Multi-stage Medical Image Captioning using Classification and CLIP

Masaki Aono[1], Hiroki Shinoda[1], Tetsuya Asakawa[1], Kazuki Shimizu[2], Takuya Togawa[2] and Takuyuki Komoda[2]

[1]*Toyohashi University of Technology, 1-1 Hibarigaokam Tempakucho, Toyohashi, Aichi, 441-8580, Japan*
[2]*Toyohashi Heart Center, 21-1Gobutori, Ohyamacho, Toyohashi, Aichi, 441-8071, Japan*

Abstract

We have participated in ImageCLEFmedical2023 caption prediction task alone as team "Bluefield-2023". In this paper, we propose a multi-stage medical image captioning, based on image classification as the early stage and CLIP (Contrastive Language-Image Pre-Training) as the final stage. In order to take advantage of the image classification problem, we have done automatic and semi-automatic grouping of both training and validation images into 7 groups (CT, MRI, Echo, Chest X-ray, X-ray Misc, Special, and Misc groups) analogous to ROCO dataset's predefined classes. The idea is to attempt to utilize CLIP model's image-text matching ability by separating given medical images into similar groups so that we try not to miss the fundamental terms that appear often inside each group. For instance, "MRI" group often includes terms such as "magnetic resonance" and "t1". We avail ourselves of these specific terms in the captions of training dataset and divide images into 7 groups in the first two stages. Then, we apply image classification for unknown (test) images into 7 groups. At the same time, we produce 7 different best CLIP models. In the fourth stage, we load the best CLIP model for each group associated with the captions, which are generated during the third (i.e., classification) stage. Although the final result with test dataset does not end up with what we have anticipated, we believe our approach would shed some light in this type of research.

Keywords

image classification, CLIP, image captioning

1. Introduction

In this paper, we focus on one of the tasks in Image Captioning; i.e., Caption Prediction Task [1]. There are several other tasks in ImageCLEFmedical2023 [2].

Image captioning has been studied almost for decades. It is interpreted as describing the content of a given image in words through natural language processing. Popular benchmark datasets for image caption include COCO [3] and Flickr30k [4]. Show and Tell [5], Show Attend and Tell [6] might belong to earlier approaches to image captioning where in most cases, CNN

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
masaki.aono.ss@tut.jp (M. Aono); shinoda.hiroki.vo@tut.jp (H. Shinoda); asakawa.tetsuya.um@tut.jp (T. Asakawa); shimizu@heart-center.or.jp (K. Shimizu); togawa@heart-center.or.jp (T. Togawa); komoda@heart-center.or.jp (T. Komoda)
https://www.kde.cs.tut.ac.jp/~aono/ (M. Aono)
0000-0003-1383-1076 (M. Aono); 0009-0008-2850-5015 (H. Shinoda); 0000-0002-8345-7094 (T. Asakawa); 0009-0000-3448-7986 (K. Shimizu); 0009-0006-6822-5427 (T. Togawa); 0009-0001-8302-968X (T. Komoda)

(Convolutional Neural Network) models are used as image feature extraction (a.k.a. image encoder), while RNN models such as LSTM are used for output text decoder. Later, CNN encoder has occasionally been replaced by Transformer based models (e.g., Vision Transformer ViT-32 [7]), and LSTM decoder has sometimes replaced by Transformer decoder such as Transform and Tell [8]. For the research related to “Show Attend and Tell" approach, it is noted that Ke et al [9] focused on vocabulary coherence to propose a “Reflective Decoding Network" to boost captioning performance.

On the other hand, image captioning for medical images where gray-scale images is dominant, has shorter history compared with the same task for general color images such as COCO and Flickr30k. As far as the authors can tell, medical image captioning in ImageCLEF has begun sometime around 2017, and has been one of main tasks in ImageCLEFmedical [1].

It should be noted that medical image dataset such as ROCO (Radiology Objects in COntext) [10] has been used in PCM-CLIP [11].

Examples of recent approaches to image-text matching are as follows: CLIP or Contrastive Language-Image Pre-training [12] was proposed for matching images and texts. BLIP or Bootstrapping Language-Image Pre-training [13] was introduced to outperform CLIP by filtering the noise generated by CLIP. BLIP was also used for image captioning as well as visual question answering. BLIP-2 was proposed [14] to generate a descriptive text given an image.

Popular evaluation criteria for medical image captioning include BLEU [15], METEOR [16], and CIDEr [17], typically has been used in machine translation. From year 2023, BERTScore [18] is added as one of the evaluation measures.

