iHealth-Chile-1 at RRG24: In-context Learning and Finetuning of a Large Multimodal Model for Radiology Report Generation

Diego Campanini [2], Oscar Loch [1,2,3], Pablo Messina [1,2,3],
Rafael Elberg [1,2,3], and Denis Parra [1,2,3]
[1] Department of Computer Science, Pontifical Catholic University of Chile.
[2] Millennium Institute for Intelligent Healthcare Engineering (iHEALTH), Chile.
[3] National Center for Artificial Intelligence (CENIA), Chile.
{oscar.loch,pamessina,rafael.elberg}@uc.cl,
dparra@ing.puc.cl, diego.campanini@ing.uchile.cl

Abstract

This paper presents the approach of the iHealth-Chile-1 team for the shared task of Large-Scale Radiology Report Generation at the BioNLP workshop, inspired by progress in large multimodal models for processing images and text. In this work, we leverage LLaVA, a Visual-Language Model (VLM), composed of a vision-encoder, a vision-language connector or adapter, and a large language model able to process text and visual embeddings. We achieve our best result by enriching the input prompt of LLaVA with the text output of a simpler report generation model. With this enriched-prompt technique, we improve our results in 4 of 5 metrics (BLEU-4, Rouge-L, BertScore, and F1-RadGraph,), only doing in-context learning. Moreover, we provide details about different architecture settings, fine-tuning strategies, and dataset configurations. Models parameters can be found in HuggingFace [1].

1 Introduction

The task of radiology report generation (RRG) from medical imaging through deep neural networks is an active area of research (Monshi et al., 2020; Messina et al., 2022). For one thing, addressing and solving this task can help radiologists in identifying anomalies from one or more input images, as well as save them time on administrative chores like typing text reports. Thus, doctors can spend more time with patients rather than clinical software (Topol, 2019). There have been several methods introduced in recent years to address this task but only recently the progress in open-source multimodal generative systems has opened the room for improving performance by integrating different modalities (text and images) in the same model. In this article, we describe our work leveraging the multimodal model LLaVA (Liu et al., 2023) to address this task.

There are several options to leverage LlaVA for this challenge, such as utilizing the original version LLaVA-1.0 (Liu et al., 2023), the clinically finetuned version LLaVA-Med (Li et al., 2023), as well as the newest version LLaVA-1.5 (Liu et al., 2024). Due to hardware limitations, in this challenge, we used the language model component with 7 billion parameters (LLaMA 1.0 and Vicuna) and we tested several configurations focusing on the *findings generation* task.

In this document, we describe details of several configurations tested, including different vision encoders (CLIP and BiomedCLIP), VL projector (matrix and MLP) and language model for text decoding (LLaMA1.0 and Vicuna-7b). Among our findings, we highlight that integrating the output of another method as input context for LLaVA resulted in our best version for the challenge.

2 Task Description

2.1 Datasets

The data provided by the challenge (Xu et al., 2024) consists of 5 datasets PadChest (Bustos et al., 2020), BIMCV-COVID19 (Vayá et al., 2020), CheXpert (Chambon et al., 2024), OpenI, and MIMIC-CXR (Johnson et al., 2019). All of them have a medical report with at least the finding section, in total, we have $344,394$ training samples.

In the present work, we focus only on the finding generation, in each training step we use the findings section, of the official train datasets. We do not use any extra dataset or data augmentation techniques.

The results reported in this work are measured in the challenge hidden test set which has $1,063$ samples for the generation of the finding section.

3 Methodology

3.1 Model Architecture

The architecture used in this work is known as Large Language and Vision Assistant for

[1] https://huggingface.co/dcampanini

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 608–613
August 16, 2024. ©2024 Association for Computational Linguistics

BioMedicine (LLaVA-Med Li et al., 2023), we fine-tuned this system following different approaches described in the section 3.2.

The LLaVA-Med system has 3 main blocks (Figure 1), the first is a vision encoder, then a vision-language connector to project the image features into the word embedding space, and finally, a Large Language Model (LLM) that processes visual and language tokens to generate a final answer.

