, V. Andrearczyk, , H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: CLEF2018 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France, 2018.

[8] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2019 concept prediction task, in: CLEF2019 Working Notes, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland, 2019.

[9] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. Friedrich, Radiology objects in context (roco): A multimodal image dataset, in: CVII-STENT/LABELS@MICCAI, 2018.

[10] Itseez, Open source computer vision library, https://github.com/itseez/opencv, 2015.

[11] V. Castro, P. Pino, D. Parra, H. Lobel, Puc chile team at caption prediction: Resnet visual encoding and caption classification with parametric relu, in: [19], 2021, pp. 1174–1183. URL: http://ceur-ws.org/Vol-2936/#paper-95.

[12] 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.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[14] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[15] 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, pp. 2818–2826. doi:10.1109/CVPR.2016.308.

[16] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807. doi:10.1109/CVPR.2017.195.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, pp. 4278–4284.

[18] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning transferable architectures for scalable image recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710. doi:10.1109/CVPR.2018.00907.

[19] 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. 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.