인공지능과 신약개발을 위한 파이썬

14주차 인공지능 기반 신약 개발 최신 연구 소개

홍 성 은

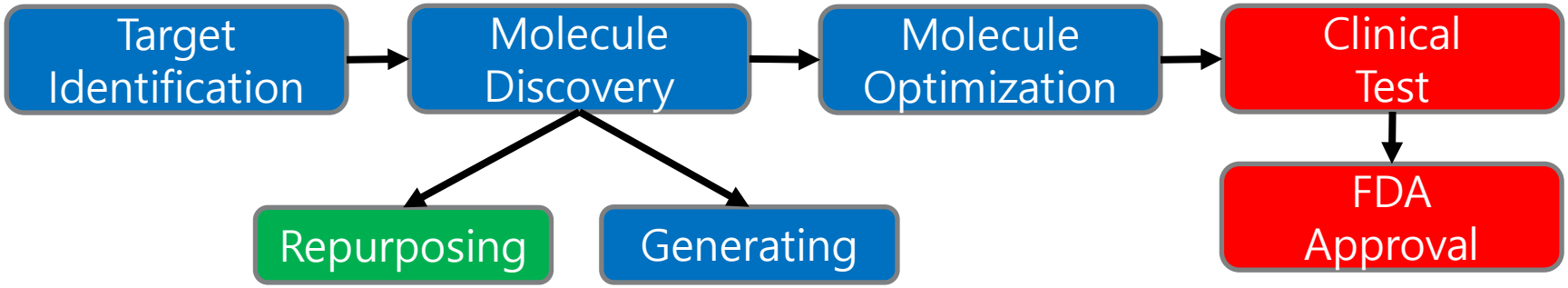
sungkenh@gmail.com

목차

- 개요
- 최신 임베딩 트렌드
- 신약 개발 인공지능 최신 연구
- Example
- GNN

개요

- Drug Discovery Process



개요

- Drug Repurposing
  - 안전함 : 이미 임상을 거친 단계의 약물임(독성 테스트, 약효 입증)
  - 비용 절감 : 임상 1상을 건너뛸 수 있음(임상 비용이 X)

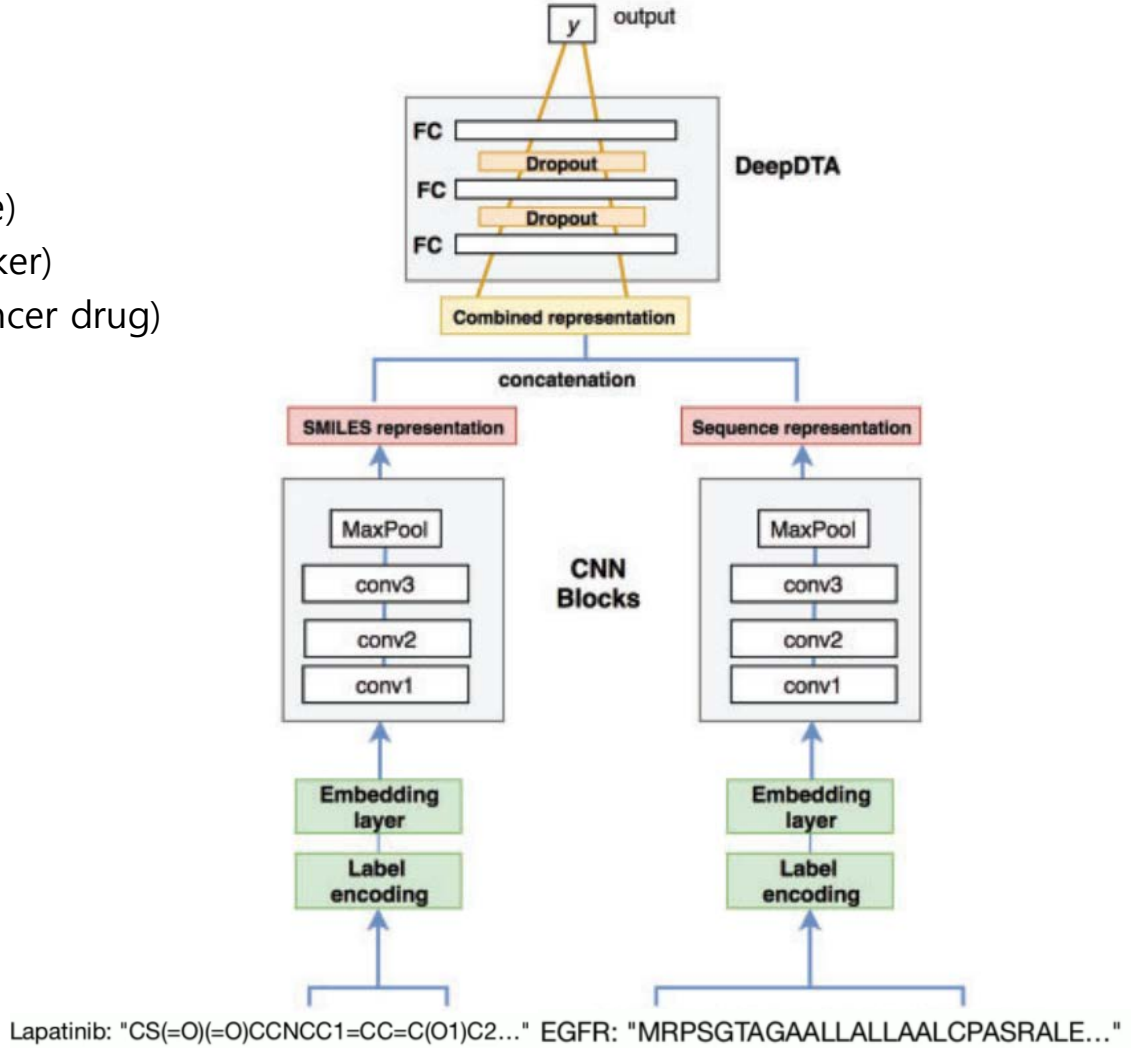
Drug	Original indication	Repurposing/ repurposed for NTD
Astemizole	Antihistamine	Malaria
Avermectin	River blindness and elephantiasis	Tuberculosis
Miltefosine	Antineoplastic	Leishmaniasis
Tamoxifen	Anticancer	Leishmaniasis
Amphotericin B	Antifungal	Leishmaniasis
Eflornithine	Anticancer	African sleeping sickness
Phosphodiesterase-inhibitor analogues	Erectile dysfunction	African sleeping sickness; Chagas's disease

Abbreviation: NTDs, neglected tropical diseases.

Allarakhia, Minna. "Open-source approaches for the repurposing of existing or failed candidate drugs: learning from and applying the lessons across diseases." *Drug design, development and therapy* 7 (2013): 753

개요

- Drug Target Interaction
  - Input : Molecule
  - Target : Protein(Biomarker)
  - Output: Interaction(Affinity Score)
  - Ex): EGFR protein(cancer biomarker)
    Lapatinib Molecule(: anti-cancer drug)

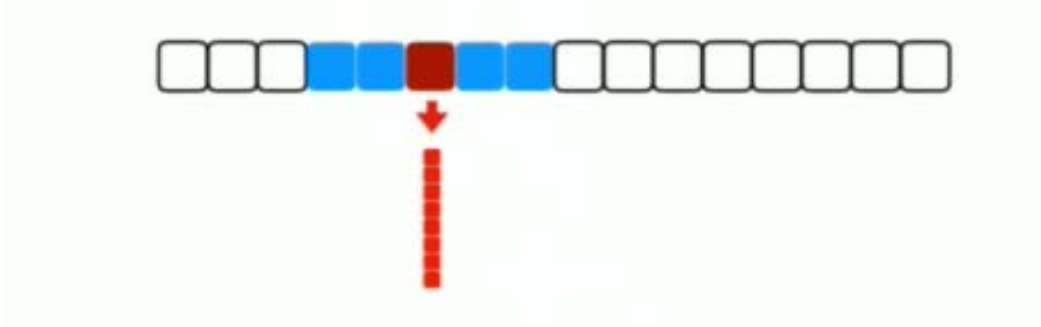


신약 개발 인공지능 최신 연구

- Drug Target Interaction
  - Sequence : SMILES, FASTA and Text
  - Vector Representation
    - One-hot vector
    - Embedding – more information
- 최신 임베딩 트렌드
  - Word2Vec: Local contextual embeddings
  - ELMO : RNN based contextual embeddings
  - Transformer: Attention RNN based contextual embeddigs
  - BERT, GPT : Transformer+ Masked LM

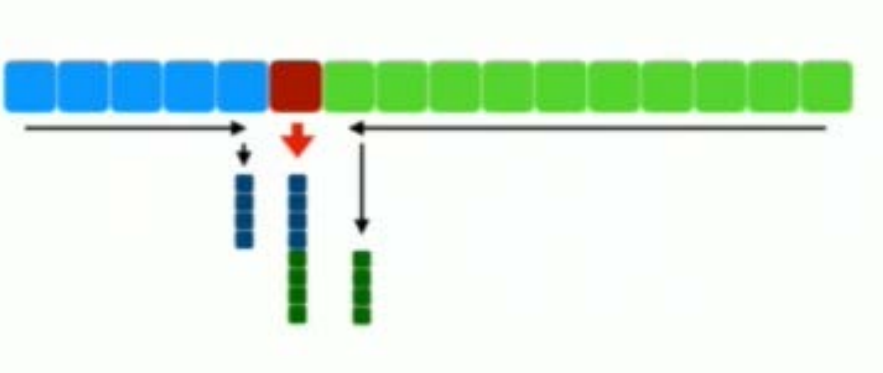
최신 임베딩 트렌드

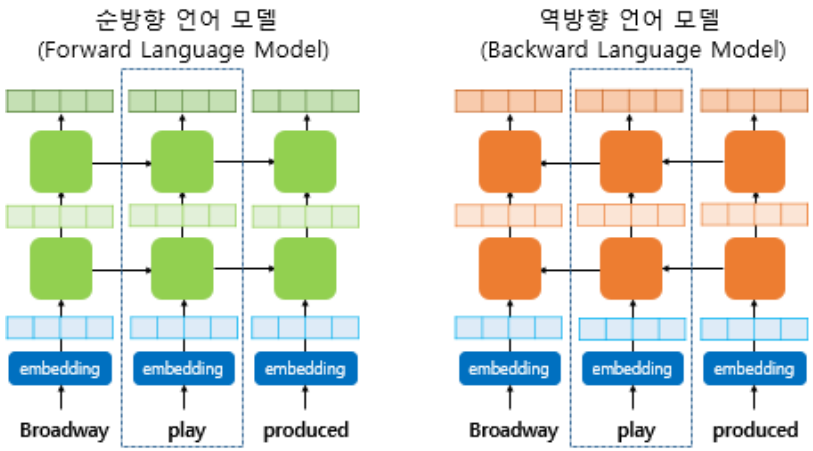
- Word2Vec
  - Word representation : word -> vector
  - 문장의 Local Context를 학습할 때 반영해서 그 정보들이 Vector에 담기도록 함
  - One-hot vector-> Dense Vector
- 단점
  - 더 많은 정보를 고려하기에 어려움이 있음(멀리 떨어진 정보들이 반영이 잘 안됨)



최신 임베딩 트렌드

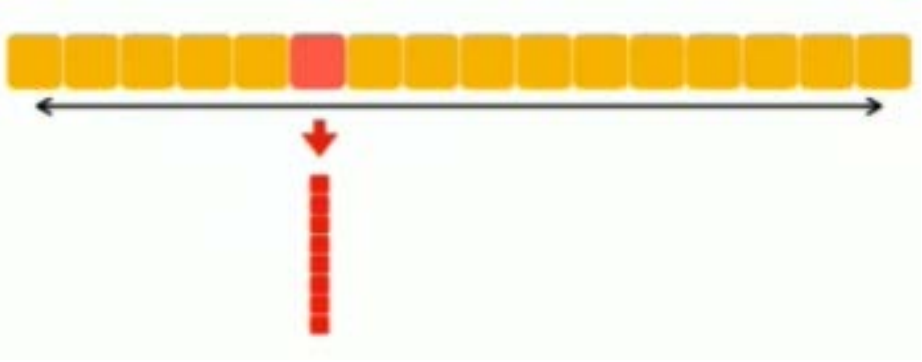
- ELMO
  - 문장의 Context를 반영할 때, 예측하려는 값 주변 뿐만 아니라 문장 전체의 Context를 모두 반영하도록 바꿔보자
  - 2018년에 제안된 새로운 워드 임베딩 방법
  - 같은 표기의 단어라도 문맥에 따라서 다르게 워드 임베딩 해보자
  - Bank라는 단어의 쓰임
    - Bank Account(은행 계좌)와 River Bank(강둑)에서의 Bank는 전혀 다른 의미를 가짐
    - Word embedding은 이를 제대로 반영하지 못함
- 단점
  - 너무 먼 곳의 정보들은 거의 반영이 되지 않음



