# Kdelab at ImageCLEFmedical 2022 Caption Prediction Task


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 Caption 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 captions describing a given medical image. We describe five approaches using image retrieval and Deep Learning . In this paper, we have adopted K-nn as image retrieval, X-VLM and Show, Attend and Tell as Deep Neural Network (DNN). Furthermore, we describe the effectiveness of a method that uses information from the CUI code as an input feature for DNN. We submitted 8 runs to the caption prediction task, and achieved the BLEU score of 0.278 and the ROUGE score of 0.158, which ranked 7th among the participating teams.



## Keywords

Image Captioning, Deep Learning, Medical Images, Image Retrieval, Concept Detection, Natural Language Processing

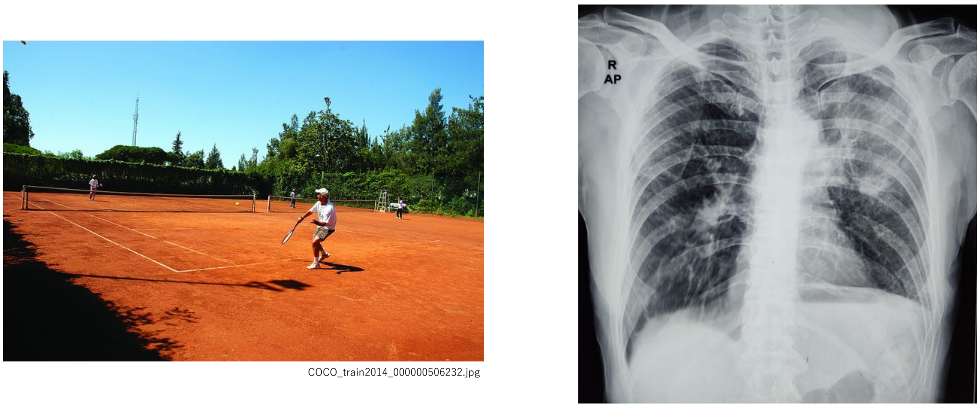

## 1. Introduction

In recent years, multimodal processing of images and natural language has attracted much attention in the field of machine learning. Image Captioning is one of these representative tasks, which aims at proper captioning of input images. As these accuracies improve, it is expected that computers will not only be able to detect objects in images, but also to understand the relationships and behaviors between objects. Image captioning is also effective in the medical field. For example, interpreting and summarizing possible disease symptoms from a large number of radiology images (e.g. X-ray images and CT images) is a time-consuming task that can only be understood by highly knowledgeable specialists. If computers could understand medical images and generate accurate captions, it would help solve the world's growing shortage of medical doctors. However, there is still the bottleneck problem that few physicians are able to give accurate annotations.

The nature of medical images are quite different from general images such as MS-COCO [1] in many aspects.


*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)



COCO_train2014_000000506232.jpg

ImageCLEFmedCaption_2022_valid_084384.jpg

A player readies for a swing during a tennis game .

chest xray posteroanterior view suggestive of a large opacity on the left side chest

**Figure 1:** Example of general (left) and medical (right) Caption Prediction data& left image : via MS-COCO, [CC BY 4.0](https://cocodataset.org/), right image:CC BY [Wadhwa et al. (2021)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8476187/).

In the following, we first describe related work on Image Captioning task and Medical Image Captioning in Section 2, followed by the description of the dataset provided for ImageCLEFmedical 2022 Caption Prediction [2] [3] 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 this section, we introduce previous studies related to this research. First, as representative studies of caption generation in general, Show and Tell by Vinyals et al. Vinyals' work uses a CNN (convolutional neural network) [4] as an image encoder and an RNN (recurrent neural network) as a decoder. Such a model is called an encoder-decoder model and is the basis of current caption generation.

Vinyals et al. achieved the highest accuracy at that time on the MS-COCO dataset. Xu et al.'s [5] work was based on Vinyals et al.'s work. By incorporating visual attention [6] into the caption generation model using CNN and LSTM (Long Short Term Memory) [7], they were able to generate more descriptive captions. This study achieved the highest accuracy at the time of publication. Anderson et al. [8] demonstrated that combining both bottom-up attention with Faster R-CNN [9] and top-down attention with weighted averaging can be used for both Image Captioning and Visual Question Answering [10]. Anderson et al. achieved SOTA on both the Image Captioning and Visual Question Answering tasks by combining both bottom-up attention with Faster R-CNN and top-down attention with weighted average. In recent years, research on caption generation using transformers has achieved SOTA.