2. Proposed Approach

Our proposed approach consists of classification of medical images, per-class CLIP model application, per-class search for most similar image, and the extraction of the caption associated with the image. Detailed processing is elaborated in the following subsections.

2.1. First stage: automatically classifying data into 6 groups

In the first stage toward medical image captioning, we have adopted classification approach. Initially, we observed ImageCLEF2023 caption data (both image and caption) carefully. Then we have decided to partition data into 6 groups; CT, MRI, Echo, Chest X-ray, X-ray Misc, and Misc groups, respectively, as shown in Figure 1 (note that “Misc" stands for Miscellaneous). The reason behind this is our strategy to “divide-and-conquer". To do this, we have looked for unique terms appearing in each group. Examples are as follows: CT group has terms such as “CT" and “computed tomography", MRI group has terms such as “magnetic resonance" and “T1", Echo group has terms such as “ultrasound" and “echocardiogram", Chest X-ray group has terms such as “chest X-ray" and “chest radiograph", X-ray Misc group has terms such as “skull X-ray" and “pelvis", while Misc group has terms such as “angiography" and “LAD". It should be noted that these 6 groups are not at all perfect. Indeed, even the sample terms described above might occur in two or more groups. Nevertheless, we believe that it is preferable to start with the above 6 groups as a rough clue to come closer to the appropriate caption, given unknown medical image with our classification strategy.

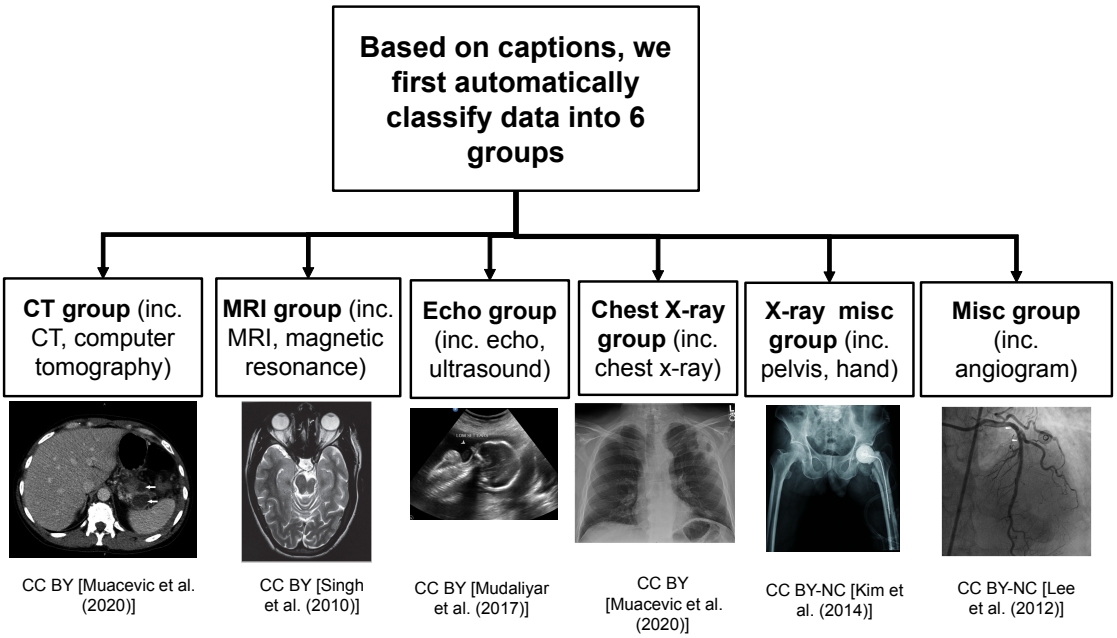


Figure 1: 6 predefined groups; both training and validation data are classified into 6 groups. Based on the terms appearing inside captions

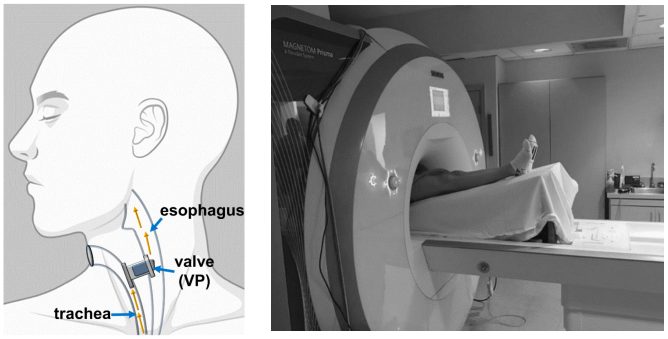


Figure 2: Examples of provided images in the training dataset that never fall into predefined 6 groups, which we put a new “Special" group.

