Polimi-ImageClef Group at ImageCLEFmedical Caption task 2022

Seyyed Ali Mir Ghayyomnia[1], Kai de Gast[2] and Mark J. Carman[3]

[1]Politecnico di Milano Piazza Leonardo da Vinci, 32 20133 Milano, Italy
[2]Politecnico di Milano Piazza Leonardo da Vinci, 32 20133 Milano, Italy
[3]Politecnico di Milano Piazza Leonardo da Vinci, 32 20133 Milano, Italy

Abstract

We present the models that PoliMi-ImageClef group developed to participate in ImageCLEFmedical Caption task [1]. The goal of this task is to identify medical concepts present in medical images with different imaging modalities, which is a milestone in automatically generating medical reports. We participated with different systems, using encoders (ResNet-50 [2], Resnext-50 [3] and Swin-Transformer [4]) combined with a feed-forward neural network to predict concepts. During development process we compared the performances of the trained models, by using a part of provided data as a test set, and the model utilizing Swin-Transformer [4] had the best performance. However submission results proved that the model based on Resnext-50 encoder [3] had the best performance on the competition test set.

Keywords

Medical Images, Concept Detection, Multi-label Classification, Deep Learning, Vision Transformer, Encoders, CEUR-WS

1. Introduction

We present the participation experience of the PoliMi-ImageClef group in ImageCLEFmedical Caption task [1] [1]. The Image Captioning is one of these research tasks, which is composed of two sub-tasks: Concept Detection and Caption Prediction. The Concept Detection task includes developing a multi-label classifier, intended for medical images, by identifying medical concepts. These concepts are assigned in terms of Unified Medical Languages System (UMLS)[5] [2] to each image. The Caption Prediction task comprises of generation of captions, which is essential in interpreting the medical images.

This paper discusses the models that were used in Concept Detection sub-task by PoliMi-ImageClef team. Our best run was ranked 6th in the competition. In this model we used a Resnext-50-32x4d encoder [3] to acquire image embeddings used for classification. Another model which showed a promising potential in validation phase, used a Swin-Transformer

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
seyyedali.mir@mail.polimi.it (S. A. M. Ghayyomnia); kai.degast@mail.polimi.it (K. d. Gast); mark.carman@polimi.it (M. J. Carman)
0000-0003-1897-0710 (S. A. M. Ghayyomnia); 0000-0001-6575-9737 (M. J. Carman)

[1] https://www.imageclef.org/2022/medical/caption
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795

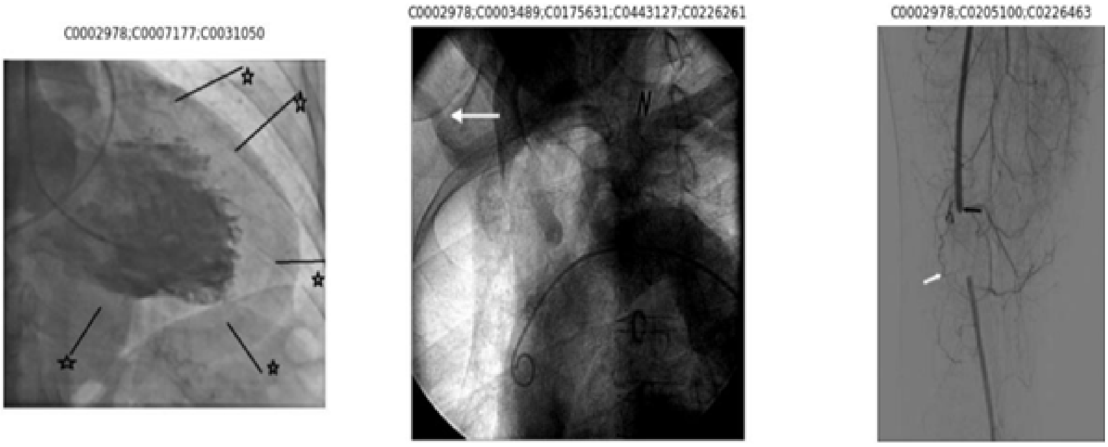
C0002978;C0007177;C0031050
C0002978;C0003489;C0175631;C0443127;C0226261
N
C0002978;C0205100;C0226463

Figure 1: Training data and their corresponding captions. From left to right, CC BY [Lambelin et al. (2014)] [9], CC BY-NC [Park et al. (2010)][10], CC BY-NC [Ã–ztÃ¼rk et al. (2015)][11]

[4] to extract the image features. Combined with a feed-forward neural network, it slightly outperformed the Resnext-50 model.

2. Data

The dataset provided for ImageCLEFmedical Caption task 2022 is an extended version of Radiology Objects in COntext (ROCO) dataset [6]. The dataset originates from biomedical articles of the PMC OpenAccess subset [7]. A similar dataset was used for the ImageCLEFmed 2020 concept detection task [8].