In 2019, Jing et al. [11] proposed a method for generating captions for chest X-ray images using an Encoder-Decoder model with Co-Attention [12]. In 2021, the Medical Caption Prediction Task was held at the international competition ImageCLEF 2021 [13]. Thus, caption generation in the medical field is a challenging research field that continues to attract attention.

## 3. Dataset

For the ImageCLEFmedical 2022 Caption Prediction task, organizers have provided us with a training set of 83,275 radiology images with the same number of captions, a validation set of 7,645 radiology images with the same number of captions, and a test set of 7,601 radiology images with the same number of captions. These images are part of ROCO dataset [14]. 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 generate automatic caption based on radiology image data.

According to our analysis, the top word frequencies were dominated by prepositions and words such as right and left that indicate position. The top 14 ranking words in terms of word frequency are summarized in Table 1.

**Table 1**
Word frequency Ranking in Caption Dataset

| Rank | Word | Freq | Rank | Word | Freq |
|---|---|---|---|---|---|
| 1 | show | 41,364 | 7 | scan | 11,655 |
| 2 | arrow | 24,555 | 8 | tomography | 10,628 |
| 3 | right | 20,340 | 9 | chest | 10,052 |
| 4 | ct | 16,495 | 10 | mass | 9,192 |
| 5 | image | 14,703 | 11 | view | 8,580 |
| 6 | left | 12,752 | 12 | radiograph | 8,025 |

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 approaches that were used in our submissions.

### 4.1. Image Preprocessing

Given that the images in the dataset are black and white images, we tried pseudo colorization to the images. Pseudo colorization is the assignment of a color map to an image. We used the Open-CV [15] JET colormap for the 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 [16] and PUC Chile Team [17] adopted this method and achieved top scores. Since the most medical images are grayscale images, retrieval methods may be more effective than DNN methods. We similarly tested the effectiveness of our image retrieval method. We illustrate our image retrieval method in Figure 3. We have tried an ensemble of image retrieval methods. The ensemble method is a majority voting of five images predicted using each feature extractor.

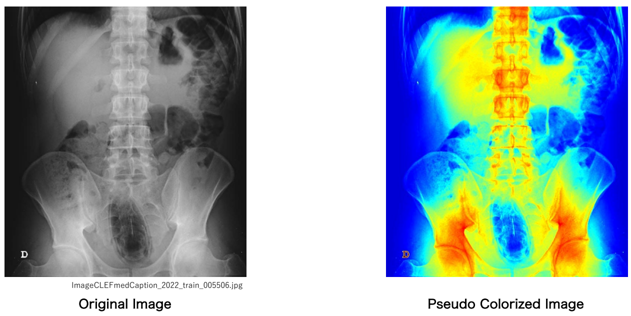



**Figure 2:** Example of Original Image (left) and Pseudo Colorization (right), CC BY [CC BY-NC-ND [Peixoto et al. (2015)]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5580006/).

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 captions to test images from the retrieval results.

We adopted DenseNet121 [18], DenseNet201 [18], ResNet-50 [19], ResNet-152 [19], EfficientNet-B0 [20], EfficientNet-B7 [20], Inception-V3 [21], Xception [22],inception ResNet-V2 [23] and Nas Net Large [24] as feature extractor.

### 4.3. DNN Approach

DNN methods were one of the orthodox methods in CLEF2021. We have adopted Show, Attend and Tell [5] and X-VLM [25] as DNN baseline methods.

