Activation functions' impact on regularization of deep neural networks application in atomic simulations

Ljubinka Sandjakoska
Faculty of Computer Scence and Engineering
UIST St Paul the Apostle
Ohrid, Republic of North Macedonia
ljubinka.gjergjeska@uist.edu.mk

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

***Abstract*—When it comes to atomic simulations, the regularization of the deep neural networks is key to its successful application. The generalization capability of deep network depends on some factors. This paper aims to show that activation function is one of the most important factors that influence to decreasing the generalization error. For that purpose, several experiments were performed. Moreover, new approach for choosing the activation function is proposed. The purpose of the activation mechanism is not to find a universal, new activation function, although this is not excluded, but the most appropriate for the given task and for the given data set. The obtained results show that using the proposed activation approach, a decreasing of the mean absolute error compared to the benchmark set is achieved.**

Keywords—activation, regularization, deep neural networks, atomic simulations

I. Introduction

Nowadays, developing advanced machine learning based approaches in many domains is more than a trend. Deep learning, as sophisticated concept of machine learning, allows solving the most specific issues related to the data. Those issues in the cheminformatics are striking and the automatic construction of complex features is crucial [1]. Due to the ability of deep learning to provide a hierarchical representation of a compound, where higher levels present more complex concepts [2], its application in atomic simulations give advantages that could not be achieved otherwise. That is noticeable in tasks such as: prediction of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties [3], prediction of molecular activity, toxicity [4], quantitative structure activity relationship (QSAR) predictive modeling [5], prediction of the drug-target interaction [6], virtual screening [7] etc.

When it comes to atomic simulations, the regularization of the deep neural networks is key to its successful application. Regularized deep neural network is capable to predict on previously unseen data with a small error. Usually regularization is viewed as modification to the learning algorithm to reduce its generalization error but not its training error [8]. In [9] more general definition is given. The authors in [9] include various properties of the loss function, the loss optimization algorithm, or other techniques and define the regularization as any supplementary technique that aims at making the model generalize better. In this research, we also follow that broader definition.

The regularization methods can be classified into several groups: *methods that affect data* (generic data-based methods [10] - [14], [15], [16] and domain-specific data-based methods [17], [18], [19]), *methods that affect the network architectures* [20], [21], [22], [23], *error terms* [24], [25], *regularization terms* [23], [27], and *optimization procedures*. The generalization capability of deep network depends on some factors. This paper aims to show that activation function is one of the most important factors that influence to decreasing the generalization gap defined as difference between the models' performance on training data and on new data.

The paper is organized as follows: in the second section the deep neural network architecture and the mechanism for the activation are presented. The third section includes the details of experimental design: description of the data sets, evaluation measures and the experimental setup. Results of the experiments followed by discussion are provided in the fourth section. At the end of this paper the concluding remarks are given.

II. Methods

A. Deep Tensor Neural Network (DTNN) architecture

The deep architecture which deploy the proposed mechanism for choosing activation function is based on the architecture in [28]. The reference DTNN use *tanh (hyperbolic tangent)* activation function, which has S-shape and take real values as inputs and outputs in range -1 to 1. For consistency, we use data division: 80/10/10, training/test/validation. The prediction of the atomic properties using DTNN (Fig.1) is realized by: computing features, generating, embedding, feedforwarding, gathering and backpropagation. Since the nature of the atomic data, there is a need for special procedures to make the data suitable for manipulation. Those procedures dos not include standard preprocessing such as denosing; elimination of redundancy; reduction; identifying and dealing with missing data; standardization; solving the problem with multicollinearity of data etc., but *featurization,* and *splitting.* Featurization is procedure for effective encoding of the molecules in the string with fix length or vectors.

The Deep Tensor Neural Network (Fig.1) does not include explicit binding information, the atomic features are updated using all other atoms based on their physical distances. In Table 1 details for the parameters of DTNN are given.

B. Activation mechanism

The proposed mechanism chose the activation function $f_A(x)$, which can have three general forms. The occurrence of each form is equally probable. The forms are:

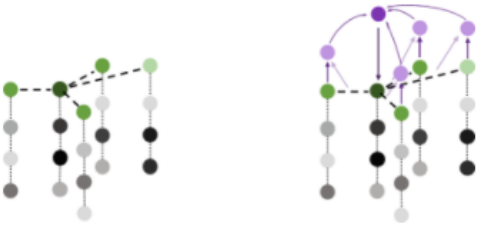

Figure 1. DTNN atomic model [29]

$$f_A(x) = U_i(x),$$
$$f_A(x) = U_i(U_{i+j}(x)) \text{ and}$$
$$f_A(x) = B(U_i(x), U_{i+j}(x))$$

where $U_i(x)$ is unary function, while $B(U_i(x), U_{i+j}(x))$ is binary function where $i \in (1,15)$ and $j \in (1,14)$.

