KDE Lab at ImageCLEFmedical Caption 2023

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

[1]Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, Japan, 441-8580
[2]Toyohashi Heart Center, 21-1 Gobutori, Oyama-cho, Toyohashi, Aichi, Japan, 441-8530

Abstract

This paper describes KDE Lab approach in ImageCLEFmedical Prediction Task 2023. ImageCLEFmedical Caption Task 2023 consists of two sub tasks, Caption Prediction and Concept Detection. Concept Detection aims to identify concepts from medical images. Caption Prediction generates description from medical images. For experiment, we applied retrieval approach and deep learning approach for tasks. In Concept Detection, we employed two methods that are fine-tuned Convolutional Neural Network (CNN) approach and retrieval approach based K-Nearest Neighbor (KNN) and cosine similarity using CNN features. In Caption Prediction, we attempted four methods that are retrieval approach based cosine similarity using Term Frequency-Inverse Document Frequency (TF-IDF) features, Show and Tell, Show, Attend and Tell, Caption Transformer. Finally, our submission with fine-tuned ResNet-152 achieved 2nd place in the Concept Detection. Additionally, our submission with Show, Attend and Tell achieved 6th place in Caption Prediction.

Keywords

Medical Images, Concept Detection, Caption Prediction, Image Captioning, CNN-RNN, Caption Transformer

1. Introduction

Image CLEF has been held as part of CLEF since 2003. Image CLEF 2023 [1] focus on multiple applications between different tasks, and ImageCLEFmedical Caption Task [2] is one of them. ImageCLEFmedical Caption Task 2023 [2] consists of two sub tasks, Caption Prediction and Concept Detection. Caption Prediction identify concepts based on the Unified Medical Language System (UMLS) [3] from medical images. Caption Prediction generates description from medical images.

In this paper, we describe the KDE Lab approach. For Concept Detection, we employed two approaches that are fine-tuned Convolutional Neural Network (CNN) approach and retrieval approach. First approach is retrieval approach that is based on AUEB model [4] in ImageCLEFmedical Caption Task 2022. We extracted features from medical images using CNN pre-trained on ImageNet [5], DenseNet-121 [6] and ResNet-152 [7], then we selected top k

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
shinoda.hiroki.vo@tut.jp (H. Shinoda); masaki.aono.ss@tut.jp (M. Aono); asakawa.tetsuya.um@tut.jp (T. Asakawa); shimizu@heart-center.or.jp (K. Shimizu); komoda@heartcenter.or.jp (T. Komoda); t.togawa0316@gmail.com (T. Togawa)
0009-0008-2850-5015 (H. Shinoda); 0000-0003-1383-1076 (M. Aono); 0000-0002-8345-7094 (T. Asakawa); 0009-0000-3448-7986 (K. Shimizu); 0009-0001-8302-968X (T. Komoda); 0009-0006-6822-5427 (T. Togawa)

concepts with cosine similarity. Second approach is fine-tuned CNN approach that is employed DenseNet-121 [6], EfficientNet-B0 [8], EfficientNetV2-M [9] and ResNet-152 [7]. We add Feed-Forward Neural Network(FFNN) as multi-label classifier to CNN pre-trained on ImageNet [5], then we fine-tune our model. Finally, we classify concepts with trained model. For Caption Prediction, we attempted four approaches that are retrieval approach, Show and Tell [10], Show, Attend and Tell [11], Caption Transformer [12]. Retrieval approach is based on AUEB model [4] in ImageCLEFmedical Caption Task 2022. We extracted features from medical images with CNN that is employed DenseNet-121 [6] and ResNet-152 [7] pre-trained on ImageNet [5]. Then we extracted Term Frequency-Inverse Document Frequency (TF-IDF) features among top k sentences in the term of feature similarity. Finally, we selected the highest similarity among TF-IDF features. Show and Tell [10] and Show, Attend and Tell [11] are Convolutional Neural Network and Recurrent Neural Network (CNN-RNN) approaches. In the architecture, we employed DenseNet-121 [6], EfficientNet-B0 [8], EfficientNetV2-M [9] and ResNet-152 [7] as CNN and Long Short-Term Memory (LSTM) [13] as Recurrent Neural Network (RNN). Furthermore, we experimented with the Caption Transformer [12], which consists of a Transformer [14] encoder and decoder.

We submitted ten submissions for both tasks. As a result, our 10th submission for Concept Detection achieved 2nd place, and our 3rd submission for Caption Prediction achieved 6th place. In the following, we will describe dataset, methods, experiment and results.

2. Dataset

In this section, we describe dataset for both tasks in ImageCLEFmedical Caption Task 2023[2]. Dataset including images, captions and concepts is extended of the Radiology Objects in COntext (ROCO) dataset [15]. The number of images is broken down into the training set, which consists of 60,918 images; the validation set, consisting of 10,437 images; and the test set, which comprises 10,473 images. The type of images in the dataset include various modalities such as CT, MRI, and X-ray.

