Video Processing and Understanding Lab

Universidad Autónoma de Madrid

SI1/PJI/2019-00414 AISEEME (2020-2022)

Aiding diagnosis by self-supervised deep learning from unlabeled medical imaging

D1.2 v1

Evaluation Datasets

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

CONSEJERÍA DE EDUCACIÓN E INVESTIGACIÓN

Comunidad de Madrid

AUTHORS LIST

Pablo Carballeira López	pablo.carballeira@uam.es
Marcos Escudero Viñolo	marcos.escudero@uam.es

HISTORY

Version	Date	Editor	Description
1.0	10/01/2021	Pablo Carballeira López	First version

CONTENTS:

1. INTRODUCTION 1
2. COLLECTION AND GENERATION OF DATASETS 3

1. Introduction

This deliverable describes the work related with the task T.1.2 “Collection and generation of datasets”: Support to other tasks by generating train and test data and associated evaluation methodologies. It includes the selection of appropriate datasets (images and associated ground-truth) and their generation if required.

2. Collection and generation of datasets

During the first months of the project, we have collected benchmarks covering each of the areas of the project. From those commonly used for image classification tasks that are here used to validate self-supervised approaches to those encompassing CT scans that are used here to validate state-of-the-art methods on lung nodule malignancy detection. In the process of selecting benchmarks, we have prioritized the selection of consolidated datasets that have been extensively used and studied in previous works.

Among those originally defined for generic image classification, we have used CIFAR10 [1], CIFAR100 [1], ImageNet [2], VOC2007 [3] and COCO2017 [4] to validate self-supervised pretext tasks. Furthermore, we have also used more specific datasets to are alike the previous ones but are focused on a particular task; hence representing a narrower domain, among these, we have explored Places205 [5] and Plant Disease [6].

We have also incorporated widely used datasets for the target applications: skin lesion and lung malignancy assessment that will allow a fair comparison at the end of the project. ISIC 2017, 2019 and 2020 datasets [7] [8] [9] for skin lesion analysis and LIDC-IDRI dataset [10] for lung nodule malignancy assessment.

Finally, for studying pretext tasks, we have also included recent datasets of different modalities to the previous ones, which data distributions are expected to be far from the rest of the benchmarks: X-Rays [11] and COVID-19 [12].

For all these datasets we have adopted the standard data sets defined and the validation measures proposed together with the datasets. We have not created specific datasets for the project as they have not been required to this point. However, we have cleaned and improved the annotations provided by LIDC-IDRI [10] by isolating the three-dimensional nodules of the CT scan, and associating them with their malignancy annotations (see Figure 1 for an example). In the original dataset, all the nodules of a patient were annotated together using the same file.

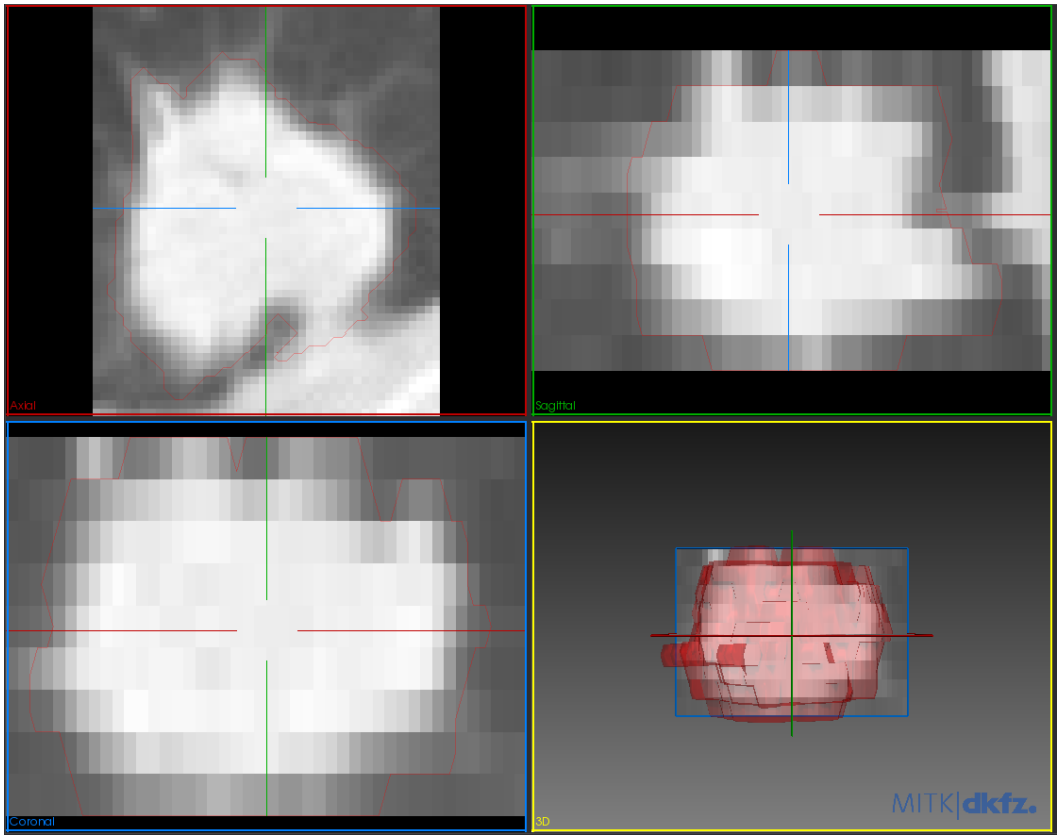

Figure 1 Isolation of a lung nodule by projecting the bidimensional mask on the 3D CT scan.

References

[1] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[2] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.

[3] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The PASCAL visual object classes challenge 2007 (VOC2007) results." (2007).

[4] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

[5] Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Places: A 10 million image database for scene recognition." IEEE transactions on pattern analysis and machine intelligence 40, no. 6 (2017): 1452-1464.

[6] Sharma, Saroj Raj, Dataset of diseased plant leaf images and corresponding labels. GitHub - spMohanty/PlantVillage-Dataset: Dataset of diseased plant leaf images and corresponding labels. 2018.

[7] Codella, Noel CF, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)." In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168-172. IEEE, 2018.

[8] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana,

Ofer Reiter, Allan C. Halpern, Susana Puig, Josep Malvehy: “BCN20000: Dermoscopic Lesions in the Wild”, 2019; arXiv:1908.02288.

[9] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J. & Soyer, P. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 8, 34 (2021). https://doi.org/10.1038/s41597-021-00815-z

[10] Armato III, Samuel G., Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." Medical physics 38, no. 2 (2011): 915-931.

[11] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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, pp. 2097-2106. 2017.

[12] Zhao, Jinyu, Yichen Zhang, Xuehai He, and Pengtao Xie. "Covid-ct-dataset: a ct scan dataset about covid-19." arXiv preprint arXiv:2003.13865 (2020).