PUC Chile team at Caption Prediction: ResNet visual encoding and caption classification with Parametric ReLU

Vicente **Castro**, Pablo **Pino**, Denis **Parra and** Hans **Lobel**

Pontificia Universidad Católica de Chile, Av. Vicuña Mackena 4860, Macul, 7820244, Chile

Abstract

This article describes PUC Chile team's participation in the Caption Prediction task of ImageCLEFmedical challenge 2021, which resulted in the team winning this task. We first show how a very simple approach based on statistical analysis of captions, without relying on images, results in a competitive baseline score. Then, we describe how to improve the performance of this preliminary submission by encoding the medical images with a ResNet CNN, pre-trained on ImageNet and later fine-tuned with the challenge dataset. Afterwards, we use this visual encoding as the input for a multi-label classification approach for caption prediction. We describe in detail our final approach, and we conclude by discussing some ideas for future work.

Keywords

Image Captioning, Medical Artificial Intelligence, Deep Learning, Perceptual Similarity, Convolutional Neural Networks

1. Introduction

ImageCLEF [1] is an initiative with the aim of advancing the field of image retrieval (IR) as well as enhancing the evaluation of technologies for annotation, indexing and retrieval of visual data. The initiative takes the form of several challenges, and it is especially aware of the changes in the IR field in recent years, which have brought about tasks requiring the use of different types of data such as text, images and other features moving towards multi-modality. ImageCLEF has been running annually since 2003, and since the second version (2004) there are medical images involved in some tasks, such as medical image retrieval. Since those versions, the ImageCLEFmedical challenge group of tasks [2] has integrated new ones involving medical images, with the medical image captioning task taking place since 2017. It consists of two subtasks: concept prediction and caption detection. Although there have been changes in the data used for the newest versions of the challenge, the goal of this task is the same: help physicians reduce the burden of manually translating visual medical information (such as radiology images) into textual descriptions. In particular, the caption prediction task within the ImageCLEFmedical challenge 2021 aims at supporting clinicians in their responsibility to provide clinical diagnoses by composing coherent captions for the entirety of a medical image.

In this document we describe the participation of our team from the HAIVis group[1] within the artificial intelligence laboratory[2] at Pontificia Universidad Catolica de Chile (PUC Chile team) in the image captioning task at MedicalImageCLEF 2021 [2]. Our team earned 1st place in this challenge, and our best submission was a combination of deep learning techniques to visually encode the medical images, followed by a traditional classification of captions that were re-ranked by statistical information obtained from the training dataset.

The rest of the paper is structured as follows: section 2 describes our data analysis, while in section 3 we provide details of our proposed methods and experiments for model training and validation. Later, in section 4 we provide details of our results, and finally in section 5 we conclude our article.

2. Data Analysis

The dataset provided for this challenge consists of two sets of 2,756 and 500 image-caption pairs for training and validation, respectively. Each caption consists of a natural language text, which is a highly technical annotation made by physicians about abnormalities and medical objects in the image it corresponds to.
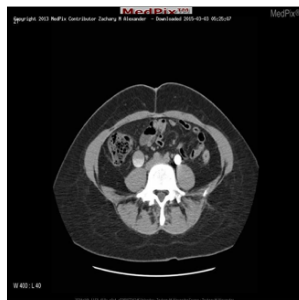
Figure 1: Dataset examples

Each caption was processed with the NLTK library[3] [3], following the evaluation methodology of the task[4]: (1) The caption is converted to lowercase. (2) All punctuation marks are removed and each caption is split into individual words. (3) Stopwords are removed using NLTK's "English" stopwords list. (4) Stemming is applied with NLTK's Snowball stemmer.

