On the importance of deep learning regularization techniques in knowledge discovery

Ljubinka Sandjakoska Atanas Hristov Ana Madevska Bogdanova

BigSkyEarth conference: AstroGeoInformatics, Tenerife, Spain,
December 17-19, 2018

Output

- Introduction
- Theory
  - Regularization techniques
  - Impact of regularization on knowledge discovery process
- Results
  - Case study
- Conclusion

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**.

Introduction

- Although the standard definition refers to KD in databases it can be extended.
- 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**
- A specific contribution in process of KD, 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.

Introduction

- 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 and video processing, natural language processing, time-series forecasting.
- Specific and interesting application in obtaining knowledge using DNNs are: automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery; road extraction from high-resolution remote sensing imagery; or targeted grassland monitoring at parcel level using sentinels, street-level images and field observations

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.**

- 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*** and ***optimization procedures***

Regularization

- 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 of a loss function , defined as:

$$\boldsymbol{w^* = minimize\ \mathcal{L}(w)} \text{ of a loss function } \boldsymbol{\mathcal{L}: W \rightarrow \mathbb{R}} \text{ defined as:}$$
$$\boldsymbol{\mathcal{L} = E_{(x,t)\sim P}[E(f_w(x), t + R(\dots)]}$$

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 $\hat{\mathcal{L}}$

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

Impact of regularization on KD 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.**
- 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.

Impact of regularization on KD process (2)

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: 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.
- For other side, regularization is supporting tool to post-processing operations that influence to the knowledge quality out of DNN.
- ***Usually regularization techniques are realized in the "black-box" but deeply affect the knowledge out the DNN***.

Impact of regularization on KD process (3)

- 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***.

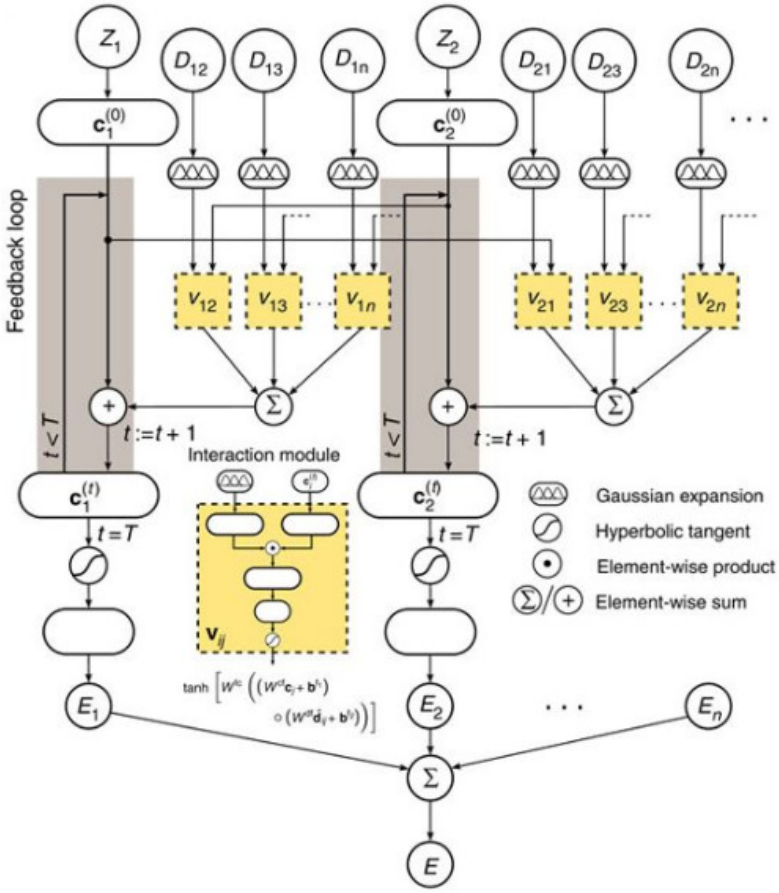
Results - Case study

The case study refers to discovering unknown relationships between molecules in atomic simulation. (Knowledge discovery in atomic simulations)

- 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.

