On the importance of deep learning regularization techniques in knowledge discovery

Ljubinka Sandjakoska ljubinka.gjergjeska@uist.edu.mk
Ss Cyril and Methodius University,
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
UIST St Paul the Apostle,
Faculty of Computer Science and Engineering
Ohrid, Republic of Macedonia

Atanas Hristov atanas.hristov@ uist.edu.mk
UIST St Paul the Apostle,
Faculty of Information and Communication Sciences
Ohrid, Republic of Macedonia

Ana Madevska Bogdanova ana.madevska.bogdanova@finki.ukim.mk
Ss Cyril and Methodius University,
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia

Abstract

Nowadays, in the era of complex data, the knowledge discovery process became one of the key challenges in the science. The evolution of the technologies imply evolution of the techniques for dealing with the data. Deep neural networks, as advanced concepts became very popular and can be viewed as tool for improvement of knowledge discovery processes. A motivation for this paper is generalization ability of deep neural network. In an attempt to better understand and to solve the problem of generalization of deep neural networks, we study several regularization techniques. Different regularization techniques, as a solution of overfitting problem, are discussed. The impact of regularization on knowledge discovery process is in the focus of this paper. In order to illustrate the effect of regularization in knowledge discovery, a case study is presented. The case study refers to discovering unknown relationships between molecules in atomic simulation. We propose a dropout method for regularization deep neural network for molecular dynamics simulations. In this paper we show that discovering high level concepts in data, during knowledge discovery, is possible with efficient training of regularized deep neural networks.

Keywords: knowledge discovery, deep learning, deep neural networks, overfitting, generalization, regularization

1. Introduction

Nowadays, in the era of complex data, the knowledge discovery (KD) process became one of the key challenges in the science. The challenges arise from the key data features: volume, velocity, variety, and veracity, which determine the dynamics of its relationships. The relationships between the data imply the growth of the field of knowledge engineering. The KD concepts are not so new, but there are new concepts in abundance that rely on the evolution of the technologies, which imply evolution of the techniques for dealing with the data. In that direction, this paper try to give different view of knowledge discovery process and aims to depict some important problems from the proposed view.
In general, knowledge discovery can be defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (William et al., 1992). Although this definition refers to KD in databases, it can be extended. Without going in details, we would like to note that in this paper we are not focused on databases but we will discuss the KD process in the context of deep learning. Deep learning, realized in advanced artificial neural networks, enables solving non-linear problems of complex systems in real times. Also, deep neural networks (DNN) is successful in dealing with the newest data issues that arise from amount and data heterogeneity (Sandjakoska and Bogdanova., 2018). A specific contribution in process of knowledge discovery, or more precisely in extracting relevant information, give the DNNs ability of learning several levels of representations, corresponding to a hierarchy of features, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts (Deng and Yu, 2014). In addition it is interesting to be mention the DNNs adaptability of modeling abstraction over multiple levels as important for KD and automated feature engineering also. DNNs, as excellent fitting tools with central task of finding a function that can well approximate a mapping from inputs to desired outputs, is essential building block of predictive models. It is proven that, the predictive models are not valid if KD is excluded. All of these advantages, make DNNs specific tool for KD. DNNs prove that are specific and effective tool for KD in various domains with different data types, such as image (Iizuka et al., 2016; Dahl et al., 2017; Cao et al., 2017; Karpathy and Fei-Fei, 2015; Ravanbakhsh et al., 2016) and video processing (Yue-Hei Ng et al., 2015; Morfi and Stowell, 2018), natural language processing (Collobert and Weston, 2008; Jaderberg et al., 2014; Ray et al., 2018), time-series forecasting (Kuremotoa et al., 2014; Längkvist et al., 2014). Specific and interesting application in obtaining knowledge using DNNs are: automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery (Zhang et al., 2018); road extraction from high-resolution remote sensing imagery (Xu et al., 2018); or targeted grassland monitoring at parcel level using sentinels, street-level images and field observations (d'Andrimont et al., 2018) etc.

The paper is organized as follows. In the second section different regularization techniques, as a solution of overfitting problem, are elaborated. Next is the discussion on the impact of regularization on knowledge discovery process. Before concluding remarks, in the fourth section a case study for discovering unknown relationships between molecules in atomic simulation is included.

2. Theory

Flexibility, as important advantage of deep learning, is based on ability to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones (Goodfellow et al., 2016). Achieving flexibility is one of the key challenging issues of DNN and it is not so easy to be realized. The challenge arise since the phenomenon of overfitting that occur when the network model works well on training data but not on test data. That mean, the predictive model is not able

to deal with previously unseen data and to give relevant knowledge for the system or/and process behaviour. Over-parametrization or obtaining knowledge for the noise and specific relationships in the training data set decrease the efficiency and increase the DNNs proneness to overfitting. In order to be solved the problem of overfitting, regularization strategies are used. The idea behind the regularization is to find how to modify the learning algorithm in procedure for obtaining knowledge, in a way that the test error will be decreased or generalization error but not its training error. It is well-known that useful knowledge cannot be obtained from the data without generalization ability of the model, especially in domain specific and case sensitive applications.