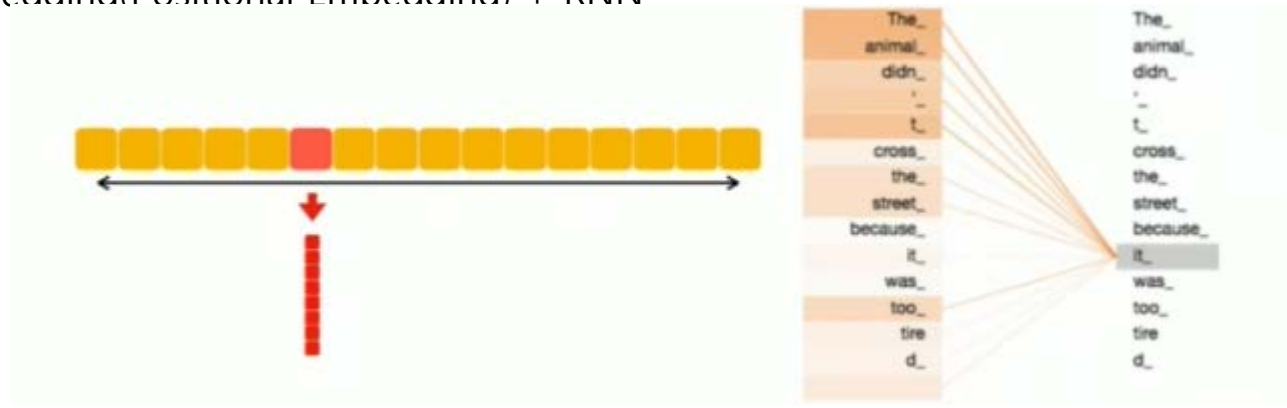


최신 임베딩 트렌드

- Transformer
  - Attention is all you need: https://arxiv.org/pdf/1706.03762.pdf
  - 기존 RNN 기반 Seq2seq모델에서 학습할 때 문장의 길이가 길어지게되면 서로 멀리 떨어진 문장에 대한 정보가 줄어들어 제대로된 예측이 불가능해짐(Long-term dependency problem)
  - 순차적으로 연산을 한다는 부분에서 연산의 병렬화가 불가능해 연산 속도가 저하된다는 단점 역시 있었음
  - 모든 토큰을 weighted sum 함



  - Replaces Embedding(Positional Embedding) + RNN



최신 임베딩 트렌드

- Transformer

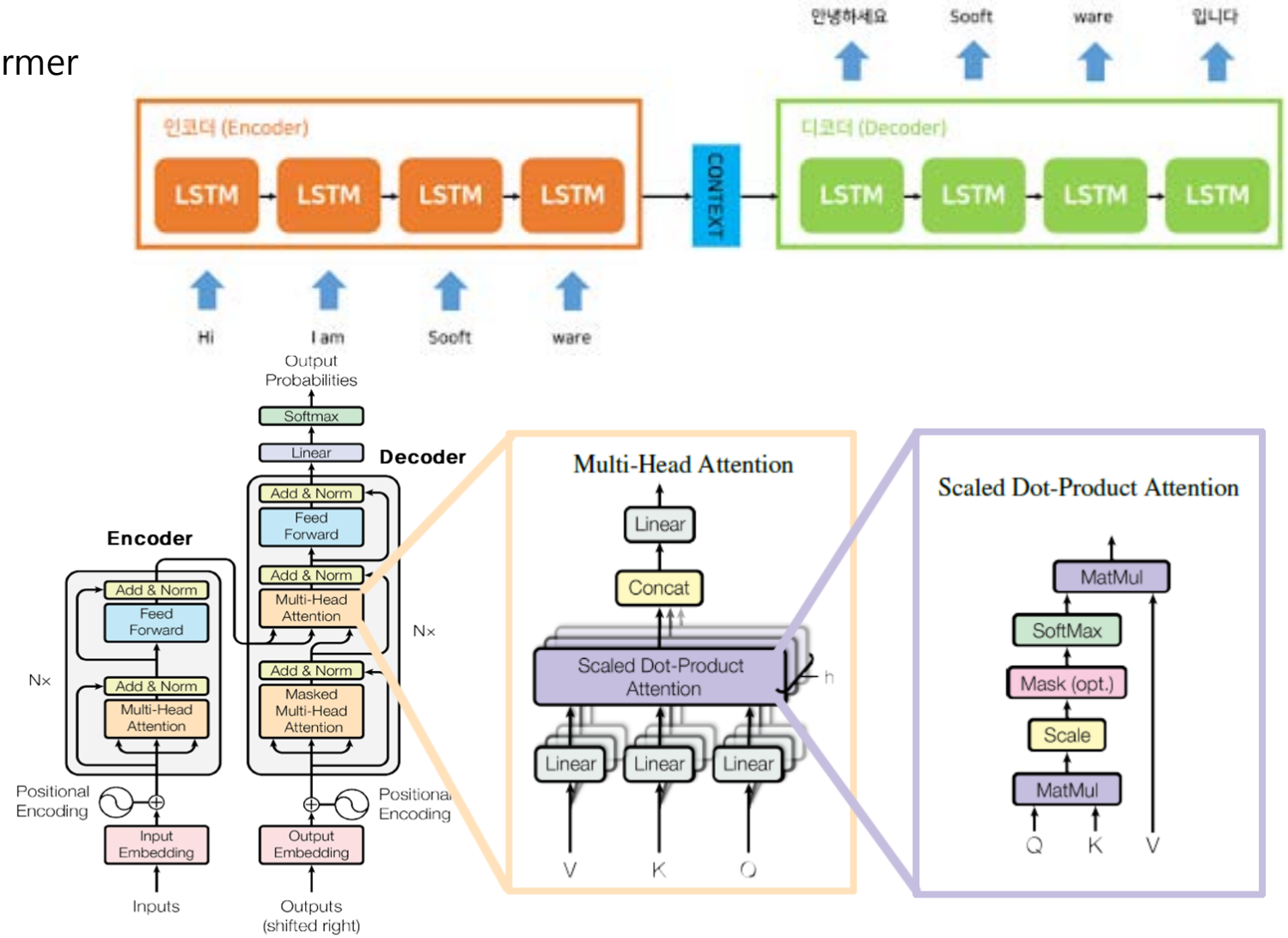
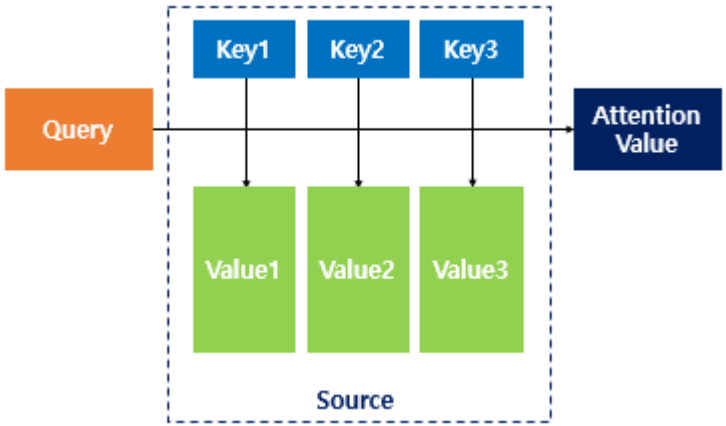


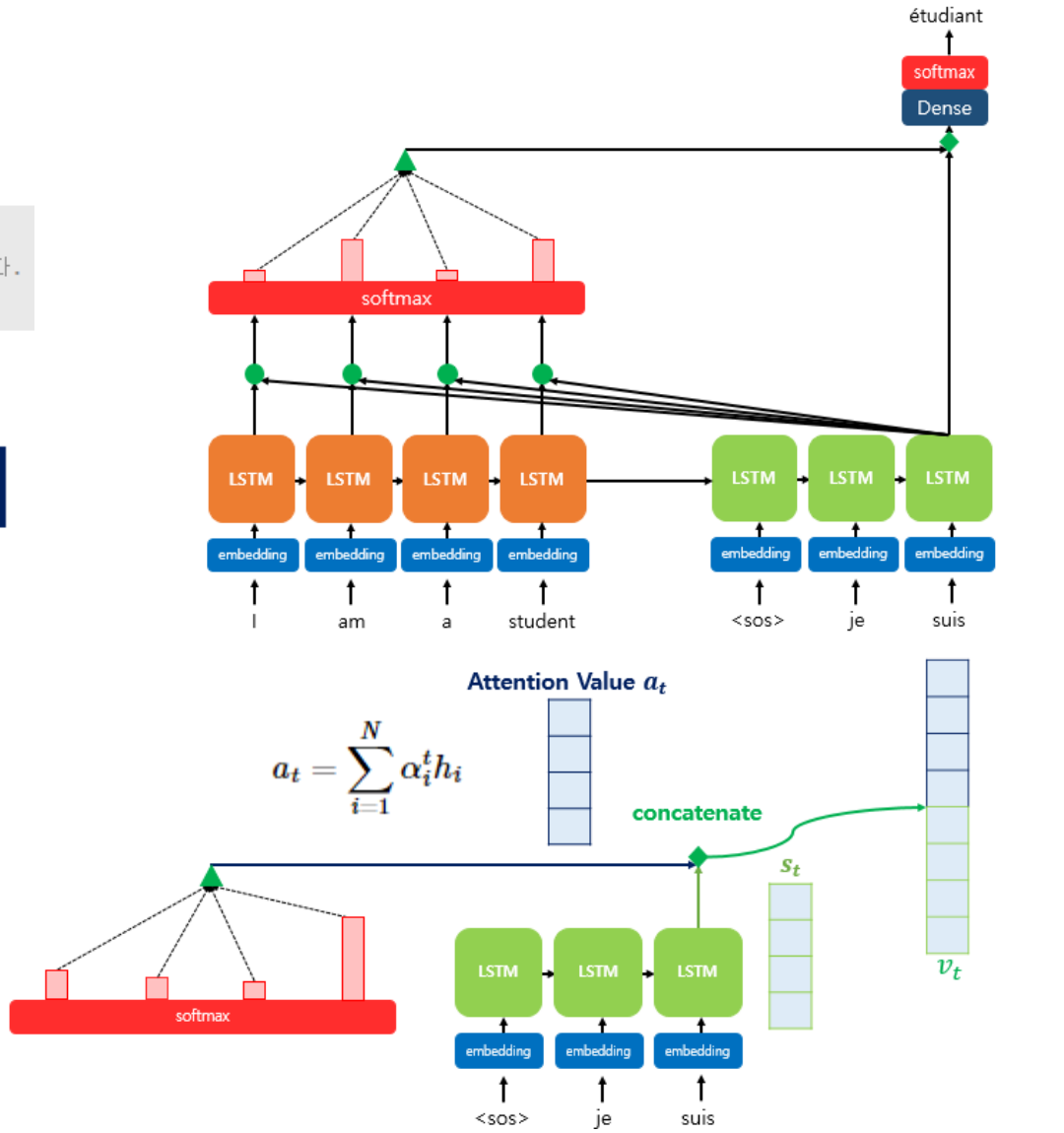
Figure 1: The Transformer - model architecture.

