计算机科学
Computer Science
ISSN 1002-137X,CN 50-1075/TP

《计算机科学》网络首发论文

题目：基于迁移学习和多视图特征融合提高 RNA 碱基相互作用预测
作者：王晓飞，樊学强，李章维
收稿日期：2021-12-16
网络首发日期：2022-11-10
引用格式：王晓飞，樊学强，李章维．基于迁移学习和多视图特征融合提高 RNA 碱基相互作用预测[J/OL]．计算机科学.
https://kns.cnki.net/kcms/detail/50.1075.TP.20221109.1822.048.html

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI: 10.11896/jsjkx.211200186

基于迁移学习和多视图特征融合提高 RNA 碱基相互作用预测

王晓飞　樊学强　李章维

浙江工业大学信息工程学院　杭州 310023
（xiaowangpluss@163.com）

摘　要 RNA 碱基相互作用对维持其三维结构的稳定具有重要作用，准确地预测碱基相互作用可以辅助 RNA 三维结构的预测。然而，用于预测 RNA 碱基相互作用的数据量少导致模型未能充分地学习到数据的特征分布，以及数据存在的特性（对称特性和类别不平衡），都影响了模型的性能。针对模型不充分学习和数据特性问题，在深度学习的基础上，提出一种高性能的 RNA 碱基相互作用预测方法 tpRNA。tpRNA 首次在 RNA 碱基相互作用预测任务中引入迁移学习以改善因数据量少产生的模型不充分学习问题，并提出高效的损失函数和特征提取模块，充分发挥迁移学习和卷积神经网络在特征学习方面的优势，以缓解数据特性问题。结果表明，引入迁移学习能降低因数据量少导致的模型偏差，提出的损失函数能优化模型的训练，特征提取模块能提取到更有效的特征。与最先进的方法相比，tpRNA 在低质量输入特征的情形下具有显著的优势。

关键词：　RNA 碱基相互作用；迁移学习；数据特性；损失函数；卷积神经网络

中图法分类号 TP301

Improving RNA Base Interactions Prediction Based on Transfer Learning and Multi-view Feature Fusion

WANG Xiaofei，FAN Xueqiang and LI Zhangwei

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract RNA base interactions play an important role in maintaining the stability of its three-dimensional structure, and accurate prediction of base interactions can help predict the three-dimensional structure of RNA. However, due to the small amount of data, the model could not effectively learn the feature distribution of the training data, and existing data characteristics (symmetry and class imbalance) affect the performance of the RNA base interactions prediction model. Aiming at the problems of insufficient model learning and data characteristics, a high-performance RNA base interactions prediction method called tpRNA is proposed based on deep learning. tpRNA introduces transfer learning in RNA base interactions prediction task to weak the influence of insufficient learning in the training process due to the small amount of data, and an efficient loss function and feature extraction module is proposed to give full play to the advantages of transfer learning and convolutional neural network in feature learning to alleviate the problem of data characteristics. Results show that transfer learning can reduce the model deviation caused by less data, the proposed loss function can optimize the model training, and the feature extraction module can extract more effective features. Compared with the state-of-the-art method, tpRNA also has significant advantages in the case of low-quality input features.

Keywords RNA base interactions, Transfer learning, Data characteristic, Loss function, Convolutional neural networks

到稿日期：2021-12-16　　返修日期：2022-05-11
基金项目：国家自然科学基金（61573317）
This work was supported by the National Natural Science Foundation of China (61573317).
通信作者：李章维（lzw@zjut.edu.cn）

1 引言

RNA 的三维结构在生命活动中具有至关重要的作用，有助于理解生物学功能，但已知结构信息的 RNA 数量非常少。同时，传统的解析技术（X 光衍射、核磁共振和冷冻电镜实验）测定 RNA 结构需要耗费大量的资源和时间[1]。因此，用于测定 RNA 结构的计算方法亟需被开发。已有研究[2]表明 RNA 碱基相互作用能作为约束条件提高其结构预测模型的准确性，准确预测 RNA 碱基相互作用成为了提高结构预测精度的可实行方法[3]。

然而，现有的 RNA 碱基相互作用预测方法[1,3]都是通过增加特征来提升模型的性能，而忽略了数据量少对模型训练产生的影响（数据量少会使模型不能充分地学习样本的特征分布，进而导致模型误差偏大）。深度学习虽然极大地推动了生物信息学领域相关任务[4-6]的发展，却极少地应用于 RNA 碱基相互作用预测任务中。同时，该预测任务所用数据存在的特性也会影响模型的训练过程，致使模型不能准确识别碱基对是否有相互作用，即类别不平衡（相互作用的碱基对数量远小于不相互作用的碱基对数量）和数据的对称特性（预测的碱基相互作用图是对称矩阵）。

