DeepDTA: modelling 1D representations with LSTM neural networks

Jorge Melo and Xavier Pinho

Universidade de Coimbra, Coimbra, Portugal

Abstract. Ozturk et al. presented a deep-learning based model (DeepDTA) that uses sequence information of drugs and targets to predict drug-target (DT) interaction binding affinities, considering a continuum of binding strength values. Their results showed that deep learning based models can be an effective approach for DT binding affinity prediction and the model which used Convolutional Neural Networks (CNNs) for modeling of protein sequences and drug 1D representations outperformed the previous state-of-the-art methods (KronRLS algorithm and SimBoost). Recurrent Neural Networks (RNNs) excel in learning from sequential data and have been widely used in studies related to time-series, natural language processing, etc Particularly, Long Short-Term Memory (LSTM) has found many applications due to its robustness against problems of long-term dependency. In this study, we tested the DeepDTA model with drug and target representations built by LSTM networks and compared the results with the original DeepDTA with CNNs. The results showed that LSTM Neural Networks could not outperform the current state-of-the-art.

Keywords: Drug-Target Interactions, Deep Learning, LSTM

1 Introduction

The prediction of drug-target interactions (DTI) is becoming more relevant in the field of drug discovery, as the cost of developing a prescription drug that gains market approval rises every year. A 2014 report published by Tufts Center for the Study of Drug Development showed that this cost is \$ 2.6 billion, increasing at an annual rate of 8.5 % since 2003.

Most of the approaches to DTI prediction have been done considering it a binary classification problem, ignoring important informations about DTI, such as binding affinity values which provide relevant information about the strength of the interaction between a DT pair. This information is usually expressed as dissociation constant (Kd), inhibition constant (Ki) or the half maximal inhibitory concentration (IC50).

Ozturk et al. followed the idea that regression-based models are better for predicting an approximate value for the strength of the DTI. Apart from this, deep learning methods enable better representations of the raw data by composing non-linear modules that transform the input data into a more abstract level. This capability makes it easier to detect hidden patterns in the data.

Different neural network architectures have been developed. Here we focus on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have been widely used in studies related to ours. CNNs were used in protein-ligand interaction scoring where they learn from 3D structures of the protein-ligand complexes (Gomes et al., 2017; Ragoza et al., 2017; Wallach et al., 2015). This approach, however, is limited by the small number of known structures reported in PDB (Rose et al., 2016).

Long-Short Term Memory Neural Networks (LSTM) is a specific Recurrent Neural Network (RNN) that is very effective to model temporal sequences and their long-range dependencies. It was proved in natural language processing studies that LSTMs architectures provide good results in the sentiment analysis field, showing that they can easily learn from sequential data, not necessarily time sequences.

It is known that the activity of any protein is influenced by its 3D structure, not only 2D or 1D structures. This means that the elements of each protein (amino acids) influence others not necessarily around them, they can be far away from them in the 1D sequence. We assumed that LSTMs could be a good way of creating compound 1D representations because they have the capability of saving information through the sequence.

The LSTM contains special units called memory blocks in the recurrent hidden layer, this unity contain memory cells with self-connections storing the temporal state of the network in addition to special multiplicative units called gates to control the flow of information. The output is modulated by the state of the cell. This is a very important property when we need the prediction of the neural network to depend on the historical context of inputs, rather than only on the very last input.

In this study we aim to predict interaction binding affinities between drug and target, while comparing the previous work of Ozturk et al. using CNNs to build compounds representation with our proposal of creating representations using LSTM Neural Networks. The results will be used to ascertain which architecture provides the best representations for drugs and targets.

2 Materials and Methods

2.1 Datasets

Our model was evaluated on two different datasets, the Kinase Dataset Davis (Davis et al., 2011) and KIBA dataset (Tang et al., 2014), which were used previously in the base model (Ozturk et al.).