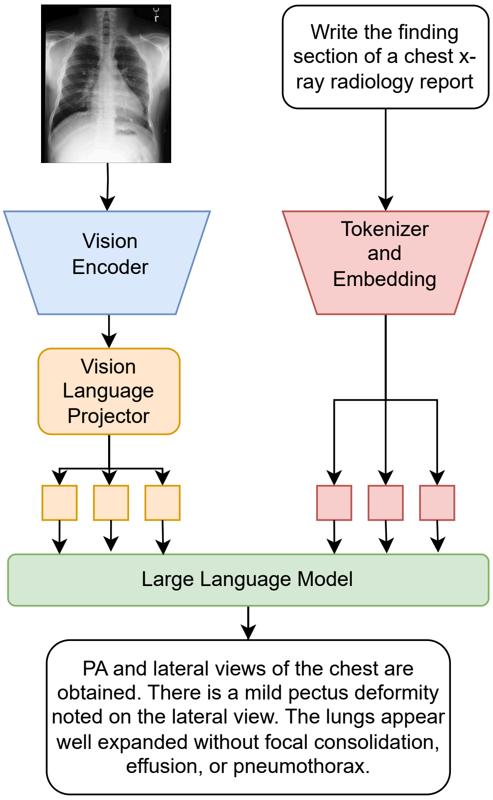
Write the finding section of a chest x-ray radiology report
Vision Encoder
Tokenizer and Embedding
Vision Language Projector
Large Language Model
PA and lateral views of the chest are obtained. There is a mild pectus deformity noted on the lateral view. The lungs appear well expanded without focal consolidation, effusion, or pneumothorax.

Figure 1: LLaVA-Med architecture used in this work. The Large Language Model (LLM) processes the features extracted from the image and the prompt.

There are different options for each block, in our case, we use 3 different vision encoders all of them based on CLIP (Radford et al., 2021), they are clip-vit-L-patch14, clip-vit-large-patch14-336, and BiomedCLIP (Zhang et al., 2023). For the connector, we choose a projection matrix and a 2 layer MLP. Finally, for the LLM we select LLaMA1.0-7B, and Vicuna-7b-v1.5. Table 1 summarizes the 3 model versions used during the challenge.

3.2 Training Strategy

We train our models in 2 stages, similar to the strategy proposed in Liu et al., 2023, but adapted for report generation, and not for instruction tuning. The stages are detailed as follows:

- **Stage 1 or alignment:** the image encoder and the LLM are frozen, and the MLP or projection matrix are trained.
- **Stage 2 or fine-tuning:** the MLP or projection matrix and the LLM are trained.

For both stages, we train with samples formed by one image and the respective finding section. Our models process one image at a time. Therefore, we manipulate the training dataset when more than one image is associated with a medical report.

For the dataset MIMIC-CXR, for each medical report we select the Anterior Posterior (AP) image or the Posterior Anterior (PA) image, and the finding section.

We use the image's name to select the frontal images for CheXpert, which indicates the view presented in the X-ray exam (frontal or lateral).

For the last 3 datasets PadChest, BIMCV-COVID19, and OpenI, we take the first image in the array of images associated with each medical report, which was, in general, a frontal view.

For the Model-1.0 (Table 1) we start the fine-tuning from a LLaVA-Med checkpoint shared in the official GitHub repository[2], and we update the linear matrix and the LLM using different combinations of the official train datasets. Stage 1 is omitted for this model since the based model was trained in biomedical data extracted from PMC-15M (Zhang et al., 2023) an image-text dataset extracted from scientific publications.

For Model-1.1 and Model-1.2 (Table 1) we train following the 2 stages strategy, for the stage 1 we use the complete challenge train dataset, considering only one image per finding section, we employ more training samples in this stage since is more general than stage 2, so more broad data can help the final model performance.

On the other hand, for stage 2, we only use MIMIC-CXR. This decision is discussed in the section 4. For these 2 models, we have to execute stage 1, since we don't have a projector specialized in medicine to connect the vision-encoder with the LLM embedding space.

For stage 1 we always use a learning rate of 1×10^{-3} and a cosine learning rate with a warmup ratio of 3%. Similarly, for stage 2 we employ the same scheduling and warmup ratio but with a learning rate of 1×10^{-4}. Every stage is performed in a GPU NVIDIA RTX A6000 with 48 GB of memory.

[2] https://github.com/microsoft/LLaVA-Med

Model version	Vision-encoder	VL Projector	LLM
Model-1.0	clip-vit-L-patch14	Matrix	LLaMA1.0-7b
Model-1.1	clip-vit-large-patch14-336	MLP	Vicuna-7b-v1.5
Model-1.2	BiomedCLIP	MLP	Vicuna-7b-v1.5

Table 1: Configuration of the 3 model architectures used during the challenge.

3.3 Text Prompt

The prompt given to the LLMs performs an important role in getting a solid model performance. In our work, the LLM can receive as input the image feature and the prompt.

For stage 1, we follow the strategy mentioned in Chaves et al., 2024 and we use only the image to train the projector, no prompt or extra information is provided to the model, in this way, we force the LLM to focus on the images.

On the other hand, for stage 2, the prompt is formed by the model context and the instruction, with the first we can control for example the model personality, asking to be polite, and in our case, we also define that the LLM does not have to provide dates, hours or text with enumeration in the report. The final instruction to the LLM is: *Write the finding section of a chest x-ray radiology report*. The complete prompt (context + instruction) is described in the following paragraph:

- **Context:** *You are LLaVA-Med, a large language and vision assistant. Write in the style of a radiologist, write one fluent text without enumeration, dates, or hours of the day, be concise, and don't provide explanations.*

- **Instruction:** *Write the finding section of a chest x-ray radiology report.*

The previously described prompt is used in stage 2 and inference.

