Predicting Chemical Properties using Self-Attention Multi-task Learning based on SMILES Representation

Sangrak Lim, Yong Oh Lee

KIST Europe, Korea Institute of Science and Technology, Campus E7 1, 66123 Saarbrücken Germany

Background – QSAR model

- Quantitative structure–activity relationship (QSAR) models extract relationships from chemical structures and predict biological activities, such as toxicity, solubility, and so on.
- Previous QSAR models utilized molecular descriptors to represent chemical properties as vectors. Such molecular descriptors require additional processes from inputs, such as the Simplified Molecular Input Line Entry System (SMILES).

Background - Contributions

- We present a Natural Language Processing (NLP) model that utilizes SMILES as direct input.
- We explored the structural differences of existing transformer-variant models and proposed a new self-attention based model.
- The representation learning performance of our self-attention module was evaluated in a multi-task learning environment using several chemical datasets.

Methods

Component 1 - CNN layer

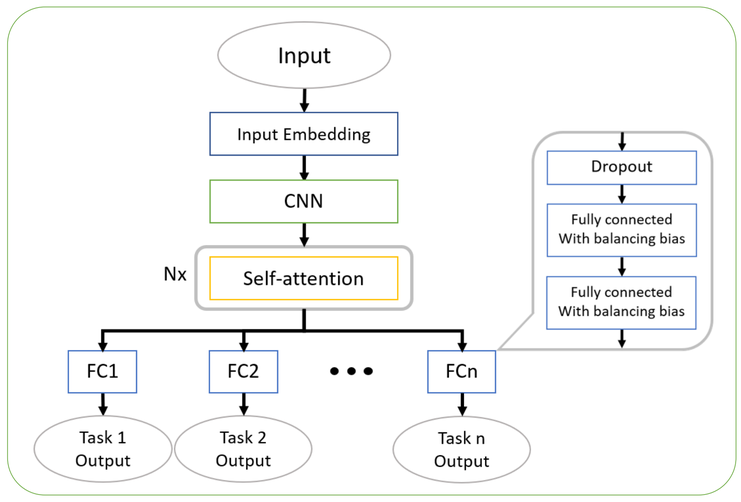
- A CNN layer serves as a shared hidden layer for multi-task learning.
- Input is a SMILES format ➔ No chemical descriptors required.

Component 2 – self-attention layer

- A Self-attention module focuses on long-range dependencies of a given input.
- No pre-training

Component 3 – discrete output layer

- Discrete output layers produce outputs for multiple tasks
- A balancing bias is applied to rectify the class-imbalance in the data



Comparison with other studies that used transformer-variants

1-b) Smiles Transformer Model [1]

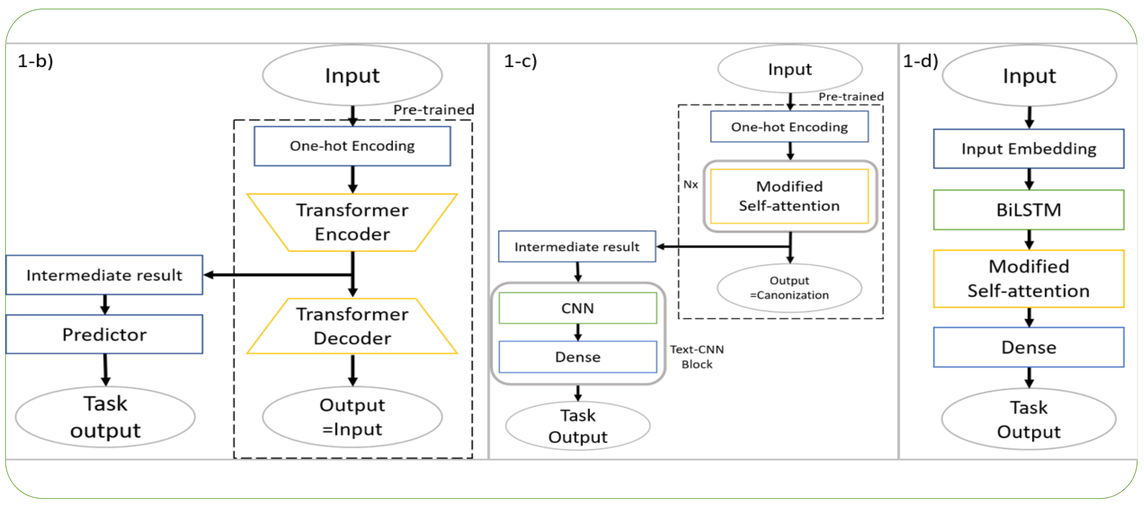
- The Smiles Transformer model uses the intermediate result obtained from pre-training
- If the pre-training objective is not closely related to the target task, the pre-training process may hurt the target task's performance.

1-c) Transformer-CNN Model [2]

- The Transformer-CNN model also implements the pre-training approach.
- The model contains text-CNN block for several CNN layers after the self-attention.