The KIBA dataset presents combined values of Ki, Kd and IC50. A model-based integration approach, termed KIBA was also introduced, and demonstrate here how it can be used to classify kinase inhibitor targets. After the original data, that contained 467 targets and 52498 drugs, were filtered (He et al. 2017) it contain a total of 229 unique proteins and 2111 unique drugs, including 246088

KIBA scores.
The Davis dataset presents the interaction of 72 kinases inhibitors with 442 kinases and the relevant inhibitors with their respective dissociation constant (Kd) values, which covers more than 80 % of the human catalytic protein kinome.

	Proteins	**Drugs**	**Interactions**
KIBA	442	68	30056
Davis	229	2111	118254

Table 1. Summary of the Datasets

2.2 Input Representation

In our study we used label encoding that converts characters for both SMILES and protein sequences to a corresponding integer (e.g. ”C”: 1, ”H”: 2, ”N”: 3, etc...).
SMILES and protein sequences have a very large range of different lengths, a maximum of 85 for SMILES length and 1200 for protein sequences length was chose for Davis dataset and for KIBA dataset a maximum of 100 for SMILES length and 1000 for protein sequences was defined. The sequences that are longer than the maximums defined are cut and shorter sequences are 0-padded.

2.3 Base Model

DeepDTA was designed to treat DTI as a regression problem by aiming to predict the binding affinity scores. Ozturk et al. model consists of two separate CNN blocks, one that receives protein sequences as inputs and other that receives SMILES from drugs. The output of each block is concatenated into a single vector that is used as input for 3 fully connected (FC) layers, called DeepDTA, which are connected to the output layer. The first layer of this model is an embedding layer which creates a 128-dimensional dense vector to represent each character, followed by the CNN block. Each CNN block consists of 3 1D-convolutional layers with increasing number of filters with the objective of capturing local dependencies in the data. In the end of each CNN block is a pooling layer that down-samples the output of the previous layer and represents a way of generalizing the features learned by the filters. The output of the two blocks is concatenated and fed into the DeepDTA.
We used the same hyper-parameters used in the base model for the DeepDTA block, with 1024 nodes in the first 2 layers, followed by a dropout rate of 0.1 and 512 nodes in the last layer. For each CNN layer, we used 16, 32 and 48 filters for each layers, respectively, and stride with value 1. We maintained Rectified Linear Unit (ReLU) as activation function for each layer and mean squared error (MSE) as the loss function.
Since we had some computational and time limitations, we could only train this

model for 20 epochs. Although we could not achieve such good results as Ozturk et al., we think that using the same conditions for each model makes it possible to compare them.

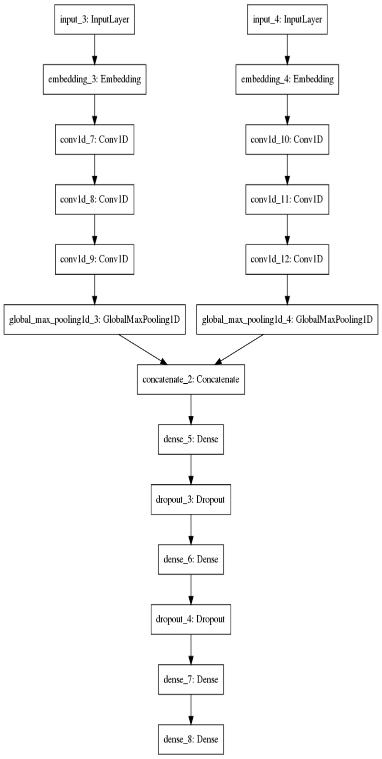

Fig. 1. Base Model Architecture

2.4 Our Model

In our study we tested LSTM Neural Networks to model both protein sequences and SMILES representations, instead of CNNs and maintaining the remaining architecture. The architecture is shown below: We defined the LSTM layer hyper-parameters with 64 units, dropout of 0.1 and recurrent dropout of 0.1. The output of each LSTM block was concatenated and fed into the DeepDTA block, maintaining the same hyper-parameters as in the first test. Our model was trained during 20 epochs, as said above.