2.1. Concept Detection

In Concept Detection, dataset consists of pair of image and concepts generated by UMLS [3]. Table 1 shows the top 10 ranking concepts in terms of frequency in training set. The highest concept is C0040405 that appears 20,955 times. In dataset, the number of unique concepts is 2,125, which is less than last year.

For experiment, we used dataset for training and tuning models. Additionally, the proportion of each set is the same as provided dataset.

2.2. Caption Prediction

In Caption Prediction, dataset consists of pair of images and captions. Table 2 shows the top 10 ranking words in terms of frequency in training set. In training set with stop-words, the is the most frequency word, which appears 86,173 times. After to remove stop-words, showing that

Table 1
Top 10 concepts by frequency in training set for Concept Detection.

Rank	Concept	Name	Frequency
1	C0040405	X-Ray Computed Tomography	20,955
2	C1306645	Plain x-ray	17,108
3	C0024485	Magnetic Resonance Imaging	10,062
4	C0041618	Ultrasonography	8,390
5	C0817096	Chest	6,805
6	C1999039	Anterior-Posterior	5,907
7	C0449900	Contrast used	4,945
8	C0002978	angiogram	4,194
9	C0037303	Bone structure of cranium	3,058
10	C1996865	Postero-Anterior	2,911

Table 2
Top 10 words by frequency in Caption Prediction.

All words (include stop-words)		
Rank	Word	Frequency
1	the	86,173
2	.	74,743
3	of	59,286
4	)	35,865
5	(	35,770
6	,	31,015
7	and	28,532
8	in	23,855
9	with	21,789
10	a	21,522

Exclude stop-words		
Rank	Word	Frequency
1	showing	16,849
2	right	13,475
3	arrow	13,383
4	left	13,250
5	CT	12,836
6	image	8,397
7	The	8,123
8	scan	7,960
9	tomography	7,006
10	shows	6,801

appears 16,849 times is the most frequency word. According our analysis, the longest length of captions in dataset is 469 words and the shortest is 1 word.

For experiment, we used dataset for training and tuning models. Additionally, the proportion of each set is the same as provided dataset.

3. Methods

In this section, we describe the methods used for the submission to Concept Detection and Caption Prediction.

3.1. Concept Detection

We employed two approaches that are fine-tuned CNN approach and retrieval approach for Concept Detection. Additionally, we attempted two preprocessing types to images.

3.1.1. Preprocessing

Images in the dataset have two types that are color and grayscale. Therefore, we attempted grayscale transform and colorization, as below.

- Grayscale Transform : We converted to grayscale(Y) from color(RGB), as below.

$$\mathrm{Y} = 0.299 \times \mathrm{R} + 0.587 \times \mathrm{G} + 0.114 \times \mathrm{B} \tag{1}$$

- Colorization : Grayscale images consist of one channel, hence we stacked channel to increase to three from one.

For training images, we applied random cropping, resizing and horizontal flip as data augmentation. First, we applied random cropping. We calculated cropping height ch_{train} and width cw_{train} from original image height h and width w:

$$ch_{\text{train}} = \sqrt{\frac{w \times h \times s}{r}} \tag{2}$$

$$cw_{\text{train}} = \sqrt{w \times h \times s \times r} \tag{3}$$

where s denotes scaling, and r is aspect ration. The parameter of scaling s is assigned random value from 0.08 to 1.0, and the parameter of aspect ration is assigned random value from 3/4 to 4/3. Then, we applied resizing. The size of resizing is assigned height rh_{train}=224 and width rw_{train}=224. Finally, we applied horizontal flip. The probability of horizontal flip is assigned p=0.5. For test images, we applied resizing and center cropping. The parameters for operations, the resizing size is assigned height rh_{test}=256 and width rw_{test}=256 and center cropping size is assigned height ch_{test}=224 and width cw_{test}=224.

3.1.2. Retrieval Approach

Retrieval approach achieved the best result in ImageCLEFmedical Caption Task 2021 [16]. Furthermore, many teams [4, 17, 18] attempted retrieval approach in ImageCLEFmedical Caption Task 2022. Thus, we experimented retrieval approach employed CNN and K-Nearest Neighbor (KNN).