---

[1]http://haivis.ing.puc.cl/

[2]http://ialab.ing.puc.cl/

[3]https://www.nltk.org/

[4]Evaluation Methodology at https://www.imageclef.org/2021/medical/caption

Figure 2 shows the distribution words per number of appearances in the dataset, for example, 28% of words have only one occurrence. Figure 3 shows the distribution of caption lengths.
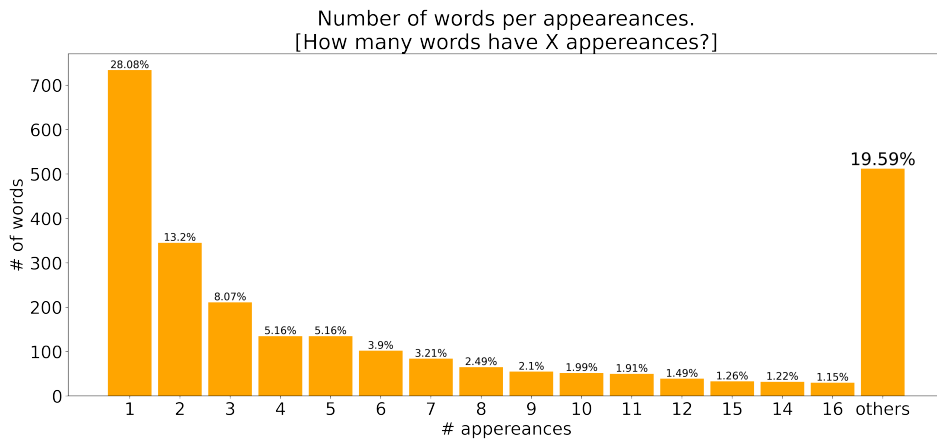
Number of words per appeareances.
[How many words have X appeareances?]

Figure 2: Distribution of number of words per number of appearances.

Distribution of caption lengths
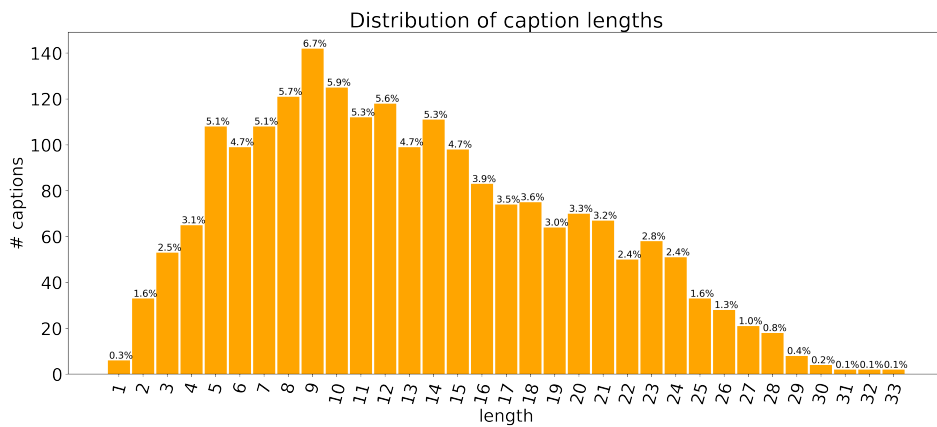
Figure 3: Caption length distribution.

Figure 4 shows the most common words in the dataset and their number of appearances, showing that some words are very common in the dataset, appearing in about 40% of all training captions. From a semantic analysis, these words seem to have broader and more descriptive meanings of the different elements in the images. This may be a direct cause of the fact that simpler and naive methods, that are based on more statistical approaches, can outperform more complex models.

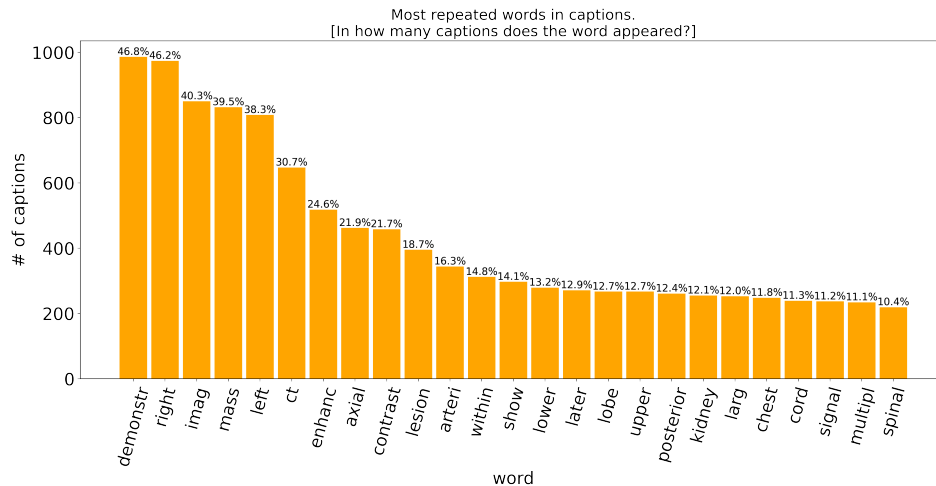**Figure 4:** Most common words in the dataset.