최신 임베딩 트렌드

- Transformer
  - Attention 메커니즘

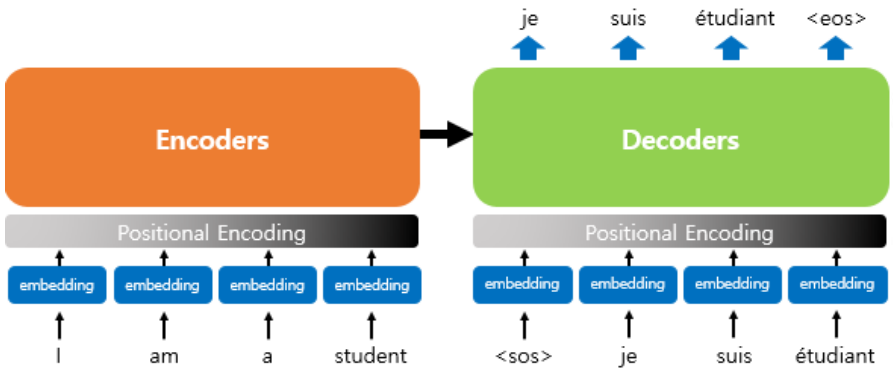
```
# 파이썬의 딕셔너리 자료형을 선언
# 키(Key) : 값(value)의 형식으로 키와 값의 쌍(Pair)을 선언한다.
dict = {"2017" : "Transformer", "2018" : "BERT"}
```





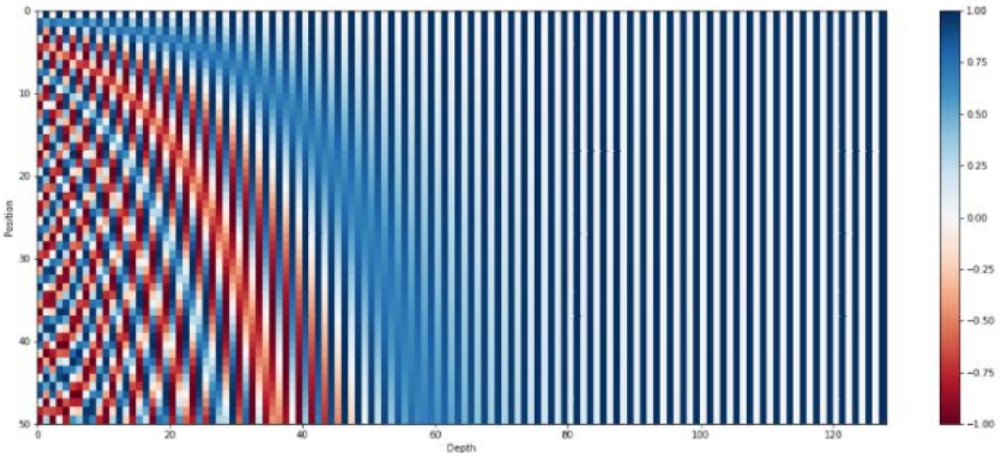
최신 임베딩 트렌드

- Transformer
  - 순차적인 정보를 넣어주기 위한 Positional Embedding



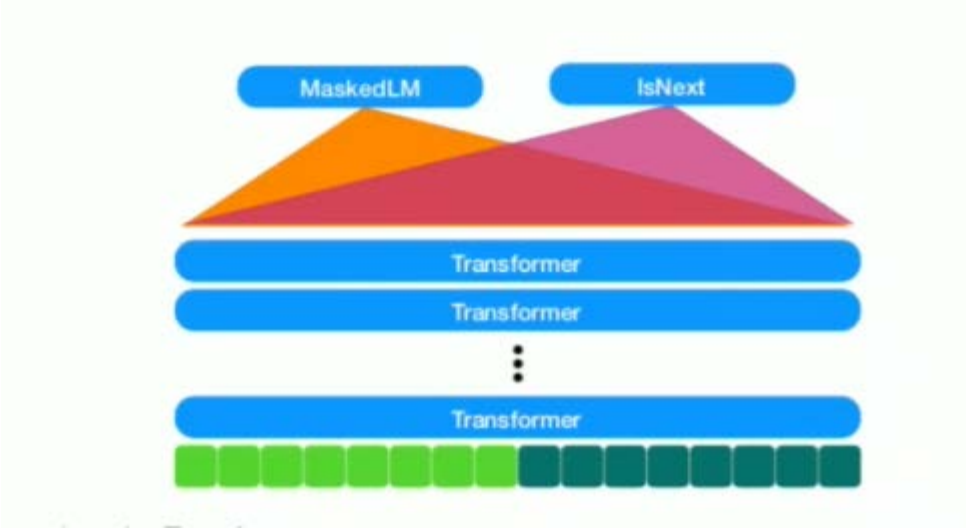
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

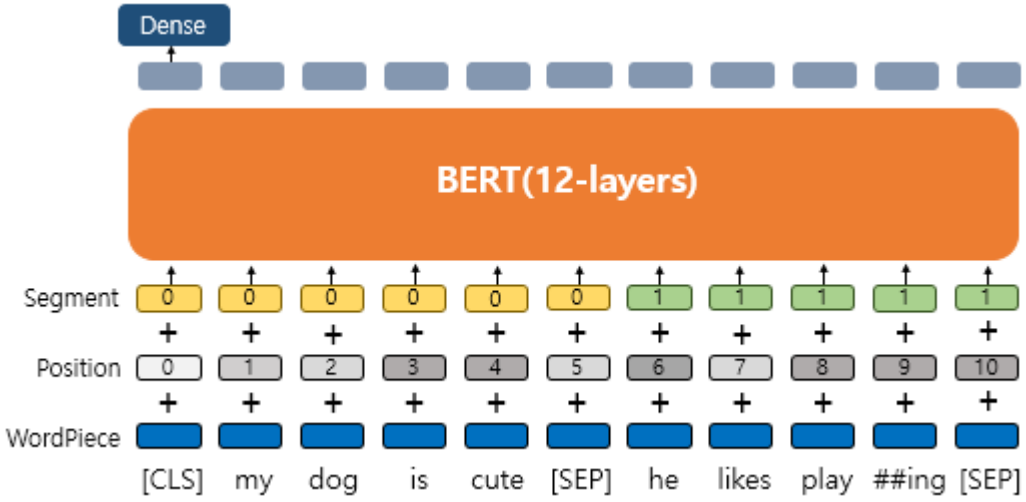
출처 : https://jalammar.github.io/illustrated-transformer/



최신 임베딩 트렌드

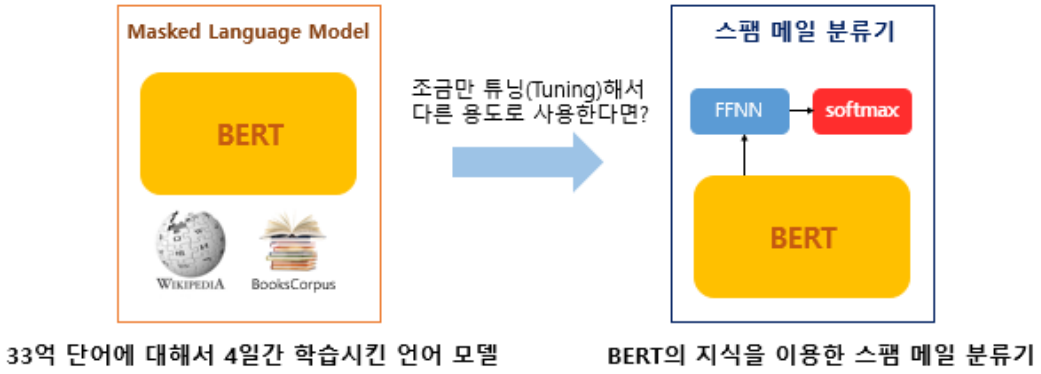
- BERT
  - Transformer + new language model
  - 2개의 문장을 입력 받음





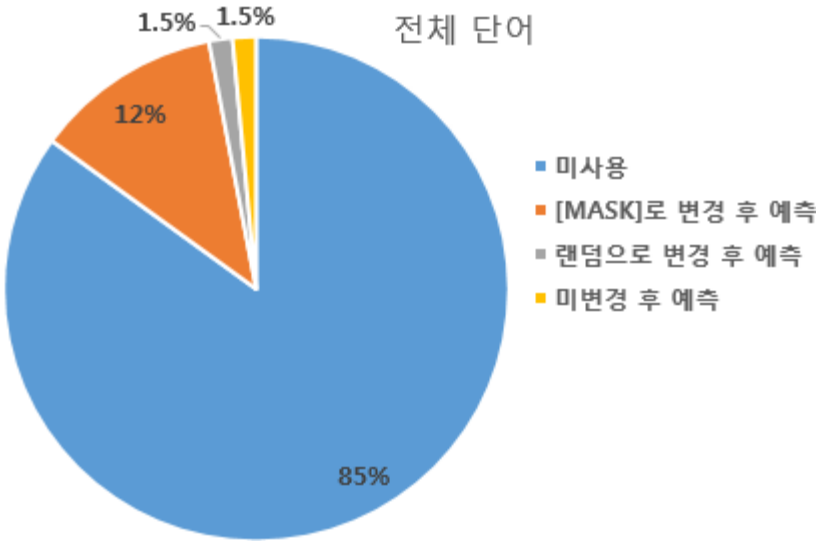
최신 임베딩 트렌드

- BERT
  - BERT는 2018년에 공개되어 등장과 동시에 수많은 NLP 태스크에서 SOTA 달성(11개로 기억)
  - BERT는 사전 훈련을 위해서 인공 신경망의 입력으로 들어가는 입력 텍스트의 15%의 단어를 랜덤으로 마스킹(Masking)
  - Masked LM Task
    - [MASK]만 사용할 경우에는 [MASK] 토큰이 파인 튜닝 단계에서는 나타나지 않으므로 사전 학습 단계와 파인 튜닝 단계에서의 불일치가 발생하는 문제가 있기 때문



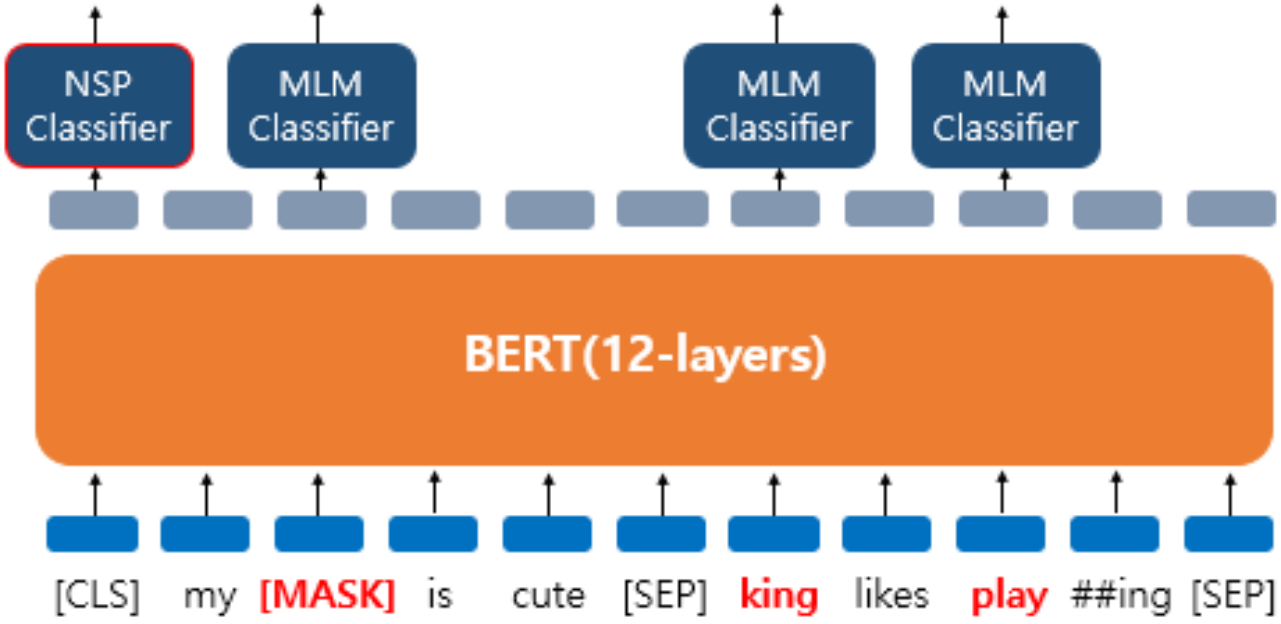
33억 단어에 대해서 4일간 학습시킨 언어 모델

BERT의 지식을 이용한 스팸 메일 분류기



최신 임베딩 트렌드

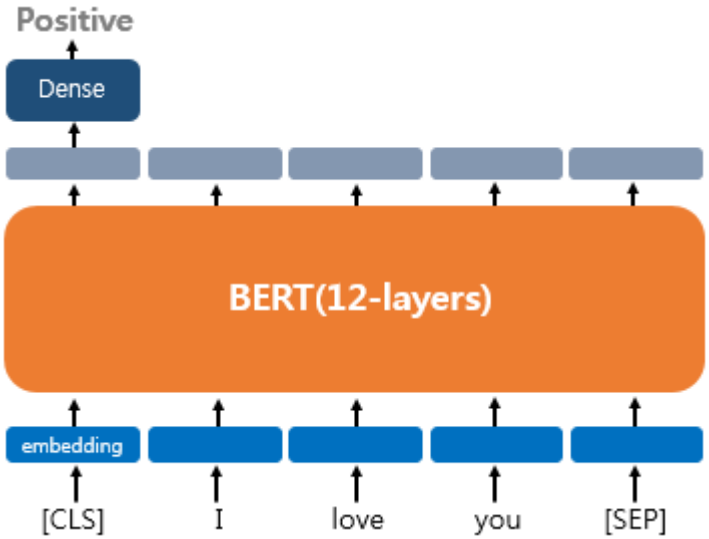
- BERT
  - IsNext Task
    - 50:50 비율로 실제 이어지는 두 개의 문장과 랜덤으로 이어 붙인 두 개의 문장을 주고 훈련