Our approach employed CNN and KNN based on AUEB Lab's approach [4] in ImageCLEFmedical Caption Task 2022. First, we fine-tuned CNN pre-trained on ImageNet [5] with medical images. We employed DenseNet-121 [6] and ResNet-152 [7] as CNN. Then, we extracted features from medical images with CNN excluding final feed-forward layer. Additionally, we calculated cosine similarity between feature extracted from test image and features extracted from train images. Finally, we selected concepts by weighted majority decision among top k data with high similarity. Weights are assigned as the reciprocal of the ranking. In ImageCLEFmedical Caption Task 2022, some teams applied KNN assigned $k = 1$. Additionally, if the parameter k is assigned the larger value, the more time it takes. Therefore, we explored the parameter k in KNN from 1 to 50 in increments of 10. Consequently, we assigned $k = 10, 20$, because F1 Score is the highest in validation set. Figure 1 shows the architecture of our retrieval approach for Concept Detection.

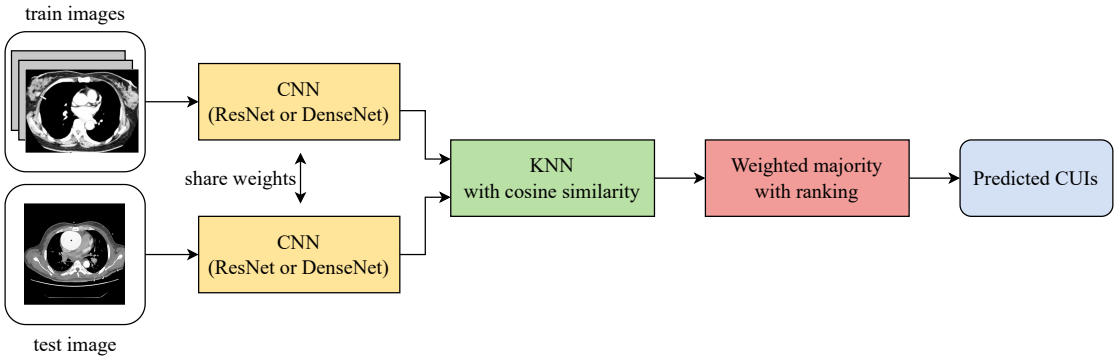

Figure 1: Flow of our retrieval approach(CNN + KNN). CNN for test image and train images shared weights. CC BY [Ng et al. (2015)], CC BY-NC [Al Mulhim et al. (2022)].

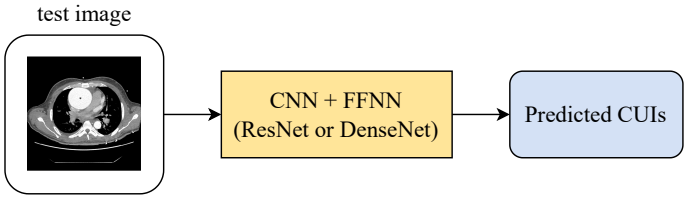

Figure 2: Flow of our fine-tuned CNN approach. CC BY-NC [Al Mulhim et al. (2022)].

3.1.3. Fine-tuned CNN Approach

In ImageCLEFmedical Caption Task 2022, many teams applied various deep learning approach [4, 17, 18, 19].

We attempted fine-tuned CNN approach with four different CNN that employed ResNet-152 [7], EfficientNet-B0 [8], EfficientNetV2-M [9] and DenseNet-121 [6]. Our model consists of fine-tuned CNN and FFNN. First, we fine-tuned CNN with medical images in training set. Then, we extracted features from medical images in test set. Finally, we identified concepts using our models. For Experimental details, Adam optimizer is used with learning rate 10^{-4}. Additionally, Binary Cross Entropy is used as loss function. Using validation data, we evaluated combination between model and preprocessing, and submitted six models with high F1 Score. Submitted models are two combinations. First one is colorization and EfficientNet-B0 [8], EfficientNetV2-M [9] or DenseNet-121 [6] as encoder. Second one is grayscale transform and ResNet-152 [7], EfficientNet-B0 [8] or DenseNet-121 [6] as encoder. Figure 2 shows the architecture of our fine-tuned CNN approach for Concept Detection.

3.2. Caption Prediction

In Caption Prediction, we attempted four approaches that are retrieval approach, Show and Tell [10], Show, Attend and Tell [11] and Caption Transformer [12].

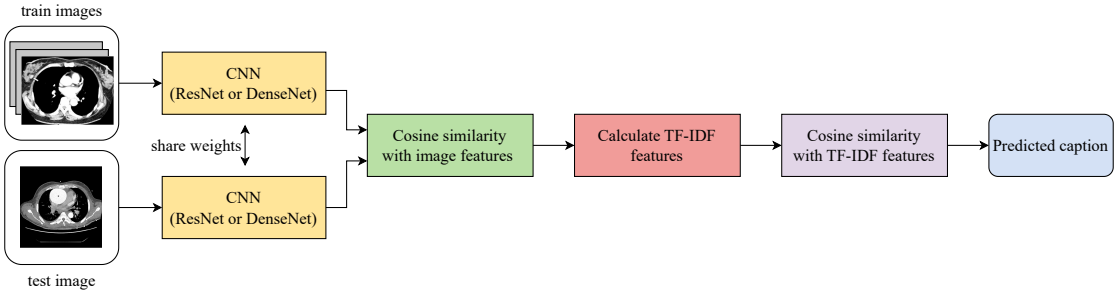

Figure 3: Flow of our retrieval approach for caption prediction. CC BY [Ng et al. (2015)], CC BY-NC [Al Mulhim et al. (2022)].

