CS230

Modeling of Odor Prediction from Chemical Structures

Chengling Qiu
Department of Materials Science and Engineering
Stanford University
clqiu@stanford.edu

Abstract

In this project, a set of different modeling methods were experimented for the multi-label classification task of predicting odor of molecules from chemical structures. The dataset was investigated and different methods of featurization of SMILES string were considered. The performance of fully connected neural network, graph convolutional network, Chemception, ChemBERTa, and classic machine learning model of random forest and label powerset transform were compared, where the convolutional network with the input of circular fingerprint and ChemBERTa model presented most promising prediction results with Jaccard index around 0.330.

1 Introduction

The task of predicting properties of a molecule from its chemical structure is crucial for a lot of applications such as drug discovery and material design. The accurate prediction can potentially avoid a long and costly discovery process by providing useful property of a molecule such as its bioactivity, toxicity or melting point. Recently, deep learning has become a powerful tool for modeling molecules, especially with certain models such graph convolutional network and its derivatives. However, little attention was been given to the fragrance and flavor industry. It is potentially very promising to predict the smell from the odorant molecules, the actual building blocks of all fragrance. Historically, people tried to predict the odors based on the functional group of the molecules but there is only a very limited number of odors can be predicted this way and some odorants having the same functional groups can smell very differently.[1]

Conventionally, most of the public accessible data sets presented the compounds as SMILES string (Simplified Molecular-Input Line-Entry System),[2] a standard method for encoding structures into sets of ASCII characters (string) to be digitally recognized, and provide the multiple odors of a particular molecule in the form of a sentence with tags of manually classified tags. Therefore, the task could be considered as a multi-label classification problem with input as SMILES string and output as the tags of odors.

For this project, a main emphasis was the experiments of different methods of representing the chemical compound from their structural information (SMILES strings). Several published different approaches for featurization of encoding chemical structures were implemented and considered, including the chemical fingerprint, graph convolution network, BERT-like model, and "grid image". Upon the completion of appropriate preparation of data, some pretrained models, including both shallow and graph models, were utilized and adapted for the classification task, where the performance was evaluated and discussed.

This was a shared project between CS230 and CS229, followed and approved by both class project policies. There were both shallow and deep models experimented in this project and the parts of this project could be arbitrarily divided into: Data augmentation, Chemception model, Graph

CS230: Deep Learning, Autumn 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

convolutionally network for CS229; ChemBERTa model, Fully connected multi-label classifer, Scikit-multilearn models for CS230. By the nature of these two classes, most of work in task setup, datasets investigation, infrastructure were shared. The report submitted for both classes followed similar outline to include all the necessary details.

2 Related work

For the featurization of chemical structures, there were several commonly adapted in the studies of prediction of molecular property. The open-source DeepChem library provided implementations of some featurizers,[3] including circular fingerprint transformers and encoding for graph convolutions. There were other approaches for representing molecules: Chemception project reported an inception-model inspired deep convoluitonal network for the prediction using the 2D images of molecules,[4] which was a much different approach without insight in the chemistry domain; ChemBERTa presented a collection of trained BERT-like models of transformers on molecular property predictions tasks.[5] The performance of these models was highly depended on the features of the chemical structures present in the dataset, such as organic molecules or complex molecules. For models with the input of some pre-processed conversion from SMILES strings, some data augmentation methods were reported with main strategies as the enumeration[6] or randomly generation SMILES string.[7] In this project, these methods of different featurizations and modeling were adapted and experimented.

3 Dataset and Features

For this project, the open-source Leffingwell PMP database was investigated for modeling where 4796 chemical molecules were represented in the SMILES string and the olfactory information was provided with manually classified sentence of odor labels. The occurrence of labels was not uniform and the number of labels for molecules was could be various. The dataset was divided into train, validation, and test set in a ratio of 70%:20%:10%.