The dataset in comprised of 83,275 radiology images in training set, 7,645 images as the validation set and 7,601 images as the test set. The total number of UMLS concepts present in the dataset is 8,374. The maximum number of concepts present in a single image is 100. In development process, we combined the training set and validation set and spliced it into training set, validation set and test set with (0.6, 0.2, 0.2) ratio. This test set was used to compare the performance of models in development process.

3. Methods

In this section we discuss different method used in development phase. In total we trained and tested 6 models. Some of these models used different encoders such as ResNet-50 [2] , DenseNet-121 [12], Resnext-50 [3], whereas others used variations of Swin-Transformer [4]. All these models were pretrained on ImageNet1k or ImageNet22k [13] in case of Swin-B [4]. Following this, we rescaled the images to 224×224 and normalized with the mean and standard deviation of ImageNet(224×224) [13]. ResNext [3] and Swin-Transformer [4] models performed remarkably well during the development phase of the model. In the following we discuss the structure of ResNext [3] encoder and Swin-Transformer [4] and their novelties.

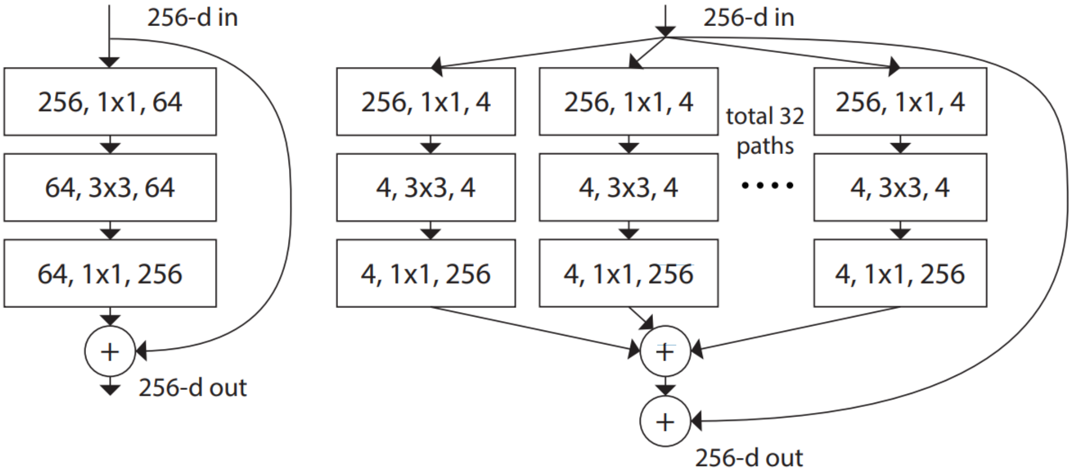

Figure 2: Comparison between structures of ResNet [2] Block and ResNext Block [3]

Common Attributes of training process: To avoid repetition, in this section we note common attributes of the training procedure for all models. We tackle this task as a multi-label classification; the target to be classified is an array of labels[3] that can be present in each image. This narrows our choice for the choice of activation function [14] to Sigmoid function for each output layer node. In addition, we need to use Binary Cross-Entropy loss function [15] to fit the model. To acquire a computationally efficient with fewer parameters, we used Adam optimizer [16].

For each encoder, we experimented with two structures :

- With no hidden layer.
- With one hidden layer : with 2048 nodes, ReLU activation function [17] and Dropout rate of 0.2 [18].

3.1. Resnext-50-based Classification

In this model, we used a Resnext-50-32x4d encoder [3], a CNN [4] with 48 layers. A ResNext repeats a building block that aggregates a set of transformations with the same topology [5]. Compared to a ResNet [2] , it exposes a new dimension, cardinality (the size of the set of transformations) C, as an essential factor in addition to the dimensions of depth and width.

3.1.1. ResNext building block

A ResNext Block [3] is a type of residual block used as part of the ResNext CNN architecture. It uses a "split-transform-merge" [19] strategy (branched paths within a single module) similar to