3.2.1. Preprocessing

Preprocessing Image is same as Concept Detection(Sec 3.1.1). As preprocessing caption, we attempted only lowercase conversion.

3.2.2. Retrieval Approach

For Caption Prediction, some teams [4, 20] attempted retrieval approach in ImageCLEFmedical Caption Task 2022. Therefore, we experimented retrieval approach based on AUEB's approach [4] in ImageCLEFmedical Caption Task 2022.

First, we extracted features from medical images with CNN pre-trained in ImageNet [5]. We employed ResNet-152 [7] as CNN. Then, we calculated cosine similarity between feature extracted from test image and features extracted from train images. Additionally, we selected top k data with high similarity. Then, we converted to TF-IDF features from top k captions. Finally, we calculated cosine similarity among those, and the highest one is predicted. We explored parameter k with validation data. Thus, we assigned $k = 50$, since Bidirectional Encoder Representations from Transformers (BERT) score [21] is the highest in validation set. Figure 3 shows the architecture of our retrieval approach for Caption Prediction.

3.2.3. Show and Tell

This method is CNN-RNN approach based on Show and Tell [10]. In our model, we employed ResNet-152 [7], DenseNet-121 [6], EfficientNet-B0 [8] and EfficientNetV2-M [9] as CNN and LSTM [13] as RNN. For generating caption, we applied greedy decoding how selecting highest probability word. For experimental details, Adam optimizer is used with learning rate 10^{-4}. Additionally, Cross Entropy loss is used for loss function. We trained models for 20 epochs with training set. While training, we saved models with minimum loss using validation set. Using validation data, we evaluated combination between model and preprocessing, and submitted four models with high BERT Score [21]. Submitted models that employed EfficientNet-B0 [8] and EfficientNetV2-M [9] as encoder applied colorization or grayscale transform. Figure 4 shows the architecture of our Show and Tell model for Caption Prediction.

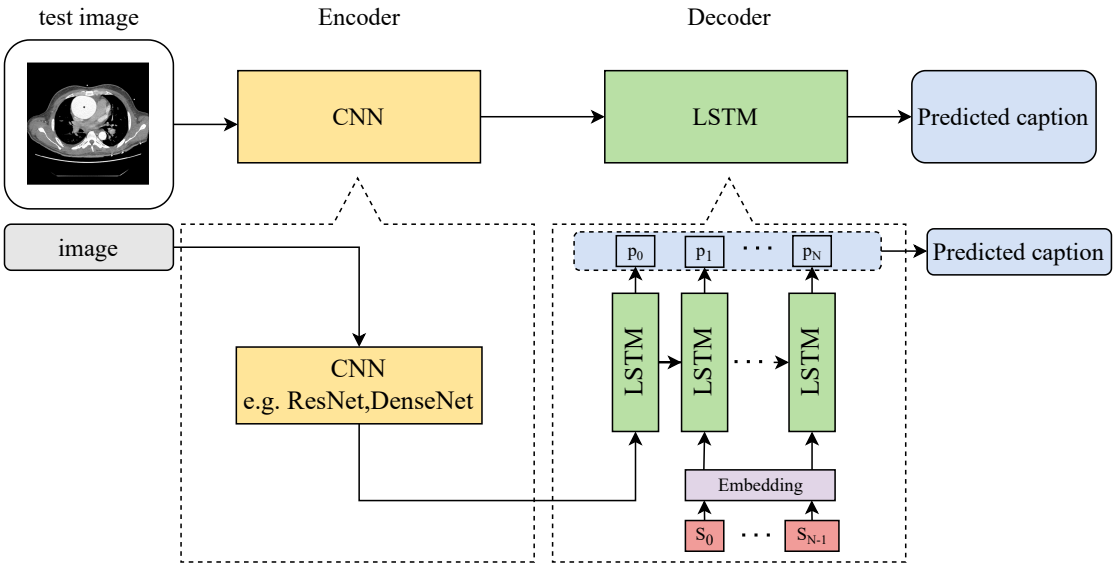

Figure 4: The architecture of Show and Tell model. CC BY-NC [Al Mulhim et al. (2022)].

