

基于改进 Attention Mask 编解码器 CPI 的研究

李大舟, 陈思思, 高巍, 于锦涛

(沈阳化工大学 计算机科学与技术学院, 辽宁 沈阳 110142)

摘要: 化合物-蛋白质相互作用(CPI)的研究对药物发现有着重要作用,它可以为药物靶标选择提供有价值的信息,在一定程度上提高先导化合物的命中率,进而加快药物发现的进程。由此提出了一种基于改进 Attention Mask 编解码器的化合物与蛋白质相互作用分类的预测模型,分别使用 RDkit 和 Item2vec 处理化合物的 SMILES 字符串和蛋白质的氨基酸序列,将得到的化合物和蛋白质低维特征表示的向量输入到该模型,通过分配权重的方式来计算蛋白质中的哪个子序列对化合物分子更重要,使用带有 Attention 机制的神经网络计算权重,模拟化合物和蛋白质之间的相互作用关系,最后作为一个二分类问题输出化合物和蛋白质是否相互作用的预测概率。模型性能测评采用 ROC 曲线下面积、准确召回率曲线作为评价指标,实验结果表明,该模型相比于 GraphDTA 和 GCN 模型而言,拥有更好的性能表现,AUC 值提高了 0.04 左右,PRC 值提高了 0.07 左右。

关键词: 深度学习;多头自注意力;化合物蛋白相互作用;Item2vec;编码器-解码器

中图分类号: TP31

文献标识码: A

文章编号: 1673-629X(2022)02-0214-07

doi:10.3969/j.issn.1673-629X.2022.02.035

Research on Compound-protein Interaction Classification Based on Improved Attention Mask Encoder-decoder

LI Da-zhou, CHEN Si-si, GAO Wei, YU Jin-tao

(School of Computer and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: The study of compound-protein interaction (CPI) plays an important role in drug discovery, which can provide valuable information for drug target selection, improve the hit rate of lead compounds to some extent, and accelerate the process of drug discovery. Therefore, a prediction model of compound-protein interaction classification based on the improved Attention Mask encoder-decoder is proposed. RDkit and Item2vec are used to process the SMILES string of the compound and the amino acid sequence of the protein, and the vector representation of low-dimensional characteristics of compounds and proteins is input into the model. The assigned weight is used to calculate which subsequence in the protein is more important for the compound molecule. The neural network with Attention mechanism is to calculate the weight and simulate the interaction between the compound and the protein. Finally as a binary classification problem, output the predicted probability of whether the compound and the protein interact. The model performance evaluation uses the area under the ROC curve and the accurate recall curve as evaluation indicators. According to the experimental results, this model has better performance than the GraphDTA and GCN models, with the AUC value increased by about 0.04, and the PRC value increased by about 0.07.

Key words: deep learning; multi-head self-attention; compound-protein interaction; Item2vec; encoder-decoder

0 引言

近年来,药物发现的技术和水平在不断进步,促进了生物制剂技术和生物制药开发的不断发展。药物发现是人类发现潜在新型药物的过程,一般是通过将化合物库、天然物质或提取物的合成小分子在完整细胞或整个生物体上进行表型筛选,从而识别在过程中具有理想治疗效果的物质^[1]。由于药物发现的进步,制

成的药剂使得许多的疾病得以预防和治疗。然而,由于目前医学水平的限制,仍有许多疾病无法得以攻克,并且不断有新型的病症出现,所以进行新型药物的研究和开发的需求十分迫切。

药物中包含了特定的化合物分子,人体内的大部分化学反应都有蛋白质的参与,因此,掌握化合物-蛋白质相互作用(compound-protein Interaction, CPIs)在

收稿日期:2021-03-21

修回日期:2021-07-21

基金项目:辽宁省教育科学技术研究项目(LJ2020033)

作者简介:李大舟(1982-),男,博士,讲师,研究方向为数据挖掘与车联网;陈思思(1997-),女,硕士研究生,研究方向为深度学习。

药物发现上有着重要的作用,研究人员可以通过 CPI 识别筛选出有效的化合物,并且可以了解药物产生副作用的原因。然而,通过生物实验的方法来确定 CPI 十分耗时且费用高昂^[2]。人类已知的蛋白质类型和化合物类型众多,若通过生物实验的方法来一一验证它们之间是否存在相互作用,这几乎很难完成的。因此,人们提出通过计算预测方法辅助 CPI 的研究,让计算机来分析数据并进行预测,进而提高药物发现的速度。