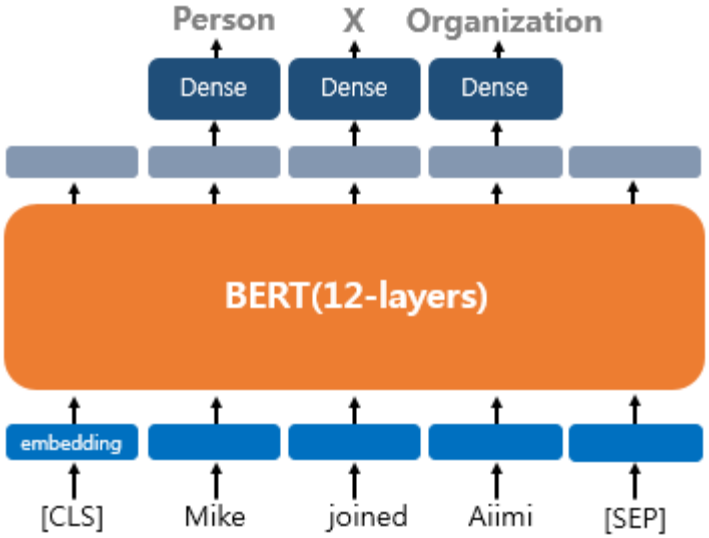
최신 임베딩 트렌드

- BERT
  - BERT 파인 튜닝 하기

Single Text Classification



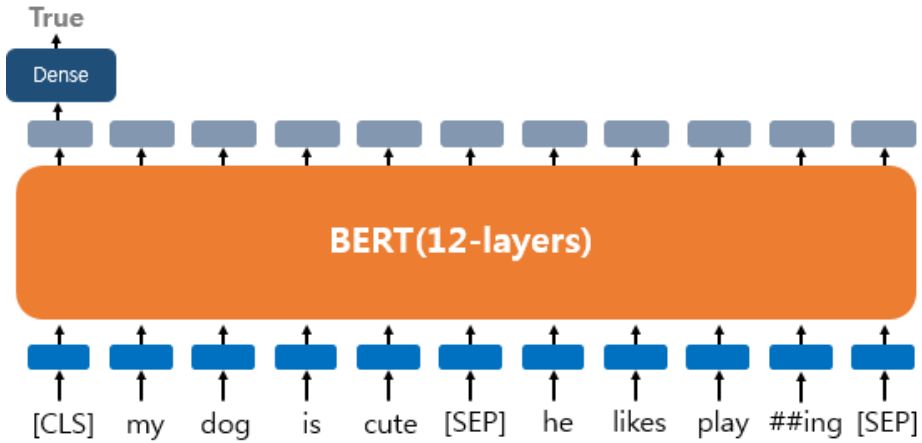
하나의 텍스트에 대한 태깅 작업(Tagging)



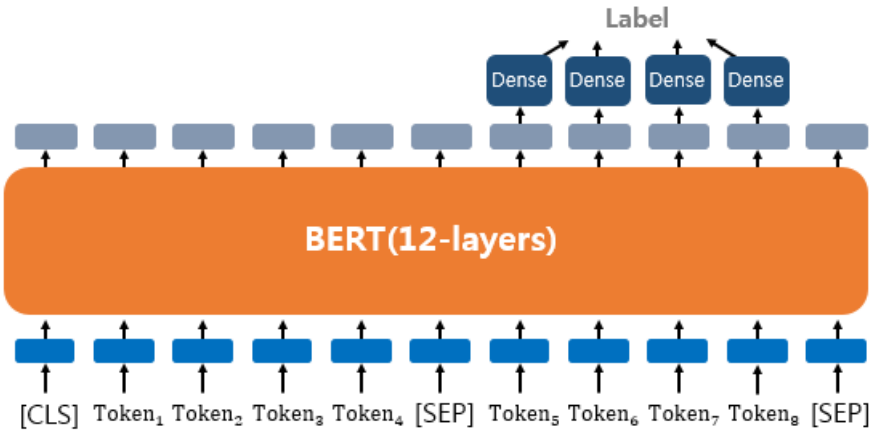
최신 임베딩 트렌드

- BERT
  - BERT 파인 튜닝 하기

텍스트의 쌍에 대한 분류 또는 회귀 문제(Text Pair Classification or Regression)



질의 응답(Question Answering)



신약 개발 인공지능 최신 연구

- Molecule Property Prediction

https://arxiv.org/pdf/2002.08264.pdf

https://dl.acm.org/doi/pdf/10.1145/3307339.3342186

Bioinformatics, 35(23), 2019, 4979–4985
doi: 10.1093/bioinformatics/btz307
Advance Access Publication Date: 9 May 2019
Original Paper

Structural bioinformatics

FP2VEC: a new molecular featurizer for learning molecular properties

Woosung Jeon and Dongsup Kim*

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Yuseong-gu, Daejeon 34141, Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 28, 2018; revised on March 28, 2019; editorial decision on April 22, 2019; accepted on April 24, 2019

Abstract

Motivation: One of the most successful methods for predicting the properties of chemical compounds is the quantitative structure–activity relationship (QSAR) methods. The prediction accuracy of QSAR models has recently been greatly improved by employing deep learning technology. Especially, newly developed molecular featurizers based on graph convolution operations on molecular graphs significantly outperform the conventional extended connectivity fingerprints (ECFP) feature in both classification and regression tasks, indicating that it is critical to develop more effective new featurizers to fully realize the power of deep learning techniques. Motivated by the fact that there is a clear analogy between chemical compounds and natural languages, this work develops a new molecular featurizer, FP2VEC, which represents a chemical compound as a set of trainable embedding vectors.

Results: To implement and test our new featurizer, we build a QSAR model using a simple convolutional neural network (CNN) architecture that has been successfully used for natural language processing tasks such as sentence classification task. By testing our new method on several benchmark datasets, we demonstrate that the combination of FP2VEC and CNN model can achieve competitive results in many QSAR tasks, especially in classification tasks. We also demonstrate that the FP2VEC model is especially effective for multitask learning.

Availability and implementation: FP2VEC is available from https://github.com/wsjeon92/FP2VEC.

Contact: kds@kaist.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, there has been a growing interest in developing machine learning methods to predict the various properties of chemical compounds (Lavecchia, 2015). Quantitative structure–activity relationship (QSAR) models represent one of the most successful methods. The principle behind the QSAR methods is that structurally similar chemicals should have similar properties (Tropsha, 2010). QSAR methods have played a vital role in drug discovery, especially in lead compound generation by virtual screening (Shoichet, 2004) and the drug's ADME (adsorption, distribution, metabolism and excretion) property optimization (Lipinski *et al.*, 2001). Another important application of QSAR methods is computational toxicity prediction, which has been attracting substantial attention recently for an attempt to replace expensive and controversial toxicology experiments on animal models (Luechtefeld *et al.*, 2018).

The prediction accuracy of QSAR models has recently been greatly improved by employing deep learning technology (Capuzzi *et al.*, 2016; Duvenaud *et al.*, 2015; Kearnes *et al.*, 2016; Mayr *et al.*, 2016; Wójcikowski *et al.*, 2018). The advent of deep learning in the drug development field occurred in 2013 when the QSAR machine learning challenge on chemical compound activity in drug discovery organized by Merck (Kaggle challenge) was won by Hinton's group, a pioneer in deep learning technology. They achieved 14% better prediction accuracy over conventional QSAR methods. Since then, deep learning methods for drug development have attracted

Molecule Attention Transformer

Łukasz Maziarka [1,2] Tomasz Danel [1,2] Sławomir Mucha [2] Krzysztof Rataj [1] Jacek Tabor [2] Stanisław Jastrzębski [3,4]

Abstract

Designing a single neural network architecture that performs competitively across a range of molecule property prediction tasks remains largely an open challenge, and its solution may unlock a widespread use of deep learning in the drug discovery industry. To move towards this goal, we propose Molecule Attention Transformer (MAT). Our key innovation is to augment the attention mechanism in Transformer using inter-atomic distances and the molecular graph structure. Experiments show that MAT performs competitively on a diverse set of molecular prediction tasks. Most importantly, with a simple self-supervised pretraining, MAT requires tuning of only a few hyperparameter values to achieve state-of-the-art performance on downstream tasks. Finally, we show that attention weights learned by MAT are interpretable from the chemical point of view.

1. Introduction

The task of predicting properties of a molecule lies at the center of applications such as drug discovery or material design. In particular, estimated 85% drug candidates fail the clinical trials in the United States after a long and costly development process (Wong et al., 2018). Potentially, many of these failures could have been avoided by having correctly predicted a clinically relevant property of a molecule such as its toxicity or bioactivity.

Following the breakthroughs in image (Krizhevsky et al., 2012) and text classification (Vaswani et al., 2017), deep neural networks (DNNs) are expected to revolutionize other fields such as drug discovery or material design (Jr et al., 2019). However, on many molecular property predic tasks DNNs are outperformed by *shallow* models suc support vector machine or random forest (Korotcov e 2017; Wu et al., 2018). On the other hand, while DNN outperform shallow models on some tasks, they tend t difficult to train (Ishiguro et al., 2019; Hu et al., 2019). can require tuning of a large number of hyperparam We also observe both issues on our benchmark (see Se 4.2).

Making deep networks easier to train has been the ce force behind their widespread use. In particular, or the most important breakthroughs in deep learning wa development of initialization methods that allowed to easily deep networks end-to-end (Goodfellow et al., 2 In a similar spirit, our aim is to develop a deep model is simple to use out-of-the-box, and achieves strong pe mance on a wide range of tasks in the field of mole property prediction.

In this paper we propose the Molecule Attention T former (MAT). We adapt Transformer (Devlin et al., 2 to chemical molecules by augmenting the self-atte with inter-atomic distances and molecular graph struc Figure 1 shows the architecture. We demonstrate MAT, in contrast to other tested models, achieves st performance across a wide range of tasks (see Figu Next, we show that self-supervised pre-training furthe proves performance, while drastically reducing the needed for hyperparameter tuning (see Table 3). In t experiments we tuned only the learning rate, testi different values. Finally, we find that MAT has i pretable attention weights. We share pretrained wei at https://github.com/gmum/MAT.

2. Related work

Molecule property prediction. Predicting properti a candidate molecule lies at the heart of many fields as drug discovery and material design. Broadly speal there are two main approaches to predicting molecular erties. First, we can use our knowledge of the underl physics (Lipinski et al., 1997). However, despite re advances (Schütt et al., 2017), current approaches re prohibitively costly to accurately predict many properti

Ardigen, Cracow, Poland.
Jagiellonian University, Cracow, Poland.
Molecule.one, Warsaw, Poland.
New York University, New York, USA.

Correspondence to:
Łukasz Maziarka <lukasz.maziarka@ardigen.com>,
Stanisław Jastrzębski <staszek.jastrzebski@gmail.com>.

Preprint. Work in progress.

SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction

Sheng Wang
University of Texas at Arlington
Arlington, Texas
sheng.wang@mavs.uta.edu

Yuzhi Guo
University of Texas at Arlington
Arlington, Texas
yuzhi.guo@mavs.uta.edu

Yuhong Wang
National Center for Advancing Translating Sciences, NIH
Rockville, Maryland
yuhong.wang@nih.gov

Hongmao Sun
National Center for Advancing Translating Sciences, NIH
Rockville, Maryland
sunh7@mail.nih.gov

Junzhou Huang*
University of Texas at Arlington
Arlington, Texas
jzhuang@uta.edu

ABSTRACT

With the rapid progress of AI in both academia and industry, Deep Learning has been widely introduced into various areas in drug discovery to accelerate its pace and cut R&D costs. Among all the problems in drug discovery, molecular property prediction has been one of the most important problems. Unlike general Deep Learning applications, the scale of labeled data is limited in molecular property prediction. To better solve this problem, Deep Learning methods have started focusing on how to utilize tremendous unlabeled data to improve the prediction performance on small-scale labeled data. In this paper, we propose a semi-supervised model named SMILES-BERT, which consists of attention mechanism based Transformer Layer. A large-scale unlabeled data has been used to pre-train the model through a Masked SMILES Recovery task. Then the pre-trained model could easily be generalized into different molecular property prediction tasks via fine-tuning. In the experiments, the proposed SMILES-BERT outperforms the state-of-the-art methods on all three datasets, showing the effectiveness of our unsupervised pre-training and great generalization capability of the pre-trained model.

CCS CONCEPTS

• **Theory of computation → Semi-supervised learning**; **Structured prediction**; • **Applied computing → Molecular sequence analysis**; **Natural Language Modeling**; *Bioinformatics*.

KEYWORDS

Unsupervised Pre-training; Semi-supervised Learning; Molecular Property Prediction; Natural Language Modeling

*Corresponding author.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342186

ACM Reference Format:
Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3307339.3342186

1 INTRODUCTION

The capability of accurate prediction of molecular properties is an essential key in the chemical and pharmaceutical industries. It benefits various academic areas and industrial applications such as improvement to rational chemical design, reducing R&D cost, decreasing the failure rate in potential drug screening trials, as well as speeding the process of new drug discovery [4]. The key problem of introducing Deep Learning into this area lies on embedding graph-like molecules onto a continuous vector space. Then the representations, as named molecular fingerprints, could be used for various applications such as molecular properties classification, regression, or generating new molecules. Instead of computing a basic property, traditional molecular fingerprints provide a description of a specific part of the molecular structure [27]. However, traditional molecular fingerprints require intensive manual feature engineering and strong domain knowledge. Besides, this kind of fingerprints is highly task-dependent, not general enough for other property prediction tasks [10].

The current success of deep learning in various areas and applications, e.g., image classification [12, 33], video understanding [1, 31, 34], medical imaging [15, 35, 42], and bioinformatics [39, 41], demonstrates that deep learning is a powerful tool in learning feature from data and good at task-related prediction. An increasing number of publications have introduced deep learning into molecular fingerprint learning [3, 39–41]. The models being introduced rely on two main deep learning structures: Recurrent Neural Networks (RNNs) [30] and Graph Convolutional Networks (GCNs) [17, 18]. For RNNs-based methods, molecules are represented as strings by Simplified Molecular-Input Line-Entry system (SMILES). In this way, the current successful models in natural language modeling could be utilized to extract high-quality

신약 개발 인공지능 최신 연구

- Molecule Property Prediction
  - FP2VEC
  - MAT
  - SMILES-BERT

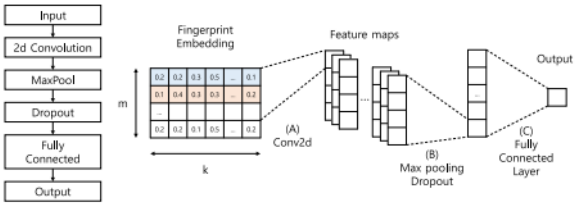
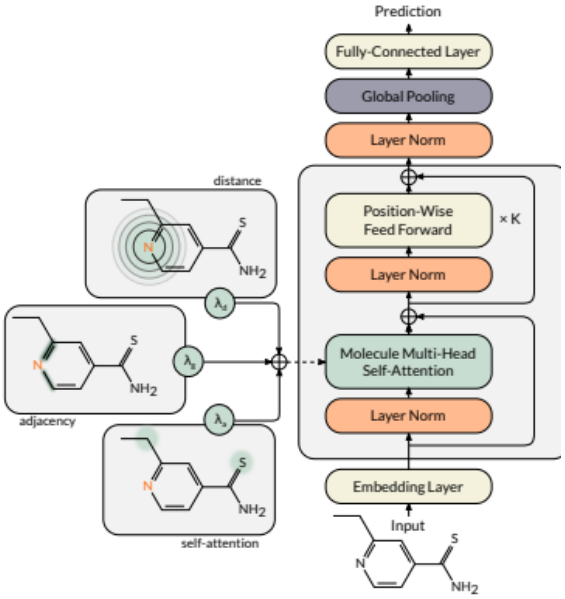


Fig. 2. (Left) The network structure and (right) data flow of the QSAR model. The model uses a padded fingerprint embedding matrix as input data. The model consists of (**A**) a two-dimensional convolutional layer, (**B**) a max pooling and dropout layer and (**C**) a fully connected layer.



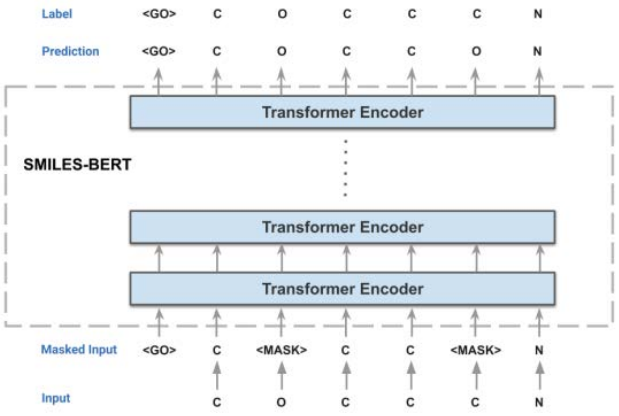


Figure 3: SMILES-BERT: pre-training stage.

신약 개발 인공지능 최신 연구

- Drug Target Interaction

https://arxiv.org/abs/1801.10193

Bioinformatics, 34, 2018, i821–i829
doi: 10.1093/bioinformatics/bty593
ECCB 2018

OXFORD

DeepDTA: deep drug–target binding affinity prediction

Hakime Öztürk[1], Arzucan Özgür[1,*] and Elif Ozkirimli[2,*]

[1]Department of Computer Engineering and [2]Department of Chemical Engineering, Bogazici University, Istanbul 34342, Turkey

*To whom correspondence should be addressed.

Abstract

Motivation: The identification of novel drug–target (DT) interactions is a substantial part of the drug discovery process. Most of the computational methods that have been proposed to predict DT interactions have focused on binary classification, where the goal is to determine whether a DT pair interacts or not. However, protein–ligand interactions assume a continuum of binding strength values, also called binding affinity and predicting this value still remains a challenge. The increase in the affinity data available in DT knowledge-bases allows the use of advanced learning techniques such as deep learning architectures in the prediction of binding affinities. In this study, we propose a deep-learning based model that uses only sequence information of both targets and drugs to predict DT interaction binding affinities. The few studies that focus on DT binding affinity prediction use either 3D structures of protein–ligand complexes or 2D features of compounds. One novel approach used in this work is the modeling of protein sequences and compound 1D representations with convolutional neural networks (CNNs).

Results: The results show that the proposed deep learning based model that uses the 1D representations of targets and drugs is an effective approach for drug target binding affinity prediction. The model in which high-level representations of a drug and a target are constructed via CNNs achieved the best Concordance Index (CI) performance in one of our larger benchmark datasets, outperforming the KronRLS algorithm and SimBoost, a state-of-the-art method for DT binding affinity prediction.

Availability and implementation: https://github.com/hkmztrk/DeepDTA

Contact: arzucan.ozgur@boun.edu.tr or elif.ozkirimli@boun.edu.tr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The successful identification of drug–target interactions (DTI) is a critical step in drug discovery. As the field of drug discovery expands with the discovery of new drugs, repurposing of existing drugs and identification of novel interacting partners for approved drugs is also gaining interest (Oprea and Mestres, 2012). Until recently, DTI prediction was approached as a binary classification problem (Bleakley and Yamanishi, 2009; Cao *et al.*, 2014, 2012; Cobanoglu *et al.*, 2013; Gönen, 2012; Öztürk *et al.*, 2016; Yamanishi *et al.*, 2008; van Laarhoven *et al.*, 2011), neglecting an important piece of information about protein–ligand interactions, namely the binding affinity values. Binding affinity provides information on the strength of the interaction between a drug–target (DT) pair and it is usually expressed in measures such as dissociation constant (K_d), inhibition constant (K_i) or the half maximal inhibitory concentration (IC_{50}). IC_{50} depends on the concentration of the target and ligand (Cer *et al.*, 2009) and low IC_{50} values signal strong binding. Similarly, low K_i values indicate high binding affinity. K_d and K_i values are usually represented in terms of pK_d or pK_i, the negative logarithm of the dissociation or inhibition constants.

In binary classification based DTI prediction studies, construction of the datasets constitutes a major step, since designation of the negative (not-binding) samples directly affects the performance of the model. As of last decade, most of the DTI studies utilized four major datasets by Yamanishi *et al.* (2008) in which DT pairs with no known binding information are treated as negative (not-binding) samples. Recently, DTI studies that rely on databases with binding affinity information have been providing more realistic binary datasets created with a chosen binding affinity threshold value (Wan and Zeng, 2016). Formulating the DT prediction task as a binding affinity prediction problem enables the creation of more realistic datasets, where the binding affinity scores are directly used.

i821

https://arxiv.org/pdf/2107.06099.pdf

Bioinformatics
Manuscript Category

OXFORD

Subject Section

Drug-Target Interaction Prediction with Graph Attention networks

Haiyang Wang [†,1], Guangyu Zhou [†,2], Siqi Liu [2], Jyun-Yu Jiang [2] and Wei Wang [2,*]

[1]Zhiyuan College, Shanghai Jiao Tong University, Shanghai, 200240, China and
[2]Department of Computer Science, University of California, Los Angeles, 90095, USA

* To whom correspondence should be addressed. † These authors contributed equally to this work.

Abstract

Motivation: Predicting Drug-Target Interaction (DTI) is a well-studied topic in bioinformatics due to its relevance in the fields of proteomics and pharmaceutical research. Although many machine learning methods have been successfully applied in this task, few of them aim at leveraging the inherent heterogeneous graph structure in the DTI network to address the challenge. For better learning and interpreting the DTI topological structure and the similarity, it is desirable to have methods specifically for predicting interactions from the graph structure.

Results: We present an end-to-end framework, DTI-GAT (**D**rug-**T**arget **I**nteraction prediction with **G**raph **AT**tention networks) for DTI predictions. DTI-GAT incorporates a deep neural network architecture that operates on graph-structured data with the attention mechanism, which leverages both the interaction patterns and the features of drug and protein sequences. DTI-GAT facilitates the interpretation of the DTI topological structure by assigning different attention weight to each node with the self-attention mechanism. Experimental evaluations show that DTI-GAT outperforms various state-of-the-art systems on the binary DTI prediction problem. Moreover, the independent study results further demonstrate that our model can be generalized better than other conventional methods.

Availability: The source code and all datasets are available at https://github.com/Haiyang-W/DTI-GRAPH

Contact: wanghaiyang@stu.pku.edu.edu, weiwang@cs.ucla.edu

1 Introduction

Detecting drug–target interactions (DTIs) potentially facilitates therapeutic target identification (Xia *et al.*, 2010; Petta *et al.*, 2016) and novel drug design (Skrabanek *et al.*, 2008; AY *et al.*, 2007; Janga and Tzakos, 2009; Kuhn *et al.*, 2008). Until quite recently, pharmacological effects were often discovered using primitive trial and error procedures, such as applying plant extracts on living systems and observing the outcomes (Singh *et al.*, 2016). However, experiment-based methods remain expensive, labor-intensive and time-consuming (Dickson and Gagnon, 2004; Kola and Landis, 2004; Kapetanovic, 2008). Evidently, there is an immense need for reliable computational approaches to identify and characterize DTIs, hoping to accelerate the pace and reduce the cost of drug development.

