Opposition on written thesis

Thesis

Label-Efficient Multi-Objective Machine Learning for e-Commerce

Mattias Arro

mattias.arro@gmail.com

Reviewing student

Erik Droh

droh@kth.se

Contents of the opposition

Overall assessment

Do you see a logical presentation in the report, so called “red thread”? Are the individual sections logically connected to each other?

Yes, I think that the thesis is coherent and follows a good red thread. Information is hinted at and presented in theory and background and is then explained further in the method chapter of how it was used. Extensive analysis is then applied to the different sections to provide a final result, discussion and conclusion.

Is the thesis understandable? Is something missing?

Yes, there’s a lot of information to digest however. Even if you are acquainted with some areas presented already. Images could help to explain some of the different models and information presented in the background which would give the reader a visual representation of a model going forward in the thesis as well as a break in between all the reading.

The only thing I feel was missing was a distinct results chapter that visually informed and convinced the reader of which model was best at labeling. Contrary to the visual similarity part which was very easy to understand and digest the results of.

Is the thesis and its content consistent and overall trustable? (evaluated, verified, tested)

I think it is. References are good, and I feel that a lot of critical thinking has been applied when evaluating sources for information, theory and application in the system. A good example of this is presented when discussing labeling of e-commerce items such that can be deemed correct but is panelized by the labeling of the item. Jacket and coat can sometimes be very similar and deemed the same thing depending on who you ask. Labeling is done by one person on a dataset and can thus be seen as incorrect if another person were to label the same item. Especially for low-level categories.

Abstract

Does it present the background, problem & purposes, what has been studied, results of the study and conclusions. Is the abstract understandable by its own?

Very good abstract, felt as if I got a complete summary of the thesis and that I dove into the thesis know what I would expect in terms of methodology, result and conclusions. I felt intrigued to continue reading how Mattias came these results and conclusions.

Introduction

Does the introduction give the big picture of the topic?

The introduction brings up a lot of information regarding many different types of models and techniques and when and where they might be used.

- Might be a bit too much information to grasp in the beginning. Many directions on where we are going.

+ Question at the end: How to achieve good neural net performance when labeled data is limited. Which is good and invites the reader to want to know more. Got a "cliffhanger" feeling.

Does it have relevant references?

The introduction is reference heavy in the beginning citing multiple areas of where deep learning has been applied. The references are relevant and sound.

References are however lost after the first paragraph which seems a bit off-balanced. Maybe add some more to the later parts of the text if possible?

Subject and problem area

Is the chosen topic and problem area of interest?

Yes, very much so. If solved has broad application on the entire e-commerce industry which today suffers from extreme time consuming manual labor in order to be relevant and meet the standards which we expect from online retailers today. This technology would be great for large companies with massive databases but also for smaller companies with limited resources and focus.

Does the topic have a clear introduction, theory and background?

Yes, well documented.

How are the theories chosen?

Theories are chosen with thorough consideration and the authors seem to have done their research on all different areas of their report to explore the possible ways of moving forward. They bring up different approaches to theories and their choices explaining why they choose a certain choice for their problem statement and why it will be beneficial for them.

Is there a relevant theory that has not been included in the report?

I think everything relevant has been covered.

How are theories described?

As described earlier, very thoroughly and with a lot of background information.

Problem Discussion and Aim

Do the authors manage to attract the reader to the area?

Yes, partly because I also did research within the same area of image classification applied to e-commerce datasets for clothing. The breadth of the research is interesting in how you can combine all possible inputs to form a better classification in terms of labels on items when input is vastly different.

Do they motivate the choice of topic?

Yes, through the client's company business model and what benefits and impact it can have on a company's economic result and resource effectiveness since they can save a lot of time by implementing this system on their existing and future product database.

Is it clear what problem statements the authors had from the beginning?

The problem statement was not explicitly written as a statement but more of list of things to be accomplished in the purpose section:

(1) "assess the relative strengths of different kinds of models and their combinations."
(2) "to determine whether an active labelling strategy reduces label complexity on a real-world dataset."
(3) "obtain a model with powerful predictive capabilities"
(4) "reduce costs by using an efficient labelling strategy"
(5) "obtain a high-quality product similarity score."