随着人工智能的快速发展,机器学习(machine learning)已经应用于生活中的不同领域。使用传统机器学习识别 CPI 的研究在不断进步。2004 年, Bredel 和 Jacoby^[3]提出了一种从化学基因组学角度开发的预测方法,在统一的模型中同时考虑化合物和蛋白质的信息。在此之后,各种基于此想法的 CPI 预测模型不断被提出。例如,在 2008 年, Jacob 和 Vert^[4]利用化学结构和蛋白质家族之间的张量积作为特征,应用成对核的支持向量机来预测 CPI。在 2009 年, Bleakley 和 Yamanishi^[5]提出二部局部模型(BLM),利用化学结构和蛋白质的氨基酸序列之间的相似性度量,应用具有已知相互作用的支持向量机来预测 CPI。为了降低化学基因组学空间的维度,在 2012 年, Cheng^[6]提出使用特征选择技术,使用选择后的特征训练支持向量机。在 2013 年, Tabei 和 Yamanishi^[7]提出使用哈希算法改进线性支持向量机的预测性能,一次获得化合物-蛋白质对的指纹。

传统的机器学习往往由多个独立的模块组成,需要多个处理步骤,并且每一步的结果会影响下一步骤的好坏,而端到端的深度学习模型可以自动学习特征,且拥有学习海量数据的能力和强大的拟合能力,只需在输入端输入原始数据,模型自动在中间层提取数据的特征,最后在输出端得到预测结果。在 2016 年, Kipf 等人^[8]提出图神经网络(graph convolutional network, GCN),该网络能够处理具有广义拓扑图结构的数据,目前主要应用于图分类^[9]、文本分类^[10]、推荐系统^[11]、疾病预测^[12]等。在 2018 年, Öztürk 等人^[13]提出 DeepDTA 模型,利用卷积神经网络(convolutional neural network, CNN)提取化合物和蛋白质的特征,然后将两个特征向量拼接起来,经过全连接层输出 CPI 二分类结果。在 2019 年, Öztürk 等人^[14]提出 WideDTA 模型,该模型类似于 DeepDTA 模型,不同之处是利用了两个额外的特征以改善模型的性能,两个特征分别是配体最大公共结构(LMCS)和蛋白质基序和结构域(PDM)。同年, Tsubaki 等人^[15]和 Nguyen^[16]分别提出 CPI-GNN 模型和 GraphDTA 模型,分别使用图神经网络(graph neural network, GNN)和图卷积网络(graph convolutional network, GCN)学习化合物分

子图的表示。在 2019 年, Schwaller 等人^[17]提出 Transformer 可用于化学反应预测,但是,仍局限于 seq2seq 任务。2019 年, Yang 等人^[18]提出了 XLNET 模型,其基于自回归(autoregressive, AR)语言模型实现了新的双向编码,考虑到在训练过程中屏蔽的单词与未屏蔽的单词之间的关系。受 XLNET 在两个序列之间获得特征的强大能力的启发,该文提出基于改进 Attention Mask 编解码器模型,将化合物和蛋白质当作两种类型的序列输入到该模型中,最终得到化合物和蛋白质是否相互作用的预测结果。

1 算法设计

该文提出的基于改进 Attention Mask 编解码器的化合物与蛋白质预测模型的主体结构如图 1 所示。

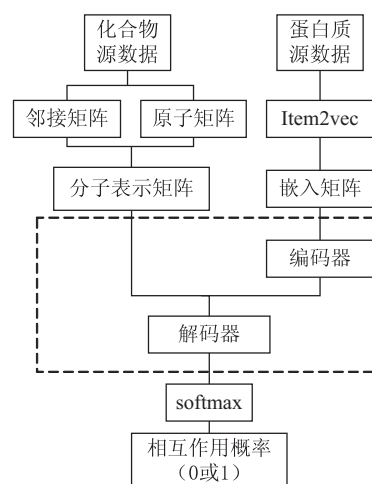


图1 模型的基本框架

首先对原始化合物数据进行处理,得到原子矩阵和邻接矩阵,然后根据关系矩阵得到化合物的分子表示矩阵;同时对原始蛋白质数据的氨基酸序列进行处理,使用 Item2vec 技术得到蛋白质嵌入矩阵;将蛋白质嵌入矩阵输入到编码器,最后将化合物和经编码器处理的蛋白质表示矩阵输入到解码器中,得到相互作用的向量,经过线性变换,最终得到化合物和蛋白质是否相互作用的预测结果。