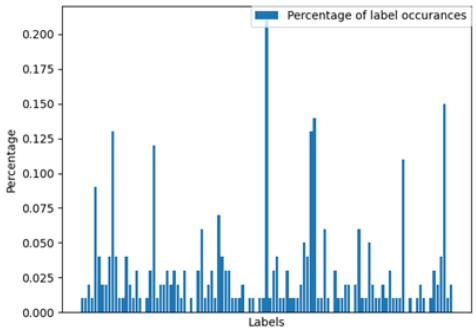

Figure 1: Label Occurrence Distribution

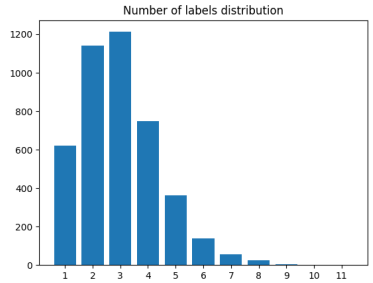

Figure 2: Number of labels for each molecule

For the evaluation for the prediction, it should be admitted that the description of odor could be subjective and heterogenous based on personal experience and other factors. In this case, the Tanimoto Similarity Score (Jaccard Index) was considered as an appropriate performance description by calculating the average of sentence matching in the proposed 3 sentences (list of labels) between the prediction and the ground truth sentences. In fact, the Tanimoto similarity score is also commonly used to describe the chemical similarity between molecules.

$$\text{Jaccard Index} = \frac{\text{number of labels matched in sets}}{\text{number of labels of union of predicted and truth sets}}$$

The labels in from the dataset were further converted into tokenized index to be recoginized by the model. The labels of each molecule were accordingly converted into a one-hot-encoded format based on the index-to-label map.

Figure 3: Index-to-label map

4 Methods

4.1 Featurization

4.1.1 Circular (Morgan) fingerprint

Circular fingerprint representation discovered local structural properties within a molecule by considering the neighboring atom relations. The experiment of multi-task classification with the circular fingerprint featurizer originated based on the chemical insight that the most odor of molecules was associated or provided with particular functional groups or molecular structures in these molecules. For example, the aromatics compounds possessed distinctive perfumed smell from the conjugated system. Circular fingerprints could be considered as an analog to convolutional networks because the identical operations were applied locally to all the atoms and the global information was gathered and combined in a pooling step. Therefore, the circular fingerprint was explored in this project because it was generally more representative for structural information than other methods. However, a main limit for this representation was that the circular fingerprint was unique for molecules so that data augmentation would be difficult to apply. The implementation from Deepchem library of circular fingerprint was used where the SMILES strings were converted and hashed into bit vectors, with the algorithm described by David Duvenaud et al.[8]

4.1.2 2D molecular structure graph conversion

Converting the SMILES encoding strings of molecules into graphs was a common strategy in molecular machine learning to exploit certain well-developed methods on fully connected network training. With the open-source cheminformatics software RDKit, the SMILES strings were converted a molecular descriptor of molecular graph and further encoded to an image with 3 channels. The conversion was based on the Chemception method:[4] given the RDKit mol. target, the atom number, Marsilli-Gasteiger partial charge,[9] and hybridization state were calculated and the 2D drawing coordinates were computed and extracted to a coordinate matrix. The 2D molecular structure was then mapped onto (80×80) grids (resolution of 0.5 Å), following the Chemception protocol.

