

Kdelab at ImageCLEF 2021: Medical Caption Prediction with Effective Data Pre-processing and Deep Learning

Riku Tsuneda¹, Tetsuya Asakawa² and Masaki Aono³

¹Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan

Abstract

ImageCLEF 2021 Caption Prediction Task is an example of a challenging research problem in the field of image captioning. The goal of this research is to automatically generate accurate captions describing a given medical image. We describe our approach to captioning medical images and illustrate the text and image pre-processing that is effective for our task dataset. In this paper, we have applied sentence-ending period removal as text pre-processing and histogram normalization of luminance as image pre-processing. Furthermore, we present the effectiveness of our text data augmentation approach. Submission of our kdelab team on the task test dataset achieved a BLEU evaluating of 0.362.

Keywords

Image Captioning, Deep Learning, Medical Images

1. Introduction

In recent years, multimodal processing of images and natural language has attracted much attention in the field of machine learning. Image Captioning is one of these representative tasks, which aims at proper captioning of input images. As these accuracies improve, it is expected that computers will not only be able to detect objects in images, but also to understand the relationships and behaviors between objects.

Image captioning is also effective in the medical field. For example, interpreting and summarizing possible disease symptoms from a large number of radiology images (e.g. X-ray images and CT images) is a time-consuming task that can only be understood by highly knowledgeable specialists. If computers could understand medical images and generate accurate captions, it would help solve the world's growing shortage of medical doctors. However, there is still the bottleneck problem that few physicians are able to give accurate annotations.

In this paper, we describe our approach to general Image Captioning task in medical domain at Image Captioning such as Fig. 1(right).

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


CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ tsuneda.riku.am@kde.cs.tut.ac.jp (R. Tsuneda); asakawa@kde.cs.tut.ac.jp (T. Asakawa); aono@tut.jp (M. Aono)

🆔 0000-0002-3063-7489 (R. Tsuneda); 0000-0003-1383-1076 (T. Asakawa); 0000-0002-8345-7094 (M. Aono)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



COCO_train2014_000000506232.jpg

Caption : A player readies for a swing during a tennis game .



Caption : Partial fusion of capitate-hamate and fusion of lunate-triquetrum.
Normal bony mineralization.

Figure 1: Example of general (left) and medical (right) Caption Prediction data& left image : via MS-COCO, [CC BY 4.0](<https://cocodataset.org/>).

In the following, we first describe related work on Image Captioning task and Medical Image Captioning in Section 2, followed by the description of the dataset provided for ImageCLEF2021 [2] Medical Image Captioning [3] dataset in Section 3. In Section 4, we describe details of the method we have applied, and then of our experiments we have conducted in Section 5. We finally conclude this paper in Section 6.

2. Related Work

In the field of image recognition, convolutional neural networks (CNN), including VGG [4], and ResNet [5], have been widely used. In the field of natural language processing for text understanding, encoder-decoder models (seq2seq) [6] have been the mainstream, but in recent years Transformers [7] such as BERT [8] have become common. The Image Captioning task is a fusion of image recognition and sentence generation, and lies in the middle of these two.

For example, Oriol Vinyals et al. proposed caption generation using an encoder-decoder model [9], and Kelvin Xu et al. proposed Show, Attend and Tell, which adds visual attention to the encoder-decoder model [10]. Recently, P. Anderson et al. presented a model using Bottom-Up Attention obtained by pre-training a Faster-R-CNN used for object detection [11].

In addition, the Caption Prediction Task is the first time of its kind to be held at an ImageCLEF conference. However, a similar task, the VQA-Med task [12], has been contested at ImageCLEF2018, 2019, and 2020.

Table 1
Word frequency Ranking in Dataset

Rank	Word	Freq	Rank	Word	Freq
1	right	824	7	axial	372
2	left	672	8	images	327
3	mass	616	9	image	326
4	ct	534	10	within	272
5	demonstrates	442	11	lesion	246
6	contrast	373	12	demonstrate	244

pre-processing the images and text in the dataset. The second is the encoder part. In the encoder part, the features of the image are extracted. The third is the decoder part. In the decoder part, words are predicted recursively using LSTM [13] and attention mechanism.