With the rapid development of machine learning techniques, various computational prediction approaches have been proposed to predict drug-target interactions. Yamanishi *et al.* (2010) integrated the relationship among the pharmacological space, the chemical space, and the topology of drug–target interaction networks to predict the associations between drugs and targets, and their experimental results have demonstrated that drug–target interactions are more correlated with pharmacological effect similarity than with chemical structure similarity. According to the similarity of chemical information, Keiser *et al.* (2009) proposed a method to explore the associations between drugs and targets. They selected 30 of predicted results for biological experiments and finally confirmed 23 with interrelationships. Wang *et al.* (2010) used supervised machine learning methods to predict the relationship between drugs and targets. To solve the problem of sample imbalance, they are collecting the positive samples from the database, and the negative samples using the random selection method. The input features of the classifier consist of the chemical structure of the drug and the sequence information of the protein. Chen *et al.* (2012) developed a novel method of Network-based Random Walk with Restart on the Heterogeneous (NRWRH) network to predict potential drug–target interactions on a large scale. The excellent experimental results show that the proposed method can discover new potential drug–target interactions for drug development. These approaches provide feasible solutions to the problem. However, the extracted features used in these approaches only

1

https://openreview.net/pdf?id=Zqf6RGp5lqf

Under review as a conference paper at ICLR 2021

MODELLING DRUG-TARGET BINDING AFFINITY USING A BERT BASED GRAPH NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the interactions between novel drugs and target proteins is fundamentally important in disease research as discovering drug-protein interactions can be an exceptionally time-consuming and expensive process. Alternatively, this process can be simulated using modern deep learning methods that have the potential of utilising vast quantities of data to reduce the cost and time required to provide accurate predictions. In this paper, we seek to leverage a set of BERT-style models that have been pre-trained on vast quantities of both protein and drug data. The encodings produced by each model are then utilised as node representations for a graph convolutional neural network, which in turn models the interactions without the need to simultaneously fine-tune both protein and drug BERT models to the task. We evaluate the performance of our approach on two drug-target interaction datasets that were previously used as benchmarks in recent work. Our results significantly improve upon a vanilla BERT baseline approach as well as the former state-of-the-art methods for each task dataset. Our approach builds upon past work in two key areas; firstly, we take full advantage of two large pre-trained BERT models that provide improved representations of task-relevant properties of both drugs and proteins. Secondly, inspired by work in natural language processing that investigates how linguistic structure is represented in such models, we perform interpretability analyses that allow us to locate functionally-relevant areas of interest within each drug and protein. By modelling the drug-target interactions as a graph as opposed to a set of isolated interactions, we demonstrate the benefits of combining large pre-trained models and a graph neural network to make state-of-the-art predictions on drug-target binding affinity.

1 INTRODUCTION

Developing personalised medicine has been at the forefront of recent disease research, which has been accelerated with vast quantities of data being generated and refined in laboratories across the world. As new drugs and proteins are regularly being produced and discovered, it is becoming ever more challenging to utilise this data correctly and gain an understanding of the biological systems that operate within complex diseases. Modern disease research and drug discovery require new methods that can capitalise on the information that is available within these vast resources and in turn, channel this knowledge towards improving drug-target interaction simulations.

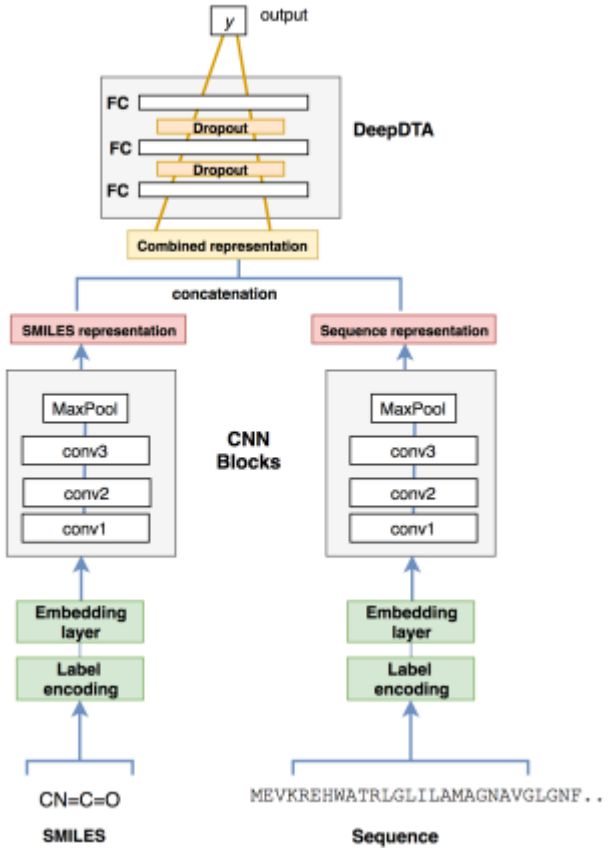
Deep learning has the potential to address the concerns of modelling such complex heterogeneous data. In recent years, deep learning has been used to model Drug-Target Interactions (DTIs) as it is ideally suited to handle large datasets without requiring feature engineering. By using deep learning to map out the drug-target landscape, one can quickly identify the proteins that are targeted by each drug – thereby accelerating drug discovery during clinical trails (Santos et al., 2017). Initial applications of machine learning models posed this as a classification problem due to the variability between each interaction pair (Bleakley & Yamanishi, 2009; Cao et al., 2014; Öztürk et al., 2016). However, these early approaches do not provide enough information about the actual binding affinity value, which is troublesome when one seeks to learn the potency of a particular drug-target pair. Deep learning now plays an important role in determining patterns in complex drug-target systems. Applications of deep learning are becoming ubiquitous in drug-drug interaction modelling (Ryu et al., 2018; Liu et al., 2016), as well as forming predictions for protein-protein interactions (Sun et al.,

20

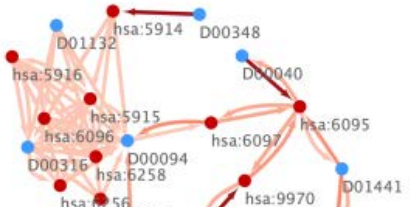
신약 개발 인공지능 최신 연구

- Drug Target Interaction

Deep DTA



DTI-GAT



BERT

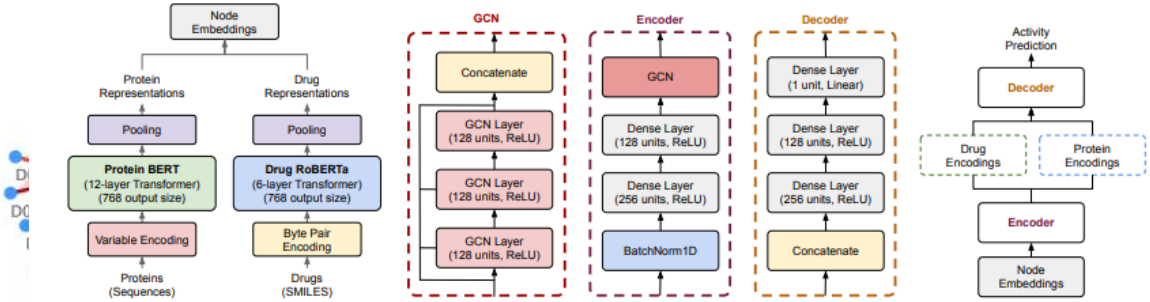


Figure 1: Overview of the BERT-GCN Approach.

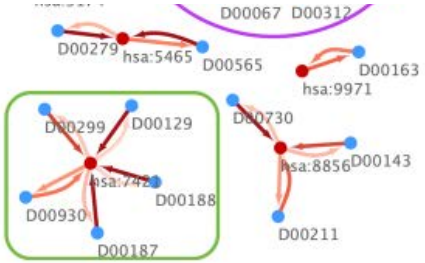
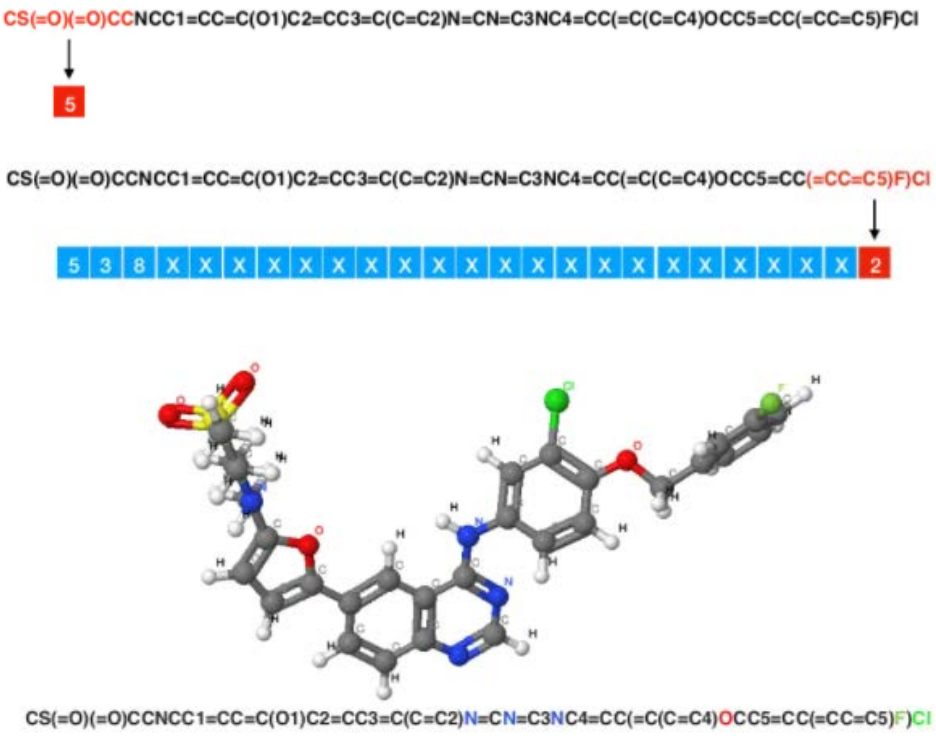
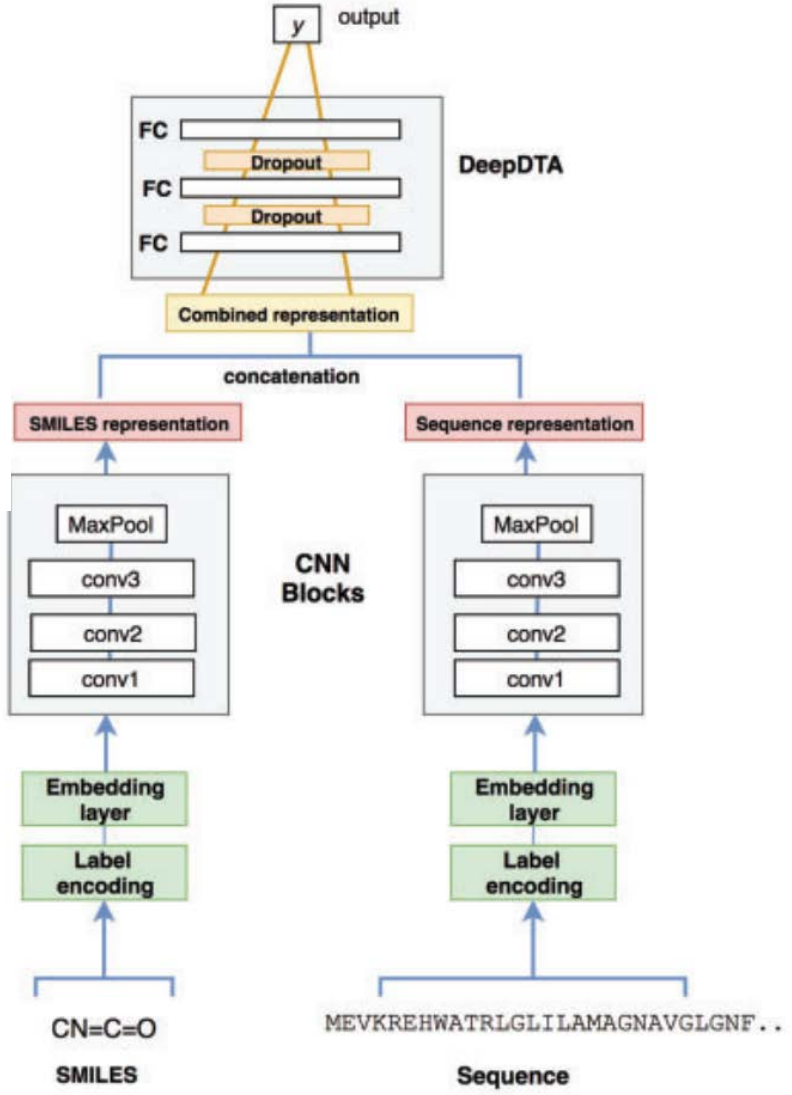


Fig. 5: Attention graph of DTI-GAT on NUC.