鉴于此，提出一种 RNA 碱基相互作用预测方法 tpRNA。tpRNA 分为 3 个部分：1）引入迁移学习[7]缓解数据量少造成的影响，训练蛋白质残基接触模型 pre_protein 作为 tpRNA 的预训练模型；2）提出 MFF（Multi-view Feature Fusion）提升特征提取能力，并以 ResNet[8]的形式搭建 MFFnet 卷积神经网络；3）特制损失函数 tpLoss 降低数据特性产生的影响，在 MFFnet 中加载 pre_protein 的参数，以 tpLoss 训练预测模型 tpRNA。基准测试表明，引入迁移学习明显地改善了模型不充分学习问题，tpLoss 优化了模型的训练，MFF 模块提高了对此类数据的特征提取能力。与最先进的方法 RNAcontact[3]相比，tpRNA 在 $L/10$ 的长程预测上具有显著的优势。

2 RNA 碱基相互作用预测

RNA 碱基相互作用预测与蛋白质残基接触预测相似，都是预测一张 contact map（残基接触图或碱基的相互作用图），且具有相似的特征分布，如图 1 所示的二维矩阵，黑色表示接触，白色表示非接触。两者均是利用多序列比对（Multiple Sequence Alignment，MSA）计算残基或碱基的进化信息[9-12] 以确定残基或碱基是否有相互作用，即 RNA 中的碱基对相互作用时，一个碱基发生变化而另一个碱基在不干扰 RNA 功能的前提下也发生变化。此外，碱基相互作用按碱基之间的碱基数量 N 划分为长程相互作用（$N\geq24$），中程相互作用（$12\leq N<24$）和短程相互作用（$6\leq N<12$）。

（a）1em8D 蛋白质　　（b）1f7u_B RNA

图 1 残基接触图和碱基相互作用图

Fig. 1 Residue contact map and base interaction map

两者的不同点在于：1）蛋白质残基有 20 种类型，而 RNA 碱基只有 4 种；2）残基接触是残基间的 C_β（甘氨酸为 C_α）原子间小于 8Å[9]，而碱基的相互作用是指它们的任意原子之间的最小距离小于 8Å[3]；3）已知结构的蛋白质数量远多于 RNA 数量。

深度学习虽然能断崖式地提升此类预测任务模型的性能，如 RNAcontact[3]，Mappred[5]和 ResPRE[13]，但这类方法均忽略了数据特性对模

1) https://github.com/xiaowang121/RNA

型训练的影响。

考虑两者之间的联系以及存在的数据特性问题，利用迁移学习将具有类似特征分布的蛋白质残基接触模型作为预训练模型，以缓解模型不充分学习问题，并设计损失函数和卷积神经网络降低数据特性对模型训练的影响，提高 RNA 碱基相互作用预测任务的精度。

3 基于迁移学习和特征融合的 RNA 碱基相互作用预测方法

tpRNA 的整体结构如图 2 所示，图中 A 和 B 分别详细地给出了蛋白质残基接触与 RNA 碱基相互作用任务的数据获取和特征提取阶段，C 和 D 清晰地显示了两个任务的训练模型。

