Generating captions for medical images with a deep learning multi-hypothesis approach: MedGIFT–UPB Participation in the ImageCLEF 2017 Caption Task

Liviu-Daniel Ştefan[1], Bogdan Ionescu[1], and Henning Müller[2,3]

[1] University Politehnica of Bucharest, 061071 Romania;
[2] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland
[3] University of Geneva, Switzerland

Abstract. In this report, we summarize our solution to the ImageCLEF 2017 caption detection task. ImageCLEF's concept detection task provides a testbed for figure caption prediction oriented systems using medical concepts as sentence level descriptions for images, extracted from the Unified Medical Language System (UMLS) dataset. The goal of the task is to efficiently identify the relevant medical concepts from medical images as a predictor of figure captions. For representing the images we used a very deep Convolutional Neural Network, namely ResNet–152 pre-trained on ImageNet and a binary annotation of the concepts. In the concept detection subtask, MedGIFT–UPB group occupied the 3rd place out of 9 groups. the proposed approach obtained the 12th position according to the f1 score (0.89) out of 20 participant runs (runs without external resources). The paper presents the procedure employed, and provides an analysis of the obtained evaluation results. The results were showed the difficulty of the task when not using any other external resources.

Keywords: ImageCLEF, medical images analysis, biomedical concepts, deep learning.

1 Introduction

ImageCLEF[4] is an image classification and retrieval benchmark that was founded in 2003 and has been run every year since then [1, 2]. A medical task was added in 2014 and has been held every year since then in varied forms [3].

The aim of the ImageCLEF [4] concept detection task is to assign to a biomedical image extracted from scholarly articles of the biomedical open access literature (PubMed Central), a set of medical concepts taken from a list of 20.463 possible pre-defined medical concepts. In this task, the participants are given a large-scale noisy dataset of 164,614 training images and 10,000 validation images associated with their corresponding labels, each of which corresponds to

[4] `http://www.imageclef.org/`

a UMLS (Unified Medical Language System) concept. In the test phase, the participants are requested to give to each of the 10,000 test images the labels of all the medical concepts describing the image. The performance evaluation is done at image level by computing the F-Measure for each image. For more details on task setup and full participant results, please refer to [5].

The challenge we addressed in our participation was the semantic gap between captions (text) and the image. Thus, the scientific challenge was to exploit these two types of information in order to automatically describe new medical images without any other resources than the training images provided. Having this in mind, we propose a flexible deep CNN that takes an arbitrary number of hypotheses as the inputs and produce at output the ultimate multi-label predictions. Our approach is based on visual features learned using a very deep Convolutional Neural Network, namely ResNet–152 [6]. We pretrained the ResNet–152 using the ImageNet model as the initialization of the network with a Sigmoid Cross Entropy Loss Layer in the train phase and Sigmoid layer in the test phase. The labels were binarized using 1 for the relevant concepts and 0 for the other concepts. Finally, we used a threshold over the scores to construct the concept list for the test images.

The remainder of this paper is organized as follows: In Section 2, we investigate the network from a model-selection and optimization perspective. Section 3 reports the experimental setup. In Section 4 we report the results of our runs with respect to the top scores. Finally in Section 5 we present conclusions and discuss current and future directions.

2 Method

To take full advantage of CNNs for multi-label image classification, in this paper we propose a flexible deep CNN structure that takes an arbitrary number of hypotheses (concepts) as input. The proposed method uses the ResNet–152 CNN and the ImageNet model as the initialization of the network. In the training phase we used the Sigmoid Cross Entropy Loss Layer. For training we binarized the labels using 1 for the relevant concepts and 0 for the rest. In the test phase we used a Sigmoid layer which produces a probability distribution over the class labels. Finally the network output gives the final multilabel scores that are thresholded to construct the concept list for the test images.

