Joseph Heenan

Udacity Machine Learning Engineer Capstone Project Proposal

Domain Background

Computational chemistry is the study of chemical reactions using computer methods. This project focuses on applications of machine learning algorithms, in particular deep learning algorithms, to problems in computational chemistry.

Bioassays produce much of the data used by computational chemists; they are a form of high-throughput experiment in which many small molecule compounds are tested against a target biological system to determine which of them exhibits 'bioactivity' in that system. The definition of 'bioactivity' may vary from assay to assay but often means increased or decreased expression of a particular gene or protein.

Problem Statement

Researchers would like to be able to predict, without actually running a physical assay, whether some imaginary compound is likely to be bioactive with regards to the target of some assay. Deep learning models which learn the "features" of compounds that are generally bioactive (e.g., "compounds with a sulfonyl ground") are highly desirable.

Historically such predictors have been closed-source programs, however in the past several years machine learning researchers and computational chemists have produced series of benchmarks for a deep learning-focused system called MoleculeNet (Wu, et al., 2017), integrated in the open-source software package DeepChem (Ramsundar, et al., 2015).

In particular the MoleculeNet-PCBA benchmark shows that the multitask deep neural network model attains an Area Under Curve-Receiver Operating Characteristic Score (AUC-ROC) mean test set score of .781. While such an accuracy is best-in-class, it still means a large amount of manual screening and analysis may be needed by researchers.

Solution Statement

The goal of this project is to investigate two strategies to improve the performance of the Multitask Neural Network on the MoleculeNet-PCBA test set benchmark:

1) Optimization of hyperparameters – via additions to the source code for the TensorFlow based deep neural network at https://github.com/deepchem/deepchem/blob/master/examples/pcba/pcba_tf.py

2) Increase in size of dataset from 128 bioassays to 256 bioassays, taken in years 2016 and 2017, since original publication of (Ramsundar, et al., 2015), to see if performance on the held-out test set improves, via modifications to: https://github.com/deepchem/deepchem/blob/master/examples/pcba/pcba_datasets.py

The authors of (Ramsundar, et al., 2015) note that performance is "still climbing…" generally with the addition of further datasets, so it will be interesting to see whether to model continues to learn when given additional assay data.

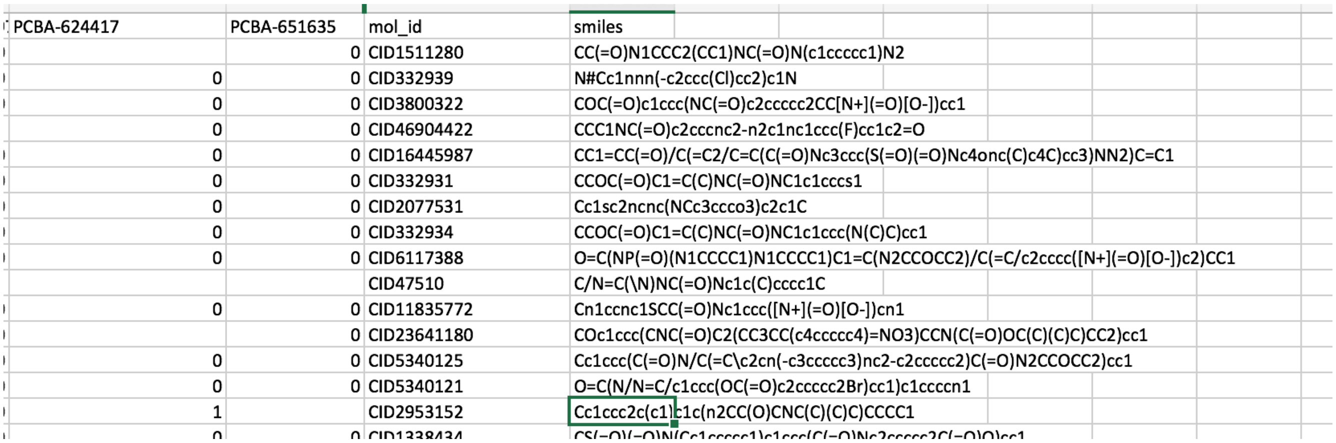
Datasets and Inputs

The PCBA test dataset is available at:

http://deepchem.io.s3-website-us-west-1.amazonaws.com/datasets/pcba.csv.gz

Its construction is documented in both (Ramsundar, et al., 2015) and in (Wu, et al., 2017). The source PCBA data is taken from the PubChem BioAssay (PCBA) website at: https://pubchem.ncbi.nlm.nih.gov/