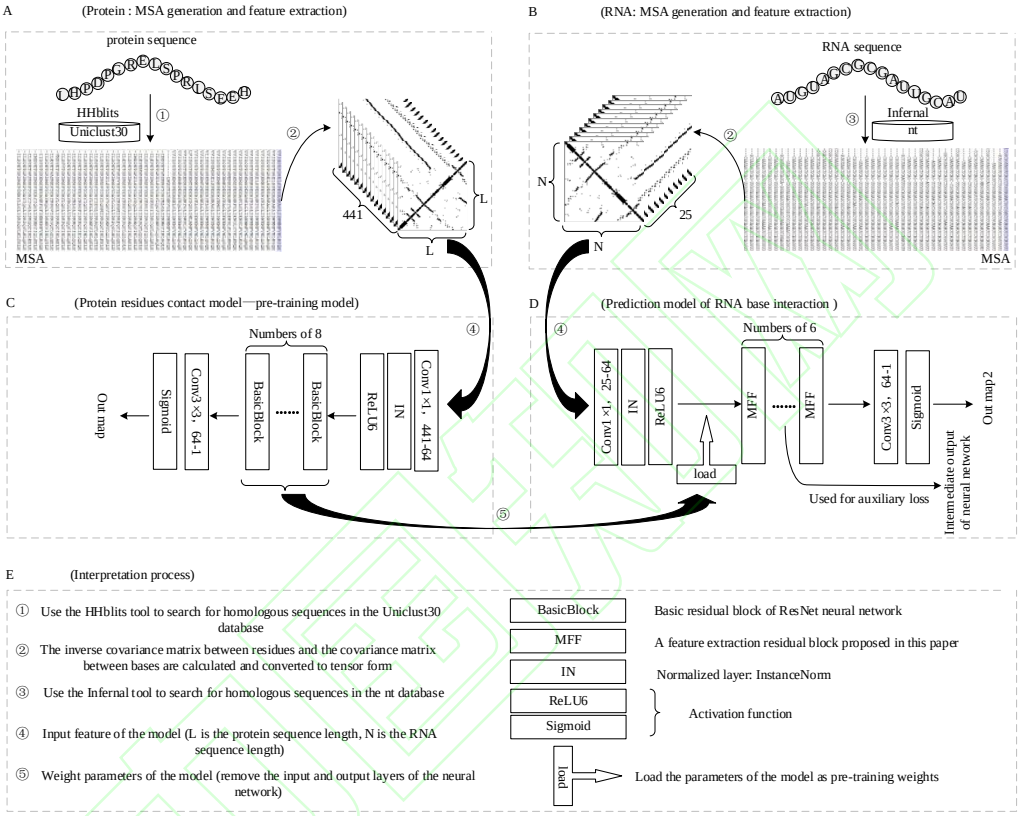

图 2　tpRNA 的整体结构

Fig. 2　Overall structure of tpRNA

3.1　特征提取模块 MFF

相比于 BasicBlock， MFF 利用两个分支的形式进行特征融合（一个分支用空洞卷积[14]扩大感受野得到与另一分支不同的空间特征信息，另一个分支通过分组卷积[15]减少参数量），再利用 SE 注意力机制[16]对融合后的特征进行加权，突出重要的特征。MFF 能更好地捕捉空间特征信息，进而提高模型对特征的提取能力。其结构如图 3 所示。

DOI: 10.11896/jsjkx.211200186

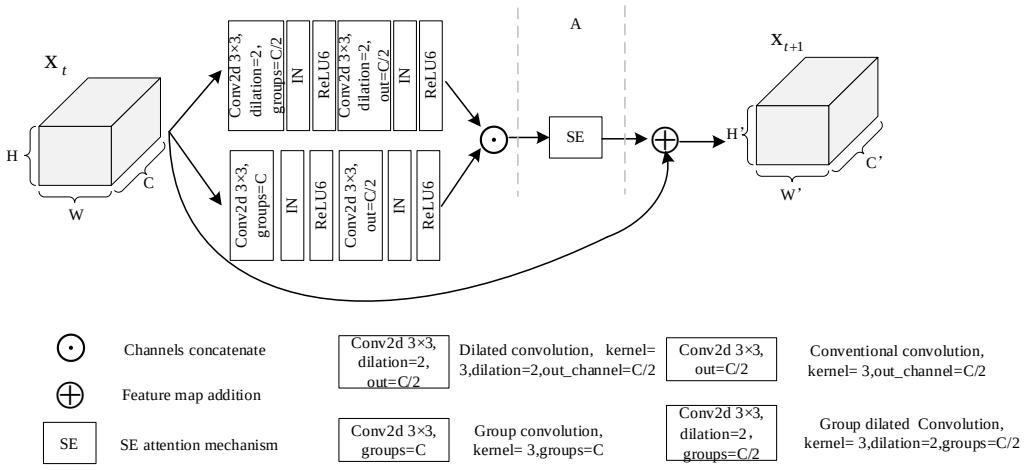

图 3　MFF 的结构图

Fig. 3　Structure diagram of MFF

图 3 中，A 部分用于提高融合特征的有益信息同时抑制产生的背景噪音，再用残差连接缓解层数过多导致的梯度消失问题。即给定输入$X_t \in R^{W\times H\times C}$，得到输出$X_{t+1} \in R^{W'\times H'\times C'}$，其过程可分为：

$$F_1 = concat\left(W_1(X_t), W_2(X_t)\right) \quad (1)$$

$$F_2 = SE(F_1) \quad (2)$$

$$X_{t+1} = \delta(F_2 \oplus X_t) \quad (3)$$

其中，$F_1 \in {R_1}^{W\times H\times C}$为两个分支融合后的特征图；*concat* 表示输出通道拼接；W_1 和 W_2 表示两个分支对输入 X_t 的特征提取；*SE* 表示注意力机制加权 F_1 输出$F_2 \in {R_2}^{W\times H\times C}$；$\delta$ 为 ReLU6 激活函数；$\oplus$表示特征图的对应位置元素相加。

RNA 碱基相互作用预测相当于医学图像分割[17]任务，对特征信息的完整度要求较高。因此，MFF 的分支设计原因在于：1）过多使用分组卷积会降低特征表达能力，故只在分支的第一层利用分组卷积减少参数量；2）分组卷积的分组数过多会减少特征通道之间的依赖关系，同时空洞卷积会增加背景噪音，故空洞卷积的分支只采用分组数为通道数一半的分组卷积；3）大量使用空洞卷积会出现棋盘效应，致使原有的特征信息被破坏，故只采用一个分支是空洞卷积的方式来扩大感受野，捕获大尺度的特征。

在低模型复杂度的水平下，通过丰富空间特征信息的方式，MFF 在训练此类数据特性的任务时显著地优于经典卷积神经网络 ResNet[8]的残差块 BasicBlock 以及轻量型网络 MobileNetV2[18]的倒残差块 InvertResidual。

3.2　损失函数

为进一步提升模型的性能，基于以下两个方面设计高效的损失函数 tpLoss。

（1）考虑正负类别不平衡问题会降低模型的精度，受 Focal Loss[19]启发，使用加权交叉熵损失函数 WCE 平衡正负样本的计算损失，优化训练过程的梯度更新，进而提升模型的性能。

