Master Thesis Review - Opponent

Thesis Author: Bc. Jakub Bureš
Thesis Title: Hybrid Discriminative-Generative Training for Set data
Reviewer: Ing. Michal Najman (Avast Research Lab)

Multiple instance learning is known to be fruitful in settings where objects inherently consist of cardinality-varying sets – which is a common scenario in many real-world use cases. However, such data formats have been explored mostly in the framework of discriminative learning, lacking behind the recent achievements in machine learning.

The thesis written by Bc. Jakub Bureš acknowledges the gap between multiple instance learning and the state-of-the-art generative approaches, taking rigorous steps towards generative multiple instance learning. Hence, the topic has been selected well as it follows the most-recent advances in machine learning and certainly develops new ideas relevant to the scientific community.

The work starts with a thorough overview of commonly used machine learning methods giving a detailed explanation of their foundational basis. This yields solid theoretical grounds for generative techniques described in further chapters.

The backbone of the thesis are two existing generative methods applied to multiple instance learning: a hybrid discriminative-generative model and a hybrid variational autoencoder. As those methods come from fixed-sized vector learning scenarios, the reader learns about necessary modifications to the original versions of the methods and their experimental evaluations where those are compared to standard discriminative machine learning.

Although struggling with high-variance noise, the well-designed experiments signal that the methods beat vanilla discriminative learning on the given data sets, suggesting that model performance in multiple instance learning settings can be improved using the proposed or extended methods.

The quality of English is decent although a couple of typos, missing articles and a few poor-structured phrases occasionally limit readability.

As far as the structure of the thesis is concerned, the thesis develops concepts gradually building on previous chapters. This is seen especially in the first half where equations are derived in detail, from definitions to optimisation tasks. For example, the derivation of the hybrid generative-discriminative model is well put. The second half, however, lacks the consistency in

paying attention to detail. In addition, some concepts are not delivered properly lacking explicit connection to those already defined. This is seen especially in final sections of the thesis where it negatively impacts text cohesion and comprehensibility. For instance, the thesis does not derive the variational autoencoder for multiple instance learning in comprehension levels set by previous chapters. Since the text in the first half of the thesis is well-readable, I assume the second half has been written under tight deadlines. Nevertheless, an expert reader will understand all concepts and methods well.

Although one may find themselves a little confused when distinguishing the thesis contribution from prior work, the contribution is clear given the overall topic of the thesis. To give two examples: first, impressive analytical thinking is shown in sections discussing author-designed experiments. Second, in addition to the final section experiments, the polynomial toy experiment and the corresponding discussions stand out.

Throughout the thesis, the derivation of the used mathematical concepts is rigorous and sound and reveals the author's deep understanding of the examined concepts and their mathematical building blocks. Also, the experiments are evaluated and discussed to the highest scientific standards.

My questions to the author are listed below:

1. In the polynomial toy problem, you state that the generative model (alpha=0) fits the data better than the discriminative model. What is your intuition behind this observation? Would this be observed with a different shape of training data, e.g. more than two clusters?
2. To construct a generative multiple instance model, you assume instances in a bag are independant. How would you modify the method if we assumed there is an unknown partial dependency between bag instances?
3. In Sec. 4.5.1, why is the constant ν so large (in the order of 10^10)? What would be the impact on target metrics (average AUC) if ν=1 or ν=10^20?
4. Why do you think the variance of average AUC is large on the given datasets?
5. What are the next steps to a fully generative multiple instance model?

To conclude, the author fulfills the thesis assignment completely and correctly. The thesis proves the author is capable of building on existing work and extending it to new concepts, while carefully designing experiments and scientifically evaluating their results. There is a space for improvements especially in the level of comprehension, delivery and the quality of English writing but the key substance makes perfect sense and the message is eventually delivered.

The work is certainly above average and should be accepted as a Master Thesis. I evaluate the thesis as Very Good, B. I believe the author has great potential in scientific disciplines and I am looking forward to his future contributions to computer science.

Ing. Michal Najman
Senior AI Researcher
michal.najman@avast.com

Avast Research Lab
Avast Software s.r.o.
Pikrtova 1737/1a
140 00 Praha 4

May 19th, 2022
in Prague