3.2.4. Show, Attend and Tell

This method is CNN-RNN approach based on Show, Attend and Tell [11]. In our model, we employed ResNet-152 [7], DenseNet-121 [6], EfficientNet-B0 [8] and EfficientNetV2-M [9] as CNN and LSTM [13] as RNN. Experimental settings such as optimizer, loss function and parameters are same as Show and Tell (Sec 3.2.3). Using validation data, we evaluated combination between model and preprocessing, and submitted four models with high BERT Score [21]. Submitted models that employed ResNet-152 [7] and DenseNet-121 [6] as encoder applied colorization or grayscale transform. Figure 5 shows the architecture of our Show, Attend and Tell model for Caption Prediction.

3.2.5. Caption Transformer

In ImageCLEFmedical Caption Task 2022, some teams [19, 17] used approach based on transformer [14] and BERT [22]. Therefore, we attempted Caption Transformer [12] based on transformer [14] encoder and decoder. In our model architecture, the encoder extracts features from patch images, while the decoder outputs word probabilities based on previous words and features extracted from images using the attention mechanism. Experimental settings such as optimizer, loss function and parameters are same as Show and Tell (Sec 3.2.3) and Show, Attend and Tell (Sec 3.2.4). We submitted the lowest loss model with validation data. Additionally, we attempted only preprocessing grayscale transform. Figure 6 shows the architecture of our Caption Transformer model for Caption Prediction.

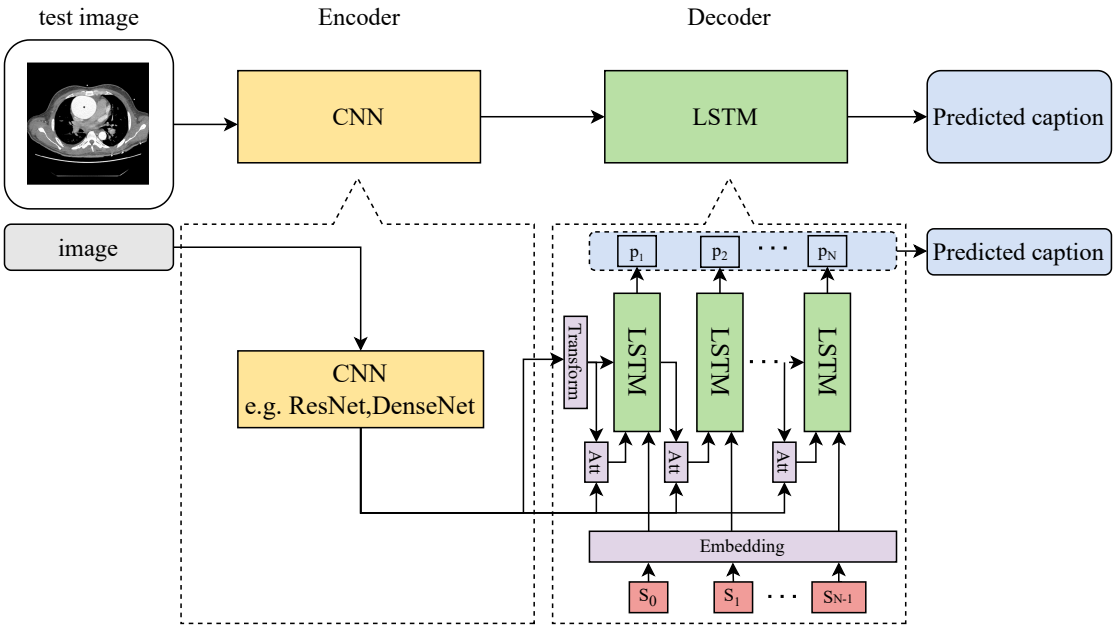

Figure 5: The architecture of Show, Attend and Tell model. CC BY-NC [Al Mulhim et al. (2022)].

4. Results & Discussion

In this section, we describe submissions, results and discussion for both tasks.

4.1. Concept Detection

We submitted ten predictions that were identified by models for Concept Detection. Submitted models are fine-tuned CNN approaches and retrieval approaches. Fine-tuned CNN approaches, a total of six models were submitted. We attempted colorization (with color images) and grayscale transform (with grayscale images) for fine-tuned CNN. In detail, we submitted ResNet-152, DenseNet-121 and EfficientNetV2-M with grayscale images, and submitted DenseNet-121, EfficientNet-B0 and EfficientNetV2-M with color images. Retrieval approaches, we submitted four models in different condition. In detail, we submitted ResNet-152 and DenseNet-121 as encoder with grayscale images, and the parameter k in KNN is $k = 10, 20$. For selecting models, we compared models under various conditions with validation data, and submitted ten models with the highest score. Evaluate metrics is used F1 Score that is calculated between y_{pred} (predicted binary arrays) and y_{true} (correct binary arrays). Additionally, this task used F1 Score Manual that is calculated using manually concepts.