An example view of the dataset is below:



PCBA-624417	PCBA-651635	mol_id	smiles
	0	CID1511280	CC(=O)N1CCC2(CC1)NC(=O)N(c1ccccc1)N2
0	0	CID332939	N#Cc1nnn(-c2ccc(Cl)cc2)c1N
0	0	CID3800322	COC(=O)c1ccc(NC(=O)c2ccccc2CC[N+](=O)[O-])cc1
0	0	CID46904422	CCC1NC(=O)c2cccnc2-n2c1nc1ccc(F)cc1c2=O
0	0	CID16445987	CC1=CC(=O)/C(=C2/C=C(C(=O)Nc3ccc(S(=O)(=O)Nc4onc(C)c4C)cc3)NN2)C=C1
0	0	CID332931	CCOC(=O)C1=C(C)NC(=O)NC1c1cccs1
0	0	CID2077531	Cc1sc2ncnc(NCc3ccco3)c2c1C
0	0	CID332934	CCOC(=O)C1=C(C)NC(=O)NC1c1ccc(N(C)C)cc1
0	0	CID6117388	O=C(NP(=O)(N1CCCC1)N1CCCC1)C1=C(N2CCOCC2)/C(=C/c2cccc([N+](=O)[O-])c2)CC1
		CID47510	C/N=C(\N)NC(=O)Nc1c(C)cccc1C
0	0	CID11835772	Cn1ccnc1SCC(=O)Nc1ccc([N+](=O)[O-])cn1
	0	CID23641180	COc1ccc(CNC(=O)C2(CC3CC(c4ccccc4)=NO3)CCN(C(=O)OC(C)(C)C)CC2)cc1
0	0	CID5340125	Cc1ccc(C(=O)N/C(=C\c2cn(-c3ccccc3)nc2-c2ccccc2)C(=O)N2CCOCC2)cc1
0	0	CID5340121	O=C(N/N=C/c1ccc(OC(=O)c2ccccc2Br)cc1)c1ccccn1
1		CID2953152	Cc1ccc2c(c1)c1c(n2CC(O)CNC(C)(C)C)CCCC1
0	0	CID1338434	CS(=O)(=O)N(Cc1ccccc1)c1ccc(C(=O)Nc2ccccc2C(=O)O)cc1

The first assay, 624417, tests bioactivity for 'class B1 G protein-coupled receptors'. The second assay, 651635, tests 'Inhibitors of ATXN expression'. As can be seen, the dataset is highly sparse, with an empty cell indicating a compound (represented by a smiles string) wasn't tested, a 0 indiciating it tested non-bioactive, and a 1 indicating it tested bioactive.

Benchmark Model

Using the index splitter and ECFP4 featurization, the PCBA dataset task 0.781% accuracy on held-out test set as noted in (Wu, et al., 2017). This is the benchmark accuracy which I will attempt to improve upon for this capstone project.

Evaluation Metrics

The AUC-ROC score will be the evaluation metric used for this project.

Project Design

I will approach the problem by first benchmarking the existing performance of the open-source model and dataset (my results approximately match the published benchmark results). Then I will fork this source repository and attempt to improve the neural network and data loading process. If relevant, I will submit any improvements as a pull request back to the maintainers of this project.

Citations:

Wu, MoleculeNet: A Benchmark for Molecular Machine Learning.

Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. arXiv preprint arXiv:1502.02072 2015

DeepChem: Deep-learning models for Drug Discovery and Quantum Chemistry. https: //github.com/deepchem/deepchem, Accessed: 2017-08-04.