Overview of the ImageCLEFmed 2021 Concept & Caption Prediction Task

Obioma Pelka[1,2], Asma Ben Abacha[3], Alba G. Seco de Herrera[4], Janadhip Jacutprakart[4], Christoph M. Friedrich[1,5] and Henning Müller[6,7]

[1]*Department of Computer Science; University of Applied Sciences and Arts Dortmund, Dortmund, Germany*
[2]*Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Germany*
[3]*National Library of Medicine, USA*
[4]*University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK*
[5]*Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany*
[6]*University of Applied Sciences Western Switzerland (HES-SO), Switzerland*
[7]*University of Geneva, Switzerland*

Abstract

The 2021 ImageCLEF concept detection and caption prediction task follows similar challenges that were already run from 2017–2020. The objective is to extract UMLS-concept annotations and/or captions from the image data that are then compared against the original text captions of the images. The used images are clinically relevant radiology images and the describing captions were created by medical experts. In the caption prediction task, lexical similarity with the original image captions is evaluated with the BLEU-score. In the concept detection task, UMLS (Unified Medical Language System) terms are extracted from the original text captions and compared against the predicted concepts in a multi-label way. The F1-score was used to assess the performance. The 2021 task has been conducted in collaboration with the Visual Question Answering task and used the same images. The task attracted a strong participation with 25 registered teams. In the end 10 teams submitted 75 runs for the two sub tasks. Results show that there is a variety of used techniques that can lead to good prediction results for the two tasks. In comparison to earlier competitions, more modern deep learning architectures like EfficientNets and Transformer-based architectures for text or images were used.

Keywords

Concept Detection, Computer Vision, ImageCLEF 2021, Image Understanding, Image Modality, Radiology

1. Introduction

This paper sets forth the approaches for the caption prediction task: automated cross-referencing of medical images and captions into predicted coherent captions implying Unified Medical Language System® (UMLS) concept detection in radiology images as a first step. This task is a part

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
obioma.pelka@fh-dortmund.de (O. Pelka); asma.benabacha@nih.gov (A. Ben Abacha); alba.garcia@essex.ac.uk (A. G. Seco de Herrera); j.jacutprakart@essex.ac.uk (J. Jacutprakart); christoph.friedrich@fh-dortmund.de (C. M. Friedrich); henning.mueller@hevs.ch (H. Müller)
0000-0001-5156-4429 (O. Pelka); 0000-0001-6312-9387 (A. Ben Abacha); 0000-0002-6509-5325 (A. G. Seco de Herrera); 0000-0001-7906-0038 (C. M. Friedrich); 0000-0001-6800-9878 (H. Müller)

of the ImageCLEF benchmarking campaign, which has proposed medical image understanding tasks since 2003; a new suite of tasks is generated each subsequent year. Further information on the other proposed tasks at ImageCLEF 2021 can be found in Ionescu et al. [1].

This is the 5th edition of the ImageCLEFcaption task. Although in 2020 the format of the task was the single task of concept detection, this year the task has expanded to include both concept detection sub task and bring back a caption prediction sub task, as the caption prediction sub task was included in the ImageCLEFmed Caption task in 2016 [2] (as a pilot sub task), 2017 [3], and 2018 [4]. In this edition, ImageCLEF 2021 uses actual radiology images annotated by real doctors, which means that the results achieved are highly relevant within a medical context.

Manual generation of the knowledge of medical images is a time-consuming process prone to human error. As this process is yet necessary assisting for the better and easier diagnoses of diseases that are susceptible to radiology screening, it is important that we better understand and refine automatic systems that aid in the broad task of radiology-image metadata generation. The purpose of the ImageCLEFmed 2021 concept detection and caption prediction tasks is the continued evaluation of such systems. Concept detection and caption prediction information is applicable for unlabelled and unstructured data sets and medical data sets that do not have textual metadata. The ImageCLEF caption task focuses on the medical image understanding in the biomedical literature and specifically on concept extraction and caption prediction based on the visual perception of the medical images and medical text data such as medical caption or UMLS® Unique Identifiers (CUIs) paired with each image (see Figure 1).

For the development data, the same data set used in the ImageCLEFVQA task [5] where radiology images were selected from the MedPix[1] database. To make the task more realistic and linked to the real-world data, the curated annotated data was used in contrast to earlier years where images were extracted from medical publications. The test set used for the official evaluation was obtained from the same source as proposed in [1].

This paper presents an overview of the ImageCLEF caption task 2021 including the task and participation in Section 2, the data creation in Section 3, and the evaluation methodology in Section 4. The results are described in Section 5, followed by conclusion in Sections 6.