（2）由于数据具有对称特性，为规范网络过深时模型对数据特性的学习，引入辅助损失。如图 2 的 D 部分所示，使用 WCE 分别计算模型中间层输出的辅助损失和模型最终输出的全局损失，记作 Loss1 和 Loss2。将 WCE 损失函数和辅助损失 Loss1 作用于梯度更新，优化模型的训练过程，进而缓解因数据特性产生的偏差问题。式（4）—式（6）分别给出了交叉熵损失函数 CE、加权交叉熵损失函数 WCE 和本文使用的损失函数 tpLoss 的计算公式：

$$CE = -\sum(p_t * \log(p_r) + (1-p_t) * \log(1-p_r)) \quad (4)$$

$$WCE = -\sum(\alpha * p_t * \log(p_r) + (1-\alpha) * (1-p_t) * \log(1-p_r)) \quad (5)$$

$$tpLoss = \beta * Loss1 + (1-\beta) * Loss2 \quad (6)$$

其中，p_t 为真实的 RNA 碱基相互作用图；p_r 为模型输出的 RNA 碱基相互作用图；$\alpha \in [0,1]$控制正负类别在计算损失时占有的权重；$\beta \in [0,1]$调节辅助损失 Loss1 与全局损失 Loss2 的关系。

4 实验与分析

4.1 实验细则

4.1.1 数据集

本文使用了两个数据集：训练蛋白质残基接触预测模型作为预训练模型的 Pdataset 和 RNA 碱基相互作用预测任务的 Rdataset。

Pdataset：先从 PDB 数据库[20]下载并去除蛋白质之间相似度大于 30%的蛋白质，再去除序列长度大于 300 的蛋白质，然后随机从中选取 2000 个蛋白质作为训练集，200 个蛋白质作为测试集。对训练集和测试集中的每一条蛋白质序列，使用多序列比对工具 HHblits[21]（参数设置：*E-value* = 0.01，*coverage* = 50%，*cutoff* = 30%）在 Uniclust30[22]数据库中进行 3 次迭代搜索，生成对应蛋白质序列的MSA_i和MSA_j；再通过残基接触定义计算真实的标签信息$Label_i$和$Label_j$，得到数据集 Pdataset{train set：$(MSA_i, Label_i)$，test set：(MSA_j ， $Label_j$)，i = 1,2,3…2000; j = 1,2,3…200 }。

Rdataset：采用 Sun 等[3]提供的数据集，由于计算资源有限，本文只选取了序列长度 $L \leq 300$ 的用于模型的训练与测试，即 Rdataset{train set：TR199，test set：TE79}。

4.1.2 特征选取

在蛋白质残基接触预测任务[23-24]中， 计算 MSA 的协方差矩阵被证明能更好地作为此类任务的输入特征。因此，本文同样利用 MSA 的协方差矩阵作为模型的输入特征，其计算公式如下：

$$COV(A_x, B_y) = f(A_x B_y) - f(A_x) f(B_y) \quad (7)$$

其中，x 和 y 表示 MSA 的第 x 列和第 y 列；A, B 表示残基或碱基的类型；$f(A_x B_y)$表示 MSA 中第 x 列出现 A 同时第 y 列出现 B 的频率；$f(A_x)$为 MSA 中第 x 列出现 A 的频率，则得到的输入特征维度为$L \times L \times 21 \times 21$或$L \times L \times 5 \times 5$（21 表示 20 种残基类型和 MSA 中未比对上而补充的“-”，5 表示 4 种碱基类型和“-”,L 为序列的长度）

4.1.3 评价指标

每个 RNA 序列长度 L 的不一致性，造成了不易评价方法之间的优劣性。本文采用 CASP（Critical Assessment of Structure Prediction）给出的蛋白质残基接触预测的评价方法，计算以序列长度而划分的不同范围的精度，分为 top L, $L/2$, $L/5$, $L/10$（即选择序列前端数量作为预测的接触集合，如 $L/2$ 为序列的前一半）。因此，RNA 碱基相互作用预测比较 4 个范围的短程、中程、长程相互作用的预测精度，精度计算公式如下：

$$precision = \frac{TP}{TP+FP} \quad (8)$$

其中，TP 表示模型判定的相互作用碱基对是正确的数量，FP 表示模型判定的相互作用的碱基对是错误的数量。

4.1.4 实验环境

所有实验均在配置为 Intel CPU i5 9600KF 和 GPU NVIDIA GTX 1660Ti 的 Windows10 操作系统中装有 Pytorch=1.5.1 版本的深度学习框架下完成。实验采用相同的参数设置，如训练的次数 *epoch*=200；学习率 *lr* 随 *epoch* 变化，即 *epoch* 为 0~100 时 *lr*=0.01，*epoch* 为 100~200 时 *lr*=0.001；优化器为 Adam；Batch size 按序列长度 L 设置为 1（L>150），2（150>L>100）和 3（L<100）。

实验所用代码与数据集已上传于 GitHub[1)]。

4.2 MFF 模块的效果

为验证 MFF 的有效性，在 Rdataset 数据集上进行消融实验，比较卷积神经网络 MFFnet 与 ResNet 和 MobileNetV2 的预测精度和模型复杂度，结果如表 1 和表 2 所列。