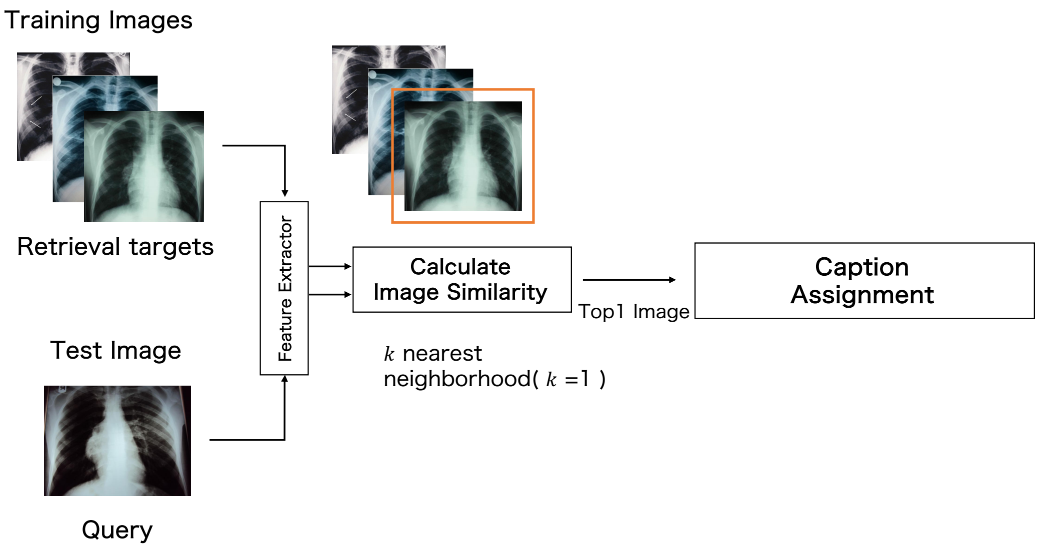



**Figure 3:** Image Retrieval System's Flow, via CC BY-NC [Hekmat et al. (2016)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4835740/), CC BY [Abidi et al. (2015)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4769046/), CC BY [Apaydin et al. (2018)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6202798/),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

#### 4.3.1. Show, Attend and Tell

This model is capable of highly accurate captioning without using object detection. Our team achived 3rd in CLEF2021 with this model and Effective Image Preprocessing. The architecture of the models is almost the same, while our model differs in that we employ ResNet-101 [19] instead of VGG16 [4] as the CNN encoder. Furthermore, we have adopted DenseNet201 as the CNN encoder. In decoder part, words are predicted by LSTM with Attention based on the image features. The output captions are the best alignment of the predicted words by Beam Search. We illustrate our image retrieval method in Figure 4.

#### 4.3.2. X-VLM

Most modern caption generation models use object detectors. X-VLM [25] achieved SOTA in the Image Captioning task without using an object detector. We have adopted this model directly to this task. We illustrate X-VLM method in Figure 5. The X-VLM consists of an image encoder, a text encoder, and a cross-modal encoder. All encoders are based on Transformer. The cross-modal encoder fuses visual and linguistic features through cross-attention at each layer. In image encoder part, after images (224×224) are divided into 32x32 patches (called patches), they are reshaped while retaining its positional information. In bounding box prediction part, the model predicts the box for the text and input CLS tokens from Cross-model Encoder into MLP. In contrastive learning part, Contrastive Learning is performed using cosine similarity

