Medical Image Captioning using CUI-based Classification and Feature Similarity

Masaki Aono[1], Tetsuya Asakawa[1], Kazuki Shimizu[2] and Kei Nomura[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 the ImageCLEFmedical2024 caption prediction subtask as team "KDE-MED-CAPTION". Although the caption detection subtask is not a direct target of our team, we have taken advantage of the CUI (Concept Unique Identifier) codes attached to each piece of data, provided for the caption detection subtask. In this paper, we propose a method for medical image captioning, using CUI-based classification, followed by feature similarity. Specifically, according to the frequency of concepts (CUIs) of the training data, we perform multi-class classification in the first step, where we use several different Deep Neural Networks (DNNs), varying the number of classes. Once the classification is done, we partition images and captions into the specified number of classes. We then extract features from trained DNNs for training, validation, and test datasets provided by organizers this year from ROCOv2-2023 data, where ROCO is an acronym for Radiology Objects in COntext. Given a test image, we apply our pre-trained DNN to predict the most likely class. Finally, we conduct similarity computation between the feature of a test image and the features of the class predicted, ranking the result by sorting in descending order, so that the caption from the top 1 ranked data, which we believe should be the most appropriate caption for the image. We repeat this process for all the test images, varying the DNNs and the number of classes we have adopted, ending up with 5 runs in our team.

Keywords

image classification, Concept Unique Identifier, image captioning

1. Introduction

ImageCLEF has a long history in the development of technologies for annotation, indexing, classification and retrieval of multimodal data, applicable to large volumes of multimodal data across multiple application scenarios and domains. This year, among several application scenarios in ImageCLEF2024 [1], we have participated in the ImageCLEFmedical2024 caption prediction subtask as team "KDE-MED-CAPTION".

In this paper, we focus on one of the subtasks in Image Captioning [2]; i.e., Caption Prediction Task [3]. There are several other tasks in ImageCLEFmedical2024 including Visual Question Answering, MEDIQA-MAGIC, and Controlling the Quality of Synthetic Medical Images created by GANs.

Here we propose a method for medical image captioning, using CUI-based classification, followed by feature similarity. Specifically, according to the frequency of concepts (CUIs) of the training data, we perform multi-class classification in the first step, where we employ several different Deep Neural Networks (DNNs), varying the number of classes. Once the classification is done, we partition images and captions into the specified number of classes. We then extract features from the trained DNNs for training, validation, and test datasets provided by the organizers this year from ROCOv2-2023 data [3], where ROCO is an acronym of Radiology Objects in COntext. Given an unknown set of test images, we apply our pre-trained DNN to predict the most likely class. Finally, we perform similarity computation between the feature of a test image and the features of the class predicted, ranking the result by sorting in descending order, yielding the caption from the top 1 ranked data, which we believe should be most

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
masaki.aono.ss@tut.jp (M. Aono); asakawa.tetsuya.um@tut.jp (T. Asakawa); shimizu@heart-center.or.jp (K. Shimizu); kein312@gmail.com (K. Nomura)
https://www.kde.cs.tut.ac.jp/~aono/ (M. Aono)
0000-0003-1383-1076 (M. Aono); 0000-0002-8345-7094 (T. Asakawa); 0009-0000-3448-7986 (K. Shimizu); 0000-0003-2838-7844 (K. Nomura)

appropriate caption for the image. We repeat this process for all the test images, varying the DNNs and the number of classes we have adopted.

2. Related Work

Image captioning has been studied almost for decades. It is interpreted as describing the content of a given image in words ordinarily through natural language processing. Popular benchmark datasets for image caption include COCO [4] and Flickr30k [5]. Show and Tell [6], Show Attend and Tell [7] might belong to earlier approaches to image captioning where in most cases, CNN (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 [8]), and LSTM decoder has sometimes been replaced by Transformer decoder such as Transform and Tell [9]. For the research related to “Show Attend and Tell" approach, it is noted that Ke et al [10] 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 are 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.

It should be noted that medical image dataset such as ROCO (Radiology Objects in COntext) [3] 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. Last year, our team used CLIP-based approach to ImageCLEF2023 medical image captioning [15].

3. Evaluation Measure

Popular evaluation criteria for medical image captioning include BLEU [16], ROUGE [17], METEOR [18], and CIDEr [19], typically has been used in machine translation.