Case study

- We propose computationally cheap and efficient **dropout based method for regularization**, implemented in molecular energy prediction.
- We evaluate proposed approach with benchmark of quantum-machine dataset.
- The experiments are conducted using Keras – the Python deep learning library, because it is minimalistic, modular, and awesome for rapid experimentation.



Proposed approach

- The proposed approach differs from standard dropout in assigning not a constant probability of omitting hidden units in the training.

- The hidden units are divided according to the group of atomic descriptors.

- Since, each group of hidden units has different contribution to the network performance, different probability for each group is assigned.

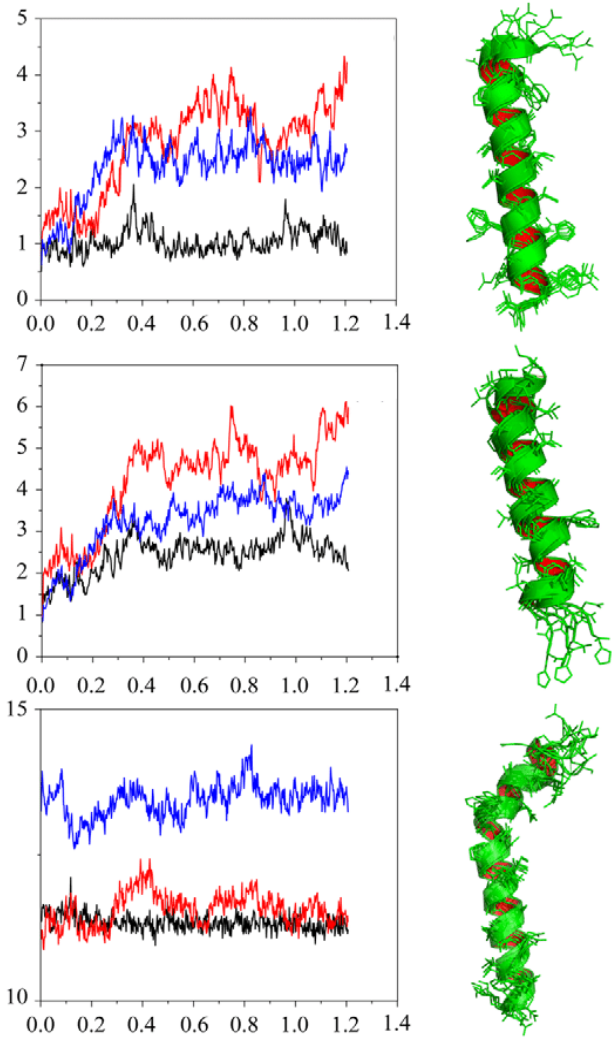
- The need of different probability is implied by possible losing information for specific relationship between the descriptors, if we exclude unit with a high activation value. Usually high activation value indicates important feature.

Proposed approach

- The probability of dropping out units depend on the molecular energy rang.
- Another difference is related to the behavior of hidden nodes that are organized in feature maps.
- In feature maps all units share the same set of weights in order to extract a certain feature by performing the same operation on all different parts of the input.
- The strong correlation between the adjacent units is avoided although there is exist a correlation between the feature maps.

Data

- This data set consists of molecular dynamics trajectories of 113 randomly selected $C_7O_2H_{10}$ isomers.
- 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.



Results

	Benzen	Saliylic acid	Malonaldehyde	Toluene
DTNN	1.7	21.7	8.2	7.8
dropAD	1.5	19.8	8.1	6.4

- DNN with the proposed dropout approach (*dropAD*) offer a higher accuracy performance.
- The good accuracy performance is due to including the relationship between the descriptors, forming groups of it.
- We can see that the mean absolute error for energy prediction is decreased using *dropAD* method.
- The difference between the compared models is in the range from 0.1 to 1.9. Even this small improvements has big contribution especially in molecular dynamics simulations.

Conclusion

- 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).