表 1　不同网络结构的预测结果

Table 1 Prediction results of different network structures

Model				*L*/2			*L*/5			*L*/10		
	short	mid	long	short	mid	long	short	mid	long	short	mid	long
ResNet34	0.37	0.431	0.509	0.523	0.564	0.657	0.655	0.703	0.75	0.698	**0.746**	0.785
ResNet36	**0.379**	0.431	0.508	0.536	0.577	0.658	0.685	0.692	0.768	0.707	0.716	0.789
ResNet38	0.368	0.429	0.517	0.536	0.57	0.650	0.66	0.678	0.752	0.705	0.72	0.791
ResNet40	0.37	0.439	0.5	**0.547**	**0.6**	0.657	**0.69**	0.715	0.75	0.729	0.741	0.789
ResNet42	0.365	0.439	0.517	0.533	0.598	0.665	0.687	0.714	0.756	0.716	0.735	**0.804**
[1]MobileNet33	0.344	0.397	0.468	0.493	0.553	0.627	0.631	0.696	0.726	0.664	0.733	0.774
[2]MobileNet33	0.353	0.411	0.485	0.512	0.567	0.655	0.659	0.692	0.748	0.714	0.724	0.773
MFFnet30	0.37	**0.443**	0.504	0.532	0.596	0.648	0.671	0.705	0.751	0.731	0.736	0.789
MFFnet36	0.37	0.431	0.51	0.54	0.582	0.652	0.68	0.709	0.757	0.725	0.731	0.791
MFFnet42	0.37	0.431	**0.523**	0.53	0.576	**0.672**	0.679	**0.72**	**0.771**	**0.748**	0.741	0.803

表 2　不同网络结构的模型复杂度

Table 2 Model complexity of different network structures

Model	*Flops* / G	*Params* / M
ResNet34	50.31	1.26
ResNet36	53.27	1.33
ResNet38	56.22	1.4
ResNet40	59.17	1.48
ResNet42	62.13	1.55
[1]MobileNet33	**7.26**	**0.179**
[2]MobileNet33	7.86	0.194
MFFnet30	7.86	0.198
MFFnet36	9.41	0.237
MFFnet42	10.97	0.276

表 1 中，*L*，*L*/2，*L*/5 和 *L*/10 为 4.1.3 节划分的指标计算范围，short，mid 和 long 则表示在计算范围内的短程、中程、长程碱基相互作用。表 2 中，*Flops* 表示计算量，即时间复杂度，计算量越大，消耗的时间越多，单位为 G；*Params* 表示模型参数量，即空间复杂度，参数量越多，占用的计算资源越多，单位为 M。ResNet34 表示卷积层数是 34 且通道数保持 64 的 ResNet；[1]MobileNet33 是卷积层数为 33 的 MobileNetV2，InvertResidual 的中间层通道数是输入通道数的 2 倍，整体通道数变化为 64-64-80；[2]MobileNet33 则表示整体通道数变化为 64-64-96；MFFnet30 表示卷积层数为 30 且通道数保持 64 的 MFFnet。

采用输入特征为 $C\times W\times H = 25\times200\times200$ 计算模型的复杂度。从表 2 中可以发现，当网络层数相同时，MFFnet 的模型复杂度远低于 ResNet 且相当于轻量型网络 MobileNetV2，如 MFFnet36 分别在计算量 *Flops* 与参数量 *Params* 上仅有 ResNet36 的 17.66%和 17.82%，MFFnet42 是 ResNet42 的 17.66%和 17.81%；[2]MobileNet33 与 MFFnet30 具有相同的模型复杂度，而 MFFnet36 分别只比 [2]MobileNet33 高了 1.55G 和 0.043M。

从表 1 可以发现，通过增加通道数，可以提高模型的性能，如 [2]MobileNet33 的预测精度在整体上高于 [1]MobileNet33。MFFnet 的预测精度远高于 MobileNetV2，同时 MFFnet36 的预测精度整体高于 ResNet36，在 *L*/10 的短程相互作用上提高了 2.5%。MFFnet42 获得了更好的结果，并在 *L*/10 的短程相互作用上比 ResNet42 提高了 4.4%。

总体上，MFFnet 的预测精度高于 ResNet，并且模型复杂度要远低于 ResNet，表明分支融合特征的形式有助于提升模型的性能；同时其与轻量型网络的模型复杂度相近，但预测精度远高于 MobileNetV2。MFFnet 在处理此类数据特性的预测问题上可以获得更优异的结果，具有计算成本低、特征提取能力强等优点。

4.3　损失函数的效果

首先在 Rdataset 数据集使用 MFFnet36 作为

1) https://github.com/xiaowang121/RNA

神经网络骨架进行损失函数的比较，确定式（5）中调节因子 α 的大小和式 6 中 β 的大小。最后证明 tpLoss 在不同神经网络结构中预测 RNA 碱基相互作用的有效性。