1.1 原始数据处理

1.1.1 蛋白质数据处理

蛋白质是构成细胞的基本有机物,氨基酸脱水缩合组成多肽链,多肽链经过折叠后组成的具有空间结构的物质就是蛋白质。蛋白质序列可以通过其物理性质或其氨基酸序列进行编码表示^[19]。文中蛋白质原始数据表现形式为氨基酸首字母缩写字符串,根据生物信息学信息可知,蛋白质序列由 20 种基本氨基酸组成,部分氨基酸英文名、中文名称和首字母缩写实例如人免疫球蛋白,其蛋白质氨基酸序列表示为“MEF-GLSWVFLVAILEGVQCEVQLVESGGGLVQPGGSLRL

SCAASGFTFSSHWMTWVRQTPGKRLEWVANVKQD
GSARYYADSVRGRFTISRDNANKSLYLQMDSLRADD
TAVYYCARSTGIDYWGQGTLVTVSS”。

Item2vec 是由 Barkan^[20]提出的一种用于学习和描述复杂句法和语义单词关系的分布式向量表示技术,借鉴于 Word2vec^[21]的 skip-gram with sampling (SGNS) 的思路,将其运用于基于物品的协同过滤 (item-based CF) 上。Item2vec 把原来蛋白质数据的高维稀疏的表示方式映射到低维稠密的向量空间中,然后用这个低维向量来表示该蛋白质,对于大量的蛋白质序列数据,可以通过 Item2vec 学习蛋白质序列的嵌入式表示,大大简化下游建模。

基于前人的工作,该文将 UniProt^[22]中的人类蛋白质氨基酸序列进行预处理,作为一个语料库,然后使用 Item2vec 训练语料,设置蛋白质嵌入向量维度为 128 维,经过 20 轮迭代训练了蛋白质嵌入模型。例如人免疫球蛋白中氨基酸序列长度为 132,将其带入训练后的蛋白质嵌入模型中,通过嵌入算法将每一个氨基酸转换为向量,对应一个长度为 128 的向量,最终人免疫球蛋白表示为大小为 (132, 128) 的矩阵形式。

PCA 是由 Pearson^[23]提出的一种统计方法,主要思想是将原始数据沿最大方差方向投影,得到原始数据的低维特征表示,从而实现数据的降维。通过 PCA 方法得到蛋白质嵌入向量实现,实现蛋白质特征维度的转变,化合物的特征维度变换同理。以人免疫球蛋白为例,输入的代表矩阵大小为 (132, 128),经 PCA 处理后的表示矩阵大小为 (132, 64)。

1.1.2 化合物数据处理

化合物是由两种或两种以上的元素组成的纯净物。简化分子线性输入规范 (simplified molecular-input line-entry system, SMILES) 是一种用于输入和表示分子的线性符号,使用 ASCII 字符串来描述分子结构。文中化合物原始数据表现形式为 SMILES 字符串,例如吩噻嗪,其 SMILES 格式为 C1=CC=C2SC3C=CC=CC3NC2=C1。依据化学特性划分原子特征,原子特征列表如表 1 所示,每种原子的特征可以使用 34 维的向量表示。

表 1 原子特征列表

特征种类	表示形式
原子类型	C, N, O, F, P, S, Cl, Br, I, 其他
原子度	0, 1, 2, 3, 4, 5, 6
形式电荷	0 或 1
自由基电子数	0 或 1
杂交型	sp, sp ² , sp ³ , sp ³ d, sp ³ d ² , 其他
芳香族	0 或 1

续表 1

特征种类	表示形式
附着的氢原子数	0, 1, 2, 3, 4
空间的螺旋特性	0 (假) 或 1 (真)
配置	R, S

RDKit 是开源化学信息学与机器学习的工具包,支持机器学习方面的分子描述符的产生。该文通过使用 RDKit 封装的函数对 SMILES 格式的化合物数据进行读取和处理,得到化合物的原子矩阵和带自环的邻接矩阵,然后利用关系矩阵,得到分子的矩阵表示。

1.2 改进 Attention Mask 编解码器架构

该文提出的模型沿用了经典的 Encoder-Decoder 结构,使用到并行化计算的自注意力机制,极大地缩短了训练时间。该整体架构如图 2 所示,其中编码器部分主要由多头自注意力层和前馈神经网络层组成,解码器部分主要由 Attention Mask 层、编码器-解码器注意力层和前馈神经网络层组成。