Maybe a general problem statement would have been good to keep in the back of the head while reading.

Do you think any important issues have been overlooked?

Maybe remove the interface part listed in the goal section since it was not something Mattias was actually tasked with doing. More so the infrastructure used is very company based and more focus could have been put into the actual classification and experimenting with different outcomes.

Are the goals, objectives, and/or deliverables achievable?

Yes, however the building of the interface is redundant since it was tasked to someone in the client company.

Boundaries

Does the work have reasonable boundaries? Does delimitation connect to the thesis?

Absolutely, covering all models and stating that the search was not to find the most accurate model is clearly stated since this was not the main objective but merely to find a good method to train different models without much labeled data.

Are the boundaries justified?
Yes

Disposition

Does it guide through the material? Is it well-balanced? Does it cover all parts?

I believe the disposition was good and I could feel a guiding theme which we in Swedish call a "red thread" throughout the thesis. I knew where I was going in the next chapter and after reading a section. Looking at the content table I think that it has a good structure. One thing that I will be discussing later is the results section.

Method

Based the problem statements, have the authors made a conscious choice of method? Is the choice of method properly motivated? Has the method been a conscious choice? Are there feasible alternatives to the chosen method? Are the methods described and discussed? Well referenced?

The method is very in-depth and describes many different possible methods and how they chose the methods that they used. The authors clearly did a lot of testing which I think is perfect in terms of robust methodologies of testing and experimenting in new environments.

More references could have been implemented here to add to similar problems faced by others, though I feel that the method mostly consists of own work and experience with the problem tackled by the authors.

Do the authors discuss validity, reliability, replication, or dependability, ethics and sustainability?

Ethics and sustainability was introduced in the introduction with a very well written thoughts and scenarios that opens up the reader to think about possible complications with technology and even possibilities it has for sustainability.

Are all these parts needed for the investigation?

Maybe some information can be removed. Certain technical specifications such as:

"Each of the three datasets is represented by a separate set of tfrecords files; these were produced by Dataflow during the preprocessing stage described in section 3.2.2."

Clearer to state the results were logged during preprocessing and not specify directly where everything comes from. Could lead to easier understanding of the method without the need to know what tfrecords are.

Data collection

How did the author perform the data collection?

Data collection was done by sampling data from the client company's databases with millions of products at random to be able to parallelize the process since they used very large datasets.

Question: Even distribution over classes? some classes could have been under and over represented? Also noted that datasets were not created in advance but different from model test to model test. What are your thoughts on using different data to assess and compare model performance? Shouldn't data be consistent when comparing?

Do the authors discuss the methods (for data collection) conformity with the method?

Yes, in the beginning of chapter 3, data preparations with a clear background information with the different types of input in the background chapter.

Results

Are the goals, objective, and deliverables achieved?

Yes, although Active learning was missing when the report was sent to me.

If testing is used, are test results used as proof (are they convincing)?

Yes, clear results of different iterations of combinations of the earlier described models are shown in figures. Maybe show misclassification examples? Show how it looks when things don't go as planned?

Overall, results chapter feels like a combination of method, results and discussion, was this the intent? I don't feel like I can comment much on structure of this has already been discussed and agreed upon with the supervisor, but I felt a bit confused reading this part since at times it felt like you were discussing the results, presenting them and explained how the results were created all at once.

Get a feeling that a lot of new information is presented all the time which makes it a bit confusing to follow along in the report.

Conclusions and Discussion

Is the purpose, goal mentioned in the Introduction achieved?

In my opinion a good multi-model performance is achieved and thus the problem with inconsistent labeling and unlabeled data is also achieved.

Is the problem solved (validity)?

Yes, I believe so. It would be interesting to hear what the company thought of the final results.

How well the problem is solved (reliability)?

I feel that the study is convulsive but would like to comment on the fact that the data used changed depending on what data the company had received from third-parties, implying that the data changed in between model comparisons which I would consider as a negative. Maybe extract datasets with the same data to do a comparison? What was the reason for not doing this?

Is there any evaluation of the methods that have been used and is the author(s) true to the data?

Yes, the whole of chapter 3.3 mostly resembles evaluation of the results and methods used to attain them.

Are the limitations in the study discussed?