A) Regularization techniques

First approach of categorization results with seven groups of regularization techniques: *parameter norm penalties*; *norm penalties as constrained optimization*; *dataset augmentation*; *injecting noise; early stopping; parameter tying and parameter sharing; bagging and other ensemble methods* (Goodfellow et al., 2016). All of this techniques refer to knowledge discovery process.

The second approach considered different properties of the learning process actually different features of the knowledge discovery and result with five groups of methods: methods that affect *data* (generic data-based methods and domain-specific data-based methods), methods that affect the *network architectures*, *error terms regularization terms*, and *optimization procedures*.

DNN defined as function $f_w: x \rightarrow y$ with changeable weights $w \in W$, should be trained and find a configuration w^*. The weight configuration should be a result of minimization procedure $w^* = \text{minimize } \mathcal{L}(w)$ of a loss function $\mathcal{L}: W \rightarrow \mathbb{R}$, defined as:

$$\mathcal{L} = E_{(x,t)\sim P}[E(f_w(x), t) + R(\dots)]$$

The above representation of the loss function is known as *expected risk*. It is composed of two generic terms: an *error function E* and a *regularization term R.* The error function is formed by consistency with the targets influence to the penalty assigned to model predictions. The regularization term do not depend on targets. It depends on some other criteria which can also be assigned on a penalty to the model. There are several different criteria approaches. Every criteria approach is included in the designing of the strategy *how the regularization will be done.* Since the data distribution *P* is unknown, the expected risk cannot be minimized directly, hence the training set *D* sampled form the distribution is given. This approach helps in minimizing of the expected risk by minimization of the *empirical risk* $\hat{\mathcal{L}}$:

$$\boldsymbol{minimize}_w \frac{1}{|D|} \sum_{(x_i,t_i)\in D} E(f_w(x_i), t_i) + R(\dots)$$

In this type of minimization of the empirical risk, the elements that have direct influence to the value of the learned weights that contributes to the regularization, needs to be identified.

B) Impact of regularization on knowledge discovery process

The KD process should result with valid, novel and ultimately understandable patterns in data. That is achievable if knowledge obtained by DNN allows predictive modelling on new previously unseen data. The quality of the predictive modelling depends on the ability of DNN's generalization. Good generalization results with accurate, consistent, and complete data and is

allowed if and only if the DNN is well regularized. It is obvious the high impact of regularization on KD process but it cannot be easy measured. In this context it is interesting to answer the question*: what could happen with the knowledge if the DNN is not regularized?* This question has two aspects of answering. In addition, we will make distinction between two types of knowledge, which form the aspects of answering the question mentioned previously: *i) knowledge **in** the "black-box" of DNN* and *ii) knowledge **out** the DNN.* The knowledge in the black box of DNN is result of the learning process, allowed by activation function. Before obtaining knowledge form learning process, performing data pre-processing operations is required. In this context, data pre-processing can be viewed as supporting tool for regularization. For other side, regularization is supporting tool to post-processing operations that influence to the knowledge quality. Usually regularization techniques are realized in the "black-box" but deeply affect the knowledge out the DNN.

Finally, if the DNN is not regularized than the knowledge would not be dynamic, but static, since the model will be trained with the data specific relations including noise, redundancy or predefined data patterns that are domain specific and other which decrease the training error but increase the generalization error. Also, without regularization the knowledge would not be sustainable. Here should be mention that regularization improves the non-depleting state of the knowledge and its acquisition.

3. Results

A) Case study

In order to illustrate the effect of regularization in knowledge discovery, we present a case study. The case study refers to discovering unknown relationships between molecules in atomic simulation. In atomic simulations, the dependencies between the molecular dynamics descriptors are too complex and we need technique in the predictive model that will prevent the pursuit of hard probabilities without discouraging correct classification. The classification depend on the effectiveness of obtaining and using the knowledge. First depend on the raw knowledge consisted in the dataset, second depend on the knowledge obtained during the learning process. The dropout regularization method affect the learning process and the knowledge that depend on the excluding some of the nodes. The key is in the excluding the nodes in each iteration and the assembling principle applied to the networks with different configuration.

For that purpose we propose computationally cheap and efficient dropout method for regularization, implemented in molecular energy prediction. We evaluate proposed approach with benchmark of quantum-machine dataset. (http://quantum-machine.org/datasets/). The experiments are conducted using Keras – the Python deep learning library.