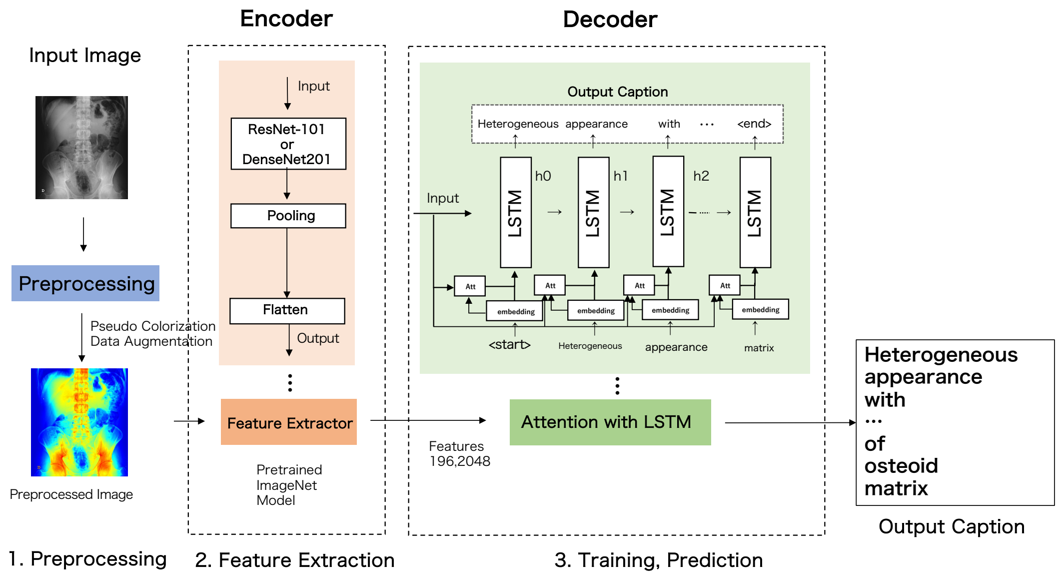



**Figure 4:** Model of Show, Attend and Tell [CC BY-NC-ND [Peixoto et al. (2015)]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5580006/)]

**Table 2**
The hyperparameters of X-VLM

| Name | Parameter |
|---|---|
| Epochs | 120 |
| Batch Size | 32 |
| Patch Size of ViT | 32 |
| Num of Decoder Layers | 6 |
| Optimizer | Adam [26] |

in vision-to-text and text-to-vision. Texts that are not paired are farther apart. In Matching Prediction, the visual and textual features are checked for a match. We have adopted a 16M parameter model (trained using CC-12M) as our pre-training model. We used this pre-trained model for downstream (Image Captioning) training. We used beam search, with the beam size for each step equal to 8. The hyperparameters we used for each model, after light tuning, can be seen in Table 2.

## 4.4. Combination of Retrieval and DNN Approach

This method adds a mechanism to the baseline method to predict CUI (Concept Unique Identifier) codes contained in images and use them as new features. Concept detection is not performed during training, but only during inference. During training, the CUI codes assigned to images are directly used for training. Figure 6 shows an overview of the proposed method during

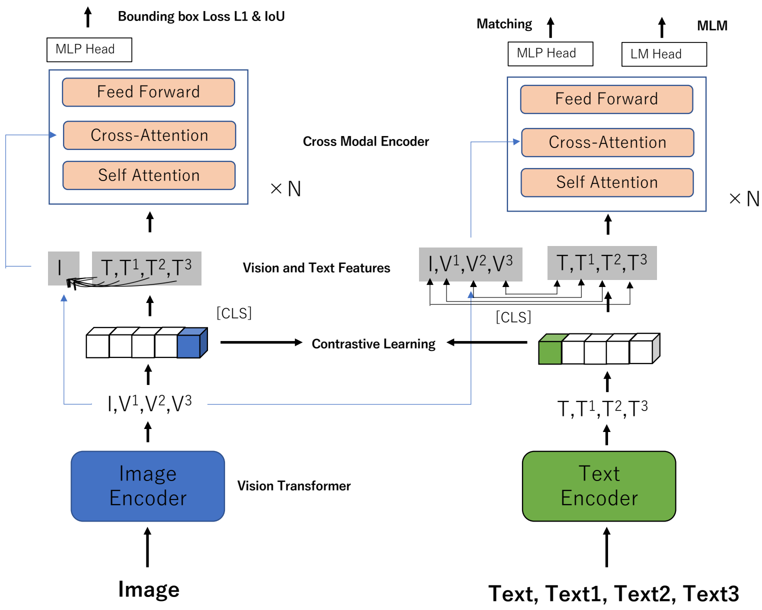



**Figure 5:** Pre-training model ardhitecture of X-VLM

training and inference, respectively. The details of this model are described next.

The procedure of creating CUI features is as follows. First, the number of predicted CUI codes is set to be 50 (the maximum number of CUI codes in the dataset), and the missing codes are compensated by assigning “None” to them. Next, for each CUI code (including “None” ), perform Word Embedding with Embedding Dim=32. Finally, the final CUI features are the flattened ones. Using this concatenated CUI and image features, Show, Attend and Tell performs caption prediction.

## 5. Submission and Results

We have submitted eight runs using the three methods and pre-processing described in the previous section. Since the official evaluation metric for caption prediction is BLEU-4, we have evaluate models using this metric in the development set to determine which models to submit (each participant was allowed a maximum of 10 submissions). Tables 3 and 4 shows the scores for the development set, and Table 5 shows the final scores for our model on the unknown test caption.