[3] Array size is 8374

[4] Convolutional Neural Network

[5] https://paperswithcode.com/method/resnext-block

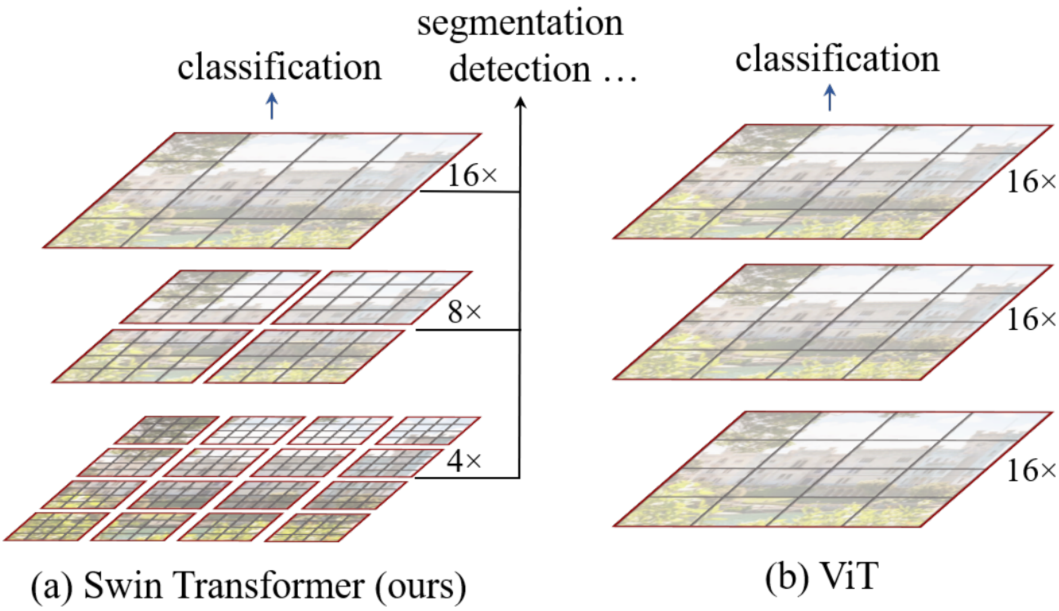

Figure 3: Comparison of the feature maps in ViT [20] and Swin-Transformer [4]

an Inception module [6] [19], i.e. it aggregates a set of transformations. The effect of this strategy on the performance is further discussed in "Aggregated Residual Transformations for Deep Neural Networks" [3].

In our validation process we experimented with different variations of FFNN for classification task, however the simplest model, with only one hidden layer, had the best performance.

3.2. Swin-Transformer

The Swin-Transformer [4] is a type of Vision Transformer [7] [20] that constructs hierarchical feature maps by merging image patches in deeper layers. It can thus serve as a general-purpose backbone for feature extraction which can be used for a variety of tasks including Image Classification, Semantic Segmentation and Dense Recognition. The architecture of Swin-Transformer and its differences with respect to previous generation of Vision Transformers, is discussed in detail in "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" [4].

In our development we used an iteration of a Swin-Transformer [4] that was pretrained on ImageNet22K [13]. The features extracted from the images were then fed to an FFNN. Thorough experimentation we chose an iteration in which the FFNN had 1 hidden layers.

[6] https://paperswithcode.com/method/inception-module
[7] https://paperswithcode.com/method/vision-transformer

Table 1
Performance results of our best systems

Model Name	Structure	Learning rate	Weights	Parameters	F1-Score(Dev)	F1-Score(Test)
Resnext-50	Resnext-50-32x4d + FFNN(0 hidden layer)	$1e-4$	ImageNet1k	22M	0.401	0.432
Swin-Base	Swin-B-224 + FFNN(1 hidden layer)	$1e-4$	ImageNet22k	88M	0.403	0.428
ResNet-50	ResNet-50 + FFNN(0 hidden layer)	$1e-4$	ImageNet1k	23M	0.399	0.425
Swin-Tiny	Swin-T-224 + FFNN(1 hidden layer)	$1e-4$	ImageNet1k	28M	0.396	0.426
DenseNet-121	DenseNet-121 + FFNN(1 hidden layer)	$1e-3$	ImageNet1k	6M	0.393	0.423
ResNet-152	ResNet-152 + FFNN(1 hidden layer)	$1e-3$	ImageNet1k	58M	0.391	0.420

4. Results

In this section we further explain the details of each run:

ResNet-50 RUNS : In this runs we used the ResNet-50 encoder [2] pretrained on ImageNet1K [13] and with 48 convolutional layers and 23M parameters. The network is trained on the training set for 5 epochs. We trained two instances of this model with learning rates of $1e-3$ and $1e-4$. The model trained with learning $1e-4$ had the best performance.

ResNet-152 RUNS : In this runs we used the ResNet-152 encoder [2] pretrained on ImageNet1K [13] and with 150 convoltional layers and 58M parameters. The encoder coupled with an FFNN with one hidden layer was trained on training set for 5 epochs. The learning rate for this training procedure was set to $1e-3$ after experimenting.

DenseNet-121 RUNS : For this run we used the DenseNet-121 encoder [12] pretrained on ImageNet1K [13] and with 120 convolutional layers and 6M parameters. The encoder coupled with an FFNN with one hidden layer was trained on training set for 5 epochs with learning rate of $1e-3$.