신약 개발 인공지능 최신 연구

- DeepDTA
  - CNN의 약점
    - 시퀀스 상에는 인접하게 표현되나
    - 3D표현상에서는 거리가 먼데 그런 부분이 표현이 안됨

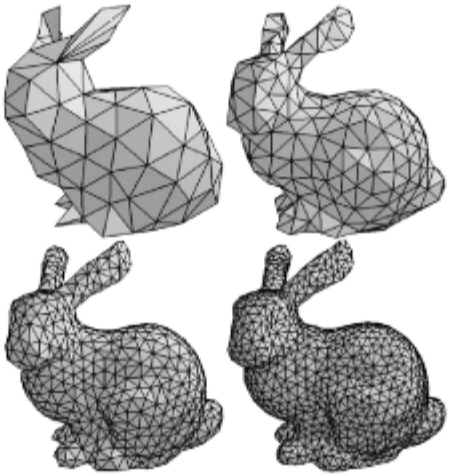




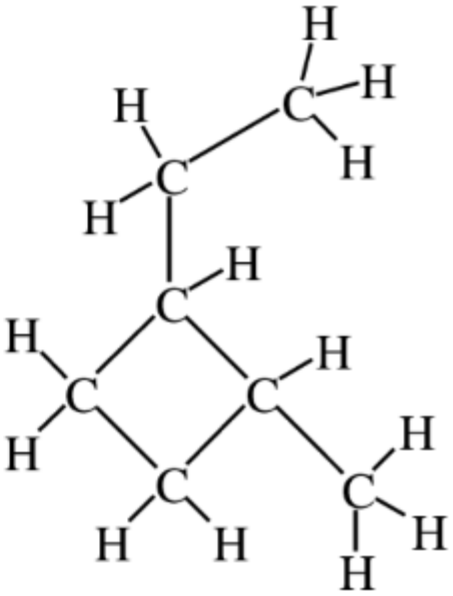
GNN

- Graph란?

Social Graph



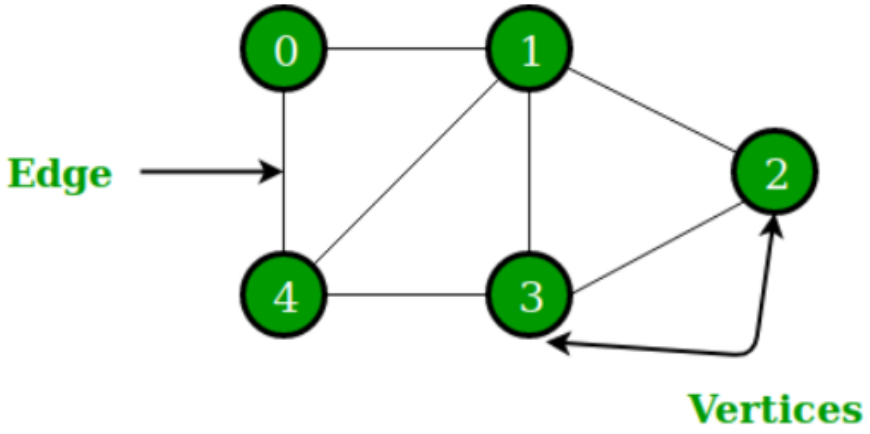
3D Mesh



Molecular Graph

GNN

- Graph란?

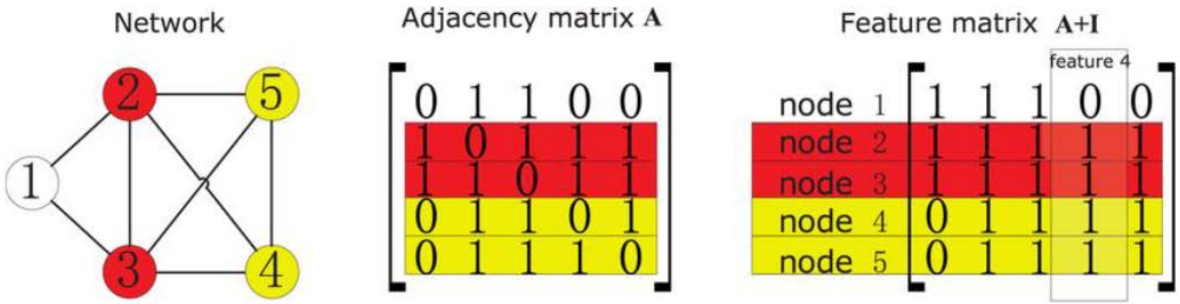


Vertex (Node)

Edge

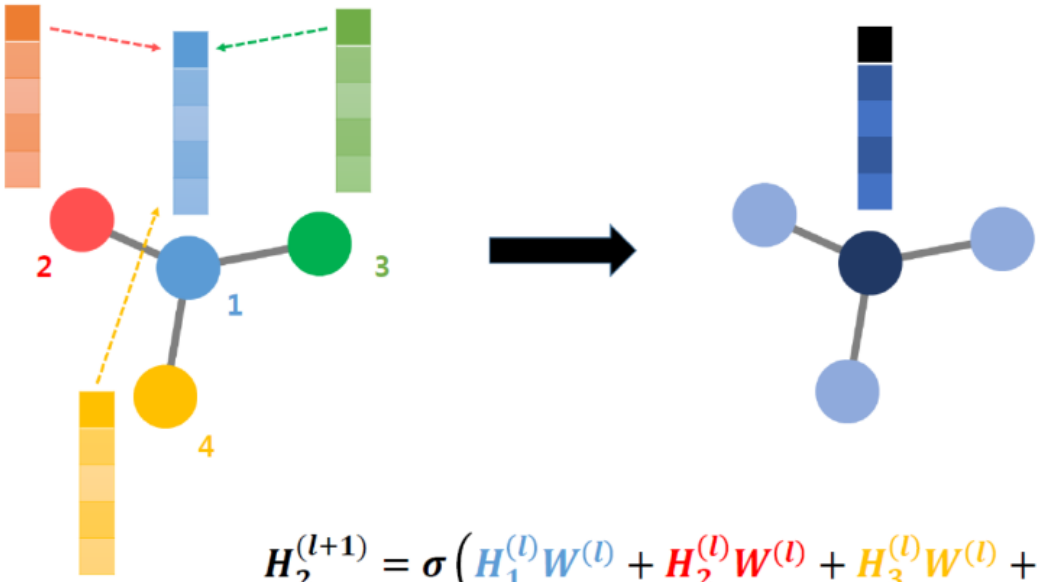
Vertex (Node)
Node Feature Matrix

Edge
Adjacency Matrix



GNN

- GCN의 연산 과정
  - Weight sharing
  - Graph의 Local feature를 하나의 뉴런이 학습함

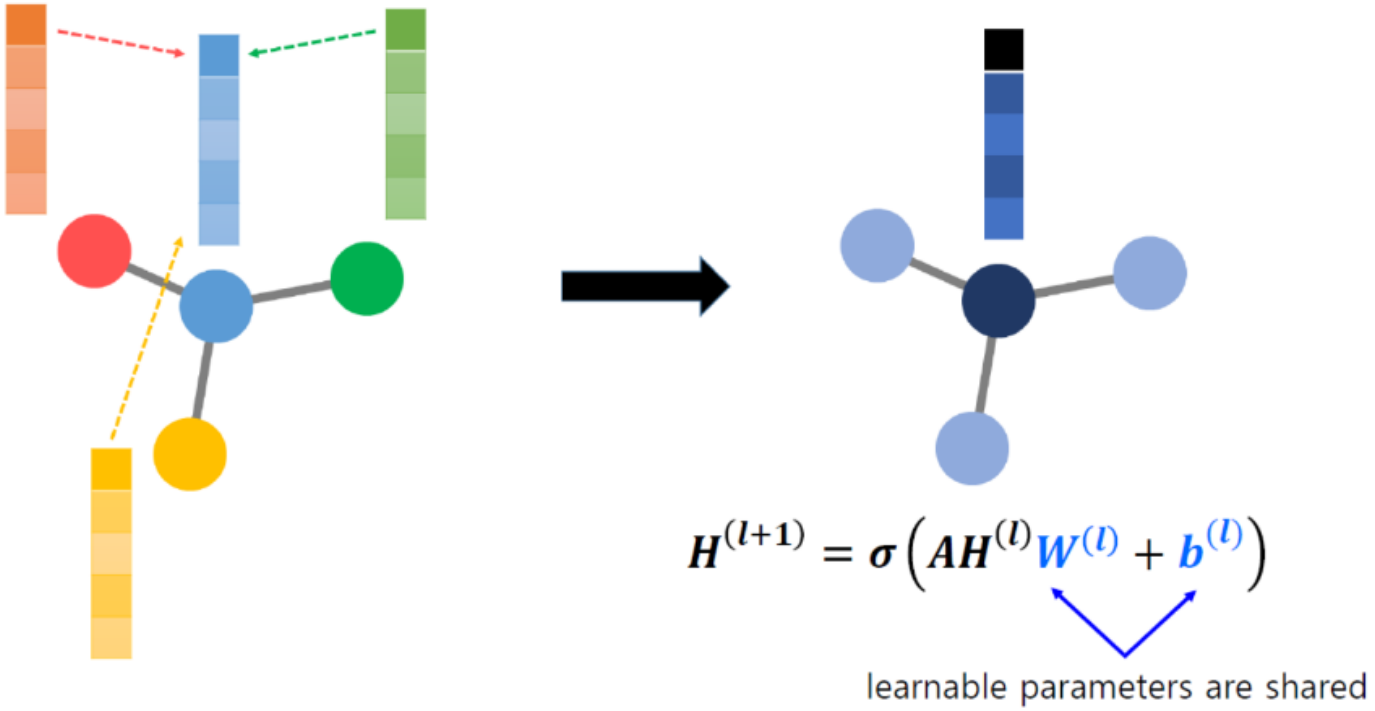


$$H_2^{(l+1)} = \sigma\left(H_1^{(l)}W^{(l)} + H_2^{(l)}W^{(l)} + H_3^{(l)}W^{(l)} + H_4^{(l)}W^{(l)} + b^{(l)}\right)$$

$$\rightarrow \quad H_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} H_j^{(l)}W^{(l)} + b^{(l)}\right)$$