2.2. Second stage: adjusting data and re-classifying them into 7 groups

In the first stage, we have obtained 6 groups of medical images with their captions as the first order approximation. During the first stage, we have realized that the initial crude grouping worked out to some extent, so that it makes it possible to construct answer dataset for deep learning classification. Careful observation of the crude grouping result makes us feel that we

Table 1
Training and validation support for 7 classes

Group name	Training	Validation
CT	21,609	3.924
Echo	8,292	1,554
MRI	8,563	1,420
Misc	6,349	1,047
Special	551	115
Chest X-ray	4,367	815
X-ray Misc	11,187	1,562

need further elaboration on the “Misc" group. In particular, we have found that there are a small group of images that are not falling into any of the 6 groups. Indeed, our initial attempt to make a “Misc" group was to classify medical diagnostic images of real human subjects which are not belonging to other groups. Typical examples of “Misc" group include diagnostic images such as “angiography". However, we have found that there are certain amount of non-diagnostic images including illustrations, graphs, plottings, animals being operated, scenes of an operation room, and photos of medical apparatuses. A couple of examples of such images are shown in Figure 2. Consequently, we thus have constructed a new “Special" group for keeping these non-diagnostic images. well. This ends up the construction of 7 groups from given data for both training and validation datasets.

2.3. Third stage: deep neural network to classify given medical images

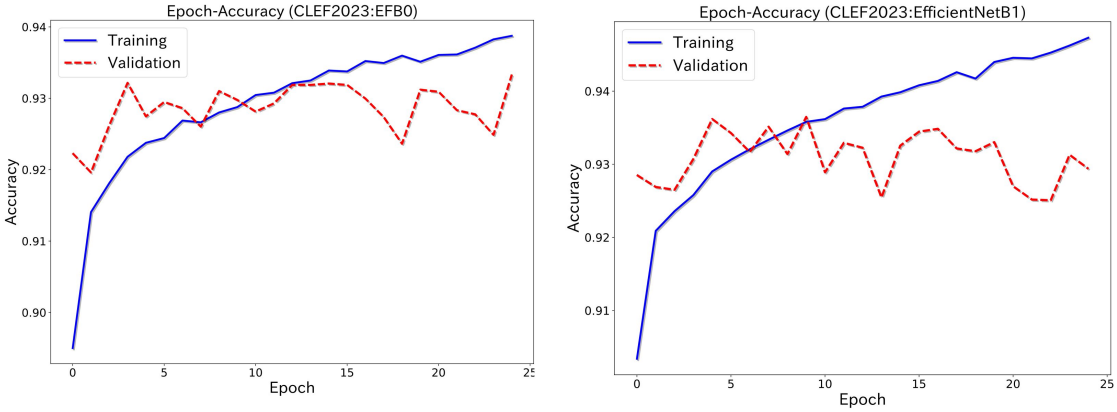
Based on the 7 groups for both training and validation datasets, we formulated a deep neural network to classify images from ImageCLEF2023 caption prediction data. Specifically, we have tested EfficientNetB0 and EfficientNetB1 for our 7-group classification [19] after trial and error of other EfficientNet group CNNs. The reason behind the adoption of these EfficientNet DNNs lies in our relevant research toward the stenosis detection for cardiac CT images where EfficientNet turned out to perform very well among several CNNs and Transformer-based DNNs [7].

Figure 3 (a) shows the Epoch-Accuracy graph of EfficientNetB0, while (b) depicts the Epoch-Accuracy graph of EfficientNetB1 with the data tagged in Table 1. Figure 4 exhibits the confusion matrices of EfficientNetB0 (a) and EfficientNetB1 (b).

The best validation accuracy for the classification with EfficientNetB0 and EfficientNetB1 are 0.9333 and 0.9365, respectively. Hyperparameters for the classifications are the same and they are as follows: 25 epochs, cross entropy loss for the loss function, AdamW optimizer [20] with learning rate 0.001. Our run2 and run3 correspond to these CNNs. Table 2 demonstrates accuracy for each class, where both validation (open) and training (closed) accuracies are shown.