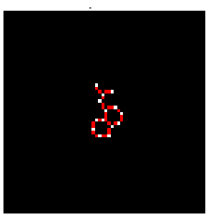

Figure 4: 2D Molecular Grid Image

4.2 Data augmentation

A noticeable challenge for machine learning study on molecular property was the limited size of available data sets. Training most conventional deep neural networks usually would require a relatively vast number of data examples, such as for computer vision or language recognition. However, most publicly available datasets for the properties of chemical molecules have samples from 4 thousands (PDBBind 2017) to 22 thousands (QM8).[10] The limited size of high-quality scientific data provides much restriction on the choice of learning models. From this perspective, data augmentation method on these datasets could be very promising for the training performance. For this project and dataset, even though chemical fingerprint ensured a one-to-one correspondence between the fingerprint description and the molecule which would be difficult to generate more examples for training, one molecule can however have multiple SMILES strings representation. Such fact had been explored as a technique for data augmentation of a molecular dataset. Specifically, the SMILES for a given molecule was obtained by traversing the atoms within the molecule with certain restriction7. Different choice of atom orderings could generate different SMILES string for the same molecule. For example, "CC(CC(=O)OC1CC2C(C1(C)CC2)(C)C)C" and "C1C2CCC(C)(C2(C)C)C1OC(=O)CC(C)C", "O1C(C)OC(C)OC1C" and "C1(C)COC(CC)O1" represented identical pairs of molecules.
Therefore, a new SMILES string representation was generated for each examples in the training and validation set by: (i) Ranking and randomizing the index for atoms; (ii) Traversing the molecules in the randomized index order; (iii)Iterating until a different SMILES string was generated. Although the generation of a greater number of SMILES strings for a given molecule or even the enumeration on randomized SMILES string could provide a much larger size of dataset, there were highly structural symmetric molecules possessing only two SMILES strings representation. Furthermore, the number of different randomized SMILES strings for each molecule could be different, thus possibly introducing a bias towards the model due to the under-representation for molecules that had a smaller number of SMILES strings. The augmented dataset was pre-processed by the 2D image conversion for Chemception model or transformer in the ChemBERTa model.

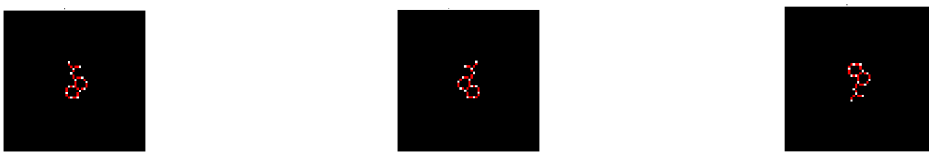

Figure 5: 2D Molecular Grid Images of Randomized SMILES String

4.3 Modeling

4.3.1 Multi-label classifier

A fully connected network from DeepChem library was experimented for the task of multi-label classification, with the input as circular fingerprints. There were several hyperparameters for the architecture of the network, including the number of layers, size of each layer, dropout rate, learning rate and choice of activation functions. To optimize the hyperparameters, an implementation of grid search optimization algorithm from DeepChem library was utilized to exhaustively compute the optimum combination of hyperparameters. The performance metric for the optimization was evaluated based on the Jaccard Index, considering the size of the intersection of the training set and the validation set. The default Adam optimizer was used for the training and cross-entropy was set as the loss function.

4.3.2 Graph convolutional networks

For graph convolution network modeling with input of circular fingerprints, the implementation of the classical graph convolution model from Duvenaud, et al from DeepChem was used for this project. These graph convolutions with a per-atom set of descriptors for each molecule and combine the descriptors over convolutional layers. The architecture from the original study was followed and adapted with experiment on hyperparameters. Batch normalization with batch size of 100 was applied to the model, and ReLU was the activation function for all the hidden layers.
For convolutional networks with input of 2D "Grid" image (Chemception), a model with vgg19

architecture was experimented and trained with implementation from fastai library.[11] The optimal learning rate was determined with the metrics as the multi-label Jaccard index. The models were trained on the original dataset, and the double-size augmented dataset.

4.3.3 Scikit-multilearn Random forest Label Powerset

The performance of classic ML algorithms of Random Forest and Laber Powerset on the multi-label classification task was discovered as a baseline comparison to the deep models. The implementation from scikit-multilearn for these two models was experimented.[12] For the random forest classifier, the binary relevance method was applied to transform the multi-label classification into several single-label separate binary classification using the random forest classifier. For the Laber Powerset method, the multi-label problem was transformed to a multi-class problem with a single multi-class classifier trained on all the combinations of labels in the training set (power set). A Gaussian Naïve bayes was used as the base classifier.