Additionally, we have considered making another test, improving the prompt using as extra information other findings sections, generated by a multilabel classifier and a group of templates (Pino et al., 2021). This different system consists of a DenseNet-121 CNN trained to classify 13 pathologies for chest X-ray images, and then using the output labels, we generate the finding section based on a group of template sentences. The LLM receives as input the image features and the new prompt with extra information, which we call enriched-prompt. The new prompt instruction looks as follows:

- **Instruction:** *Write the finding section of a chest x-ray radiology report using the image, and the following information: the lungs are clear. heart size is normal the cardiomediastinal silhouette is normal. there is noted left sided or right sided , small, moderate, or large pneumothorax in the lung no pleural effusions. there is no evidence of fibrosis no displaced fracture is seen there is a noted right sided or left sided picc or tube*

4 Experiments and Results

In Table 2 we report the results of the 3 model versions trained with different dataset configurations and performing or not stage 1. All results are only for the finding generation task.

The metrics outline in Table 2 are BLEU4 (B4 Papineni et al., 2002), ROUGEL (RL Lin, 2004), Bertscore (BS Zhang et al., 2019), F1-cheXbert (F1-cXb Smit et al., 2020), and F1-RadGraph (F1-RG Delbrouck et al., 2022a), the last column represents the average between these metrics. All the values are calculated using the official leaderboard web page with the framework VilMedic (Delbrouck et al., 2022b).

The first result in Table 2 is for the Model-1.0 without any posterior finituning or training, which is the original LLaVa-Med shared in the official repository, it has a poor performance generating finding. It is by far our worst model, so it should be fine-tuned to get good results even in tasks inside the biomedical domain.

From our experiments with Model-1.0, we see that considering only MIMIC-CXR we have good enough results comparable to using MIMIC-CXR + CheXpert, and consistently outperforming the same Model-1.0 trained with the complete challenge datasets (Table 2). For this reason, the training of the other models is performed only employing MIMIC-CXR for stage 2.

When we make use of BiomedClip (Zhang et al., 2023) we see a clear improvement in 6.31 percentual points for F1-cheXbert in comparison with the second-best model in this metrics (29.37 vs

Stage1	Ep	Stage2	Ep	B4	RL	BS	F1-cXb	F1-RG	Avg
Model-1.0 Clip LLaMA1.0-7B									
None	0	None	0	0.95	11.69	27.79	15.46	4.15	12.01
None	0	MIMIC-CXR	1	4.67	19.58	48.74	18.00	16.29	21.46
None	0	MIMIC-CXR	3	5.05	19.13	47.51	23.06	15.77	**22.10**
None	0	MIMIC-CXR + CheXpert	1	4.78	19.19	47.57	20.25	15.47	21.45
None	0	All	1	3.24	15.45	42.12	18.11	11.79	18.14
Model-1.1 Clip-336 Vicuna1.5-7B									
All	1	MIMIC-CXR	1	5.16	19.68	47.92	11.61	16.78	20.23
All	1	MIMIC-CXR	3	3.92	19.75	48.06	5.92	16.06	18.74
Model-1.1 Clip-336 Vicuna1.5-7B Enriched-prompt									
All	1	MIMIC-CXR	1	**6.46**	**20.51**	**49.23**	9.35	**18.59**	20.83
Model-1.2 BiomedClip Vicuna1.5-7B									
All	1	MIMIC-CXR	1	3.48	16.31	35.49	**29.37**	15.51	20.03

Table 2: Results on the hidden test set, for all 3 model versions without applying enriched-prompt, and the Model-1.1 improved through the enriched-prompt technique. All numbers are calculated using Vilmedic on the official challenge web page.

Model	B4	RL	BS	F1-cXb	F1-RG	Avg
Model-1.1 Clip-336 Vicuna1.5-7B	5.16	19.68	47.92	11.61	16.78	20.23
DenseNet-121 classifier + templates	4.81	15.96	44.03	**33.69**	18.41	23.38
Model-1.1 Clip-336 Vicuna1.5-7B + enriched prompt from DenseNet-121 classifier	**6.46**	**20.51**	**49.23**	9.35	**18.59**	20.83

Table 3: Efects of the enrich-prompt technique. The last row represents the metrics of the resulting system, which is the Model-1.1 but enhanced with the enriched prompt coming from the DenseNet-121+templates system.