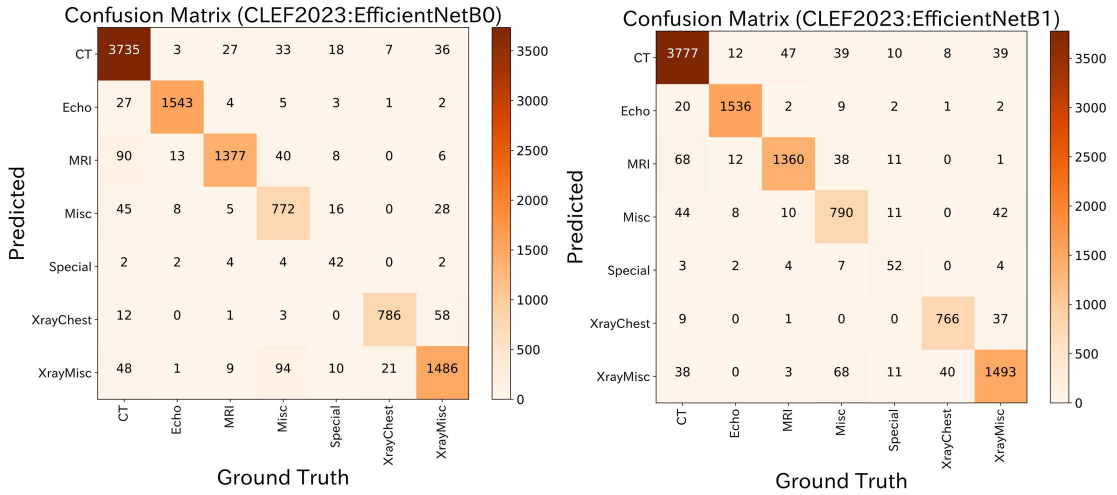
2.4. Fourth stage: Application of CLIP to each class

Once a given unknown image is classified into one of the 7 groups during the third stage, we proceed to the fourth stage. In this stage, we have adopted CLIP, as a multi-modal model, to



(a) EfficientNetB0 (b) EfficientNetB1

Figure 3: Epoch-Accuracy graphs for EfficientB0 (a) and EfficientNetB1 (b)



(a) EfficientNetB0 (b) EfficientNetB1

Figure 4: Confusion matrices for classification with EfficientB0 (a) and EfficientNetB1 (b)

predict caption from image. According to the data shown in Table 1, we applied CLIP training process 7 times for each group. As far as we know, we can choose either Vision Transformer (ViT-32) [7] or ResNet50 as the backbone of the CLIP model [12]. Based on our past experience of the success of ResNet CNN family [21] on medical images where black and white images are dominant, we have chosen ResNet50 for our backbone inside CLIP.

The overall process in the fourth stage is illustrated in Figure 5. Best CLIP model can be found independently for each group as enclosed in dotted rectangle in Figure 5. Given an unknown image from test dataset, we apply our trained classifier mentioned in the third stage to predict

Table 2
Training and validation accuracies (%) for each class with EfficientNetB1

Group name	Training	validation
CT	96.8	95.4
Echo	98.6	97.8
MRI	95.3	95.3
Misc	85.1	83.1
Special	62.6	53.6
Chest X-ray	94.5	94.0
X0ray Misc	91.6	92.3

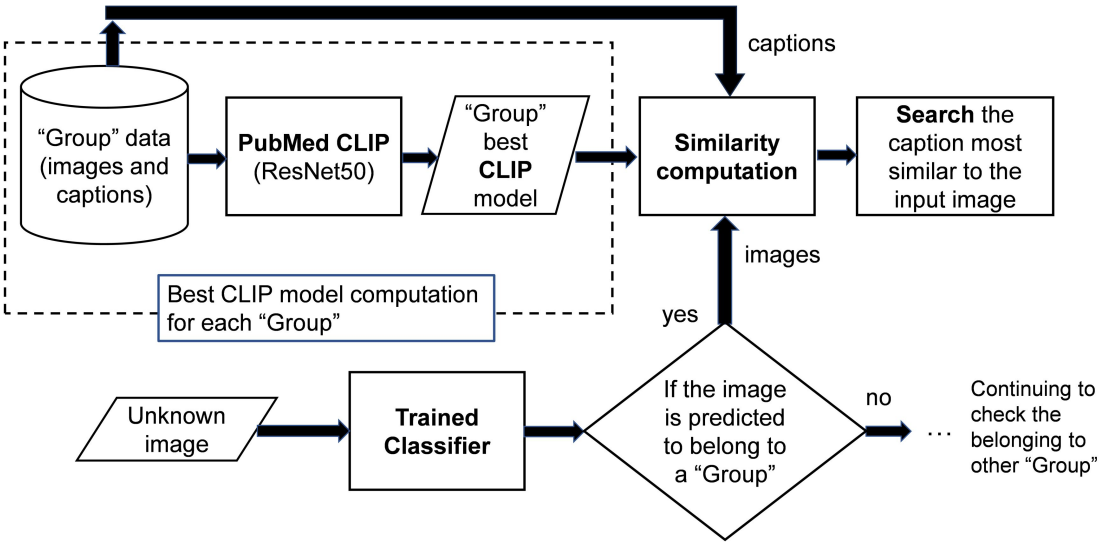


Figure 5: Overall image captioning flow using group-based CLIP model with trained classifier The evaluation result provided by organizer is shown in Table 3.