TABLE I DTNN ARCHITECURE DETAILS

DTNN_model
n_tasks (int)
n_embedding (int, optional)
n_hidden (int, optional)
n_distance (int, optional)
distance_min (float, optional)
distance_max (float, optional)
mode (str)
compute_features_on_batch(X_b)
default_generator
dataset (Dataset)
epochs (int)
mode (str)
deterministic (bool)
pad_batches (bool)
DTNNEmbedding
n_embedding: int, optional
periodic_table_length: int, optional
init: str, optional
DTNNStep
n_embedding: int, optional
n_distance: int, optional
n_hidden: int, optional
init: str, optional
activation: str, optional
DTNNGather
n_embedding: int, optional
n_outputs: int, optional
layer_sizes: list of int, optional(default=[1000])
init: str, optional
activation: str, optional

The values of i and j are determined by the number of unary functions that are used. The set of unary functions used in this research contains the following functions $0, 1,\ x, -x,\ |x|,\ x^{-1},\ x^2,\ x^{\frac{1}{2}},\ e^x,\ e^x - 1, log(e^x + 1),\ x \cdot \sigma(x)$ (known as Swish), where $\sigma(x) = (1 + e^{(-x)})^{-1}$ is sigmoid function, $x \cdot \tanh(\ln(1 + e^x))$ (known as Mish) and $\frac{x}{\frac{x}{\alpha}+e^{-\frac{x}{\beta}}}, \alpha, \beta > 0$ (as Soft-Root-Sign). The set of binary function include the following functions: $U_1 + U_2, U_1 - U_2,\ U_1 \cdot U_2, \frac{U_1}{U_2}, U_1{}^{U_2}, \max(U_1, U_2), \min(U_1, U_2)$.

First of all, the search space is formed by the specified unary and binary functions as the sum of the permutations for all three general forms (Fig.1) in which the activation function occurs. The activation mechanism automatically selects one of the possible activation functions from the search space, formed

a) For the first general form of the activation function, defined with $f_A(x) = U_i(x)$, we have to choose one from 15 unary functions. Or permutations $P(15,1) = \frac{15!}{(15-1)!} = 15$;

b) For the second general form, defined with $f_A(x) = U_i(U_{i+j}(x))$, where $i \neq j$, we chose 2 from 15 unary functions, or $P(15,2) = \frac{15!}{(15-2)!} = \frac{15 \cdot 14 \cdot 13!}{13!} = 15 \cdot 14$ functions;

c) For the third general form, $f_A(x) = B(U_i(x), U_{i+j}(x))$, we use 15 unary и 7 binary functions and obtain $7 \cdot \frac{1}{15} P(15,15) = 7 \cdot \frac{1}{15} \frac{15!}{(15-15)!} = 7 \cdot \frac{1}{15} \frac{15 \cdot 14!}{0!} = 7 \cdot 14!$ functions.

The total number of candidate functions is $15 + 15 \cdot 14 + 7 \cdot 14!$ i.e. 610 248 038 625. Due to computer resources, the number of used unary and binary functions is limited. The large number of possible activation functions further increases the complexity and causes difficulties in the activation mechanism. It should be emphasized that the purpose of the activation mechanism is not to find a universal, new activation function, although this is not excluded, but the most appropriate for the given task and the given data set. The activation mechanism solves a classic optimization problem, with iterative search - attempts to find the best of the candidates solutions according to the selected criterion.

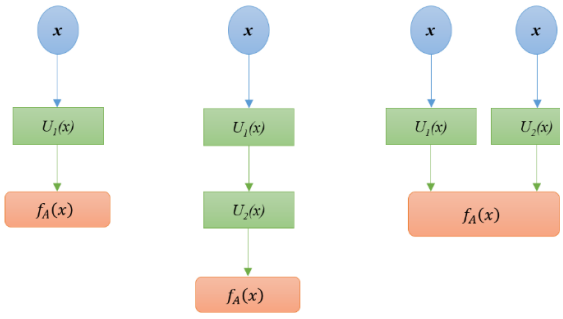

Figure 2. Combining functions
a) First general form b) Second general form c) Third general form

III. EXPERIMENTAL DESIGN

We use DeepChem[1] package, which is result of project that aims to create high quality, open source tools for drug

[1] https://deepchem.io/

discovery, materials science, quantum chemistry, and biology. In the implementation are used Scikit-learn, TensorFlow [30] and Keras [31]. For experiment tracking wandb.ai tool is deployed. In order to show the influence of choosing the activation function, well established activation functions are applied in different experiment runs.