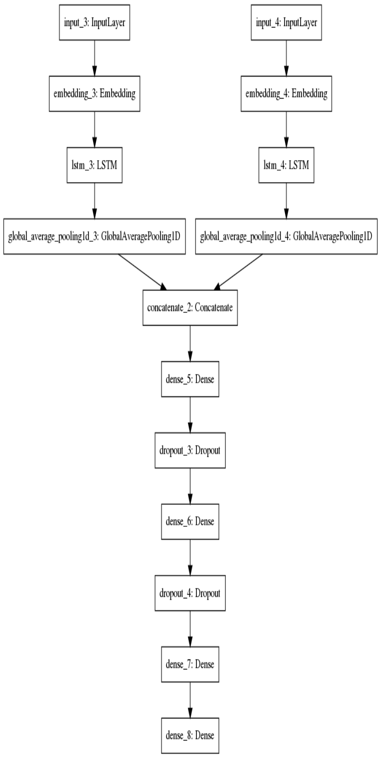

Fig. 2. Our Architecture

3 Results

To compare the two models we used the same method used in the previous work, Concordance Index - this metric tells us whether the predicted binding affinity values of two random drugtarget pairs were predicted in the same order as their true values were. We used paired-t test for the statistical significance tests with 95 % confidence interval. We also used MSE to evaluate our results. The obtained values are shown in Table 3.1.

Dataset	Compounds	CI	MSE
KIBA	CNN	0.795	0.354
	LSTM	0.763	0.397
Davis	CNN	0.821	0.602
	LSTM	0.773	0.834

Table 2. Results for CI and MSE for both representation modelling methods and datasets.

Our results represent lower values for CI below than Ozturk et al. obtained with CNNs, but we blame that on the small number of epochs during which our model trained. Since they're close to the results obtained by the previous work, we consider them valid.

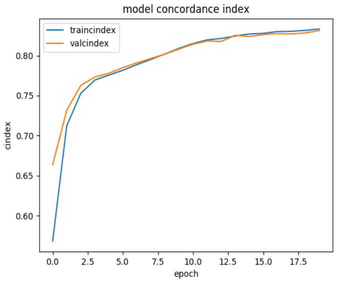

Fig. 3. Evolution of CI in the model with CNNs and data from the Davis dataset while training.

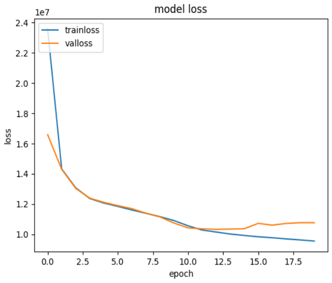

Fig. 4. Evolution of loss (MSE) in the model with CNNs and data from the Davis dataset while training.

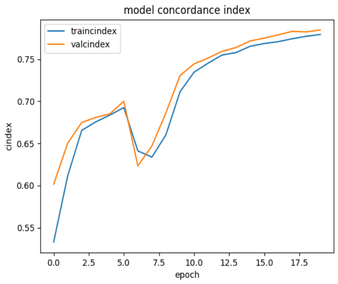

Fig. 5. Evolution of CI in the model with LSTM and data from the Davis dataset while training.

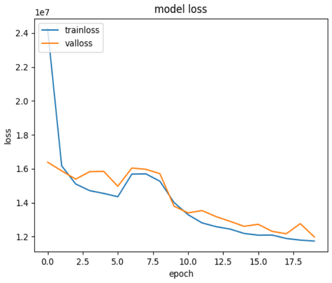

Fig. 6. Evolution of loss (MSE) in the model with LSTM and data from the Davis dataset while training.

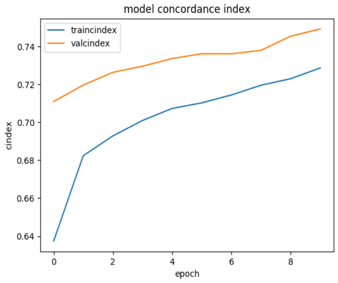

Fig. 7. Evolution of CI in the model with CNNs and data from the KIBA dataset while training.