4.3.4 ChemBERTa

ChemBERTa presented a collection of trained BERT-like models of transformers on molecular property predictions tasks.[5] To present a basic performance of the model, a pretrained tokenizer and RoBERTa model from ZINC-250k dataset was fine-tuned for this task. The attention patterns produced by the tokenizer could be visualized by the Bertviz tool.[13]

5 Results

The performance for the models described was measured in the top-3 Jaccard index and summarized in the table below. The presented performance was resulted from the optimized best model for a particular architecture.

Model	Original Dataset	Augmented Dataset
Fully Connected Classifier	0.286	-
Graph Convolutional Network Classifier	0.325	-
Chemception Model	0.247	0.268
Random Forest	0.221	-
Label PowerSet (Naïve Bayes)	0.242	-
ChemBERTa Model	0.317	0.318

6 Discussion

Among the modeling methods experiments, the graph convolutional networks with input of circular fingerprint achieved the best performance. Such observation was consistent with the conclusion from several molecular property prediction modeling studies that the graph models generally outperformed the shallow models (RF, SVM). The nature of graph models provided opportunities for including more details of molecular information in the featurization. However, the Chemception models with vgg19 architecture did not perform well as expected. This could be resulted from the limited size of input image examples even with the augmentation. Noticeably, the BERT-like model ChemBERTa suggested very promising performance on the classification. Given the fact that the tokenizer and model were transferred from pretrained model on much larger dataset, the performance would probably be more accomplished if hyperparameters could be better fine-tuned.
From the analysis on the dataset, the very imbalanced distribution of label occurrence suggested a need for data resampling. However, the techniques of resampling on imbalanced multi-label datasets had only been address recently and the performance of most of these methods greatly depended on the traits of the multi-label datasets. For example, the synthetic minority over-sampling (SMOTE) technique required insight on local data point distribution for considering the synthetic data point generation sources.[14] For this dataset, there was seemingly no reasonable method of quantitative description for SMILES strings data point distribution with the experimented featurization methods (e.g. circular fingerprint). The task of resampling on the imbalanced SMILES string dataset could be

an emphasis for future work.
Another possible direction for the future work could be taking more molecular features into consideration at the stage of the 2D image encoding. For this project, only the Gasteiger charges and hybridization types were computed for molecules. Based on the nature of the dataset, there could be other useful features such as the types of chirality for stereochemistry. For example, the inputs of enantiomers and diastereomers could not be distinguished with the current featurization, where bias will be introduced in the training. With more molecular details encoded, the classification results can be better achieved.

7 Conclusion

In this project, a set of different modeling methods were experimented for the multi-label classification task of predicting odor of molecules from chemical structures. The dataset was investigated and different methods of featurization of SMILES string were considered. The performance of fully connected neural network, graph convolutional network, Chemception, ChemBERTa, and classic machine learning model of random forest and label powerset transform were compared, where the convolutional network with the input of circular fingerprint and ChemBERTa model presented most promising prediction results. The approach for data resampling on the imbalanced SMILES string dataset and encoding more molecular features can be directions for future work.

Appendix

	SMILES	SENTENCE
0	C/C=C/C(=O)C1CCC(C=C1C)(C)C	fruity,rose
1	COC(=O)OC	fresh,ethereal,fruity
2	Cc1cc2c([nH]1)cccc2	resinous,animalic
3	C1CCCCCCCC(=O)CCCCCCC1	powdery,musk,animalic
4	CC(CC(=O)OC1CC2C(C1(C)CC2)(C)C)C	coniferous,camphor,fruity