2. Task and Participation

In 2021, the ImageCLEFcaption task consisted of two sub tasks: concept detection and caption prediction.

The concept detection sub task follows the same format proposed since the start of the task in 2017. Participants are asked to predict a set of concepts defined by the Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [6] (UMLS-CUI) based on the visual information provided by the radiology images.

The caption prediction sub task follows the original format of the sub task used between 2017 and 2018. The task was run again because of participant demand. This sub task aims to define automatic captions for the radiology images provided.

In 2021, 25 teams registered and signed the End-User-Agreement that is needed to download the development data. 10 teams submitted 75 runs for evaluation (8 teams submitted working

[1]https://medpix.nlm.nih.gov/home

notes) attracting more attention than in 2020. Each of the groups was allowed a maximum of 10 graded runs per sub task.

Table 1 shows all the teams who participated in the task and their submitted runs. 5 teams participated in the concept detection sub task this year, two of those teams participated also in 2020. 8 teams submitted runs to the caption prediction sub task. However, two teams decided not to submit working notes describing the used techniques.

3. Data Creation

Figure 1 shows an example from the data set provided by the task.

UMLS CUI	**UMLS Meaning**
C0228134	Spinal epidural space
C0223491	Structure of lumbar spinal canal
C0205400	Increased Thickness
C0150312	Present
C0000833	Abscess
C0085222	Psoas Abscess
C0024485	Magnetic Resonance Imaging

Caption :

Enhancing epidural collection lumbar spinal canal.

Dural sac is compressed by collection.

Dural thickening present.

Abscess is primarily posterior in location.

Paraspinal and psoas muscles abscesses noted on right.

Figure 1: Example of a radiology image with the corresponding UMLS®CUIs and captions extracted from the ImageCLEFcaption 2021 task.

In the previous editions, the data set distributed for the task originates from biomedical articles of the PMC Open Access[2] [15]. To make the task more realistic, in this fifth edition, the collection contains real radiology images annotated by medical doctors. The data set used is the same as the ImageCLEFVQA task [5] where radiology images were selected from the MedPix[3] database. Only cases where the diagnosis was made based on the image were selected and their annotations were used as a basis for the extraction of the concepts and captions. A semi-automatic text pre-processing was applied to improve the quality of the data and to extract the concepts (UMLS-CUI) using the captions, location, and diagnosis as filters. The curated data

[2]https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/[last accessed: 27.06.2021]
[3]https://medpix.nlm.nih.gov/home

Table 1
Participating groups in the ImageCLEF 2021 caption task and their runs submitted to both sub tasks: T1-Concept Detection and T2-Caption prediction. Teams with previous participation in 2020 are marked with an asterisk.

Team	Institution	Runs T1	Runs T2
AEHRC-CSIRO [7]	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Australia	-	9
AUEB NLP Group* [8]	Information Processing Laboratory, Department of Informatics, Athens University of Economics and Business, Athens, Greece	10	7
ayushnanda14	Department of Computer Science and Engineering, Siva Subramaniya Nadar College of Engineering, Kalavakkam, India	-	1
IALab PUC [9]	Department of Computer Science, Pontificia Universidad Católica de Chile, Región Metropolitana, Chile	-	7
ImageSem [10]	Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China	9	9
IALab PUC [11]	Department of Computer Science, Pontificia Universidad Católica de Chile, Región Metropolitana, Chile	2	-
jeanbenoit_delbrouck	Laboratory of Quantitative Imaging and Artificial Intelligent, Department of Biomedical Data Science, Stanford University, Stanford, United States	-	3
kdelab [12]	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	-	10
NLIP-Essex*-ITESM [13]	School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK and Instituto Tecnologico y de Estudios Superiores de Monterrey, Monterrey, Mexico	6	-
RomiBed [14]	The Center for machine vision and signal analysis, University of Oulu, Oulu, Finland	2	1

included radiology images categorised into seven sub-classes indicating the image acquisition technique with a corresponding set of concepts.

We have also validated all the captions manually and checked the coherence of the generated concepts in the training, validation, and test sets.

The following subsets were distributed to the participants where each image has one caption and multiple concepts (UMLS-CUI):

- *Training set* including 2,756 images and associated captions and concepts.
- *Validation set* including 500 images and associated captions and concepts.
- *Test set* including 444 images (and associated reference captions and concepts).