3. Method and Experimentation

While addressing the task we tried three main approaches: a pure statistical method, a multi-label classification approach (*MLC*) and a perceptual similarity based model (*Sim*).

3.1. Statistical approach

Our initial approach to the challenge tried to leverage the statistics related to the composition of each caption. This first model was a naive algorithm that randomly selected a caption length from the training set, created a list of this length with the most popular words in the dataset, and shuffled them to get a random order. This simple method obtained a mean BLEU score of 0.357 on the validation set and, when submitted, scored 0.378 points in the test set.

This first approach helped us gain an intuition about how the BLEU score varied and how susceptible it was to different components of the caption. Our initial hypothesis was that more relevant than the order of words in the caption, the correctness of them was the most significant element for the metric. To test this assumption, we explore the alternative of a multi-label classification approach that, given an image, predicted the most relevant words in the caption.

3.2. Multi-label classification approach (MLC)

In this approach we consider each word as a class, and trained a Convolutional Neural Network (CNN) to predict the words of a caption given an image. Then, the top classified words are selected and ordered by a statistical rule to produce a final caption. Figure 5 shows the full pipeline for caption generation with our approach, and we give more details next[5]:

---

[5]For all our implementations we used PyTorch as our main DL framework: https://pytorch.org/

3.2.1. Preprocessing

We process each image-caption pair to reduce the number of target classes (words), and to prepare the image to pass it through the network. The following steps were applied:

1. **Caption processing**: We processed each caption according to the evaluation methodology described in the previous section, transforming each caption into a list of stemmed words (labels). The vocabulary is composed by all the words in the training data with four appearances or more. We did not perform any special handling for words in the validation set that were not present in the training vocabulary. After filtering, the training vocabulary size was reduced to 1,075 words (1,189 when using training and validation set).

2. **Image processing**: Each image is transformed to have pixel values within a [0, 1] range (in each RGB channel) and then is normalized by the mean and standard deviation (over each channel), according to `torchvision` documentation [6]. As a data augmentation method, a crop of 300x300 pixels is taken from the image. For the training set, this crop is selected from a random location, whereas for validation and testing, the central crop of the image is always taken. This is a common training setup and has been used for similar purposes in past versions of the challenge [4].

3.2.2. Classification training

Several ResNet [5] and DenseNet [6] model architectures were tested, with and without fine tuning from ImageNet [7] pre-trained weights. Fine tuning of a DenseNet121 model pre-trained on the ChestX-ray14 dataset [8] was also tested. Different layers of the network were frozen during fine-tuning, as a measure to avoid over-fitting.

In addition, the last layer of the network was replaced with a fully connected layer that matched the dimensionality of the training vocabulary size. Furthermore, we added a dropout layer and passed the resulting values by a Parametric ReLU (PReLU) [9] activation function. With this, the output of our model was a vector of dimension vocabulary size and unbounded range.

In training, we sought to minimize the Binary Cross Entropy loss between the vector predicted by the model and the one-hot encoded ground truth, calculated as:

$$L = \sum_c^C -w_c[y_c \cdot \log \sigma(x_c) + (1 - y_c) \cdot \log(1 - \sigma(x_c))]$$

where C is the number of labels to classify. In code, this loss was calculated with `BCEWithLogits`[7] function from `pytorch`. As optimizer, we used Adam[10] with no weight decay and an initial learning rate of 5e-4, after epoch 15 this last hyper-parameter was reduced to 1e-4.

