Review of Master Thesis:
Semi-Supervised Learning of Heterogenous Structured Data

Thesis author: Bc. Michaela Mašková
Reviewer: RNDr. Petr Somol, Ph.D.
20 May 2022

There are two orthogonal areas of particular interest in current Machine Learning. First, the modeling of structural and/or hierarchical data - where transformation to vector form is nontrivial or impractical - has been recognised as key in building models of entities in systems like computer networks, or social networks on the internet. Second, while there has been enormous advances in deep learning in recent years, it remains notably harder to build more expressive models than fully supervised discriminative ones. Yet higher model expressivity and/or less dependence on availability of labels is crucial for better understanding of true properties of data, with applications spanning from anomaly detection to identification of inner structure of classes.

This thesis aims at combining both of these areas and taking steps towards more expressive models capable of modeling structural data with incomplete labels. This is a field with little (if any) practical results available yet, but it is a field of great interest in industrial applications. Given my current involvement in Avast Research I can confirm that any improvement in structural modeling would lead to notably improvements in detection capability in cybersecurity, but also in other applied areas. The problem of limited availability of labels is also significant. In cybersecurity, when building training sets representing variations of malware and benign software, obtaining labels can be a costly and tedious process. In industry it is now of utmost importance to find ways of learning from data with limited or no label information.

In the first section of the thesis the author reviews a selection of known techniques that would be used in modified form later in the thesis to solve the problem of learning from hierarchical structural data. All of the existing techniques are applicable to vector spaces, none is trivially usable to solve the problem at the core of this thesis. Their modifications and interplay, however, will be considered in later chapters. From the discussed Semi-Supervised methods I confirm that Self-Supervised Learning in particular is gaining traction in industry. Metric learning has also shown a lot of promise lately, with Triplet Loss being reported from multiple domains as significant. In the area of clustering the problem of evaluation metrics is discussed, which I find important as well. Mapping methods and latent space representation is discussed as well.

The second section covers the so-called HMill framework proposed by Pevny et al, as a practical form of neural networks suitable for modeling structural hierarchical data of mixed types. The framework can be seen as less general than Graph Neural Networks, but easier and more efficient in use for many industrial applications given its ability to directly learn from generic JSON type of data.

The third section is the core of the thesis in terms of novelty. The author proposes a number of models enabling semi-supervised learning from hierarchical structural data. These models are applicable under limiting conditions (e.g. the supported depth of hierarchy is limited). For the purpose of comparing and evaluating a number of novel concepts the selection of methods is good, and is more than sufficient for a MSc thesis. The list of considered models covers a variety of approaches to the problem: self-supervised variant of HMill, for comparison also its ArcFace based variant, its variant with Metric learning and clustering (in input versus latent space), a variational auto-encoder based generative model.

The fourth section defines the experimental setup, taking use of three various data sets of different types. It explains in more detail the interplay between models as feature extractors and subsequent use of clustering on the latent space. Section five summarises key achieved results in tables and graphs.

The key observations cover the non-trivial observation, that metric learning with standard HMill model on reduced number of labels tends to give more consistent and competitive results than self-learning and generative models. That said, some of the reasons are discussed, and some of the generative model results appear to show the way forward towards better exploiting their potential. As part of presenting the results the thesis gives useful illustration of the various aspects that may affect the success of learning. All results are presented in scientifically sound and fair way.

The thesis overall presents sound scientific non-trivial results and I consider it a success for that reason. That said, it is not without issues. Overall, the structural clarity of writing can be improved. At some moments it takes effort for the reader to recognise whether a different term is used to describe the same thing as in previous section. The overall storyline takes effort to recognise, as the thesis is more a collection of individually written focused passages, than a clearly understandable coherent study that would navigate the user for easier understanding. Some of the questions or comments below that come to mind are positively motivated by inspiring read, others are the result of the hard-to-read text:

- Chapter 1 jumps to the topic quick; it would be easier for readers to be a bit more verbose about the definition of the problem being solved (classification through learned model on labeled data)
- it would help to get some lead towards discussion of what is or is not possible if assumptions in page 11 do not hold (which may be the case with realistic data)
- example of a hard to read explanation : paragraph discussing augmented data in page 16.
- in page 18 the additive angular margin penalty m is also a hyper parameter? As there is a number of hyper parameters that may affect the results of methods like ArcFace, it would be good to include discussion in the results section about how robust the results are w.r.t. hyper parameters
- does “self-supervised model” in page 35 mean the same as “self-learning” in page 12?
- In section 3.1 the description of the self-supervised HMIll variant it would help to have less ambiguity in method description… does the algorithm need to start from some subset of labeled data?
- In section 4.3 a projection to 2D is used. I thought this would be to allow graphical illustration but there are no such illustrations included. Is there a different reason to not use a different dimensionality of the embedding space? Actually, illustrations showing the projection of latent spaces would be nice to see, possibly with some of the labels included to visually illustrate achieved clustering quality
- Also in section 4.3 - all considered clustering algorithms seem to require explicit setting of the number of clusters. It would be nice to consider also algorithms capable of estimating the number of clusters (for practical reasons in applied setting)
- When reading about results in Chapter 5, it is not explicitly clear how exactly “classifier” and “classifier+triplet” use the training data. I assume they are trained in fully supervised manner on the respective downsampled training data, but I better ask. This is a crucial setting, essential for all further discussions in text, and would deserve better introduction in text
- In page 47 “The results are on par with the classification accuracy, sometimes the clustering accuracy achieves slightly better results.” Feels confusing. The reader has to think twice to realise which experiments are which and to what results the sentence applies.
- In last paragraph in page 48 you suggest that “One of the solutions that could help would be to make the generative model more dependent on a given label and remove the instance latent space….”. The suggestion sounds interesting and would deserve further details to be discussed.
- The bottom paragraph in page 49 inspires me to ask a fundamental question: In self-supervised learning, the initial iterations (when the model is not yet precise!) seem to be crucial to determine the ultimate result. Isn’t this a dangerous robustness issue with self-supervised learning as studied here?
- in tables 5.4 and 5.6, do I understand it right that “self-classifer” means the same thing as “self-supervised classifier”? Same with “self-ArcFace” and “self-supervised ArcFace”
- regarding the ArcFace limited success - can it be that the random search of hyper parameters was insufficient to find the best s and m parameters? I should note that similar questions would deserve attention in the text in other contexts as well.

- in section 5.3 on JSON dataset you note that “”The self-supervised models seem to have an edge over the classifier and classifier with triplet regularisation”. This is encouraging. More details would be helpful to get about the exact setup of the experiment - e.g., which parts of the JSON schema have been used, as features or as lablels, has there been any augmentation and if so, how was it defined.
- looking at Fig 5.4, I realise that it would be beneficial to see also the comparison of training times. Is not the SSL accuracy edge paid for by a significantly higher complexity of training?
- page 55 last row in the bottom: you note that k has been chosen “mainly 1-5”. This is confusing. So the kNN results are consistent for all k=1,..,5 ?
- section 5.4 summary - the first paragraph answering the first question does not seem to reflect the results of SSL on JSON?. Third paragraph - “It shows that the learned embedding space is not discriminative enough….” Feels confusing. Is it meant to claim that this is true for both discriminative and generative?
- in page 57 “The model is designed to be a collection of generated submodes for leafs in the JSON structure and joined with a classifier to create a hierarchical version of the semi-supervised M2 model”. Just to better understand the model limitations - am I right that the possible variability in the number of leafs is not generatively modelled? (I see the constant of 18 models).
- in page 57 “”Unfortunately, the M2 generative semi-supervised model fell short of expectations. …. The problem might be that the generative model learns an instance embedding and can reconstruct the data well even when ignoring the label provided with them”. Good observations, I tend to agree. Just for interest - can you comment on the possibility of utilising GANs in this context? I also agree on next page with “Further research should focus on designing a better generative model for data, as well as extension to multi-level tree-structured data…”
- there is a good selection of cited works in Bibliography. For inspiration I would recommend to consider one more angle to the problem – the body of work in graph neural networks, that can possibly provide inspiration for further work on generative models in relation to structural data. Note that there have been advances in graph generation, cf. Guo, Xiaojie, and Liang Zhao. "A systematic survey on deep generative models for graph generation." arXiv preprint arXiv:2007.06686 (2020).

Conclusion

The author shows great promise to continue improving as a talented scientist. In next iterations of author’s publishing efforts I recommend to consider “wearing the uninformed reader’s hat” more thoroughly, to provide easier to follow text. Throughout the thesis there is frequently too much assumed to be clear for the reader while it is not. More explicit description of individual methods and their setup would help to make the thesis readable for a wider audience than a small circle of readers versed in this particular area. That said, I am sure we can expect inspired research from the author in the future and I encourage the author to continue with a scientific career.

I recommend to accept this thesis as a Master thesis and the author to obtain the respective engineering title. I suggest to award the thesis **with mark A**.

RNDr. Petr Somol, Ph.D.
petr.somol@gmail.com (preferred), tel 603 719 429

Research Fellow
Institute of Information Theory and Automation
Czech Academy of Sciences
Pod vodárenskou věží 4
18208 Praha 8
somol@utia.cas.cz

AI Research Director
Avast Software
Pikrtova 1737/1A
140 00 Praha 4-Nusle

petr.somol@avast.com