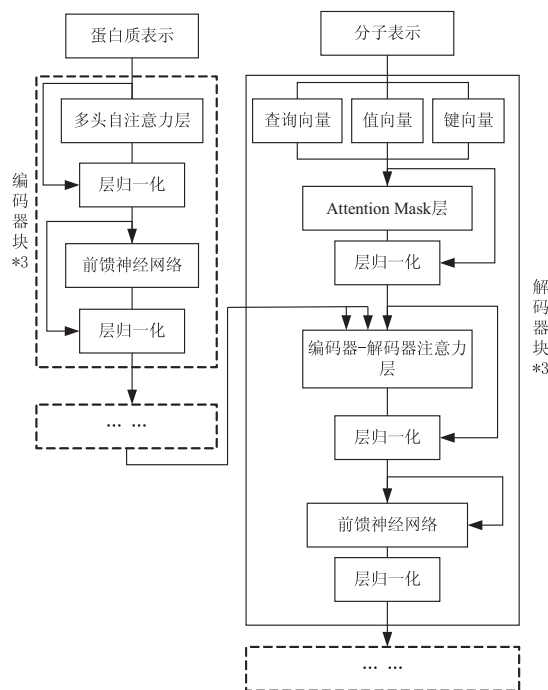


图 2 编-解码器架构

1.2.1 编码器

编码器的结构如图 2 左侧虚线框内所示,由 3 个编码器块堆叠而成,每一个编码器块都由两个子层组成,并且每一个子层之间都使用了残差连接和层归一化操作。

编码器的第一个子层是多头自注意力层。自注意力的本质是通过当前词来引入上下文的信息,以此增强对当前词的表示。首先根据输入的化合物的原子序列,通过线性变换得到 Q, K, V 的向量表示,然后根据公式 1 计算注意力值。

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (1)$$

式中, \mathbf{Q} 值、 \mathbf{K} 值和 \mathbf{V} 值分别代表注意力机制中的查询向量(Query Vector)、键向量(Key Vector)和值向量(Value Vector), d_k 表示特征向量的维度。计算过程中将 $\frac{1}{\sqrt{d_k}}$ 作为缩放因子,可以避免因为维度过高对梯度优化造成影响的情况。

多头自注意力层是包括了多个按比例缩放的自注意力层,可以在不改变参数量的情况下增强注意力的表现力,扩展模型专注不同位置的能力。多头自注意力是对 \mathbf{Q} 、 \mathbf{K} 、 \mathbf{V} 进行分组计算注意力值,如公式 2 所示,然后拼接所有注意力头,计算过程如公式 3 所示。

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (2)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_n) \mathbf{W}^o \quad (3)$$

式 2 中, $\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V$ 是由 \mathbf{Q} 、 \mathbf{K} 、 \mathbf{V} 分别乘以不同的随机初始化权值矩阵得到的,式 3 中, \mathbf{W}^o 表示附加的权重矩阵。

编码器的第二个子层是前馈神经网络层。前馈神经网络层(feed forward layer, FFL)是由两层全连接神经网络组成的,选择 ReLU 作为激活函数,如公式 4 所示。该网络层对注意力的输出进行空间交换,增加了模型的表现能力。

$$\text{FFN} = \text{Max}(0, \mathbf{X} * \mathbf{W}_4 + b_4) \mathbf{W}_5 + b_5 \quad (4)$$

式中, \mathbf{X} 表示经多头自注意力层的输出矩阵, \mathbf{W}_4 和 \mathbf{W}_5 表示权重矩阵, b_4 和 b_5 表示网络的偏置。

由于网络不断加深,数据的分布也在不断地发生变化,同时可能会带来梯度消失或爆炸等问题。加入残差连接可以从一定程度上缓解因为梯度爆炸导致的网络退化问题,而加入层归一化可以保证数据的稳定分布,同时可以加速模型的优化速度。残差连接和层归一化操作如公式 5 所示。

$$\text{Output} = \text{LN}(\mathbf{X} + (\text{SubLayer}(\mathbf{X}))) \quad (5)$$

式中, \mathbf{X} 表示每个子层的输出, $\text{SubLayer}()$ 表示子层本身的输出, LN 表示 Layer Normalization, Layer Normalization 的计算公式如下:

$$f(x) = \alpha \frac{x - \mu}{\sigma} + \beta \quad (6)$$

式中, μ 、 σ 分别表示均值和方差, α 表示缩放参数, β 表示平移参数。

1.2.2 解码器

编码器的结构如图 2 右侧虚线框内所示,由 3 个解码器块堆叠而成,每一个解码器块都由三个子层组成,与编码层一样,每一个子层之间同样使用了残差连接和层归一化操作。