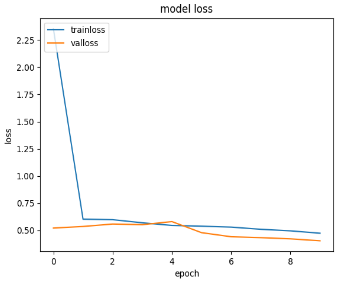

Fig. 8. Evolution of loss (MSE) in the model with CNNs and data from the KIBA dataset while training.

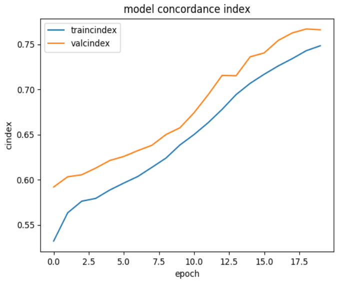

Fig. 9. Evolution of CI in the model with LSTM and data from the KIBA dataset while training.

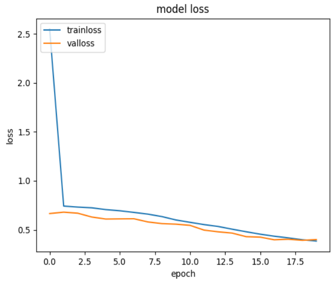

Fig. 10. Evolution of loss (MSE) in the model with LSTM and data from the KIBA dataset while training.

4 Discussion

We tested a different approach for modelling of compound sequences based on the work done by Ozturk et al. We used Long Short-Term Memory Neural Networks to learn representations from the raw data of proteins sequence and drugs and tested them with the DeepDTA architecture.
We compare the performance of the proposed model with the study of Ozturk et al. on the Davis kinase-drug dataset and the KIBA dataset. The results show that the model that uses LSTM-blocks to learn protein sequences and drugs performed poorly compared with the model that used CNN-blocks, showing that despite being a good method to learn from sequencial data, LSTM Neural Networks may not be suited for this case of study. But we can also consider that the difference is quite small, showing that, despite scoring bellow CNNs in this particularly case, LSTM Neural Networks may get better results in future work, maybe with a different approach.
The results also indicated that the model performed in the KIBA dataset performed better than the Davis dataset, which is expected because the KIBA dataset is larger than the Davis. The increase in the data enables the deep learning architectures to learn patterns better.
After analyzing our methodology we proposed as future work the use of LSTM for representation for protein sequences and drugs...

5 References

1. Ozturk, H., Ozgur, A. , Ozkirimli, E. DeepDTA: Deep drug-target binding affinity prediction. in Bioinformatics (2018). doi:10.1093/bioinformatics/bty593
2. Pahikkala, T. et al. Toward more realistic drug-target interaction predictions. Brief. Bioinform. (2015). doi:10.1093/bib/bbu010
3. Tang, J. et al. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. J. Chem. Inf. Model. (2014). doi:10.1021/ci400709d
4. He, T., Heidemeyer, M., Ban, F., Cherkasov, A. , Ester, M. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J. Cheminform. (2017). doi:10.1186/s13321-017-0209-z
5. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. (2011). doi:10.1038/nbt.1990
6. CSDD. Tufts Center for the Study of Drug Development. Briefing Cost of Developing a New Drug Innovation in the Pharmaceutical Industry: New Estimates of R , D Costs (2014).
7. Lecun, Y., Bengio, Y. , Hinton, G. Deep learning. Nature (2015). doi:10.1038/nature14539
8. Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. Neural Comput. (1997).
doi:10.1162/neco.1997.9.8.1735
8. Wallach I. et al.. (2015) Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery.arXiv preprint arXiv: 1510.02855

9. Gomes J. et al.. (2017) Atomic convolutional networks for predicting protein-ligand binding affinity.arXiv preprint arXiv:1703.10603.
10. Ragoza M. et al.. (2017) Proteinligand scoring with convolutional neural networks. J. Chem. Inf. Model., 57, 942957.