通过观察 *L*/5 和 *L*/10 的预测结果确定调节因子 α 的大小， 图 4 和图 5 分别显示了 *L*/5 与 *L*/10 范围内的短、中、长程碱基相互作用预测精度随 α 值的变化， no 表示未用 WCE 训练的结果。可以发现，$\alpha = 0.5$时的结果都要低于 no，即倍数性地降低整体 Loss 会减弱模型对正负类别的判别能力，进而影响模型的性能。而$\alpha = 0.2$与$\alpha = 0.6$的训练结果都优于 no，因为在计算 Loss 时通过增加正样本或增加负样本的权重可以有效地抑制类别不平衡问题，提高模型对此类数据的分类性能。而$\alpha = 0.2$的效果优于$\alpha = 0.6$ 的效果，因为负样本的数量远多于正样本的数量，对计算负样本 Loss 部分增加大于正样本的权重会提高模型对负样本的偏好，使模型能更准确地区分负样本，进而提高分类的精度。

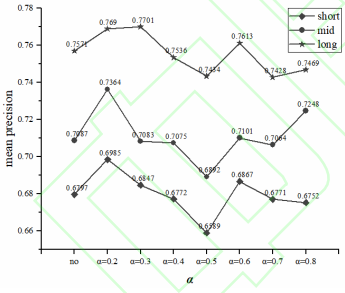

图 4　*L*/5 范围内的预测结果

Fig. 4 Prediction results in L/5 range

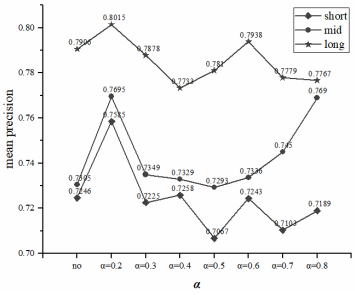

图 5　*L*/10 范围内的预测结果

Fig. 5 Prediction results in L/10 range

在此基础上添加辅助损失规范模型对数据的对称性进行学习，以$\beta = 0.5$为中点上下取值确定 β。表 3 收集了不同损失函数下的训练结果和添加不同 β 值的辅助损失训练结果，其中 CE 表示 MFFnet36 采用 CE 训练；CE[1] 为 MFFnet36 采用 CE，辅助损失和全局损失的权重都为 1 的模型训练；CE[2] 表示 MFFnet36 采用 CE 加入$\beta = 0.1$ 时的辅助损失用于训练；baseline 为 MFFnet36 采用$\alpha = 0.2$时的 WCE 进行训练；+β 则表示在 baseline 的基础上加入不同 β 值进行训练。

表 3　不同损失函数的比较

Table 3 Comparison of different loss functions

Loss	*L*			*L*/2			*L*/5			*L*/10		
	short	mid	long	short	mid	long	short	mid	long	short	mid	long
CE	0.37	0.431	0.51	0.54	0.582	0.652	0.68	0.709	0.757	0.725	0.731	0.791
CE[1]	0.364	0.419	0.516	0.518	0.586	0.664	0.664	0.710	0.769	0.728	0.748	0.805
CE[2]	**0.384**	0.44	0.514	0.545	0.588	0.659	0.698	0.703	0.749	0.739	0.736	0.793
baseline	0.38	0.432	0.522	0.535	0.58	0.663	0.699	**0.736**	0.769	0.759	0.77	0.802
+β=0.1	0.381	**0.449**	**0.525**	**0.55**	**0.608**	**0.683**	**0.702**	0.735	0.777	**0.76**	**0.773**	0.804
+β=0.2	0.382	0.429	0.522	0.54	0.591	0.66	0.688	0.711	0.749	0.744	0.738	0.793
+β=0.3	0.376	0.433	0.497	0.537	0.589	0.641	0.687	0.705	0.754	0.734	0.722	0.781
+β=0.4	0.36	0.415	0.501	0.508	0.564	0.652	0.652	0.69	0.758	0.725	0.744	0.779
+β=0.5	0.378	0.437	0.505	0.531	0.585	0.666	0.666	0.709	**0.778**	0.712	0.731	**0.809**
+β=0.6	0.376	0.417	0.506	0.548	0.57	0.666	0.693	0.71	0.759	0.753	0.757	0.789
+β=0.7	0.377	0.404	0.505	0.524	0.541	0.654	0.658	0.668	0.764	0.684	0.700	0.792

+β=0.8	0.372	0.434	0.513	0.533	0.580	0.656	0.660	0.703	0.751	0.724	0.736	0.778

对比 baseline 与不同的+β，当辅助损失的权重过大时，模型的预测精度会整体降低；较小的β，会增加 L 与 $L/2$ 的短、中程预测精度，但较大幅度降低了对 $L/5$ 与 $L/10$ 的短、中程预测；$\beta = 0.1$时会增加整体的预测精度，因为辅助损失是由模型的中间部分计算，而模型的前半部分不能充分地学习到数据的特征分布，权重过大会扰乱模型对原有的特征分布学习，故辅助损失的占有权重要低于全局损失。