解码器的第一个子层是改进的 Attention Mask 层。传统的自回归模型的缺点是不能同时利用上文或者下文的信息,而传统的自编码模型的缺点是会导致预训练阶段和微调阶段出现不一致的问题。改进的 Attention Mask 层部分避免了二者的缺点,在传统的自回归模型的模式下,引入全排列语言模型(permutation language modeling, PLM),保持当前词的位置不变,对文本中的其他词进行重新编排,使得当前中心词的下文也有可能出现在中心词的上文中,然后将句尾的一定量的词进行遮掩,使用自回归方式预测被遮掩的词。全排列语言模型的优化目标最大似然化概率如公式 7 所示。

$$\max(\theta) E_{z \sim Z_T} \left[\sum_{t=1}^T \log p_{\theta}(x_{z_t} | x_{z_{<t}}) \right] \quad (7)$$

式中, T 表示序列长度, Z_T 表示所有可能出现的排列序列, z_t 表示第 t 个元素。例如存在一个长度为 T 的序列,从序列的所有可能的排列序列中随机采样一个,然后通过计算来分解联合概率成条件概率,并加权求和得到预测当前词概率最大的参数 θ ,由此捕获双向的语境。具体的 PLM 操作是通过双流自注意力机制实现的,双流自注意力机制由内容流注意力机制和查询流注意力机制组成,同时引入了两个隐状态,分别是内容隐状态 h_{z_t} 和查询隐状态 g_{z_t} 。双流注意力机制的计算过程如公式 8 和公式 9 所示。

$$g_{z_t}^{(m)} \leftarrow \text{Attention}(Q = g_{z_t}^{(m-1)}, KV = h_{z_{<t}}^{(m-1)}; \theta) \quad (8)$$

$$h_{z_t}^{(m)} \leftarrow \text{Attention}(Q = h_{z_t}^{(m-1)}, KV = h_{z_{\leq t}}^{(m-1)}; \theta) \quad (9)$$

式 8 中,上标 m 表示层数, \mathbf{Q} 值、 \mathbf{K} 值和 \mathbf{V} 值分别代表注意力机制中的查询向量、键向量和值向量, z_t 表示 $z \in Z_T$ 的前 $t-1$ 个元素。

双流注意力机制的计算过程如图 3 所示。图 3(a)为内容流注意力机制的计算过程,查询向量 \mathbf{Q} 为 $h_1^{(0)}$,键向量、值向量 \mathbf{KV} 为 $[h_1^0, h_2^0, h_3^0, h_4^0]$ 。图 3(b)为查询流注意力机制的计算过程,查询向量 \mathbf{Q} 为 g_1^0 ,键向量、值向量 \mathbf{KV} 为 $[h_2^{(0)}, h_3^{(0)}, h_4^{(0)}]$ 。图 3(c)左侧为双流注意力的具体实现过程,将内容流注意力机制和查询流注意力机制结合,将内容流隐状态初始化为 $e(x)$,将查询隐状态初始化为 w 。图 3(c)右侧为掩码矩阵,灰色表示遮掩,白色表示不遮掩。通过遮掩某个单词,使其在预测的时候不发生作用,被遮掩的词与当前词存在位置关系。

例如原本输入的句子是“1,2,3,4”,若经过 PLM 操作后的排列序列为“3,2,4,1”,表明在预测“2”的时候,可以看到上文的“3”的信息;当预测“4”的时候,可以看到上文“3”和“2”的信息,并以此类推。内容流和查询流掩码矩阵如图 3(c)右图所示,通过掩码矩阵,将句子改成随机的排列组合,实现同时利用上下文信

息预测当前词。

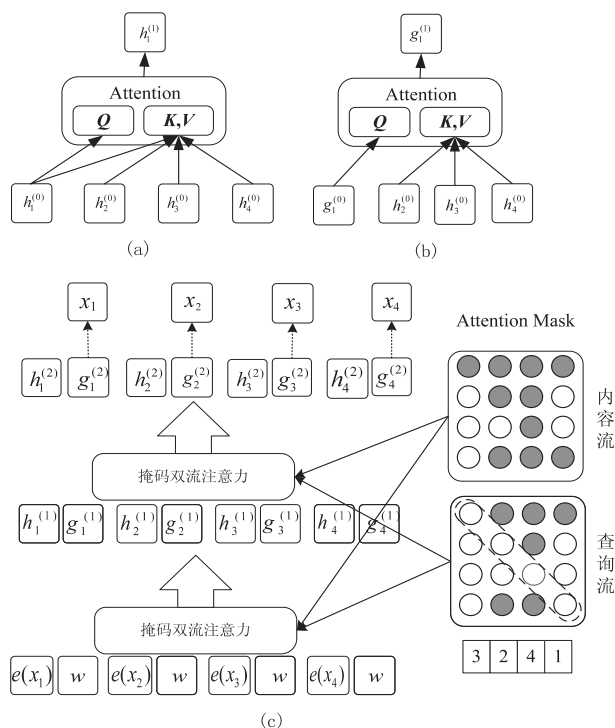


图 3 双流自注意力机制计算过程图示

解码器的第二个子层是编码器-解码器注意力层,它的输入由两部分构成,分别是掩码多头注意力层的输出 Q 和编码器的输出 K, V ,通过注意力机制增强对当前词的表示,并提取编码器和解码器间的交互信息。