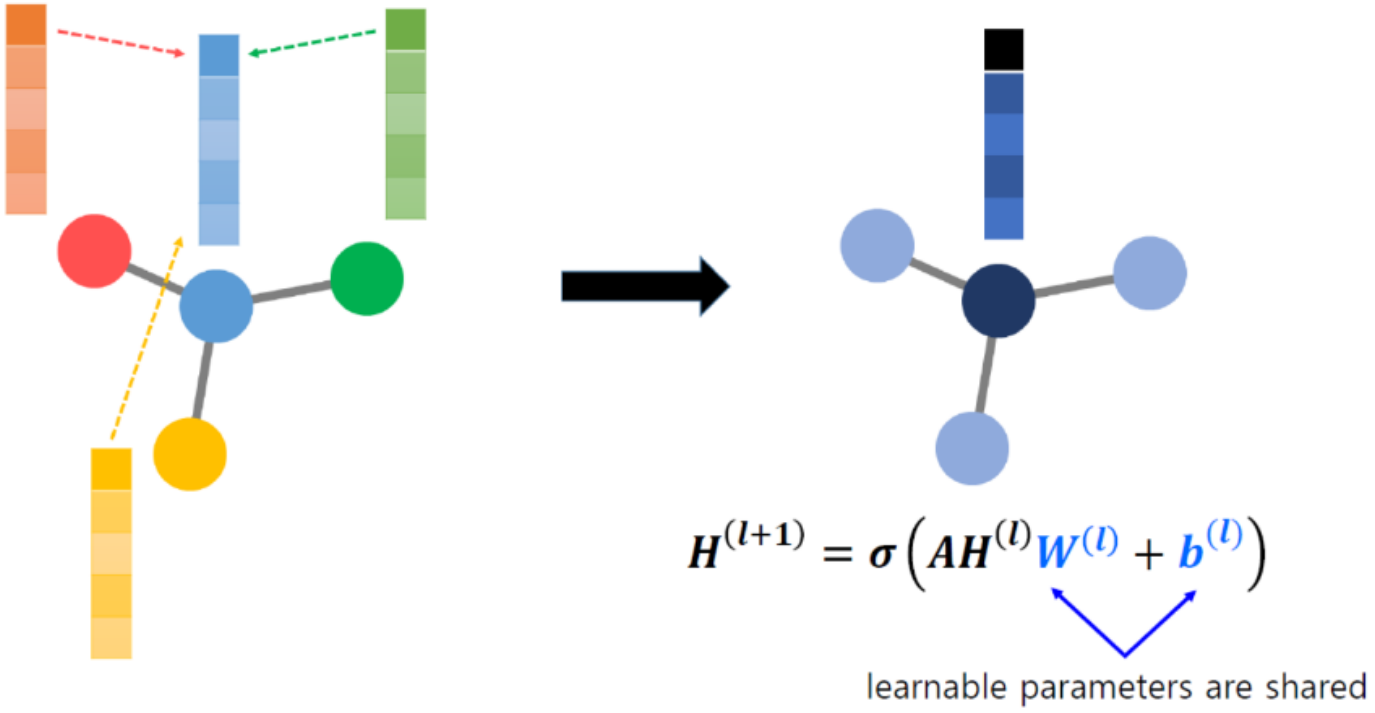
GNN

- GCN의 연산 과정
  - Weight sharing
  - Graph의 Local feature를 하나의 뉴런이 학습함



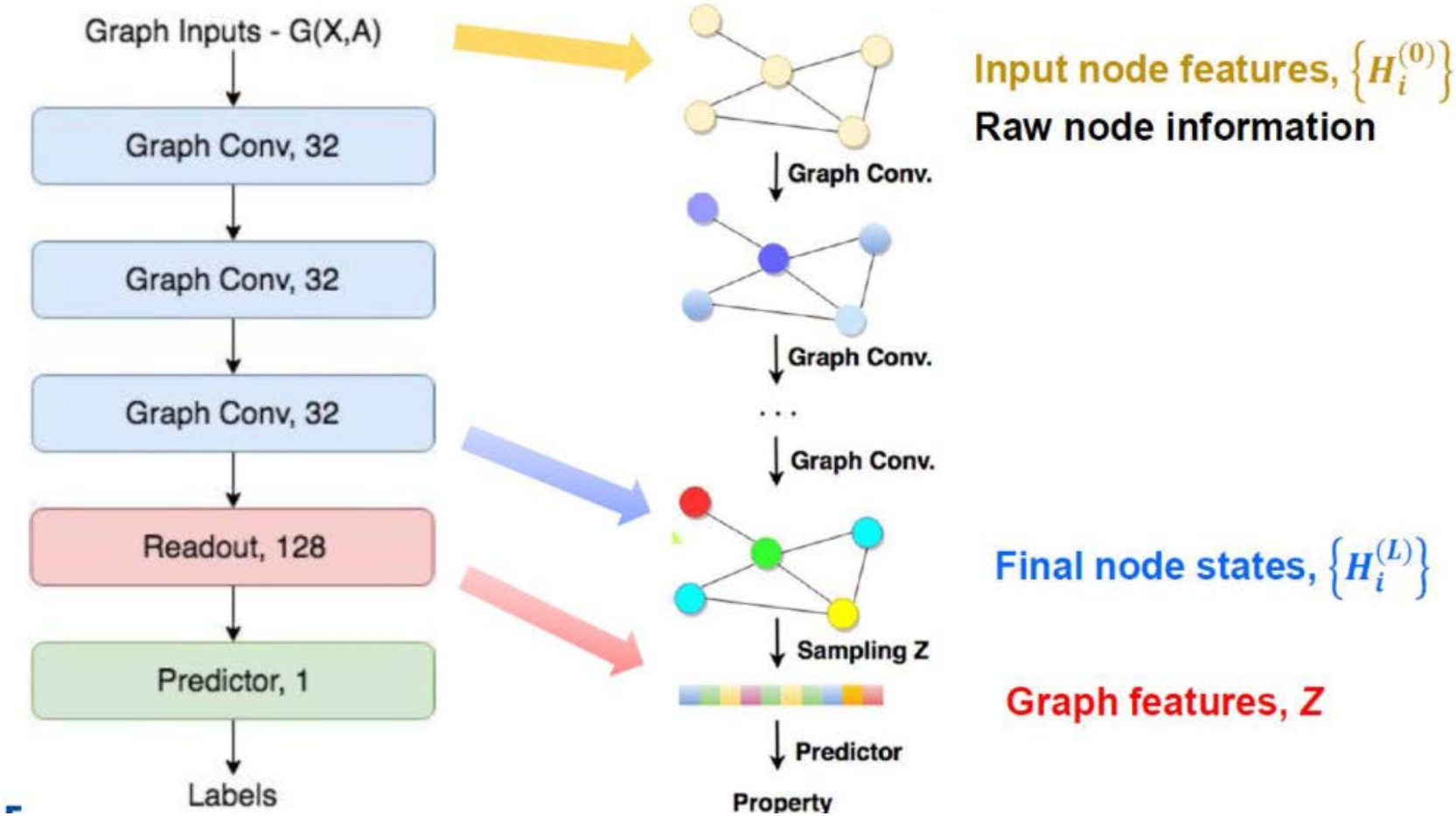
GNN

- GCN의 연산 과정



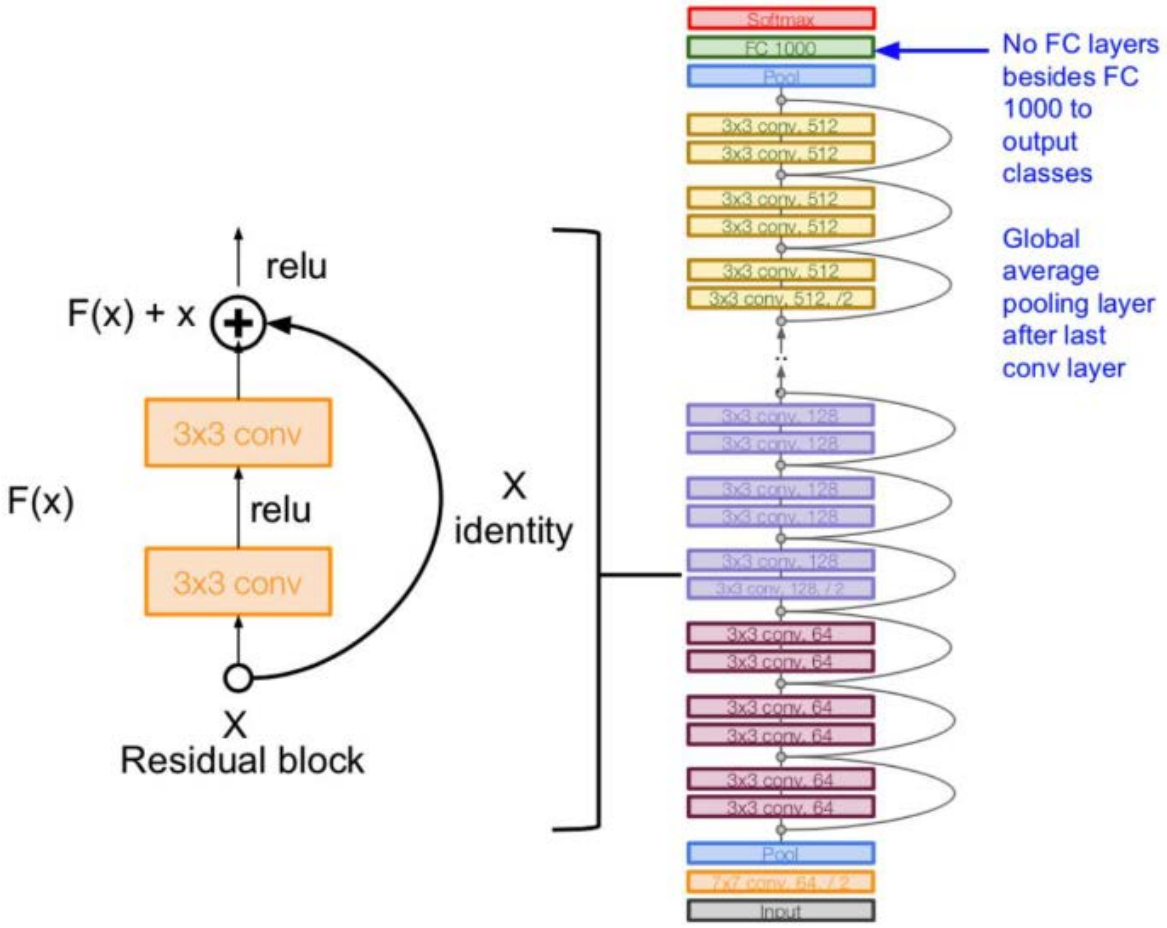
GNN

- GCN의 전체 구조



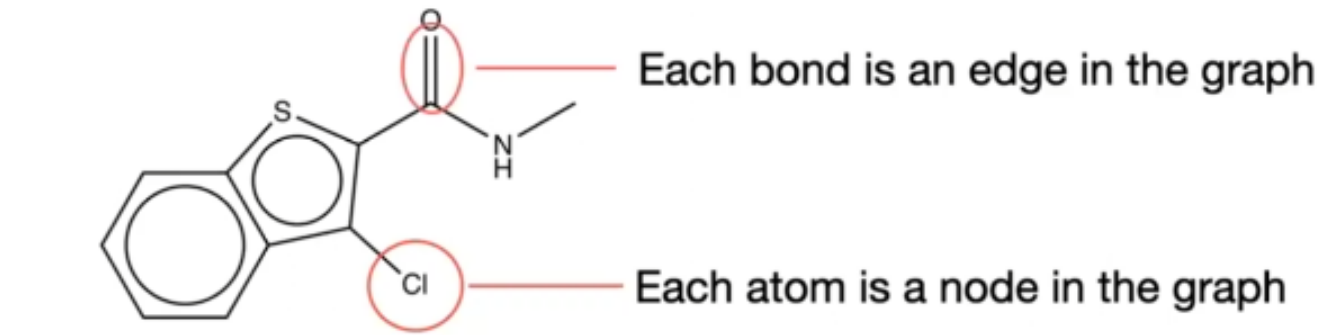
GNN

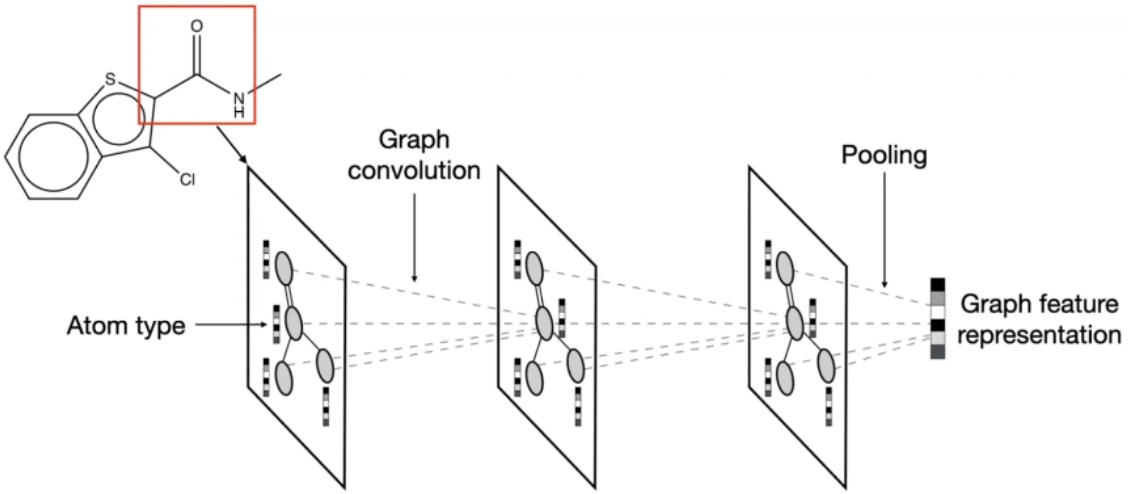
- Skip connection



GNN

- 3D 분자의 구조적 정보를 반영해보자





GNN

- 3D 분자의 구조적 정보를 반영해보자