the group most likely fitting. Then, we proceed to do similarity computation using group-best CLIP model, followed by sorting the confidence values from CLIP model, getting the most similar image, and yielding the caption corresponding to the image. Thus, the caption for the unknown input image can be retrieved from the captions belonging to the predicted group.

Our evaluation results (run2 and run3) with test data received from organizers are shown in Table 3. Although our results may not yield good performance, we believe that our proposed classification-based method shed some light in the future research. For example, it might be an idea to apply "Show Attend and Tell" to each group after classification into 7 groups to see what the results would look like.

Table 3
Test data accuracies (%) of our runs for test data. Run2 was based on EfficientNetB0, while Run3 was based on EfficientNetB1 as pretrained classifiers in the third stage.

Bluefield run	BERT Score	ROUGE	BLUERT	METEOR	CIDEr	CLIP Score
Run2	0.5779	0.15344	0.27164	0.15431	0.06006	0.10091
Run3	0.5776	0.15392	0.27136	0.15404	0.06970	0.10482

3. Conclusion

In this paper, we proposed a multi-stage image captioning, based on image classification as the early stage and CLIP (Contrastive Language-Image Pre-Training) as the final stage. For classification into 7 groups, we combined automatic grouping by means of terms specific to each group, yielding 7 groups of training and validation dataset. Our implementation with EfficientNet for the classification resulted in more than 93 % on the average. However, CLIP of the last stage needs more improvement, judging from the result from organizers. Nevertheless, we believe that our approach sheds some light on the applications with medical images and their captions.

4. Discussion

As we have demonstrated, the accuracy of classification in the third stage turned out to be quite good in terms of Figures 3 and 4, as well as Table 2 except group “Special" introduced in the second stage. It might be controversial to introduce such a separate group with non-diagnostic images including illustrations and medical apparatuses as “Special" group, apart from “Misc" (i.e., Miscellaneous) group. An improvement to identify “Special" group in the second stateg might be to classify given images of this group automatically based on specific vocabularies by taking close look at the associated captions, which should allow us to reproduce our approach better.

Using CLIP and the subsequent retrieval of captions in the training data could be improved. For instance, during similarity computation, we have trimmed the length of captions into somewhere from 70 to 80 words, while when we retrieve the real caption in the given dataset, we return the caption with the original whole length. We have examined, with the validation data, the trimming effect of the caption length after submission, turned out to improve a few percentages in terms of BERT Score. Other evaluation criteria might yield similar improvements.

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] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[2] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.

[4] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2641–2649. doi:10.1109/ICCV.2015.303.

[5] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 2048–2057. URL: https://proceedings.mlr.press/v37/xuc15.html.

[7] 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: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[8] A. Tran, A. Mathews, L. Xie, Transform and tell: Entity-aware news image captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[9] L. Ke, W. Pei, R. Li, X. Shen, Y.-W. Tai, Reflective decoding network for image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[10] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): A multimodal image dataset, in: D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi,

V. Cheplygina, D. Mateus, E. Trucco, E. Granger, P. Jannin (Eds.), Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer International Publishing, Cham, 2018, pp. 180–189.

[11] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, W. Xie, Pmc-clip: Contrastive language-image pre-training using biomedical documents, 2023. arXiv:2303.07240.

[12] M. V. Conde, K. Turgutlu, Clip-art: Contrastive pre-training for fine-grained art classification, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 3951–3955. doi:10.1109/CVPRW53098.2021.00444.

[13] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 12888–12900. URL: https://proceedings.mlr.press/v162/li22n.html.

[14] Y. Tewel, Y. Shalev, I. Schwartz, L. Wolf, Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17918–17928.

[15] 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, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[16] A. Lavie, A. Agarwal, METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 228–231. URL: https://aclanthology.org/W07-0734.

[17] G. Oliveira dos Santos, E. L. Colombini, S. Avila, CIDEr-R: Robust consensus-based image description evaluation, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, Online, 2021, pp. 351–360. URL: https://aclanthology.org/2021.wnut-1.39. doi:10.18653/v1/2021.wnut-1.39.

[18] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[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, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[20] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[21] 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.