4. Evaluation Methodology

The performance evaluation follows the approach used in the previous edition in evaluating both sub tasks separately. For the concept detection sub task, the balanced precision and recall trade-off were measured in terms of F1 scores. Caption prediction performance is assessed on the basis of BLEU-scores [16]. Candidate captions are lower cased, stripped of all punctuation and English stop words. Finally, to increase coverage, Snowball stemming was applied. BLEU-scores are computed per reference image, treating each entire caption as a sentence, even though it may contain multiple natural sentences. Average BLEU-scores across all test images was reported.

5. Results

For the concept detection and caption prediction sub tasks, Tables 2 and 3 show all the results of the participating team. The results will be discussed in this section.

5.1. Results for the Concept Detection sub task

In 2021, five teams participated in the the concept prediction sub task submitting 29 runs. Table 2 presents the results achieved in the submissions.

The AUEB NLP Group from Athens University of Economics (Greece) submitted the best performing result with an F1-score of 0.505 [8]. They submitted the best results also in previous years and extended their earlier work. They used Ensembles of classifiers based on DenseNet-121 [17] and in this year added Networks that have been trained with supervised contrastive learning [18]. These are followed by a feed-forward Neural Network (FFNN), which acts as the classifier layer on the top. Other submissions are more information retrieval oriented and use CNN encoders of recent architectures like EfficientNet-B0 [19] and create an ensemble of image embeddings. The networks were first pre-trained on the ImageNet data set [20] and then fine-tuned using the ImageCLEF 2021 concept detection data set. Several aggregation methods such as the intersection, majority voting, and union of predicted concepts were experimented. The system with the majority voting of concepts from image embeddings achieved the overall highest F1-Score.

The second best system was proposed by NLIP-Essex-ITESM, a joint team from University of Essex (UK) and ITESM (Mexico). They reached an F1-score of 0.469 and a detailed description of their work is presented in [13]. They also proposed two routes, an information retrieval based approach and a multi-label classification system. For the information retrieval approach image embeddings from ImageNet [20] pretrained DenseNet-121 [17] and EfficientNet [19] have been

Table 2
Performance of the participating teams in the ImageCLEF 2021 Concept Detection Task

Group Name	Submission Run	F1-Score
AUEBs_NLP_Group	136458	0.505
AUEBs_NLP_Group	136455	0.495
AUEBs_NLP_Group	135963	0.493
AUEBs_NLP_Group	136052	0.493
AUEBs_NLP_Group	135847	0.490
NLIP-Essex-ITESM	132945	0.469
AUEBs_NLP_Group	135870	0.466
AUEBs_NLP_Group	135862	0.459
AUEBs_NLP_Group	136307	0.456
NLIP-Essex-ITESM	136429	0.451
AUEBs_NLP_Group	135989	0.451
NLIP-Essex-ITESM	136404	0.440
NLIP-Essex-ITESM	136400	0.423
ImageSem	135873	0.419
NLIP-Essex-ITESM	133912	0.412
ImageSem	135871	0.400
ImageSem	136142	0.396
ImageSem	135858	0.380
ImageSem	136129	0.370
IALab_PUC	135810	0.360
NLIP-Essex-ITESM	136379	0.355
ImageSem	136140	0.355
AUEBs_NLP_Group	136371	0.348
ImageSem	136141	0.327
RomiBed	136011	0.143
IALab_PUC	135197	0.141
RomiBed	136025	0.137
ImageSem	136143	0.037
ImageSem	136144	0.019

tested. The multi-label classifier was based on DenseNet-121. The best submission came from the retrieval technique based on DenseNet-121 with cosine similarity.

The ImageSem Group from Chinese Academy of Medical Sciences and Peking Union Medical College (China) reached an F1-score of 0.419 and details are provided in [10]. They used multi-label classification with DenseNet [17] and Inception-V3 [21] networks. Interestingly, they submitted models for subgroups of concepts and reached the best results by predicting only the Imaging Types. The subgroup of Imaging Types contains 99 of the 1,586 concepts from the dataset.

In the concept prediction sub task the IALab PUC from Pontificia Universidad Católica de Chile reached an F1-score of 0.360. The best submission of this group, described in [11] uses image embeddings with Learned Perceptual Image Path Similarity (LPIPS) [22] based on VGG [23] models.

The RomiBed group from University of Oulu (Finland) reached an F1-score of 0.143 and described their approach in [14]. They used an image embedding from a MobileNet-v2 architecture and added a GRU layer for the prediction.

To summarize, in the concept detection sub task, the groups typically used deep learning models trained as multi-label classificators or more Information Retrieval oriented solutions. For the IR solutions, image embeddings from deep learning models are typically used. In this year, more modern deep learning architectures like EfficientNets [19] and Visual Transformers (ViT) [24] were proposed for the solutions.