2.1 Network Description

The ResNet implemented in this paper is based on He et al. [6] which achieved state–of–the–art results in the ILSVRC [7] and COCO 2015 [8] competitions obtaining first places in all main tracks. ResNet–152 is composed of a 152–layer Residual Network having the following architecture (Table 1).

The central idea of ResNets is to solve the neural network tendency to obtain higher training error as the depth increases due the gradient and training signals vanishing when they are propagated through many layers. This is done by

Table 1. ResNet-152 architecture.

Layer name	Residual-Block	No. of blocks
conv1	7×7(stride 2)	$\times$ 1
max-pool	3×3(stride 2)	-
conv1	1×1: 3×3: 1×1	$\times$ 3
conv2	1×1: 3×3: 1×1	$\times$ 8
conv3	1×1: 3×3: 1×1	$\times$ 36
conv5	1×1: 3×3: 1×1	$\times$ 3

adding skip connections that bypass a few convolutional layers at a time. Each bypass generates a residual block in which the convolution layers predict a residual that is added to the block's input tensor. The last layer of each residual block is applied with a global batch normalization, with the beta and gamma parameters being set to trainable all throught-out. Following most CNN architectures, Rectified Linear units (ReLU) are added after the BN layers.

We decided to start with ResNet–152 as a baseline due to the capacity of the network to avoid negative outcome while increasing network depth.

2.2 Multi-Label Classification

Multi–label image classification is a practical problem, due to majority of real–world images being with more than one object of usually different categories. To tackle this problem we use a cross–entropy loss layer that is defined as:

$$L(X,Y) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \ln a(x^{(i)}) + \left(1 - y^{(i)}\right) \ln\left(1 - a(x^{(i)})\right) \tag{1}$$

where, $X = \left\{x^{(1)}, \ldots, x^{(n)}\right\}$, is the set of input examples in the training dataset, and $Y = \left\{y^{(1)}, \ldots, y^{(n)}\right\}$ is the corresponding set of labels for these input examples. The $a(x)$ represents the output of the neural network given input x. Each of the $y^{(i)}$ is either 0 or 1, and the output activation $a(x)$ is restricted to the open interval (0, 1) by using a sigmoid.

3 Experimental Setup

The aim of the experiments described in this section was to not use external resources for enriching the description of the medical images. Therefore, a single kind of information was used: visual information. As the training dataset is extremely small and the concept relations are relatively complex, as it can be seen in Figure 1, training very deep CNNs is quite challenging. Deep features have demonstrated that CNN models pre-trained on large–scale single–label datasets with data diversity e.g., ImageNet, can be transferred to extract features for other image datasets without many training samples.

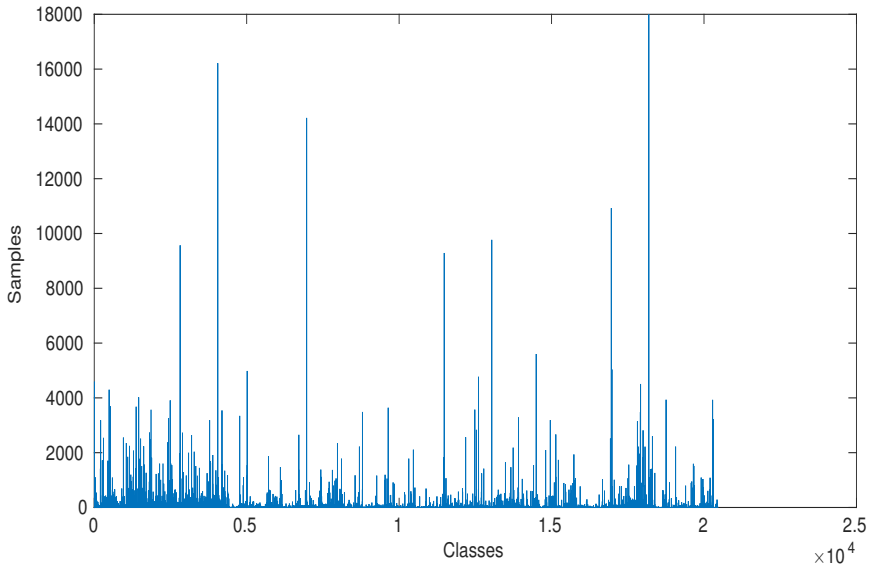

Fig. 1. Shows the number of samples per class.