Since 2023, CLIPScore [20] and BERTScore [21] have been added as evaluation measures. From year 2024, other evaluation criteria are added, including BLEURT [22], RefCLIPScore [20], Clinical-BLEURT [23], and MedBERTScore [23]. Currently BERTScore is the official main evaluation measure for caption prediction subtask.

4. Proposed Approach

Since the beginning of the Medical Image Caption task in ImageCLEF2017, the number of medical images and their associated captions has increased year by year. At the same time, medical images themselves have become more diverse and multimodal in terms of UMLS (or CUI (Concept Unique Identifier)) concepts. The volume of data itself has also increased to the extent that the total word vocabulary returned by an ordinary tokenizer (e.g., NLTK word tokenizer [24]) exceeds 100,000 even on training data alone. For example, for diversification and multimodality, it is possible to take an image of an organ with different devices such as CT, MRI, and ultrasound. If we could pre-classify medical images into different classes, it may be possible to reduce the risk of adding “t1" and “t2" in captions for CT and ultrasound images, as these words are specific to medical images taken by MRI devices. We have also experienced that a recent deep learning algorithm suffers from computational time and space, due to the larger amount of data to process in an epoch loop. The idea of our approach is rooted in the above observation.

4.1. CUI-based Partition of Data

Based on the above observation, we decided that a DAC (Divide-and-Conquer) approach would be appropriate for handling this amount of data. In implement the DAC approach, we have taken advantage of CUI codes or UMLS concepts, for the subtask of caption detection. First, we took statistics (i.e., frequency) of CUI codes for the training data. The top twelve frequencies of CUI codes are shown below, where the first column corresponds to the frequency number of each CUI code, the second column corresponds to the short description of the code, and the third column is the CUI code itself in square brackets:

```
20536 X-Ray Computed Tomography [ C0040405 ]
16679 Plain x-ray [ C1306645 ]
9841 Magnetic Resonance Imaging [ C0024485 ]
8333 Ultrasonography [ C0041618 ]
7530 Chest [ C0817096 ]
7148 Anterior-Posterior [ C1999039 ]
3965 Angiogram [ C0002978 ]
3401 Postero-Anterior [ C1996865 ]
3340 Bone structure of cranium [ C0037303 ]
3184 Abdomen [ C0000726 ]
2688 Pelvis [ C0030797 ]
2351 Lower Extremity [ C0023216 ]
```

A natural DAC approach might be to take the top of several classes according to the ranking as above. We have noticed that the top two (i.e., CT (Computed Tomography) and Plain X-ray) have a large numerical dominance compared to other CUI codes. It should also be noted that each image has one or more CUI codes. Taking this into account, we believe that by appropriately prioritizing the CUI codes, it is possible, to some extent to balance the number of data in each class. For example, a chest x-ray image should naturally be classified in “X-ray group" without fine level priorities. However, if we have “Chest group" and a higher priority than “X-ray group", we believe “Chest group" takes advantage of collecting chest x-ray images before “X-ray group" devours them. Specifically, we have developed an approach to partition the data based on two priorities using the following algorithm:

Algorithm 1: Priority-based Partition

```
input dataset, itemset, priority1, priority2
begin
    for key ∈ dataset.keys() do
        /* key is an image file name */
        datalist ⟵ dataset[key]
        done ⟵ False
        for item ∈ datalist do
            if item ∈ priority1 then
                /* if CUI code is included in priority1 */
                itemset[item].add(key)
                done ⟵ True
        if done then
            continue
        else
            item ⟵ datalist[0]/* head element */
            if item ∈ priority2 then
                /* if CUI code is included in priority2 */
                itemset[item].add(key)
```

In effect, for the case of 10 groups of training data, with the first scan by priority1 = [‘C0002978’,‘C0041618’, ‘C0037303’, ‘C0023216’, ‘C1140618’, ‘C0817096’] and the second scan by priority2 = [‘C0024485’, ‘C0040405’, ‘C1306645’], we have the following number:

```
[ C0002978 ] Angiography, 3965
[ C0041618 ] Echo (Ultrasonography), 8333
[ C0037303 ] Bone, 3323
[ C0023216 ] Lower (Extremity), 2331
[ C1140618 ] Upper (Extremity), 1126
[ C0817096 ] Chest, 7411
[ C0024485 ] MRI, 9729
[ C0040405 ] CT, 18686
[ C1306645 ] (Plain) X-ray, 4435
[ other CUIs] Misc, 824
```

In this way, introducing a priority-based partitioning algorithm makes it possible to decrease the dominance of CT and (plain) X-ray group numbers. Note that Bone, Chest, Upper, and Lower groups’ CUI might be taken from not always the first column in caption concepts data as illustrated in Figure 1.