This year's models for concept detection show again increased F1-scores in comparison to earlier years. This could partly be explained by a smaller number of potential concepts in the images. More modern architectures have been used and show improvements. Transformer-based architectures and solutions arrived at both sub tasks. This year, machine learning-based methods and information retrieval oriented solutions were used more equally by all groups. In former years the majority of proposed solutions used multi-label approaches. A few participants noticed that less complex solutions showed the best results.

Table 3
Performance of the participating teams in the ImageCLEF 2021 Caption Prediction Task

Group Name	Submission Run	BLEU-score
IALab_PUC	136474	0.510
IALab_PUC	136474	0.509
AUEB_NLP_Group	135921	0.461
AUEB_NLP_Group	135921	0.452
AUEB_NLP_Group	135921	0.448
IALab_PUC	136474	0.442
AUEB_NLP_Group	135921	0.440
AEHRC-CSIRO	135507	0.432
AEHRC-CSIRO	135507	0.430
AEHRC-CSIRO	135507	0.426
AEHRC-CSIRO	135507	0.423
AEHRC-CSIRO	135507	0.419
AEHRC-CSIRO	135507	0.416
AEHRC-CSIRO	135507	0.415
AEHRC-CSIRO	135507	0.405
AEHRC-CSIRO	135507	0.388
IALab PUC	136474	0.378
AUEB_NLP_Group	135921	0.375
IALab_PUC	136474	0.370
kdelab	134753	0.362
kdelab	134753	0.362
kdelab	134753	0.362
IALab_PUC	136474	0.354
kdelab	134753	0.352
IALab_PUC	136474	0.351
kdelab	134753	0.339
kdelab	134753	0.297
kdelab	134753	0.291
kdelab	134753	0.287
jeanbenoit_delbrouck	135533	0.285
kdelab	134753	0.280
kdelab	134753	0.267
ImageSem	136138	0.257
jeanbenoit_delbrouck	135533	0.251
jeanbenoit_delbrouck	135533	0.251
RomiBed	135896	0.243
ImageSem	136138	0.203
AUEB_NLP_Group	135921	0.199
ImageSem	136138	0.181
ImageSem	136138	0.137
ayushnanda14	136389	0.103
ImageSem	136138	0.102
ImageSem	136138	0.049
ImageSem	136138	0.038
ImageSem	136138	0.004
ImageSem	136138	0.001

5.2. Results for the Caption prediction task sub task

In this fifth edition, the caption prediction sub task attracted 8 teams which submitted 40 runs. Table 3 presents the results of the submissions. Two groups, jeanbenoit_delbrouck and ayushnanda14 decided not to submit working notes and therefore no description about the approaches is available.

The best model for the caption prediction sub task was presented by IALab PUC from Pontificia Universidad Católica de Chile. They reached a BLEU-score of 0.510 and described the methods in [9]. Three methods were tested, a statistical oriented method, a similarity based on LPIPS [22] and a multi-label classification (MLC) approach. The MLC approach used a ResNet34 [25] network and ordered the predicted caption words based on statistical analysis from the training set.

The AUEB NLP Group from Athens University of Economics (Greece) submitted the second best performing result with a BLEU-score of 0.461 [8]. Two different approaches were tested, a Show Attend and Tell [26] approach and an ensemble of different image embeddings. The best result came from the ensemble with embeddings from CNN architectures like DenseNet [17]. Interestingly, the method used general language models like GPT-2 [27].

A group from the Australian e-Health Research Centre (AEHRC-CSIRO) reached a BLEU-score of 0.432 [7]. The group used modern network architectures like Visual Transformers (ViT) [24] and tested different pre-trainings on medical datasets like ROCO [28] and CheXpert. The best model had the simplest configuration and no pre-training on medical datasets.

The kdelab group from Toyohashi University of Technology (Japan) reached a BLEU-score of 0.362 [12]. They used a standard Show Attend and Tell [26] model and focused their work on image pre-processing like Histogram Normalizations. This improved the result to 0.362 from a baseline model reaching a BLEU-score of 0.339.

The ImageSem Group from Chinese Academy of Medical Sciences and Peking Union Medical College (China) reached a BLEU-score of 0.257 and details are provided in [10]. They used an approach based on sentence patterns for the caption prediction. For this, they used the results from the first sub task on concept detection and inserted the found concepts in caption patterns like: <image> of <body> demonstrate <findings>.

The RomiBed group from University of Oulu (Finland) reached a BLEU-score of 0.243 and described their approach in [14]. They used an attention based encoder-decoder model for the caption prediction.