A. Datasets

For the purpose of the research presented here we use four datasets form MoleculeNet[2] aggregator [29]: QM7, QM7b, QM8 and QM9. The datasets belong to quantum mechanics category for solving regression task.

1) QM7 and QM7b data set

The data sets contain 3D Cartesian coordinates and electronic properties of 7165 molecules. The main goal is to predict the following electronic properties:

- Atomization energy - PBE0
- Excitation energy of maximal optimal absorption - ZINDO
- Highest absorption - ZINDO
- HOMO - ZINDO
- LUMO - ZINDO
- 1st excitation energy - ZINDO
- Ionization potential - ZINDO
- Electron Affinity - ZINDO
- HOMO - KS
- LUMO - KS
- HOMO - GW
- LUMO - GW
- Polarizability - PBE0
- Polarizability - SCS

2) QM8

This data set is consisted of low-lying singlet-singlet vertical electronic spectra of over 20 000 synthetically feasible small organic molecules. QM8 data are used for prediction of electronic spectra in a high-throughput manner across chemical space:

- E1 - CC2
- E2 - CC2
- f1 - CC2
- f2 - CC2
- E1 - PBE0
- E2 - PBE0
- f1 - PBE0
- f2 - PBE0
- E1 - CAM
- E2 - CAM
- f1 - CAM
- f2 – CAM

3) QM9

QM9 data include geometric properties and energetic, electronic and thermodynamic properties in order to predict atomic properties using geometric properties (atomic coordinates) which are integrated into features

- mu
- alpha
- HOMO
- LUMO
- gap
- R2
- ZPVE
- U0
- U
- H
- G
- Cv

B. Evaluation measures

The proposed approach is evaluated using Mean Absolute Error (MAE), as a golden standard for regression task. The MAE is calculated using the formula $MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$, where $\hat{y}_i$ denotes the predicted value for a single data using the model, while the actual value from the training set is denoted by y_i. Thus, the term *error* depict the difference between the predicted value for a variable and its actual (observed, measured) value.

C. Experimental setup

The optimization of the hyper parameters is done by Gaussian Process Optimization in maximum 20 iteration. The training of the models is limited and is not longer than 10 hours.

IV. Results and Discussion

In order to depict the effectiveness of proposed mechanism for activation of the neurons and the impact on regularization in general, several experiments were obtained. The first experiment includes prediction of the atomic properties using 3 architectures: ANI-1 [32] - Extensible neural network, DTNN [28], MPNN [33] - Message Passing Neural Network. All models are graph-based, as a more convenient for modeling molecules, since the fact that the molecules and their features are often represented by graphs. The results are given in Table II and Fig.2.

TABLE II PERFORMANCE OF THE REFERENCE VS PROPOSED APPROACH

Dataset	Best performance	MAE kcal/mol [29]	MAE kcal/mol [proposed approach]
QM7	ANI-1	2.8600	**0.80960**
QM7b	DTNN	1.7700	**0.03656**
QM8	MPNN	0.0143	**0.00129**
QM9	DTNN	2.3500	**2.14323**

[2] https://moleculenet.org/

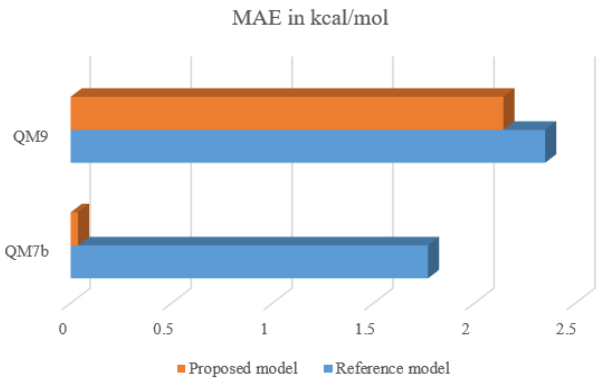

Figure 2. Deep Tensor Neural Network performance – MAE in kcal/mol: Reference vs. Proposed model

If we compere the performance of ANI-1 architecture and the DTTN architecture with applied proposed activation mechanism on QM7 dataset we can see improvement of the performance for more than 3.5 times. For the QM7b dataset the improvement is most noticeable (Fig. 2). We have decreasing of MAE value for more than 40 times. When it comes to MPNN, the obtained results with the applied activation approach also outperform the benchmark. For QM9 the difference is small, but in terms of atomic simulations small differences can have a big impact. Decreasing the MAE value, as indicator for regularized neural network, show us that activation function has a regularization effect on the network.