4.2. Caption Preprocessing

Apart from partitioning given data, in order to infer a better caption for each unknown image in the test dataset, we have applied simple filtering to remove some of the verbose expressions in the caption files in the 2023 data. Our filtering strategy is as follows:

- To remove expressions containing “Permission” or “Source” in the last part of the caption
- To remove expressions related to annotations from images such as (arrow), (arrowhead), and (circle)
- To trim too long captions
- To replace number expressions such as Figure 1 by (plain) Figure
- To remove the starting sentence’s parentheses such as (1) and (a)

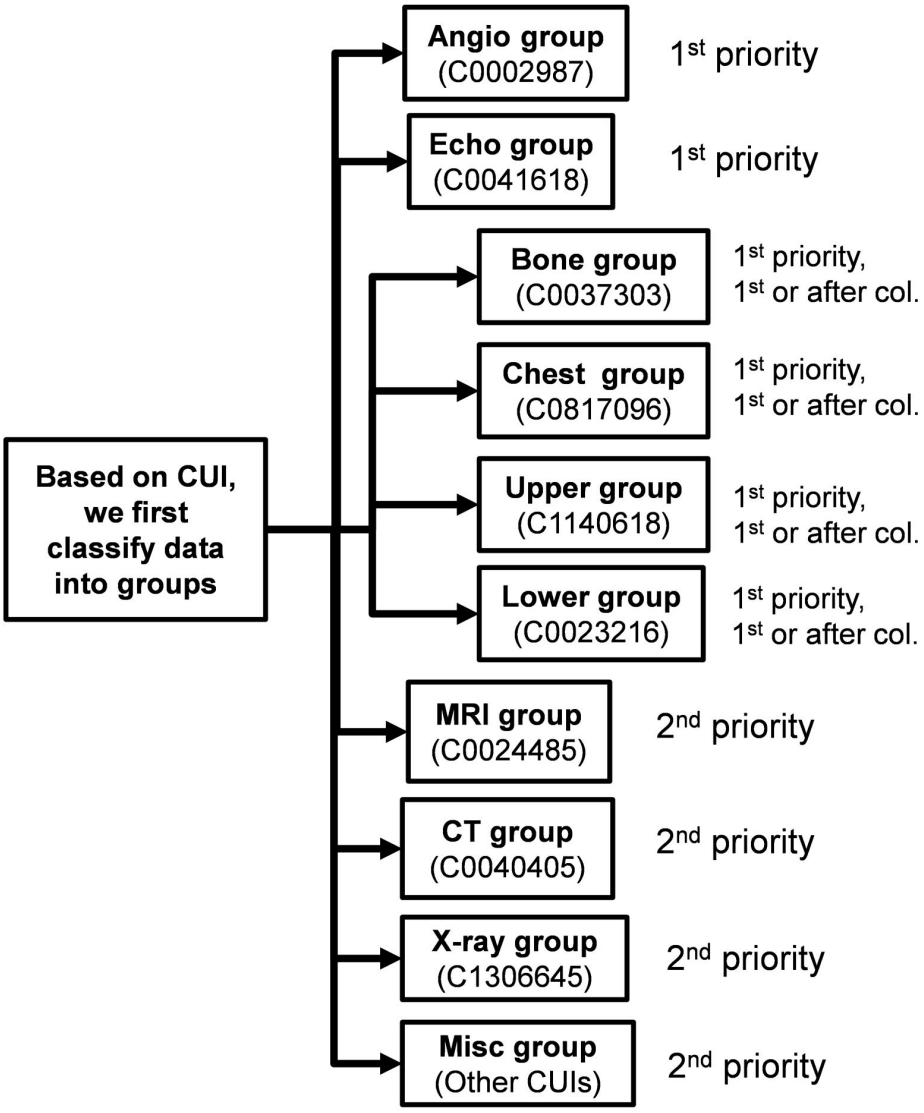

Figure 1: 10 predefined groups and when to partition; one of our typical configurations, although we allow 11 and 6 predefined versions.