The main difference between the proposed approach and the standard dropout method is in assigning not a constant probability of omitting hidden units in the training. In general, the hidden units are divided according to the group of atomic descriptors. Since the network performance highly depends from the contribution of each group of the hidden units, different probability for each group have been assigned. The molecular energy rang is crucial parameter that determinates the probability of dropping out unit. In order to achieve better performance, a matrix that includes information of molecular rang energy have been used as mask. This mask has been set as a random seed in order to avoid the randomness. Another key difference is related to the behaviour of hidden nodes. The hidden nodes are organized in feature maps. In order to extract same specific feature, all units in feature maps share the same set of weights and performs the same operation over all different parts of the input. Even that there is a correlation between the feature maps, the strong correlation between the adjacent units is avoided. Each feature map is associated to a group of molecular descriptor. Feature detectors are created by the different probability p_{di}. The index *i* of the probability refer to a group of descriptors.

This data set consists of molecular dynamics trajectories of 113 randomly selected $C_7O_2H_{10}$ isomers (20). First the encoding of the molecules is done in order to be eligible for the input of the neural network. The input vector includes molecular geometries as *xyz* trajectories, and energies valence densities, additional - consistent energy calculations of all isomers in equilibrium are included. All trajectories are calculated at a temperature of 500 K and a resolution of 0.5 fs. The molecules have different sizes and the molecular potential energy surface exhibit different levels of complexity. In order to avoid problem of data incompleteness standard preprocessing techniques are performed.

B) Summary of the results from proposed approach *dropAD* (Tab.1)

Table 1. MEAN ABSOLUTE ERRORS (IN MEV) FOR ENERGY PREDICTION

	Benzen	***Saliylic acid***	***Malonaldehyde***	***Toluene***
DTNN (20)	1.7	21.7	8.2	7.8
dropAD	1.5	19.8	8.1	6.4

4. Conclusion

In the case study, we showed that using proposed novel dropout method improves the state-of-the-art of applied deep neural networks in chemical computations on the benchmark dataset. Also, discovering high level concepts in data, during knowledge discovery, is possible with efficient training of regularized deep neural networks.

Acknowledgments

This work is partially supported by EU under the COST Program Action TD1403: Big Data Era in Sky and Earth Observation (BIG-SKY-EARTH).

References

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, Knowledge Discovery in Databases: Overview, AI Magazine Volume 13 Number 3 (1992) (© AAAI), pp: 57-70

Lj. Sandjakoska and A. M. Bogdanova, Deep Learning: The future of chemoinformatics and drug development, 15th International Conference on Informatics and Information Technologies, CIIT, 2018

L. Deng and D. Yu, “Deep learning: methods and applications”, Microsoft Research, Now publishers, 2014.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa."Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification".ACM Transaction on Graphics (Proc. of SIGGRAPH), 35(4):110, 2016

Ryan Dahl, Mohammad Norouzi, Jonathon Shlens, Pixel Recursive Super Resolution, 2017, Google Brain, https://arxiv.org/pdf/1702.00783.pdf

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017, https://arxiv.org/abs/1611.08050

Andrej Karpathy Li Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015,

Siamak Ravanbakhsh, Francois Lanusse, Rachel Mandelbaum, Jeff Schneider, Barnabas Poczos, Enabling Dark Energy Science with Deep Generative Models of Galaxy Images, 2016 https://arxiv.org/abs/1609.05796

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici: "Beyond Short Snippets: Deep Networks for Video Classification", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4694-4702

Morfi, V.; Stowell, D. Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets. *Appl. Sci.* **2018**, *8*, 1397.

Ronan Collobert, Jason Weston,"A unified architecture for natural language processing: deep neural networks with multitask learning",Proceeding ICML '08 Proceedings of the 25th international conference on Machine learning,Pages 160-167, doi: 10.1145/1390156.1390177

Jaderberg, M, Vedaldi, A. and Zisserman, A., Deep Features for Text Spotting, European Conference on Computer Vision, 2014.

Takashi Kuremotoa, Shinsuke Kimuraa, Kunikazu Kobayashib, Masanao Obayashia, "Time series forecasting using a deep belief network with restricted Boltzmann machines" Neurocomputing,Volume 137, 5 August 2014, Pages 47-56

Zhang, W.; Witharana, C.; Liljedahl, A.K.; Kanevskiy, M. Deep Convolutional Neural Networks for Automated Characterization of Arctic Ice-Wedge Polygons in Very High Spatial Resolution Aerial Imagery. *Remote Sens.***2018**, *10*, 1487.

Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461.

d'Andrimont, R.; Lemoine, G.; van der Velde, M. Targeted Grassland Monitoring at Parcel Level Using Sentinels, Street-Level Images and Field Observations. *Remote Sens.* **2018**, *10*, 1300.

Martin Längkvist, Lars Karlsson, Amy Loutfi, A review of unsupervised feature learning and deep learning for time-series modelling, Pattern Recognition Letters Volume 42, 1 June 2014, Pages 11-24.

Ray, J.; Johnny, O.; Trovati, M.; Sotiriadis, S.; Bessis, N. The Rise of Big Data Science: A Survey of Techniques, Methods and Approaches in the Field of Natural Language Processing and Network Theory. *Big Data Cogn. Comput.* **2018**, *2*, 22.

I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning", MIT Press, 2016. [Online]. http://www.deeplearningbook.org

(20) http://quantum-machine.org/datasets/