解码器的第三个子层是前馈神经网络层,其工作原理与编码器中的前馈神经网络层一样。该子层的输入为编码器-解码器注意力层的输出。最后该层的输出是化合物和蛋白质相互作用的特征向量,将其经过 softmax 函数,最终得到化合物和蛋白质是否相互作用的概率。

2 实验设计

2.1 实验环境

本实验在 windows10 系统下进行,使用 Intel@ i5-8265U 作为计算单元,内存为 8 GB。模型使用 Pytorch 框架进行搭建,版本为 1.6.0+cu101。构建模型所用的代码使用到 RDKit 库。

2.2 数据集

文中用于训练 Item2vec 模型的蛋白质数据来自于 UniProt 蛋白质数据库^[22]。选取 UniProt 蛋白质数据库中 Swiss-Prot 子库里的人类蛋白质序列作为一个语料库,源数据格式如表 3 所示,总计 20 413 条,提取蛋白质的氨基酸序列数据,使用该数据对 Item2vec 模型进行预训练,学习蛋白质的嵌入式表示。

文中化合物和蛋白质数据主要来源于 Lifan^[24] 构

建的 GPCR 标签反转数据集,据实验验证,标记反转实验可以有效地评估隐藏的配体偏差对模型的影响,降低基于化学基因组的化合物和蛋白质相互作用任务的常见风险。GPCR 数据集主要有化合物信息、蛋白质信息和表示是否相互作用的布尔值,数据集包含了 356 种蛋白质和 5 459 种化合物的 15 343 种作用对。

对于 GPCR 组,随机选择 500 个配体,并将所有涉及这些配体的 CPI 负样本汇集在一起。另外,选择了 500 个配体,并将所有相关正样本汇集在一起。在实验设计后,最终建立了 1 537 个相互作用的 GPCR 测试集,剩余的数据集被用来确定超参数。

2.3 评价指标

实验中,采用二分类交叉熵损失函数、ROC 曲线下面积 (AUC) 以及精度-召回率曲线 (PRC) 作为模型的评估指标。

二分类交叉熵 (Binary Cross Entropy): 是多分类 softmax_cross_entropy 的一种特殊情况,当只有两类标签时,即 0 或者 1,使用逻辑回归的损失函数,如公式 10 所示。

$$\text{Loss} = -[y \log \hat{y} + (1 - y) \log (1 - \hat{y})] \quad (10)$$

式中, \hat{y} 表示模型预测的概率, y 表示实际的标签值。

ROC (Receiver Operating Characteristic): 以假正例率 (FPR) 为 X 轴、真正例率 (TPR) 为 Y 轴绘制的反映模型敏感性和精确性的趋势走向的曲线。

AUC (Area Under Curve): ROC 曲线下的面积。若分类器的性能越好,则 AUC 值越接近 1。

PRC (Precision Recall Curve): 以查全率 (Recall) 为 X 轴、查准率 (Precision) 为 Y 轴绘制的图,可以对分类器的整体效果进行综合评价。该评估指标引入“平衡点” (BEP) 概念,当查全率等于查准率时取的值越大时,表明该分类器的性能越好。

2.4 参数

该模型的编码器和解码器的层数各为 3 层,多头注意力头数为 8 个,经 PCA 处理后的蛋白质表示和原子表示的维度为 64,编码器和解码器完全连接的前馈神经网络层中隐藏单元数量为 512,Dropout 为 0.2,学习率为 $1e-4$,批尺寸大小为 64。

2.5 实验结果分析及对比

该文使用 GPCR 测试集在模型上进行训练,采用接收机工作特性曲线下面积 (AUC)、准确召回率曲线 (PRC) 作为模型的评估指标。

从图 4 中可以看出,随着迭代次数的增加,模型的 Loss 值在逐渐变小,且愈加接近饱和,在迭代 50 轮前,模型的 AUC 值和 PRC 值的变化明显,随着迭代次增加,模型训练愈加接近饱和,评估指标趋于平缓 and 稳定,模型的最优 AUC 值和 PRC 值分别为 0.865 和