Yes, they are since they want to inform the reader once again of the true intent of the problem statement and research conducted. Highest accuracy measures are not the goal but to see if deep learners can work together with less data and if they outperform individual models at labeling accuracy.

Are there specified proposals for future work?

No, unfortunately not!

Is the outcome of the project evaluated by third party?

Yes I believe so, both industrial and academic supervisor seem to have been present during the project as well as input from other personnel within the client company.

Are there relevant comments and considerations on ethics and sustainability?

Not in the discussion or conclusion but in the introduction which I think were well written.

General final thoughts

I think that the problem faced is as Mattias states not a trivial one. I think they solved the problem with comparing if multi-models would outperform the solo learners. I also think that the technology stack used proves problem solving when actually applying machine learning to real world applications which is very interesting since I think that a big problem for companies today is the “Data-in”, “Data-out” step of the process. Meaning that to effectively and (as shown in this report) economically transfer data to and from the machine learning middle step and do something with the results gained form the learner. I definitely feel as you state that this thesis provides and introduces a framework or general way of thinking when it comes to getting started with GCP and workflow for ML analysis.

References

What are references the authors chose to use? Are the references useful?

The references are diverse and trustworthy. Reports and sources that are referenced are from highly credible sources and still used in correct context to give examples and strengthen their own theories!

Are the references valid? Are the references justified (age, topic and content)?

Since machine learning research is fairly new in terms of the topic discussed in this thesis I would expect the references to fairly new in regard to age. The majority of all references date from 2014-2018 which I personally deem as highly credible and relevant sources. Even though this area is new, and a lot of papers have come out the latest couple of years, some papers are older but still relevant since they are the original papers of key ideas and models which are still used today!

Language and technical performance

Does the report contain typo errors, incorrect sentence structures or other similar defects?

Did not find any spelling errors only some weird sentences. Listed at the end.

Is the report divided into logical pieces?

Yes, although results might be better off separated into an own chapter

Honesty and critical distance

Is it easy for the reader to distinguish what is taken from the literature and other sources and what the authors' own opinions?

Yes. It's easy to notice since if a thought or hypotheses is presented it tries to prove itself in regard to other references and sources. Not to copy them.

Do the authors show a critical distance to the theories and conclusions?

Yes. As explained in an earlier topic the authors show criticism towards the labeling process and that even though a labeled dataset is labeled in a certain way and results depend on this labeling it can in fact be correct. This also adds to their initial problem statement that different affiliate and brands have different views and data that are explained or presented as different. There are a lot of bias going into labeling of products since it's often done by one person and is a result of an individual experience. Say that the person labeling the data is color blind, can we then trust the labels?

Also the fact that they tried many different methods and approaches when comparing the models proves that critical thinking has been present during the full-course of the project not staying naïve to one type of solution but to try out many to find what works best!

General impression

Do the authors contribute something new in the chosen field?

I think so, there are a lot of in-depth and advanced content here which can be used by others who want to progress in the field and build something similar.

Other issues

Is the material coherent? Are the titles of the sections correct? Are the figures (tables) well selected and illustrated? Are the figures/tables described in the text? Do the figures (tables) have a clear purpose and fulfil the purpose?

Figures and tables in the report are selective and I feel that the figures and tables that are presented in the report add a lot of value to the text, in return no figures are redundant. Some figures, for example. the ones from tensorflow are a bit small in the text and can be hard to read sometimes.

Results from tensorflow: Could be that you have to wait for all graphs to render complete since they all don't complete the full steps. I can imagine that the log files must have been huge!

Appendices

List the specific details that should be corrected that do not. (All the things you cannot declare in your oral opposition but are of value for the respondent in order to improve the report)

Text Errors

- "Apache Beam - a data processing engine akin akin to Apache Spark."
- "Computing embeddings in Dataflow is slow, and me way want to enable more flexible ways of restricting…"
- **Data missing -** The train and test set sizes were as follows: rule-based (1.14M / 142K), exclusive (6K, 6.6K), exclusivised
- Figure 3.21 and 3.22 merge In the captions

Final Questions

- Future research, what do you think are reasonable future research questions in regard to extending your work?
- Was the client company happy with the results and are they thinking of implementing the system?
- Do you think you would have, in theory, reduced the costs for the company by implementing this system?