Table 3 shows result of our submissions for Concept Detection. For test data, our best approach is EfficientNetV2-M + FFNN with color images that had 0.5074 as F1 Score and 0.9320 as F1 Score Manual. Additionally, the model achieved 2nd place on the ranking of task.

We show comparing such as models and conditions. First, comparing retrieval approaches and fine-tuned CNN approaches, EfficientNetV2-M + FFNN with color images that is best

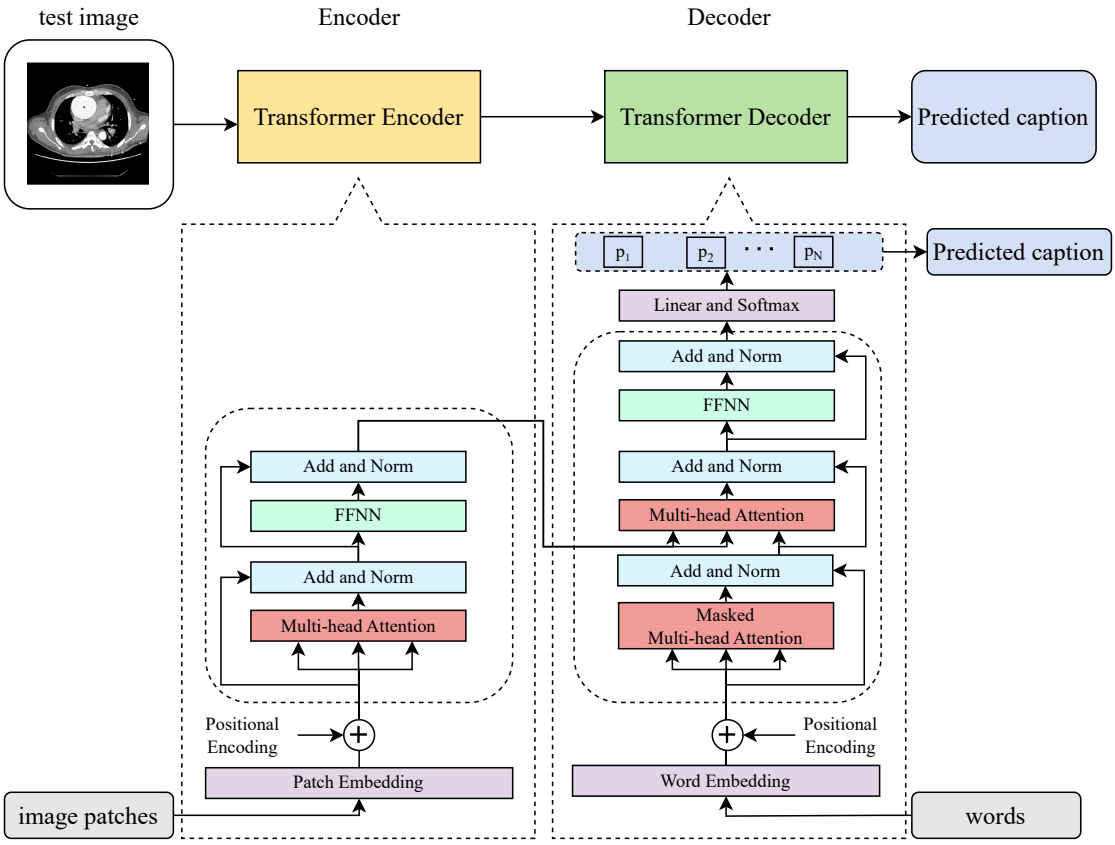

Figure 6: The architecture of Caption Transformer model. CC BY-NC [Al Mulhim et al. (2022)].

performed in fine-tuned CNN approaches is higher F1 Score than retrieval approaches such as ResNet-152 + KNN ($k = 10$). Therefore, we indicated that fine-tuned CNN approach is better performance than retrieval approach among our approaches in concept detection task. Incidentally, we attempted Vision Transformer-B16[23] as encoder. However, the model using Vision Transformer-B16[23] is lower F1 score than other CNN-based models. Therefore, we didn't submit that model. In the future work, we would like to compare CNN-based models and transformer-based models in medical images. Then, comparing image types, preprocessing colorization at EfficientNetV2-M + FFNN is higher performance than grayscale transform at same architecture. Furthermore, EfficientNet-B0 + FFNN with color images improved performance than using grayscale images. Hence, we indicated that using colorization is better performance than using grayscale transform among our approaches in concept detection task. Incidentally, we couldn't attempt retrieval approaches with colorization. In the future work, we would like to attempt retrieval approaches with colorization and compare grayscale and colorization. Additionally, we would like to attempt other colorization such as pseudo-colorization because colorization was better performance than grayscale transform.

Table 3
Submission results for Concept Detection.