We have adopted Show, Attend and Tell as the base model. This model is known to have high accuracy among Image Captioning models that do not use object detection such as Faster R-CNN [14].

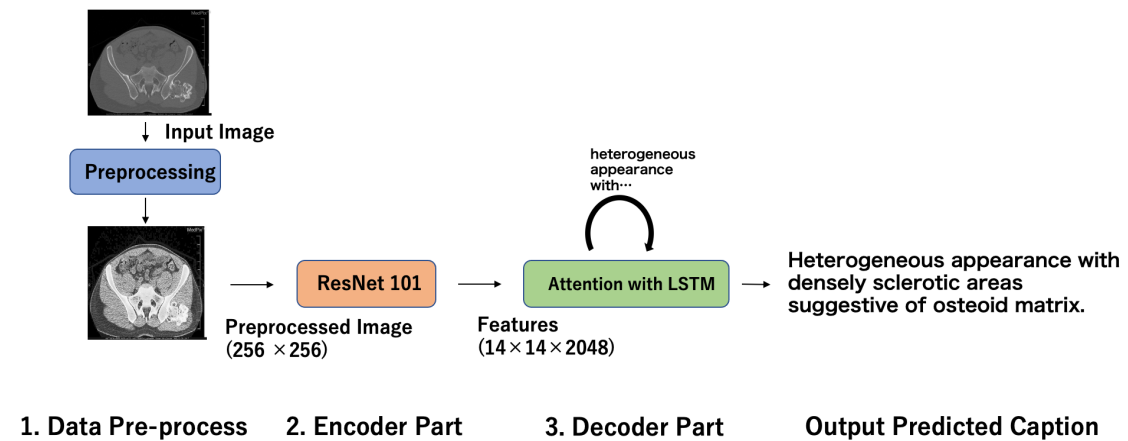


Figure 3: Over view of the our captioning framework

4.1. Input Data Pre-processing

4.1.1. Image Pre-processing

Image pre-processing includes image normalization.

The image processing consists of two steps. In the first step, we normalize images using histogram smoothing based on the luminance of the image. In the second step, we resize all images to a size of 256×256 .

We have tried two ways to normalize the luminance distribution of an image. The first is histogram flattening. Histogram flattening is a method of smoothing the luminance distribution

of the entire image. When flattened, the contrast of the image is enhanced and the image becomes clearer. The second is adaptive histogram flattening. This method performs the histogram flattening described in the first method on a small area of the image. In general, this technique can reduce the occurrence of tone jumps. A comparison of the raw image and the pre-processed image is shown in Figure 4.