---

[6] Documentation @ https://pytorch.org/vision/stable/models.html
[7] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html
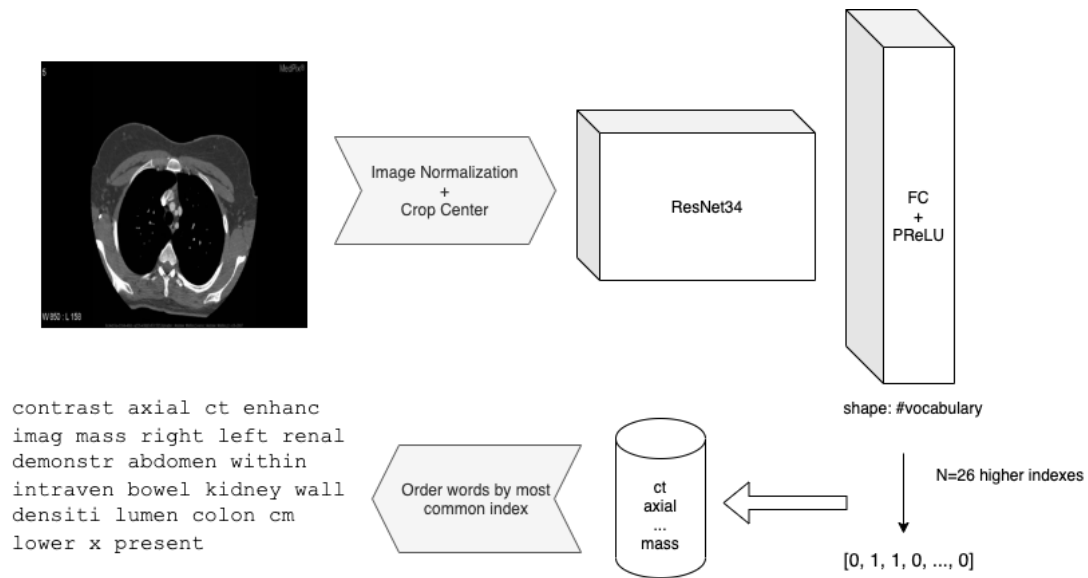
Figure 5: Model diagram. Top $N = 23$ classified words are selected for the caption.

3.2.3. Captioning

Once the classification output is obtained from our visual model, it needs to be translated into a caption. We define N as the length of the output caption, a hyper-parameter of the model and choose the N highest scoring words. Then, we used a statistical approach to order the words in a logical sentence: for each word, we define its position as the most common one it has across all training captions.

Two output examples from our model are shown next, with good (Fig. 6) and bad (Fig. 7) performance:

Figure 6: Example of caption prediction with good performance, BLEU= 0.850 (N=23)

Validation Image:

Ground Truth:
----
fractur disloc though c2 c3
disc fractur involv transvers foramenaat
c2

Predicted:
----
coron multipl axial note posterior
ct enhanc later tissu t2
imag right mass left signal
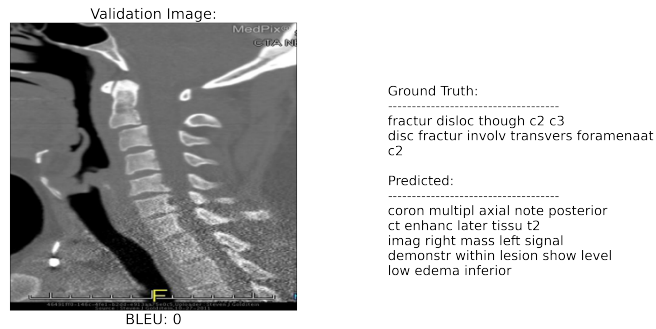demonstr within lesion show level
low edema inferior

BLEU: 0

Figure 7: Example of caption prediction with bad performance, BLEU= 0, (N=23)

3.3. Similarity-based approach (Sim)

Another method that we used and resulted in a fairly good experimental performance was a similarity-based approach. For each test image, we ranked the most similar images in the training set using the Learned Perceptual Image Path Similarity (LPIPS) [8][11], a learned metric based on the similarity between deep features from several neural network layers, in our experiments, an AlexNet[12] model. Then, the caption from the closest training image is assigned to the test image.

This approach resulted in a very good test performance and helped us to reach and maintain the top 3 in the leaderboard. Furthermore, we tested this approach for the concept detection task where we also achieved better performance.