First, we describe our results and findings in the development set. In image retrieval methods, accuracy has turned out to improve when using ensembles with simple majority voting. Ensemble 1 has a higher BLEU score than Ensemble 2. Comparing Cosine similarity and Euclidean distance, Euclidean distance provides better retrieval accuracy. The DNN approach yields higher accuracy for Show, Attend and Tell than X-VLM. We speculate that this is because the MS-COCO dataset was used in the X-VLM pre-training. The combination of Image Retrieval and DNN

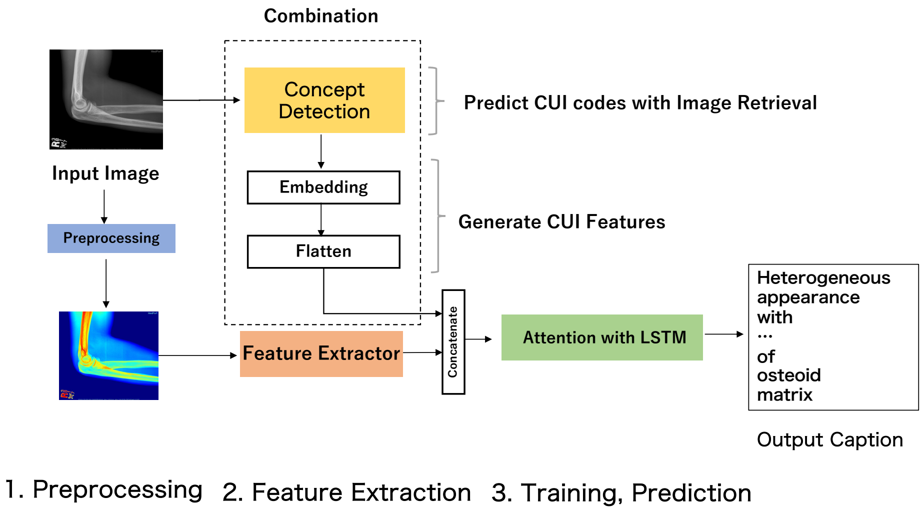



**Figure 6:** Combination of Neural Network and Image Retrieval System [CC BY-NC-ND [Peixoto et al. (2015)]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5580006/)]

method have may a negative effect on learning. On the other hand, Pseudo colorization is not effective, while it works well for X-VLM.

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. Overall, the results of the test set scored very much higher than the development set. The highest scoring submission has turned out to be an image retrieval system using Euclidean distance.

Finally, from organizer's evaluation, we have achieved a BLEU score of 0.278, a ROUGE score of 0.158 , a METEOR score of 0.073, a CIDEr score of 0.411, a SPICE score of 0.051and a BERT score of 0.600 in the ImageCLEFmedical 2022 Caption Prediction task, placing us 7th. Our submission ranked 7th in the BLEU score, but 1st in the CIDEr score. We achieved the highest CIDEr score by using image retrieval method to predict words that appear only in certain images (infrequent words).

## 6. Conclusion

We have described our system with which we submitted to the ImageCLEFmedical 2022 Caption Prediction task. In our system, we have done our own data pre-processing, and have attempted to automatically generate caption with image retrieval, DNN and combination of retrieval and neural network.

The results demonstrate that some of experiment have improved the caption prediction accuracy of the image retrieval. Pseudo colorization and combination approach turns out to be ineffective in this task.