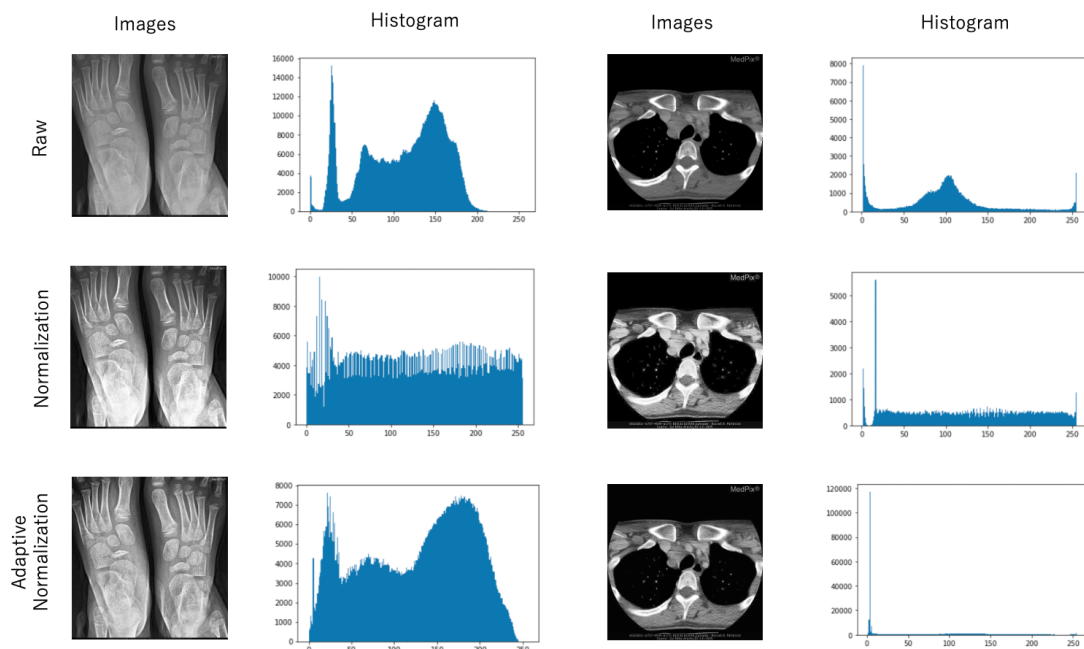


Figure 4: Raw images and normalization images

4.1.2. Text Pre-processing

We preprocess the text by removing and lowercasing periods in the captions of the training data. In general, the MS-COCO captioning task is not case-sensitive, and it is well known that symbols such as periods had better be removed. If there are multiple captions for a single image, only the period in the last caption is should be removed. As a contribution to these, the period is recognized as one of the words in the sentence, since the period is present only in the sentence.

4.2. Caption Data Expanding using EDA

We tried EDA (Easy Data Augmentation) [15] as an extension of our text dataset. EDA is a text classification task in natural language processing, and is an effective method that works well when the dataset is small. In a typical captioning task using MS-COCO, five captions are provided for one image. However, in the ImageCLEF2021 dataset, only one caption per image

is provided. We have tested the effectiveness of this approach using various data expansion methods in EDA.

4.3. Neural network model

As a base neural network model for caption generation, we have adopted "Show, Attend and Tell" model [10]. This model is capable of highly accurate captioning without using object detection. The architecture of the models is almost the same, but our model differs in that we employ ResNet-101 [5] instead of VGG16 [4] as the CNN encoder .

5. Experiments and results

5.1. Setting up hyper-parameters and performing pre-processing with validation data

We experimented with hyper-parameter adjustment and image pre-processing using training and validation data. As noted in 4.1, all characters in the train caption data are lowercased.

We have setup the following hyper-parameters as follows; batch size as 32, optimization function as "Adam" with a decoder learning rate of 0.001 , and the number of epochs 200. For the implementation, we employ PyTorch1.7.1 [16] as our deep learning framework. For the evaluation of captioning , we utilize BLEU4 [17]. Table 2 shows the results. Here we compare in terms of BLEU for data pre-processing.

Table 2

Validation BLEU-4 of two data pre-processing ("val" in the table means validation.)

Model	Pre-processing	val BLEU-4
Xu et al. [10]	None	0.432
	Histogram Normalization	0.437
	Adaptive Histogram Normalization	0.436

5.2. The results with test data

The test dataset consists of the test images distributed as described in 4.1. The test image consists of 444 medical images, without the correct answer captions. In contrast to the text pre-processing in 5.1, the captions used in the training have been all lowercased and the periods at the end of sentences were deleted.

Table 3 shows the BLEU results for the test data. In the experiments on the test data, the BLEU evaluation was the highest when Histogram Normalization was used. Example of our seemingly successful caption generation results are shown in Fig 5.

Table 4 shows the BLEU ratings for the EDA attempts. The pre-processing of the dataset uses the method that achieved the highest BLEU rating in Table 3. Using EDA's synonym substitution and other methods, we compare the case of adding one caption, two captions, and four captions. In all cases where data expansion has been performed using EDA, the BLEU rating has dropped.

Table 3

The results of experiment for Image pre-processing for test data ("val" in the table means validation.)