4. Results

To evaluate our model we measured the BLEU score [13] for each caption generated against its ground truth, following the challenge evaluation procedure[9]. It is important to emphasize that this metric must be calculated with version v3.2.2 of the NLTK library since new updates change the results considerably. Table 1 shows our methods' scores in the validation set.

Table 1
Results in the validation set.

Method	BLEU
Sim: LPIPS Similarity (from AlexNet)	0.459
MLC: ResNet	0.544

Additionally, we measured word recall as a metric for the classification method. Since BLEU is a precision-based metric, including a recall-based metric should help evaluate the performance

---

[8]Code available @ https://github.com/richzhang/PerceptualSimilarity
[9]Refer to "Evaluation methodology" @ https://www.imageclef.org/2021/medical/caption
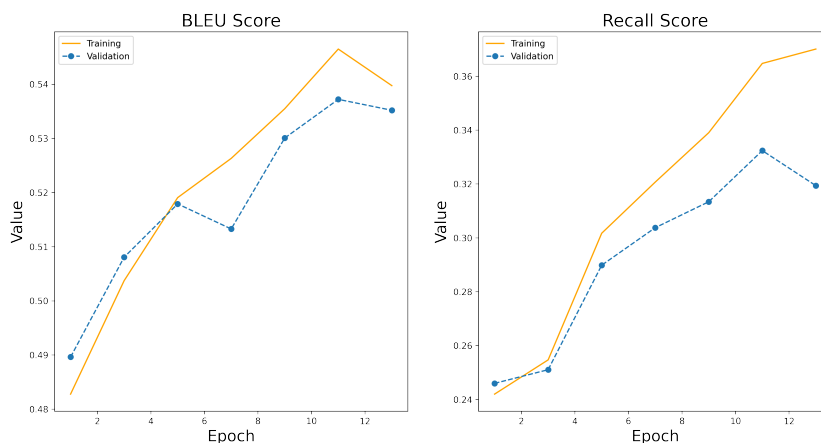
Figure 8: BLEU and Recall Score during training with (N=26)

more extensively, leading to better captions.

The best result was achieved with the multi-label classification approach, using a ResNet34 [5] model pre-trained on ImageNet and fine-tuned for 15 epochs, only with the last 5 layers with learnable parameters, whilst the other layers were frozen. The training scheme mentioned above was followed. For word selection we set N=26, value that was inferred from the distribution in Figure 3 and validated with experimental results. Figure 8 shows the development of BLEU and word recall during training.

4.1. CrowdAI Runs

Four submissions were made to crowdai.org using the methods described, the details and results are shown in Table 2.

Table 2
Submission results.

	Method	BLEU
Subm1	**Statistical:** random length + most common words + random order	0.378
Subm3	**MLC:** ResNet50, random length + fixed order	0.351
Subm4	**Sim:** LPIPS similarity approach	0.442
Subm6	**MLC:** ResNet34 and most common index for ordering, trained for 20 epochs	0.509
Subm7	**MLC:** ResNet34 and most common index for ordering, trained for 15 epochs	**0.510**

5. Conclusion

In this article we have provided details of the participation of the PUC Chile team, winners of the caption prediction task within the ImageCLEFmedical challenge 2021. In the process of building our final submission, we tested several approaches, detailed in this paper. Our final submission was based on a ResNet34 architecture to visually encode the input medical image, followed by predicting captions as a multi-label word classification task, and finally re-ranking the word order based on statistical information from the training dataset. In future work, we plan at testing other CNN architectures, perform further experiments exploiting perceptual similarity, and test other techniques for neural language modeling.

Acknowledgments

References

[1] 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. Stefan, 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.

[2] 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: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.

[3] E. Loper, S. Bird, NLTK: The Natural Language Toolkit, in: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02, Association for Computational Linguistics, USA, 2002, p. 63–70. URL: https://doi.org/10.3115/1118108.1118117. doi:10.3115/1118108.1118117.

[4] D. Lyndon, A. Kumar, J. Kim, Neural Captioning for the ImageCLEF 2017 Medical Image Challenges., in: CLEF (Working Notes), 2017.

[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] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional

Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[7] 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, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[8] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[9] K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[10] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc., Red Hook, NY, USA, 2012, p. 1097–1105.

[13] 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 on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.