The idea behind this preprocessing comes from the fact that we do not synthesize an image caption, but rather search for the best matching caption from the 2023 caption data. In other words, we think that the retrieved caption should be as simple sentences as possible, so that acknowledgment of the source (item 1), annotations (item 2), excessively long and detailed captions (item 3), figure numbers (item 4), and non-ascii symbols at the beginning of the caption (item 5) should be avoided as much as possible. As a specific example of “excessibly long and detailed caption", in the training caption dataset, ROCOv2_2023_train_046629 has more than 400 words.

4.3. Classification of Each Group by Deep Neural Networks

Based on the subdivided groups as described in the previous sections, we have formulated deep neural networks to classify images from ImageCLEF2024 caption prediction data. Specifically, we have adopted EfficientNet [25], ResNeXt [26], and Vision Transformer [8] of our deep neural networks for classification. The reason for adopting these DNN models lies in our relevant research into the detection of cardiac diseases where these DNNs turned out to perform quite well among several DNNs. The output of each image by these DNNs is a feature vector. EfficientNet (or EfficientNetB0) has 1,280 dimensions,

ResNeXt (or ResNeXt50) has 2,048 dimensions, and Vision Transformer (or ViT32) has 768 dimensions, respectively. The feature vector is extracted from one layer before the last one of the original DNN, pre-trained by ImageNet. It should be noted that we have performed each classification with a specific DNN using training data from ImageCLEF2023 to train the DNN and validation data to evaluate the DNN in PyTorch. Then, we have extracted feature vectors from all the data of each group, as well as from the test data of ImageCLEF2024.

4.4. Similarity Computation

In this stage, we compute the similarity between the feature vectors in the test dataset and feature vectors in each group. Then we sort the result in descending order. We simply select the most similar element and return the caption of the element.

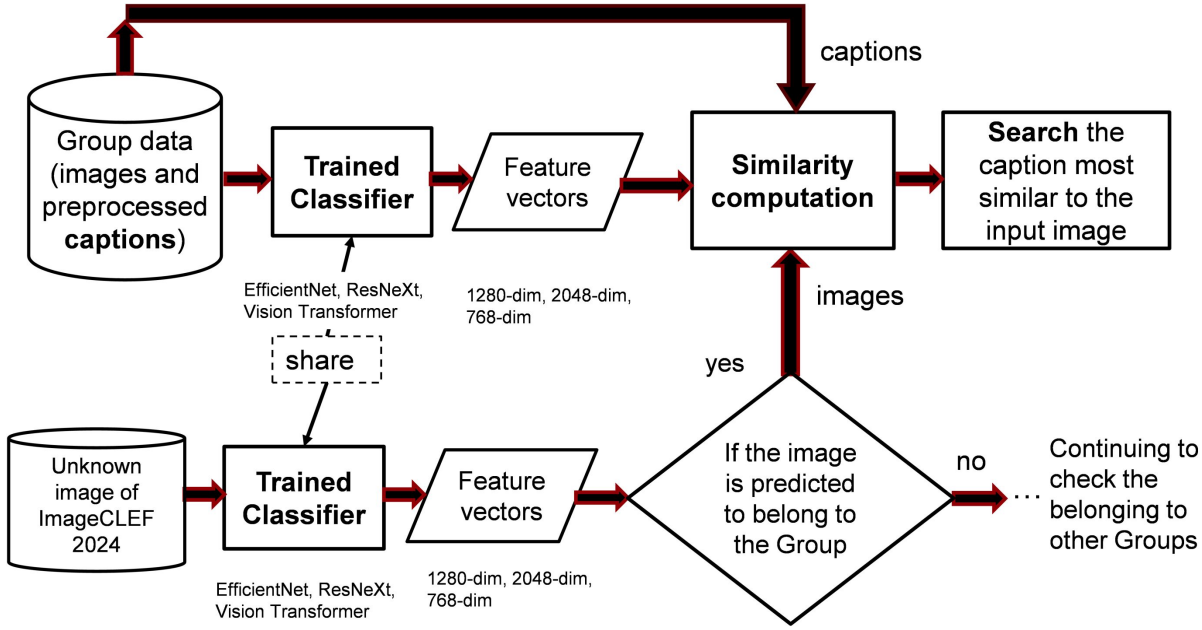

Figure 2: Overall image captioning flow using group-based method with trained classifier.