0.883。

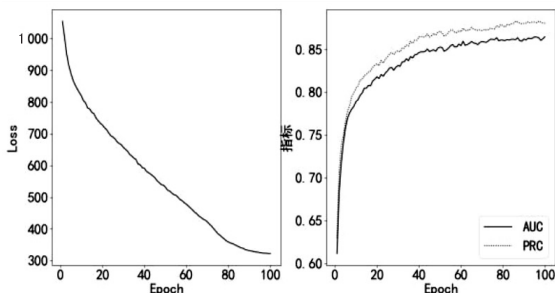


图 4 模型训练 Loss 变化和 AUC 值、PRC 值变化

该文对模型进行调参试验,采用控制变量法进行调参。实验设置如表 2 所示,实验结果如图 5 所示。

表 2 模型对比实验设置

实验设置	调节变量	AUC	PRC
其他设置如 2.4 节所述	batchsize=64	0.865	0.883
	batchsize=32	0.865	0.879
	batchsize=96	0.828	0.841

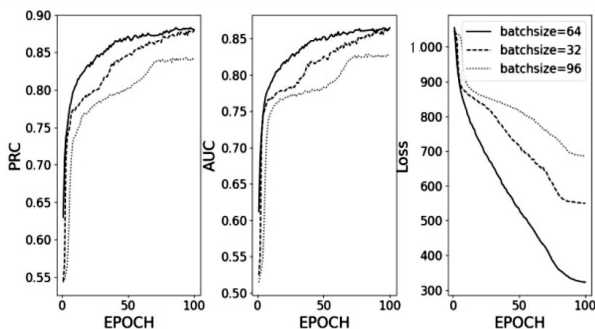


图 5 不同 batchsize 下 Loss 值、PRC 值和 AUC 值变化

从图 5 中可以看出,在同样的迭代次数下, batchsize 为 64 时,模型的 Loss 值相对于另外两个更低, batchsize 为 96 时效果较差。在 batchsize 为 64 时,模型的 PRC 和 AUC 值优于另外两种情况, batchsize 为 96 时效果较差。

该文对模型网络结构也进行了对比实验。实验设置如表 3 所示,实验结果如图 6 所示。

表 3 结构对比实验设置

实验设置	参数设置	AUC	PRC
2.4 节所述设置	num_blocks=3	0.865	0.883
	Hid_dim=64		
拓宽网络	num_blocks=3	0.862	0.869
其他设置如表 2	Hid_dim=128		
加深网络	num_blocks=6	0.816	0.839
其他设置如表 2	Hid_dim=64		

从图 6 可以看出,在相同的迭代次数下,拓宽网络的 PRC 和 AUC 相较于原始网络在一开始处于较为落后的趋势,后来逐渐接近;在相同的迭代次数下,加深网络的 PRC 和 AUC 一直处于落后的趋势。在其他设

置不变的情况下,原始网络的参数设定的 PRC 和 AUC 达到最优的情况。

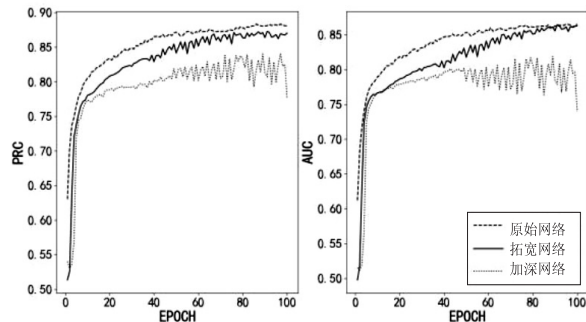


图 6 拓宽网络和加深网络的 PRC 和 AUC 变化

该文选择了经典的机器学习模型和两种流行的行业常用模型与该文提出的模型进行对比,实验结果如表 4 所示。在 GPCR 测试集上,该文提出的模型在 AUC 和 PRC 方面均优于对比的模型,在数据集上取得了较好的性能,AUC 值和 PRC 值均有提升,表明该模型具有更强的学习蛋白质和化合物之间相互作用的能力。

表 4 GPCR 测试集性能

模型	AUC	PRC
KNN	0.774	0.763
L2	0.810	0.889
SVM	0.813	0.802
GraphDTA	0.825	0.817
GCN	0.822	0.809
Ours	0.865	0.883

3 结束语

该文尝试将深度学习技术应用于 CPI 预测的研究中,将该任务转换成标签二分类的问题进行解决。在使用传统的编解码器模型的基础上,在解码器中使用到改进的 Attention Mask 层,以此来处理蛋白质和化合物二分类任务。在 AUC 和 PRC 指标测评下,与其他模型相比,该文改进的模型在实验上拥有更好的性能表现。

