Jan Čech
Center for Machine Perception
Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University in Prague

Master's Thesis Review

Prague, August 20, 2020

Title: Deep multiple-instance learning for detecting multiple myeloma in CT scans of large bones
Author: Vojtěch Mach
Supervisor: Dr. Rer. nat. Jan Hering
Date received: August 17, 2020

The thesis presents a method for detecting malignant marrow lesions from CT images of femur bones. A semi-synthetic dataset was constructed by augmenting real negative samples with lesions. Two experiments, learning convolutional network classifier, were conducted in: (1) a standard fully supervised, and (2) Multiple-instance learning (MIL) settings.

The thesis has six chapters. The first chapter presents the theoretical background including both medical imaging and machine learning techniques. The second chapter reviews related work. The third chapter describes the semi-synthetic dataset that was used for subsequent experiments. Chapter four presents the method including implementation details. Chapter five gives the experiments. Chapter six conclude the thesis.

The thesis is generally well written and comprehensible. The theoretical background is extensive and presents recent methods in machine learning. The problem is apparently challenging in several respects, e.g. a limited size of real dataset, noisy images. The MIL scenario is very interesting, however not trivial to implement. Author clearly showed a certain competence in implementing advanced machine learning models, despite the MIL experiment was rather negative. Rigorous quantitative evaluation of trained classifiers is given.

There are a few weak points in the thesis:

1. Review of related literature is limited. Chapter 2 on related work is brief and does not discuss many MIL approaches.
2. Description of the semi-synthetic dataset construction presents techniques to clear the data. The process is heuristic and is not much illustrated by examples. Perhaps due to a limited size of the real dataset, most of the cleaning processes could have been done manually.
3. The technical chapter mixes implementation details with description of the method, i.e. design of the instance and the bag classifiers. The order of presentation and level of details is unusual, but still acceptable.
4. A single MIL method, using "instance aggregation" in Eq. (4.1), is attempted. The aggregation function is often called the MIL pooling layer in literature and there are several simpler option to try, e.g. maximum, averaging, weighted averaging or exploiting attention mechanism [1]. Why baseline MIL pooling were not tried? The proposed MIL classifier did not learn from scratch. It is unclear if the classifier overfits or there is an implementation mistake. To avoid the latter, researchers usually test the method on the MNIST benchmark. The MNIST dataset [2] of hand written digits is very popular and accurate instance level classifiers exist. Positive bags are synthesised to contain at least one instance of a particular digit (e.g. 9), while negative bags do not contain the digit.
5. The size of the semi-synthetic set is never detailed. It was perhaps generated on-line when training, but how many samples were generated throughout the learning process should have been specified.
6. All the experiments were done on the semi-synthetic dataset. It is unclear how the result (for fully-supervised) classifier generalizes to the real data. Did you try to measure accuracy on unseen split of the real dataset?

In summary, I suggest assessing the thesis by

B – very good.

Ing. Jan Čech, Ph.D.

References

[1] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *CoRR*, vol. abs/1802.04712, 2018. [Online]. Available: http://arxiv.org/abs/1802.04712

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner., "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Novemebr 1998.