To summarize, in the caption prediction task, several teams used variations of the Show, Attend and Tell model [26]. New approaches were used such as Transformer-based architectures and general language models like GPT-2 [27]. Transfer Learning has frequently been used and some teams in both sub tasks tried to pretrain with more medically oriented datasets like ROCO [28] or CheXpert. Interestingly, pre-training with medical oriented datasets seem to be not helpful in this task and many groups found that the most simple architectures provided the best results.

6. Conclusion

This year's caption task of ImageCLEF included several changes in comparison to earlier years. The task was divided into two sub tasks, the concept detection sub task which is comparable to the years before and the re-introduced caption prediction sub task. Another difference was the choice of images, which no longer come from publications but from original radiology images and the captions were produced by clinicians. This change was appreciated by the participants to be more realistic. As a result more teams took part in one or both tasks. A few teams saw the concept detection as a prerequisite to the caption prediction task and provided interesting caption template-based solutions for the caption prediction from detected concepts. Others use variations of Show, Attend and Tell for the caption prediction and participated only in the caption prediction. For the concept detection, mainly multi-label classification or more information retrieval oriented solutions based on image embeddings were proposed. In this year more modern neural network architectures like EfficientNets and ViT were used for the images, and Transformers and General Language models used for the texts. Several participants found that the variation of caption texts was lower compared to earlier years. As a result, more simple solutions produced the best results. In consequence, we seek to increase the number of images and concepts for later competitions and try to increase the variation of the caption texts.

Acknowledgments

This work was partially supported by the University of Essex GCRF QR Engagement Fund provided by Research England (grant number G026).

References

[1] B. Ionescu, H. Müller, R. Peteri, 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.

[2] A. García Seco de Herrera, R. Schaer, S. Bromuri, H. Müller, Overview of the ImageCLEF 2016 medical task, in: Working Notes of CLEF 2016 (Cross Language Evaluation Forum), 2016.

[3] C. Eickhoff, I. Schwall, A. Garcia Seco De Herrera, H. Müller, Overview of ImageCLEFcaption 2017–image caption prediction and concept detection for biomedical images, CEUR Workshop Proceedings, 2017.

[4] A. Garcia Seco De Herrera, C. Eickhof, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, CEUR Workshop Proceedings, 2018.

[5] 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.
[6] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) 267–270. doi:10.1093/nar/gkh061.
[7] A. Nicolson, J. Dowling, B. Koopman, AEHRC CSIRO in ImageCLEFmed Caption 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[8] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[9] V. Castro, P. Pino, D. Parra, H. Lobel, PUC Chile team at Caption Prediction: ResNet visual encoding and caption classification with Parametric ReLU, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[10] X. Wang, Z. Guo, C. Xu, L. Sun, J. Li, ImageSem Group at ImageCLEFmed Caption 2021 Task: exploring the clinical significance of the textual descriptions derived from medical images, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[11] G. Schuit, V. Castro, P. Pino, D. Parra, H. Lobel, PUC Chile Team at Concept Detection, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[12] R. Tsuneda, T. Asakawa, M. Aono, Kdelab at ImageCLEF 2021: Medical Caption Prediction with Effective Data Pre-processing and Deep Learning, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[13] 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: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[14] D.-R. Beddiar, M. Oussalah, T. Seppänen, Attention-based CNN-GRU model for automatic medical images captioning: ImageCLEF 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[15] R. J. Roberts, PubMed Central: The GenBank of the published literature, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 381–382. doi:10.1073/pnas.98.2.381.
[16] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[17] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, July 22-25, 2017, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.
[18] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Kr-

ishnan, Supervised contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 18661–18673. URL: https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[19] 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 (ICML 2019): 10-15.06.2019; Long Beach, California, US, volume 97, Long Beach, California, USA, 2019, pp. 6105–6114. URL: http://proceedings.mlr.press/v97/tan19a.html.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (2014). doi:10.1007/s11263-015-0816-y.

[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] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595. doi:10.1109/CVPR.2018.00068.

[23] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv 1409.1556 (2014).

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021): 03-07.05.2021; Online Event, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[25] 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, Las Vegas, USA, June 26 - July 1, 2016, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2015, pp. 2048–2057. URL: http://proceedings.mlr.press/v37/xuc15.html.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Languagemodels are unsupervised multitask learners, Technical Report, Open-AI, 2019.

[28] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: Intravascular Imaging and Computer Assisted Stenting - and - Large-Scale Annotation of Biomedical Data and Expert Label Synthesis - 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, 2018, pp. 180–189. URL: https://doi.org/10.1007/978-3-030-01364-6_20. doi:10.1007/978-3-030-01364-6_20.