Model	B4	RL	BS	F1-cXb	F1-RG	Avg
Model-1.1 Clip-336 Vicuna1.5-7B	5.16	19.68	47.92	11.61	16.78	20.23
DenseNet-121 classifier + templates v1	4.81	15.96	44.03	**33.69**	18.41	**23.38**
Model-1.1 Clip-336 Vicuna1.5-7B + DenseNet-v1	**6.46**	20.51	49.23	9.35	18.59	20.83
DenseNet-121 classifier + templates v2	4.74	16.17	47.28	27.44	13.08	21.74
Model-1.1 Clip-336 Vicuna1.5-7B + DenseNet-v2	5.94	19.40	47.20	7.15	16.87	19.31
DenseNet-121 classifier + templates v3	5.50	17.11	48.97	26.26	14.47	22.46
Model-1.1 Clip-336 Vicuna1.5-7B + DenseNet-v3	5.10	19.98	48.94	8.11	17.47	19.92
DenseNet-121 classifier + templates v4	4.18	17.05	42.91	27.20	**19.42**	22.15
Model-1.1 Clip-336 Vicuna1.5-7B + DenseNet-v4	5.21	**20.80**	**50.14**	5.90	18.51	20.11

Table 4: Impact of the enriched prompt technique using different template models, and the same multimodal model highlighted in gray. The resulting model's metrics are pointed out in yellow.

23.06). This suggests that the feature extracted from the image with this vision encoder allows to the model classify properly more pathologies than the previous vision encoders, considering that F1-cheXbert is a metric focus in the classification of 14 labels.

We apply the enriched-prompt technique to the model with the best F1-RadGraph, which is the *Model-1.1 Clip-336 Vicuna1.5-7B*, the result of employing this procedure is an improvement in BLEU-4, Rouge-L, Bert-Score, and F1-RadGraph, but a big fall in F1-cheXbert (Table 2, 3), this indicates that the model is not good at classifying the 14 classes considered by the metric.

Table 3 shows the change in the metrics for 2 base models, combined across the prompt. When we apply in-context learning to the Model-1.1 Clip-336 Vicuna1.5-7B adding to the prompt the reports generated by the DenseNet-121+templates, the resulting model overcomes the metrics of both previous systems, except for F1-cheXpert.

Another consequence of implementing an enriched prompt is the generation of shorter findings in comparison with those generated by the other model versions and by the classifier plus template sentences.

In Table 4 we show more evidence of the enriched prompt technique. The most consistent effect over the two base models is observed in the F1-RadGraph metric, improving up to 1.81 percentage points (pp) in the base multimodal model, and up to 3.79 pp for the template models. For Rouge-L, and Bert-Score we can also see an enhancement in the based models, the most outstanding result is the increase of 7.23 pp in Bert-Score for the DenseNet-121 classifier + templates v4. The different versions of the template models consider distinct types of templates and classifier hyperparameters, more detail about it can be found in the paper of iHealth-Chile-3&2. On the other hand, the effect of the enriched prompt technique in F1-cheXbert is always a big fall.

5 Conclusion

In this work, we performed an analysis of different model architectures based on LLaVA-Med, we conclude that using the best possible vision-encoder, and LLM we can improve some specific aspects of the system, such as the NLP overlapping (BLEU-4 and Rouge-L) or the more classification related metrics (F1-cheXbert), nevertheless to see more consistent results we suggest that more quality data is needed, particularly for alignment (stage 1). Moreover, since the promising results in F1-cheXbert obtained with BiomedCLIP is convenient to develop a vision-encoder custom to x-ray images.

Finally, the enriched-prompt techniques show auspicious results. It can work as a guide for the LLM, it shows good metrics when we calculate BLEU-4, Rouge-L, BertScore, and F1-RadGraph, but it should be complemented with an accurate classifier system to improve the F1-cheXbert.

Limitations

There are some limitations in the system that we propose. For instance, our model is unable to use multiple images, however, the medical reports for chest x-rays are usually formed by two or three views of the patient chest, so we are missing potentially important information.

The quality of the medical report generated with the enriched prompt technique should be analyzed in more depth, especially because of the large drop in the F1-cheXbert metric.

Another limitation is that our approach is computationally expensive, which limits the quantity of experiments that we can perform. Finally, our reports are not hallucinations free, for example in some cases, the model generates findings referring to another report for the same patient, but this is a problem because the model does not know previous patient exams.

Acknowledgements

This work has been funded by Millennium Institute for Intelligent Healthcare Engineering (iHEALTH) ICN2021_004; National Center for Artificial Intelligence (CENIA) FB210017, Basal ANID; Fondecyt 1221425; and the National Agency for Research and Development (ANID) through the Scholarship Program / Doctorado Becas Chile / 2019 - 21191569.

References

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. 2024. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40.

Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In *Machine Learning in Medical Imaging*, pages 654–663, Cham. Springer International Publishing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Eric Topol. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, 1st edition. Basic Books, Inc., USA.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and “discharge me!”. In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. 2023. Large-scale domain-specific pretraining for biomedical vision-language processing.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.