Model	Image Pre-processing	val BLEU	test BLEU
Xu et al. [10]	None	0.436	0.332
	Histogram Normalization	0.451	0.362
	Adaptive Histogram Normalization	0.443	0.352



synpic100589

Hyp : Figure demonstrating positioning of hardware of lumbar and sacral.

Ref : Figure demonstrating positioning of hardware of lumbar spine.



synpic58672

Hyp : Axial ct with contrast showing hilar and mediastinal lymph nodes also labeled arrow - air in esophagus aao - ascending aorta dao - descending aorta lb - left bronchus rmb - right mainstem bronchus rul - right upper lobe bronchus svc .

Ref : There are chunky calcifications in multiple mediastinal lymph nodes. Pleural nodules and interstitial thickening in the upper lobes. Also labeled Arrow - air in esophagus AAo - Ascending aorta DAo - Descending aorta LB - left bronchus RMB.

Figure 5: Example of generated caption**Table 4**

The Results of using EDA to extend training data.

Image Pre-processing	Added Caption by EDA	val BLEU	test BLEU
Histogram Normalization	None	0.451	0.362
	one caption	0.417	0.339
	two captions	0.397	0.291
	four captions	0.384	-

The results of the submissions of the participants with the highest BLEU values are shown in Table 5. Our rank turns out to be 4th of participants.

6. Conclusions

We have described our system with which we submitted to the ImageCLEF2021 Caption Prediction task. In our system, we have done our own data pre-processing, and have attempted to

Table 5

The best participants' runs submitted for the Caption Prediction task

Group Name	Rank	BLEU
IALab_PLC	1	0.510
AUEB_NLP_GROUP	2	0.461
AEHRC-CSIRO	3	0.432
kdelab	4	0.362
jeanbenoit_delbrouck	5	0.285
ImageSem	6	0.257
RomiBed	7	0.243
ayushnanda14	8	0.103

add data augmentation with EDA. In addition, two types of luminance smoothing and period removal were applied to image and text pre-processing. The results demonstrate that these processes have improved the caption prediction accuracy of the neural network model. EDA turns out to be ineffective in this task. Finally, from organizer's evaluation, we have achieved a BLEU score of 0.362 in the ImageCLEF2021Caption Prediction task, placing us 4th.

Acknowledgment

A part of this research was carried out with the support of Grant for Education and Research in Toyohashi University of Technology.

References

- [1] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: <http://arxiv.org/abs/1405.0312>. arXiv: 1405.0312.
- [2] B. Ionescu, H. Müller, R. Péteri, A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ștefan, 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.
- [3] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, 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.

- [4] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, CoRR abs/1409.1556 (2015).
- [5] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
- [6] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to Sequence Learning with Neural Networks, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 3104–3112.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, CoRR abs/1706.03762 (2017). URL: <http://arxiv.org/abs/1706.03762>. arXiv:1706.03762.
- [8] 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.
- [9] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156–3164. doi:10.1109/CVPR.2015.7298935.
- [10] 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, CoRR abs/1502.03044 (2015). URL: <http://arxiv.org/abs/1502.03044>. arXiv:1502.03044.
- [11] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and VQA, CoRR abs/1707.07998 (2017). URL: <http://arxiv.org/abs/1707.07998>. arXiv:1707.07998.
- [12] 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.
- [13] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780.
- [14] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, CoRR abs/1506.01497 (2015). URL: <http://arxiv.org/abs/1506.01497>. arXiv:1506.01497.
- [15] J. W. Wei, K. Zou, EDA: easy data augmentation techniques for boosting performance on text classification tasks, CoRR abs/1901.11196 (2019). URL: <http://arxiv.org/abs/1901.11196>. arXiv:1901.11196.
- [16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, CoRR abs/1912.01703 (2019). URL: <http://arxiv.org/abs/1912.01703>. arXiv:1912.01703.
- [17] Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing, 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://www.aclweb>.

[org/anthology/P02-1040](https://doi.org/10.3115/1073083.1073135). doi:10.3115/1073083.1073135.