实验结果表明,该模型可以学习期望的 CPI 特征,性能更稳定且准确率更高。如果将该模型应用于实际的药物发现研究中,可以为药物靶标选择提供一定的参考价值,加快药物发现的进程。同时深度学习不要求具备生物学和药理学等专业知识,就可以得到数据背后的隐藏信息,且对于数据量特别大的数据具有明显的优势。然而,该模型构造了一个注意力矩阵,需求与输入呈平方关系,因此,对内存和算力的需求非常高。

参考文献:

- [1] 张 婷,卢 岩,陈 娟,等.我国生物医药领域技术创新态势研究[J].中国新药杂志,2020,29(22):2521-2527.
- [2] DIMASI J A,HANSEN R W,GRABOWSKI H G. The price of innovation: new estimates of drug development costs[J]. Journal of Health Economics,2003,22(2):151-185.
- [3] BREDEL M, JACOBY E. Chemogenomics: an emerging strategy for rapid target and drug discovery[J]. Nature Reviews Genetics,2004,5(4):262-275.
- [4] JACOB L, VERT J P. Protein-ligand interaction prediction: an improved chemogenomics approach[J]. Bioinformatics, 2008,24(19):2149-2156.
- [5] BLEAKLEY K, YAMANISHI Y. Supervised prediction of drug - target interactions using bipartite local models[J]. Bioinformatics,2009,25(18):2397-2403.
- [6] CHENG F,ZHOU Y,LI J,et al. Prediction of chemical - protein interactions: multitarget-QSAR versus computational chemogenomic methods[J]. Molecular BioSystems,2012,8(9):2373-2384.
- [7] TABEI Y, YAMANISHI Y. Scalable prediction of compound-protein interactions using minwise hashing[J]. BMC Systems Biology,2013,7(6):1-13.
- [8] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609.02907, 2016.
- [9] YADATI N, NIMISHAKAVI M, YADAV P, et al. HyperGCN:a new method of training graph convolutional networks on hypergraphs[J]. arXiv:1809.02589,2018.
- [10] YAO L,MAO C,LUO Y. Graph convolutional networks for text classification[C]//33rd AAAI conference on artificial intelligence (AAAI 2019). Hawaii:AAAI,2018.
- [11] 江 原. 基于图卷积与神经协同过滤的融合信息推荐模型[D]. 长春:吉林大学,2018.
- [12] 宁世琦,郭茂祖,任世军. 基于图卷积网络的癌症临床结果预测的半监督学习方法[J]. 智能计算机与应用,2018,8(6):44-48.
- [13] ÖZTÜRK H, ÖZGÜR A, OZKIRIMLI E. DeepDTA: deep drug - target binding affinity prediction[J]. Bioinformatics, 2018,34(17):i821-i829.
- [14] ÖZTÜRK H, OZKIRIMLI E, ÖZGÜR A. WideDTA: prediction of drug-target binding affinity[J]. arXiv:1902.04166, 2019.
- [15] TSUBAKI M,TOMII K,SESE J. Compound - protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. Bioinformatics,2019,35(2):309-318.
- [16] NGUYEN T,LE H, VENKATESH S. GraphDTA: prediction of drug-target binding affinity using graph convolutional networks[EB/OL]. 2019. <https://www.biorxiv.org/content/10.1101/684662v1.abstract>.
- [17] SCHWALLER P, LAINO T, GAUDIN T, et al. Molecular transformer for chemical reaction prediction and uncertainty estimation[J]. arXiv:1811.02633,2018.
- [18] YANG Z,DAI Z,YANG Y,et al. Xlnet:generalized autoregressive pretraining for language understanding[J]. arXiv:1906.08237,2019.
- [19] SALADI S M, JAVED N, MÜLLER A, et al. A statistical model for improved membrane protein expression using sequence-derived features[J]. Journal of Biological Chemistry,2018,293(13):4913-4927.
- [20] BARKAN O,KOENIGSTEIN N. Item2Vec:neural item embedding for collaborative filtering[C]//2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). Italy:IEEE,2016.
- [21] MIKOLOV T, GRAVE E, BOJANOWSKI P, et al. Advances in pre-training distributed word representations[J]. arXiv:1712.09405,2017.
- [22] APWEILER R,BAIROCH A,WU C H,et al. UniProt:the universal protein knowledgebase[J]. Nucleic Acids Research, 2004,32(D1):D115-D124.
- [23] PEARSON K. On lines and planes of closest fit to systems of points in space[J]. The London,Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901,2(11):559-572.
- [24] CHEN L,TAN X,WANG D,et al. TransformerCPI:improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments[J]. Bioinformatics,2020,36(16):4406-4414.