**Table 3**
The scores of our Image Retrieval systems on our development set

| ID | Approach | Caluculation | BLEU-4 score |
|---|---|---|---|
| ex01 | Image Retrieval with DenseNet121 | Cosine Similarity | 0.025 |
| ex02 | Image Retrieval with EfficientNetB0 | Cosine Similarity | 0.027 |
| ex03 | Image Retrieval with EfficientNetB7 | Cosine Similarity | 0.026 |
| ex04 | Image Retrieval with DenseNet201 | Cosine Similarity | 0.028 |
| ex05 | Image Retrieval with ResNet-50 | Cosine Similarity | 0.026 |
| ex06 | Image Retrieval with ResNet-152 | Cosine Similarity | 0.026 |
| ex07 | Image Retrieval with Xception | Cosine Similarity | 0.025 |
| ex08 | Image Retrieval with InceptionResNetV2 | Cosine Similarity | 0.023 |
| ex09 | Image Retrieval with NasNet Large | Cosine Similarity | 0.022 |
| ex10 | Image Retrieval with InceptionV3 | Cosine Similarity | 0.026 |
| ex11 | Ensemble1 (ex01, ex02, ex04, ex05, ex06) | Cosine Similarity | 0.031 |
| ex12 | Ensemble2 (ex01, ex02, ex03, ex04, ex07) | Cosine Similarity | 0.029 |
| ex13 | Ensemble1 (ex01, ex02, ex04, ex05, ex06) | Euclidean Distance | **0.033** |
| ex14 | Ensemble2 (ex01, ex02, ex03, ex04, ex07) | Euclidean Distance | 0.031 |

**Table 4**
The scores of our approaches on our development set

| Approach | Image Preprocessing | BLEU-4 score |
|---|---|---|
| Show, Attend and Tell with ResNet-101 | None | 0.092 |
| Show, Attend and Tell with ResNet-101 | Pseudo Colorization | 0.090 |
| Show, Attend and Tell with DenseNet201 | None | **0.095** |
| Show, Attend and Tell with DenseNet201 | Pseudo Colorization | 0.091 |
| X-VLM | None | 0.055 |
| X-VLM | Pseudo Colorization | 0.061 |
| Combination of Retrieval and DNN | None | 0.056 |
| Combination of Retrieval and DNN | Pseudo Colorization | 0.054 |

## 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)

## References

[1] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: ECCV, 2014.

[2] 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,

**Table 5**
The scores of all of systems on ImageCLEFmedical 2022 Test set

| Approach | Image Preprocessing | BLEU-4 | ROUGE | Run ID |
|---|---|---|---|---|
| Ensemble1 Retrieval with Cosine Similarity | None | 0.268 | 0.153 | 181904 |
| Ensemble2 Retrieval with Cosine Similarity | None | 0.271 | 0.141 | 181908 |
| Ensemble1 Retrieval with Euclidean Distance | None | **0.278** | **0.158** | 182351 |
| Show, Attend and Tell with ResNet-101 | None | 0.221 | 0.154 | 181901 |
| Show, Attend and Tell with ResNet-101 | Pseudo Colorization | 0.218 | 0.150 | 181903 |
| Show, Attend and Tell with DenseNet201 | None | 0.222 | 0.157 | 182347 |
| X-VLM | Pseudo Colorization | 0.191 | 0.154 | 181946 |
| Combination of Retrieval and DNN | None | 0.187 | 0.113 | 182165 |


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.

[3] 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.

[4] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, CoRR abs/1409.1556 (2015).

[5] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, pp. 2048–2057.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[7] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, 2018, pp. 6077–6086. doi:10.1109/CVPR.2018.00636.

[9] S. Ren, K. He, R. B. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks., in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), NIPS, 2015, pp. 91–99. URL: http://dblp.uni-trier.de/db/conf/nips/nips2015.html#RenHGS15.

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: 2015 IEEE International Conference on Computer Vision (ICCV),

2015, pp. 2425–2433. doi:10.1109/ICCV.2015.279.
[11] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018). URL: http://dx.doi.org/10.18653/v1/P18-1240. doi:10.18653/v1/p18-1240.
[12] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6274–6283. doi:10.1109/CVPR.2019.00644.
[13] 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.
[14] 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.
[15] Itseez, Open source computer vision library, https://github.com/itseez/opencv, 2015.
[16] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: [13], 2021, pp. 1184–1200. URL: http://ceur-ws.org/Vol-2936/#paper-96.
[17] V. Castro, P. Pino, D. Parra, H. Lobel, Puc chile team at caption prediction: Resnet visual encoding and caption classification with parametric relu, in: [13], 2021, pp. 1174–1183. URL: http://ceur-ws.org/Vol-2936/#paper-95.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] Y. Zeng, X. Zhang, H. Li, Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts, (pre-print) (2021). arXiv:2111.08276.
[26] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2015).