The overall process in this stage is illustrated in Figure 2. Thus, the caption for the unknown input image can be retrieved from the captions belonging to the predicted group.

4.5. Detailed Experiments and Results

Our evaluation results (run1 to run5) with the test data received from the organizers are summarized in Table 1. As shown in Table 1, we have tested three different numbers of groups; 6, 10, and 11. Initially we thought 10 groups seemed the best among these. Thus, in Runs 1,2, and 3, we keep DNN as EfficientNetB0. It turned out that 11 groups seemed a slightly better than the other number of groups. In practice, however, we kept 10 groups while varying DNNs (from EfficientNet to ResNeXt and ViT32). As a result, the ViT32 DNN model with 10 groups proved to be the best among our submissions.

Table 2 summarizes the numbers of all the evaluation measures returned by the roganizer.

5. Conclusion

In this paper, we proposed a method for medical image captioning, using CUI-based classification, followed by feature similarity. Specifically, we took DAC (Divide-and-Conquer) approach, where we first divided the data (training, validation, and test data of ImageCLEF2023) into 5, 10, or 11 groups (classes)

Table 1
BERT Scores of our runs for test data.

Run No.	Run ID	BERT Score	DNN model	No. of groups
Run1	423	0.5646478601	EfficientNetB0	6
Run2	424	0.5646478602	EfficientNetB0	11
Run3	460	0.5630373607	EfficientNetB0	10
Run4	544	0.5664648379	ResNeXt50	10
Run5	557	0.5673294605	ViT32	10

in advance. We then used several Deep Neural Network (DNN) models, including EfficientNet, ResNeXt, and Vision Transformer, to the training data to obtain a trained DNN model from each DNN. For images in each group, we extracted features from one layer before the last layer of the pretrained DNN. Similarly, we extracted features from the test images. Finally, we computed feature vector similarity between these two to find out the most similar image in each group. We returned the caption corresponding to the most similar image from each group.

Even though we took a simple and no natural language processing approach, the result returned by organizers showed that the BERT Score of our team seemed quite close to other teams.

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] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[2] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[3] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset, Scientific Data (2024). doi:10.1038/s41597-024-03496-6.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part VIII, Springer-Verlag, Berlin, Heidelberg, 2023, p. 525–536. URL: https://doi.org/10.1007/978-3-031-43993-3_51. doi:10.1007/978-3-031-43993-3_51.

[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] M. Aono, H. Shinoda, T. Asakawa, K. Shimizu, T. Togawa, T. Komoda, Multi-stage Medical Image Captioning using Classification and CLIP, Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), 2023. URL: https://ceur-ws.org/Vol-3497/paper-113.pdf.

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

[17] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[18] 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.

[19] 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.

[20] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for

Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. URL: https://aclanthology.org/2021.emnlp-main.595. doi:10.18653/v1/2021.emnlp-main.595.

[21] 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.

[22] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704.

[23] A. Ben Abacha, W.-w. Yim, G. Michalopoulos, T. Lin, An investigation of evaluation methods in automatic medical note generation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2575–2588. URL: https://aclanthology.org/2023.findings-acl.161. doi:10.18653/v1/2023.findings-acl.161.

[24] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly, Beijing, 2009. URL: https://www.nltk.org/book. doi:https://my.safaribooksonline.com/9780596516499.

[25] 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.

[26] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 5987–5995. URL: https://doi.org/10.1109/CVPR.2017.634. doi:10.1109/CVPR.2017.634.

Table 2
Evaluation result of our best Run returned by the organizers

Run ID	BERT Score	ROUGE	BLUE-1	BLEURT	METEOR	CIDEr	CLIP Score	RefCLIPScore	ClinicalBLEURT	MedBERTScore
557	0.567329461	0.132495659	0.106024751	0.256576245	0.038628194	0.038403904	0.765058846	0.760957958	0.502233801	0.569658518