Run 1: DET_ResNet152_SCEL_t_0.06.txt:
In this run we only used visual features learned using a very deep Convolutional Neural Network. We build a 152-layer ResNet with 50 bottleneck residual blocks. When input and output dimensions did not match, the skip connection uses a learned linear projection for the mismatching dimensions, and an identity transformation for the other dimensions.

Training In training, each image is resized into 256×256 pixels without cropping. At each iteration, a mini-batch of 4 samples is constructed by sampling 4 training images from the dataset with the gradient being accumulated over 64 batches. The batches were formed by randomly shuffling the database with the random seed parameter set to 45. We pre-trained the ResNet–152 using a starting learning rate of 10^{-7} for weights, with a decay of 9^{-10}. Optimization is performed using Stochastic Gradient Descent (SGD), with a momentum of 0.9. We executed 40 epoches in total and decreased the learning rate with a polynomial decreasing policy at a power of 3. The network was trained using the Caffe framework [9] running on an NVIDIA Quadro M4000.

Testing The output of the last fully-connected layer of the ResNet-152 is fed into a c-way softmax that produces a probability distribution over the class labels. In order to construct the concept list for each image, we defined a threshold. To

define an appropriate value for this threshold we performed an exploration by using the validation dataset. In this run we assigned the best threshold found in the exploration performed on the validation set (T = 0.06). This threshold was used for all concepts and yields an F1 score of 0.104 on the validation set.

4 Results

In our participation, we submitted one run to the concept detection task. In this section, we describe the run and compare the performance with other automated runs that did not use any external sources.

In Table 2, it can be seen the performance of the proposed and the top techniques for the caption detection subtask. The table reports scores corresponding to the test set, from the metric proposed by the organizers. Our score is reported by run1, which uses a threshold over the scores of the class labels.

Table 2. Selected subset of results from ImageCLEFcaption (2017) – concept detection task. Comparison of our submitted run to top scores with no external resources.

Group Name	**Run type**	**Mean F1 Score**
Aegean AI Lab	Automatic	0.158
Information Processing Laboratory	Automatic	0.143
Information Processing Laboratory	Automatic	0.141
MedGIFT-UPB	Automatic	**0.089**

5 Conclusions

In this paper, we present a framework to address the multilabel image classification problem. However, training a multilabel CNN is not applicable on noisy datasets with a limited number of training samples due to the large number of parameters to be learned. We show that a CNN pre-trained on single label image datasets, e.g., ImageNet, can be transferred to tackle the multi-label problem. A research direction is to introduce more aggressive data augmentation techniques designed to improve the network generalization capabilities. Aligned with new techniques to generate medical hypotheses, training a multi-label CNN on a medical dataset would become more feasible. Still, the task at hand is clearly very difficult without use of external resources.

References

1. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF – Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Springer International Series On Information Retrieval. Springer, Berlin Heidelberg (2010)

2. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., García Seco de Herrera, A., Bromuri, S., Amin, M.A., Kazi Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., Roldán García, M.d.M.: General overview of ImageCLEF at the CLEF 2015 labs. In: Working Notes of CLEF 2015. Lecture Notes in Computer Science. Springer International Publishing (2015)
3. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0) (2015) 55 – 61
4. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: CLEF 2017 Proceedings. Lecture Notes in Computer Science, Dublin, Ireland, Springer (September 11-14 2017)
5. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS.org ¡http://ceur-ws.org¿ (September 11-14 2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)
7. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3) (2015) 211–252
8. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014)
9. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)