Run ID	Model name	Image type	F1 Score	F1 Score Manual
1	ResNet-152 + FFNN	grayscale	0.4979	0.9259
2	DenseNet-121 + FFNN	grayscale	0.4992	0.9235
3	EfficientNetV2-M + FFNN	grayscale	0.5000	0.9221
4	ResNet-152 + KNN (k=10)	grayscale	0.3991	0.7417
5	ResNet-152 + KNN (k=20)	grayscale	0.3886	0.7252
6	DenseNet-121 + KNN	grayscale	0.1061	0.2263
7	DenseNet-121 + KNN	grayscale	0.0993	0.2135
8	DenseNet-121 + FFNN	color	0.5016	0.9221
9	EfficientNet-B0 + FFNN	color	0.4979	0.9223
10	EfficientNetV2-M + FFNN	color	**0.5074**	**0.9320**

4.2. Caption Prediction

We submitted ten predicted captions that were generated by models for Caption Prediction. Submitted models are one retrieval approach, one Caption Transformer, four Show and Tell and four Show, Attend and Tell. retrieval approach employed ResNet-152 as encoder, and the parameter k in the model is set 50. Caption Transformer was used only grayscale images. Show and Tell, we submitted four predictions, EfficientNet-B0 as encoder with grayscale images, EfficientNetV2-M as encoder with grayscale images, EfficientNet-B0 as encoder with color images and EfficientNetV2-M as encoder with color images. We submitted four predictions with Show, Attend and Tell. In detail, we employed ResNet-152 as encoder with grayscale images, DenseNet-121 as encoder with grayscale images, ResNet-152 as encoder with color images and DenseNet-121 as encoder with color images. For selecting models, we compared models under various conditions with validation data, and submitted ten models with the highest F1 score. In Evaluate metrics, the primary metric is used BERT Score [21] that aims to measure the quality comparing between predicted captions and correct captions. Furthermore, the secondary metric is used ROUGE [24] that measures the number of matching unigram between predicted captions and correct captions. Additionally, BLEURT [25], BLUE [26], METEOR [27], CIDEr [28] and CLIPScore [29] are used as metric for test set.

Table 4 shows result of our submissions for Caption Prediction. For test data, our best approach based on BERT Score [21] is Show, Attend and Tell : ResNet-152 with grayscale images that had 0.6145 as BERT Score [21] and 0.2223 as ROUGE [24]. Additionally, the model achieved 6th place on the ranking of task.

We show comparing such as models and conditions. First, comparing approaches, Show, Attend and Tell : ResNet-152 with grayscale images is best performance, and Show, Attend and Tell : ResNet-152 with color images is 2nd ranking in our submission. Therefore, we thought that Show, Attend and Tell is better approach than other models. Incidentally, we attempted Vision Transformer-B16[23] as encoder in Show and Tell. However, the model using Vision Transformer-B16[23] is lower BERT score for validation set than other CNN-based models. Therefore, we didn't submit that model. In addition, Caption Transformer is lowest BERT score than other model. The validity of the Transformer-based model in medical imaging

Table 4
Submission results for Caption Prediction.

Run ID	Model name	Image type	BERT Score [21]	ROUGE [24]	BLEURT [25]	BLUE [26]	METEOR [27]	CIDEr [28]	CLIPScore [29]
1	Show and Tell : EfficientNet-B0	grayscale	0.6088	0.2160	0.2979	0.1640	0.0699	0.1519	0.8043
2	Show and Tell : EfficientNetV2-M	grayscale	0.6082	0.2143	0.2911	0.1585	0.0686	0.1569	0.8027
3	Show, Attend and Tell : ResNet-152	grayscale	**0.6145**	0.2223	0.3013	0.1564	0.0724	0.1818	0.8062
4	Show, Attend and Tell : DenseNet-121	grayscale	0.6094	0.2004	0.2766	0.1249	0.0596	0.1320	0.7828
5	Retrieval approach : ResNet-152 ($k = 50$)	grayscale	0.5789	0.1838	0.2904	0.1484	0.0698	0.0837	0.7826
6	Caption Transformer	grayscale	0.4425	0.1079	0.2968	0.0709	0.0528	0.0057	0.7304
7	Show and Tell : EfficientNet-B0	color	0.6097	0.2204	0.3004	0.1694	0.0724	0.1608	0.8080
8	Show and Tell : EfficientNetV2-M	color	0.6044	0.2166	0.3011	0.1743	0.0730	0.1605	0.8066
9	Show, Attend and Tell : ResNet-152	color	0.6143	**0.2319**	**0.3063**	**0.1749**	**0.0772**	**0.1989**	**0.8083**
10	Show, Attend and Tell : DenseNet-121	color	0.6107	0.2152	0.2935	0.1577	0.0693	0.1585	0.8041

needs to be validated. Then, comparing preprocessing, we used colorization and grayscale transform. As a result, Show and Tell : EfficientNetV2-M and Show, Attend and Tell : DenseNet-121 that using grayscale transform are better performance than using colorization. On the other hands, Show and Tell : EfficientNet-B0 and Show, Attend and Tell : DenseNet-121 that using colorization are better performance than using grayscale transform. Thus, we thought that preprocessing effectiveness depends on encoder. In the future work, the effectiveness of more detailed preprocessing needs to be verified. Incidentally, we would like to attempt other colorization such as pseudo-colorization.