ResNext-50 RUNS : In this series of runs we experimented with ResNext encoder [3] capabilities. We used the pretrained encoder on ImageNet1K [13] with 22M parameters. For the two structures mentioned previously we experimented with 2 values of learning rates; $1e-3$, $1e-4$. The models were trained for 5 epochs. The structure with no hidden layers and learning rate of $1e-4$ produced the best F1-Score.

Swin-Transformer RUNS : In this series of runs we experimented with Swin Transofrmer [4] capabilities. The structure used for the FFNN with one hidden layer. We used 2 versions of the Swin-Transformer :

- Swin-T [4] , tiny version pretrained on ImageNet1K
- Swin-B [4] , base version pretrained on ImageNet22K

Learning the rate for these runs was set to $1e-4$ after experimenting. The structures were trained for 5 epochs. The Swin-B model had the best the results between these runs.

Performance Evaluation: The performance index proposed in the Caption Detection subtask is F1-score [21]. The F1-score is calculated in “binary” averaging method for each image. Then all F1-scores are summed and averaged over the number of elements in the test set (7,601), giving the final score [8]. Table reports the details of the models used and their performances during development and testing. The F1-score during development was computed by processing the result of the predictions over the test set generated in development process from splicing the merged set of training data and validation data.

5. Conclusions and Future work

We investigated the performance of Resnext-50 [3] and Swin-Transformer [4] in a multi-label classification task. In our development phase as seen in table Swin-Transformer outperformed Resnext-50 by slight advantage. In future work, we aim to further investigate the potential of Swin-Transformers, in Concept Detection and Caption Prediction contexts, to improve the performance of medical image captioning systems.

References

[1] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[3] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, CoRR abs/1611.05431 (2016). URL: http://arxiv.org/abs/1611.05431. arXiv:1611.05431.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, CoRR abs/2103.14030 (2021). URL: https://arxiv.org/abs/2103.14030. arXiv:2103.14030.

[5] O. Bodenreider, The unified medical language system (umls): Integrating biomedical terminology, Nucleic acids research 32 (2004) D267–70. doi:10.1093/nar/gkh061.

[6] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, 2018, pp. 180–189. doi:10.1007/978-3-030-01364-6_20.

[7] PMC Open Access Subset, Bethesda (MD): National Library of Medicine, 2003. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[8] B. Ionescu, H. Müller, R. Péteri, A. B. Abacha, V. Datla, S. A. Hasan, D. Demner-Fushman, S. Kozlovski, V. Liauchuk, Y. D. Cid, V. Kovalev, O. Pelka, C. M. Friedrich, A. G. S. de Herrera,

[8] https://www.imageclef.org/2022/medical/caption

V.-T. Ninh, T.-K. Le, L. Zhou, L. Piras, M. Riegler, P. l Halvorsen, M.-T. Tran, M. Lux, C. Gurrin, D.-T. Dang-Nguyen, J. Chamberlain, A. Clark, A. Campello, D. Fichou, R. Berari, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, Overview of the ImageCLEF 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 12260 of *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*, LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece, 2020.

[9] CC BY [Lambelin et al.(2014)], National Institutes of Health's National Library of Medicine, 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4083777/.

[10] CC BY-NC [Park et al. (2010)], National Institutes of Health's National Library of Medicine, 2010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2966712/.

[11] CC BY-NC [Ã–ztÃ¼rk et al. (2015)], National Institutes of Health's National Library of Medicine, 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5779162/.

[12] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, CoRR abs/1608.06993 (2016). URL: http://arxiv.org/abs/1608.06993. arXiv:1608.06993.

[13] J. Deng, R. Socher, L. Fei-Fei, W. Dong, K. Li, L.-J. Li, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), volume 00, 2009, pp. 248–255. URL: https://ieeexplore.ieee.org/abstract/document/5206848/. doi:10.1109/CVPR.2009.5206848.

[14] T. Szandala, Review and comparison of commonly used activation functions for deep neural networks, CoRR abs/2010.09458 (2020). URL: https://arxiv.org/abs/2010.09458. arXiv:2010.09458.

[15] U. Ruby, V. Yendapalli, Binary cross entropy with deep learning technique for image classification, International Journal of Advanced Trends in Computer Science and Engineering 9 (2020). doi:10.30534/ijatcse/2020/175942020.

[16] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014. URL: http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[17] A. F. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375 (2018).

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.

[19] A. Khan, A. Sohail, U. Zahoora, A. S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, CoRR abs/1901.06032 (2019). URL: http://arxiv.org/abs/1901.06032. arXiv:1901.06032.

[20] 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.

[21] Z. C. Lipton, C. Elkan, B. Narayanaswamy, Thresholding classifiers to maximize f1 score, 2014. URL: https://arxiv.org/abs/1402.1892. doi:10.48550/ARXIV.1402.1892.