对比 CE，CE^2，baseline 和+β=0.1 可以看出，加权交叉熵和辅助损失的引入都不同程度地提升了模型的性能，辅助损失能大幅提升模型在 L 和 $L/2$ 的短、中程的预测精度，而加权交叉熵则可以全面提高精度，并且二者的结合具有 1+1>2 的效果。对比 CE，CE^1 和 CE^2 可以看出，CE^1 提升了 $L/5$ 和 $L/10$ 的中、长程预测精度，但低于 CE^2 的 L 和 $L/2$ 的短、中程预测精度，即在大的计算范围内和对角线周围点的预测上，CE^1 的效果逊于 CE^2。为了更好地在整体计算范围内提升 RNA 碱基相互作用模型的性能，采用+β=0.1 的形式计算 Loss 作用于梯度更新，进而达到对训练过程的优化目的。

为了验证 tpLoss 在处理此类数据特性问题上的普遍性，采用不同的损失策略训练 ResNet36 和 [1]MobileNet33，结果如图 6 和图 7 所示。其中，+α 表示采用$\alpha = 0.2$时的 WCE 训练，+α+β 表示采用$\alpha = 0.2$时的 WCE 并加入$\beta = 0.1$的辅助损失进行训练。

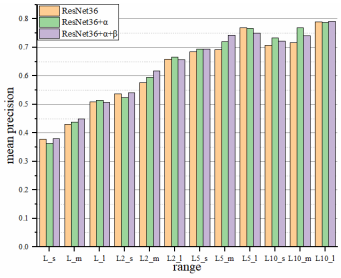

图 6 ResNet：不同损失计算策略

Fig. 6 ResNet: different loss calculation strategies

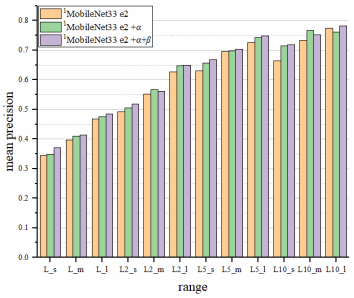

图 7 MobileNetV2：不同损失计算策略

Fig. 7 MobileNetV2: different loss calculation strategies

可以发现，两种损失策略对 ResNet 在中程的碱基相互作用预测精度上都有显著的提升。虽然在 L 与 $L/2$ 的短程预测上+α 的策略低于基础的收益，但+α+β 策略能在整体上提高 ResNet 对此类数据特性的预测性能。图 7 清晰地显示了+α 和+α+β 策略都大幅度提升了 MobileNetV2 的预测性能，并且在总体上+α+β 产生的效果大于单一策略+α。表 3、图 6 和图 7 说明， tpLoss 能更好地适用于此类数据特性的任务，优化卷积神经网络的训练过程，提高模型的准确性。

4.4 迁移学习的效果

本文引入迁移学习是为了处理因数据量少造成的模型不充分学习问题，将相似数据分布的蛋白质残基接触模型作为预训练模型，补充 RNA 碱基相互作用模型对特征的学习能力。如图 1 所示，首先采用 ResNet18 在 Pdataset 数据集中训练蛋白质残基接触模型，确定预训练模型 pre_protein；再在 Rdataset 数据集上训练加载了 pre_protein 权重参数的 RNA 碱基相互作用预测模型，分别比较 MFFnet，ResNet 和 MobileNetV2 加入预训练模型后的预测精度，并对 TE79 的预测精度进行逐一的比较，结果如图 8 和图 9 所示。