5. Conclusion

In this paper, we described our approach in Concept Detection and Caption Prediction in ImageCLEFmedical Caption Task 2023. For Concept Detection, we employed retrieval approach and fine-tuned CNN approach. Furthermore, we conducted parameter tuning and attempted two preprocessing, colorization and grayscale transform. As a result of the submission, we achieved 2nd place with fine-tuned CNN approach using EfficientNetV2-M. For Caption Prediction, we attempted retrieval approach, CNN-RNN approaches and Caption Transformer. As a result of the submission, we achieved 6th place with Show, Attend and Tell using ResNet-152 as CNN. In feature work, we will compare transformer-based model and CNN-based model. Furthermore, we will attempt pseudo-colorization to outperform our current approach.

Acknowledgments

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, 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.

[2] 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.

[3] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[4] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2022, in: [30], 2022, pp. 1355–1373. URL: http://ceur-ws.org/Vol-3180/#paper-101.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[6] 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), 2017, pp. 4700–4708.

[7] 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, pp. 770–778.

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

[9] M. Tan, Q. Le, EfficientNetV2: Smaller Models and Faster Training, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (ICML), volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10096–10106. URL: https://proceedings.mlr.press/v139/tan21a.html.

[10] 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, pp. 3156–3164.

[11] 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 (ICML), volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 2048–2057. URL: https://proceedings.mlr.press/v37/xuc15.html.

[12] W. Liu, S. Chen, L. Guo, X. Zhu, J. Liu, CPTR: Full Transformer Network for Image Captioning, CoRR abs/2101.10804 (2021). URL: https://arxiv.org/abs/2101.10804. arXiv:2101.10804.

[13] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[14] 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 (NeuralIPS), volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[15] 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.

[16] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, Aueb nlp group at imageclefmed caption tasks 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (CLEF 2021), number 2936 in CEUR Workshop Proceedings, Aachen, 2021, pp. 1184–1200. URL: http://ceur-ws.org/Vol-2936/#paper-96.

[17] F. D. Serra, F. Deligianni, J. Dalton, A. Q. O'Neil, CMRE-UoG team at ImageCLEFmedical Caption 2022: Concept Detection and Image Captioning, in: [30], 2022, pp. 1381–1390. URL: http://ceur-ws.org/Vol-3180/#paper-103.

[18] R. Tsuneda, T. Asakawa, K. Shimizu, T. Komoda, M. Aono, Kdelab at ImageCLEFmedical 2022: Medical Concept Detection with Image Retrieval and Code Ensemble, in: [30], 2022, pp. 1608–1618. URL: http://ceur-ws.org/Vol-3180/#paper-123.

[19] L. Lebrat, A. Nicolson, R. S. Cruz, G. Belous, B. Koopman, J. Dowling, CSIRO at ImageCLEFmedical Caption 2022, in: [30], 2022, pp. 1455–1473. URL: http://ceur-ws.org/Vol-3180/#paper-109.

[20] R. Tsuneda, T. Asakawa, K. Shimizu, T. Komoda, M. Aono, Kdelab at ImageCLEFmedical 2022 Caption Prediction Task, in: [30], 2022, pp. 1596–1607. URL: http://ceur-ws.org/Vol-3180/#paper-122.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, CoRR abs/1904.09675 (2019). URL: http://arxiv.org/abs/1904.09675. arXiv:1904.09675.

[22] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[23] 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, CoRR abs/2010.11929 (2020). URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[24] 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.

[25] T. Sellam, D. Das, A. P. Parikh, BLEURT: Learning Robust Metrics for Text Generation,

CoRR abs/2004.04696 (2020). URL: https://arxiv.org/abs/2004.04696. arXiv:2004.04696.

[26] 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.

[27] M. Denkowski, A. Lavie, Meteor Universal: Language Specific Translation Evaluation for Any Target Language, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 376–380. URL: https://aclanthology.org/W14-3348. doi:10.3115/v1/W14-3348.

[28] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-Based Image Description Evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575.

[29] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, CoRR abs/2104.08718 (2021). URL: https://arxiv.org/abs/2104.08718. arXiv:2104.08718.

[30] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF), number 3180 in CEUR Workshop Proceedings, Aachen, 2022. URL: http://ceur-ws.org/Vol-3180/.