Figure 6: Data Preview

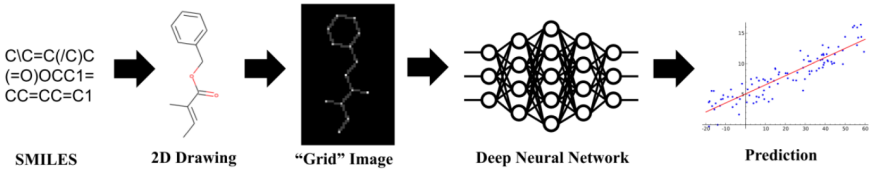

Figure 7: Chemception Outline.[4]

Algorithm 1 Circular fingerprints

Input: molecule, radius R, fingerprint length S
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \dots, \mathbf{r}_N]$ ▷ concatenate
    $\mathbf{r}_a \leftarrow \text{hash}(\mathbf{v})$ ▷ hash function
    $i \leftarrow \text{mod}(r_a, S)$ ▷ convert to index
    $\mathbf{f}_i \leftarrow 1$ ▷ Write 1 at index
Return: binary vector $\mathbf{f}$

Figure 8: Circular fingerprints generation algorithm[8]

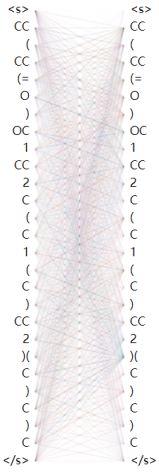

Figure 9: Attention head view for tokenized SMILES string

The Hyperparameters of the best-model by grid search:

Hyperparameter for multi-label classifier	
Size of each layer	256, 512, **1024**
Number of layers	1,2,3,4,5,**6**,7,8,9,10
Activation function (hidden layer)	**ReLU**, sigmoid
Dropout rate	0.2, **0.5**
Learning rate	0.01,0.001,**0.0001**

Hyperparameter for Fingerprint CNN	
Width of channels for convolution layers	[64,64], **[128, 128]**
Width of channels for atom level dense layer	64, **128**
Dropout rate	**0.2**, 0.5

References

1. Genva, Manon, et al. "Is It Possible to Predict the Odor of a Molecule on the Basis of its Structure?." International Journal of Molecular Sciences 20.12 (2019): 3018.
2. Jastrzębski, Stanisław, Damian Leśniak, and Wojciech Marian Czarnecki. "Learning to smile (s)." arXiv preprint arXiv:1602.06289 (2016).
3. Ramsundar B, Eastman P, Walters P, et al. Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more[M]. " O'Reilly Media, Inc.", 2019.
4. Goh, Garrett B., et al. "Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models." arXiv preprint arXiv:1706.06689 (2017).
5. Chithrananda, Seyone, Gabe Grand, and Bharath Ramsundar. "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction." arXiv preprint arXiv:2010.09885 (2020).
6. Bjerrum E J. SMILES enumeration as data augmentation for neural network modeling of molecules[J]. arXiv preprint arXiv:1703.07076, 2017.
7. Arús-Pous, Josep, et al. "Randomized SMILES strings improve the quality of molecular generative models." Journal of cheminformatics 11.1 (2019): 1-13.
8. Duvenaud D K, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints[C]//Advances in neural information processing systems. 2015: 2224-2232.
9. Gasteiger, Johann, and Mario Marsili. "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges." Tetrahedron 36.22 (1980): 3219-3228.
10. Feinberg, Evan N., et al. "PotentialNet for molecular property prediction." ACS central science 4.11 (2018): 1520-1530.
11. Howard J, Gugger S. Fastai: A layered API for deep learning[J]. Information, 2020, 11(2): 108.
12. Szymanski, Piotr, and Tomasz Kajdanowicz. "Scikit-multilearn: a scikit-based Python environment for performing multi-label classification." The Journal of Machine Learning Research 20.1 (2019): 209-230.
13. Vig, Jesse. "A multiscale visualization of attention in the transformer model." arXiv preprint arXiv:1906.05714 (2019).
14. Charte F, Rivera A J, del Jesus M J, et al. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation[J]. Knowledge-Based Systems, 2015, 89: 385-397.