V. Conclusions

In this paper a different approach to regularization is presented. The main idea is to show that – the selection process of the activation function has a big influence to the need for regularization of the network. The new approach, named as an activation mechanism, don not try to find a universal, new activation function, although this is not excluded, but the most appropriate one for the given task and for the given data set. The obtained results show that the activation function can be viewed as an implicit regularizer.

References

[1] Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), pp.1798-1828.

[2] Bengio, Y., 2013, July. Deep learning of representations: Looking forward. In *International Conference on*

[3] Hughes, T.B., Miller, G.P. and Swamidass, S.J., 2015. Modeling epoxidation of drug-like molecules with a deep machine learning network. *ACS central science*, *1*(4), pp.168-180.

[4] Unterthiner, T., Mayr, A., Klambauer, G. and Hochreiter, S., 2015. Toxicity prediction using deep learning. *arXiv preprint arXiv:1503.01445*.

[5] Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E. and Svetnik, V., 2015. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, *55*(2), pp.263-274.

[6] Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J.K., Ceulemans, H. and Hochreiter, S., 2014. Deep learning for drug target prediction. *Work. Represent. Learn. Methods complex outputs*.

[7] Hamanaka, M., Taneishi, K., Iwata, H., Ye, J., Pei, J., Hou, J. and Okuno, Y., 2017. CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. *Molecular informatics*, *36*(1-2), p.1600045.

[8] I. Goodfellow, Y. Bengio and Aaron Courville (2016). Deep learning. (MIT Press.)

[9] J. Kukačka, V. Golkov, and D. Cremers. "Regularization for deep learning: A taxonomy". [Online]. Available: ArXiv:1710.10686

[10] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, (2012). "Improving neural networks by preventing co-adaptation of feature detectors". ArXiv: 1207.0580.

[11] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, (2015). "Dropout as data augmentation." ArXiv: 1506.08700.

[12] P. Morerio, J.Cavazza, R. Volpi, R. Vidal, and V. Murino, (2017).Curriculum dropout. ArXiv:1703.06229.

[13] S. Maeda, (2014). A Bayesian encourages dropout. ArXiv:1412.7003.

[14] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, (2016a). "Densely connected convolutional networks." ArXiv: 1608.06993

[15] T. DeVries, and G. W. Taylor, (2017). "Dataset augmentation in feature space". In Proceedings of the International Conference on Machine Learning (ICML), Workshop Track.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, (2016). "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 2818–2826.

[17] A. Fawzi, S. Horst, D. Turaga, and P. Frossard, (2016). "Adaptive data augmentation for image classification." In Proceedings of the IEEE International Conference on Image Processing (ICIP), pp:3688–3692

[18] J. Salamon, and J. P. Bello, (2017). "Deep convolutional neuralnetworks and data augmentation for environmental sound classification." *IEEE Signal Processing Letters*, 24(3) pp: 279–283.

[19] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, (2010). "Deep big simple neural nets excel on handwritten digit recognition." *Neural Computation*, 22(12) pp: 1–14.

[20] V. Dumoulin, and F. Visin (2016). "A guide to convolution arithmetic for deep learning." ArXiv: 1603.07285.

[21] F. Yu, and V. Koltun (2015). Multi-scale context aggregation by dilated convolutions. ArXiv: 1511.07122.

[22] J. Long, E. Shelhamer, and T. Darrell, (2015). "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 3431–3440

[23] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, (2011c). Contractive autoencoders: Explicit invariance during feature extraction. In Proceedings of the International Conference on Machine Learning (ICML), pp: 833–840

[24] S. Ruder, (2017). An overview of multi-task learning in deep neural networks. ArXiv: 1706.05098

[25] F. Milletari, N. Navab, and S. A. Ahmadi, S. A. (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation." In Proceedings of the International Conference on 3D Vision (3DV), pp: 565—571. IEEE

[26] M. Sajjadi, M. Javanmardi, and T. Tasdizen, (2016). "Regularization with stochastic transformations and perturbations for deep semi-supervised learning." In *Advances in Neural Information Processing Systems* (NIPS), pp: 1163–1171

[27] B. Neyshabur, "Implicit Regularization in Deep Learning," Ph.D. dissertation, TOYOTA Technological Institute at Chicago, 2017.

[28] Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R. and Tkatchenko, A., 2017. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, *8*(1), pp.1-8.

[29] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K. and Pande, V., 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, *9*(2), pp.513-530.

[30] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. and Ghemawat, S., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[31] https://keras.io/

[32] Smith, J.S., Isayev, O. and Roitberg, A.E., 2017. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science*, *8*(4), pp.3192-3203.

[33] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. and Dahl, G.E., 2017. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.012*