1-d) BiLSTM-SA Model [3]

- The concept of the BiLSTM-SA model implements a self-attention module without the multi-task learning scheme.



Results – Tox21, BBBP, and CLINTOX dataset

Our SA-MTL model exhibited the state-of-the-art performance in the Tox21 and several other datasets.

Tox21

- For the Transformer_CNN model, placing the CNN layers after the self-attention component was not an appropriate option to enhance performance.
- Both of the Transformer_CNN model and the Smiles_Transformer model used the self-attention component for pretraining. The objective of the pre-training approach should resemble the target task.

BBBP/CLINTOX

- Our SA-MTL model could achieve AUC score of 0.966. One of the reasons for the high score is the positive to negative ratio. The positive-to negative ratio of this BBBP and CLINTOX dataset are different from the other datasets.

TABLE III
TOX21 EVALUATION RESULTS COMPARED TO OTHER MODELS

Comparison results on Train and Test Data		
Model	Notes	Average AUC
SA-MTL(OURS)	random split	**0.9**
SCFP	cross-validation	0.877
FP2VEC	random split	0.876
BiLSTM-SA	stratified random split	0.842
GC	random split	0.829
Transformer_CNN	cross-validation & augmented	0.82
Smiles_Transformer	random split	0.802
Comparison results on Score Data		
Model	Notes	Average AUC
SA-MTL(OURS)	without ensemble	0.806
SA-MTL(OURS)	with ensemble	**0.842**
DeepTox[27]	with ensemble	0.837
SCFP	without ensemble	0.813

Note: The best results on the test set are highlighted in bold.

TABLE I
DATASET STATISTICS

Dataset	Num of Classes	Ave. Instances	Pos/Neg ratio†
Tox21	12	7831	1:13.4
BBBP	1	2031	1:0.25
CLINTOX	2	1478	1:0.06|1:12.25*

† The positive to negative ratio is the total sum value of the training data if the number of classes is more than one except CLINTOX.
* The CLINTOX has two classes that have different Pos/Neg ratio.

TABLE VI
BBBP AND CLINTOX EVALUATION RESULTS COMPARED TO OTHER MODELS

Dataset	Model	Notes	Average AUC
BBBP	SA-MTL(OURS)	scaffold split	**0.954**
	SA-MTL(OURS)	random split	0.945
	Transformer_CNN	CV & augmented	0.92
	KernelSVM*	scaffold split	0.729
	FP2VEC	random split	0.713
	Smiles_Transformer	scaffold split	0.704
CLINTOX	SA-MTL(OURS)	random split	**0.992**
	SA-MTL(OURS)	scaffold split	0.99
	Smiles_Transformer	scaffold split	0.954
	Weave*	random split	0.832
	Transformer_CNN	CV & augmented	0.77

Note: The best results on the test set are highlighted in bold.

Ablation study

- We performed to evaluate the effectiveness of several features in SA-MTL.
- The self-attention module and the multi-task learning scheme are two essential components of our model.
- The first component is just a CNN layer, however, we showed that the CNN layer has significant role for learning the shared factors of multiple tasks.
- The multi-head and the position encoding features did not have a significant impact on chemical compound prediction. More than one multi-head seems to have an over-parameterization issue for a certain task. And the position encoding has limited effects because an atom'position does not convey grammatical meanings.

* Github Repository : https://github.com/arwhirang/sa-mtl

TABLE VII
PERFORMANCE CHANGES BY MODIFYING SEVERAL FEATURES OF OUR MODEL IN THE TOX21 DATASET.

	Modified Features	Average AUC
SA-MTL		0.9
SA-MTL	- Two-Character Embedding	0.888
SA-MTL	- Multi-task Learning	0.871
SA-MTL	- Self Attention Module	0.798
SA-MTL	- CNN	0.824
SA-MTL	CNN<>RNN*	0.895
SA-MTL	Discrete Output Layer<>Max Pooling**	0.865
SA-MTL	+ Multi-head (5)	0.892
SA-MTL	+ Position encoding	0.892

* We experimented by replacing the first CNN layer of our model with an RNN layer.
** We experimented by replacing the discrete output layer of our model with a max pooling layer.

References

1) Honda, S., Shi, S., & Ueda, H. R. : SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. In: arXiv preprint arXiv:1911.04738. (2019).
2) Karpov, P., Godin, G., & Tetko, I. V. : Transformer-CNN: Fast and Reliable tool for QSAR. arXiv preprint arXiv:1911.06603. (2019).
3) Zheng, S., Yan, X., Yang, Y., & Xu, J. : Identifying Structure–Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. In: Journal of chemical information and modeling, 59(2), 914-923. (2019).

Acknowledgement

This study is supported by National Research Council of Science & Technology (NST) grant by the Korea government (MSIP) (No. CAP-17- 01-KIST Europe) and the Basic research grant (12001).