[1)] https://github.com/xiaowang121/RNA

DOI: 10.11896/jsjkx.211200186

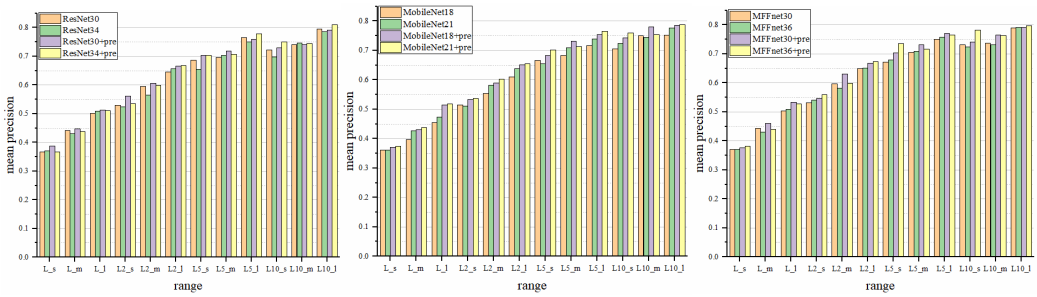

图 8　引入迁移学习的效果

Fig. 8　Effect of introducing transfer learning

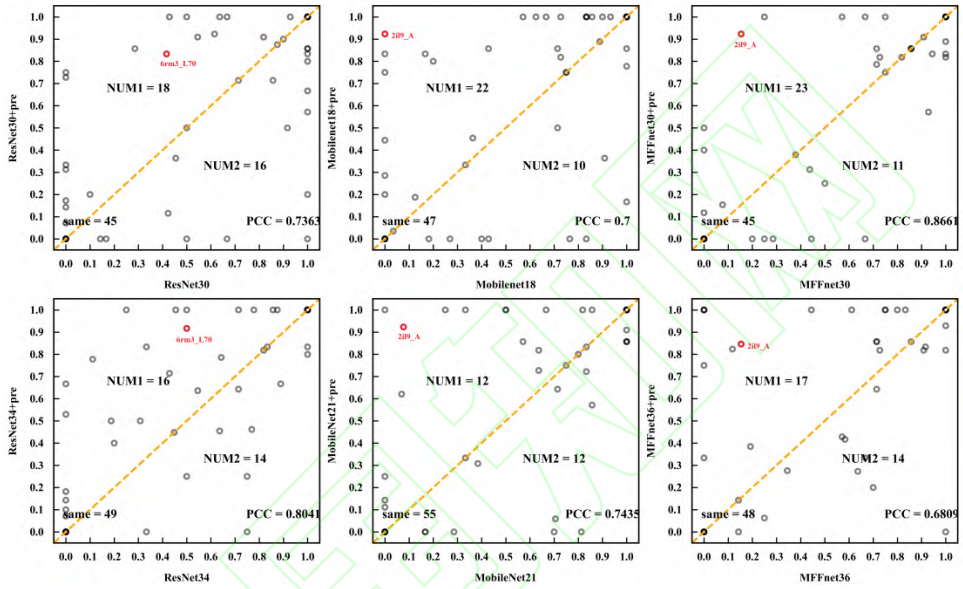

图 9　TE79 数据集在 *L*/10 范围内的短程预测精度的分布（电子版为彩图）

Fig. 9　Distribution of short-range prediction precision of TE79 dataset in L/10 range

图 8 显示，引入迁移学习后，能不同程度地提升 RNA 碱基相互作用的预测精度，并且对 MobileNetV2 具有大幅度的提高；卷积层数低的 ResNet 高于卷积层数高的预测精度，尤其在短程预测上较为明显；而且 *L* 和 *L*/2 上的中程以及 *L*/5 和 *L*/10 的长程预测也较低，但迁移学习的引入改善了 ResNet 结构在卷积层数较高时对 *L*/5 和 *L*/10 的预测缺陷。MFFnet 和 ResNet 的结构类似，可以发现它们预测的结果呈现相似的分布，但在 12 个不同范围的预测中 MFFnet 要整体优于 ResNet。

图 9 中，NUM1 表示引入迁移学习后的预测精度大于未引入迁移学习时的 RNA 数量，NUM2 则为前者少于后者的数量，same 表示预测结果相同的数量；PCC 表示模型之间的相关性。可以发现，当 ResNet 和 MobileNetV2 的卷积层数增加时，PCC 值随之增加；而 MFFnet 则相反，并且 MFFnet36 的 PCC 只有 0.6809，即迁移学习对 MFFnet36 的提升相似度不高，具有差异性，可以更好地应用于 RNA 碱基相互作用预测。同时，迁移学习提高了难预测的 RNA 碱基相互作用预测精度，如图 9 中红色标出的 6rm3_L70 和 2il9_A。迁移学习的引入会降低个别 RNA 在 *L*/10 的短程预测精度，但总体趋势优于未引入迁移学习的效果。

图 8 和图 9 表明，引入迁移学习可以缓解模

型因数据量少而学习不充分的问题，提高难预测样本的精度，但并非全面地提高预测精度，原因在于：1） pre_protein 的训练是利用 ResNet18 结构，可能未完全学习到数据的特征分布；2）蛋白质残基有 20 种类型，而 RNA 碱基只有 4 种，因此加载 pre_protein 会使得模型更偏好于预测 RNA 中数量较多的短、中程碱基相互作用。

4.5 方法比较

Zhang 等[25]指出，MSA 的有效性会影响利用 MSA 的预测任务精度。通过计算有效的 MSA 序列条数确定 MSA 质量，计算公式如下：

$$Neff = \frac{1}{\sqrt{L}} \cdot \sum_{n=1}^{N} \frac{1}{1+\sum_{m=1,m\neq n}^{N} d[\partial_{m,n} \geq 0.8]} \quad (9)$$

其中，L 为 RNA 序列长度；N 为 MSA 中的序列条数；$\partial_{m,n} \geq 0.8$表示 MSA 中第 m 条与第 n 条序列的相似度大于或等于 0.8；d[*]表示当*成立时为 1，否则为 0。

通常计算出 *Neff* 小于 1 的 MSA 是低质量 MSA，该样本是难预测样本[26]。因此本文根据计算的 *Neff* 值，以 1 和 10 为界限，将 TE79 划分为 3 个子集（*Neff* 为 0~1 的 TE43，*Neff* 为 1~10 的 TE20，*Neff* 大于 10 的 TE16），进一步比较模型在不同 MSA 质量下的预测精度。图 10 给出在 TE16 和 TE43 上的 L 和 $L/2$ 范围的预测结果，可以看出 MobileNetV2 的预测精度低于 ResNet 和 MFFnet，说明 MSA 的有效序列数过多或过少都会影响该结构的性能。在 TE20 中，ResNet 与 MobileNetV2 大致趋于领先，但 ResNet 对短程预测的效果会低于 MFFnet，并且在 TE43 的预测精度大幅度低于 MFFnet，尤其在中程相互作用的预测上，而且 tpRNA 在低输入特征下的预测精度有绝对的优势。

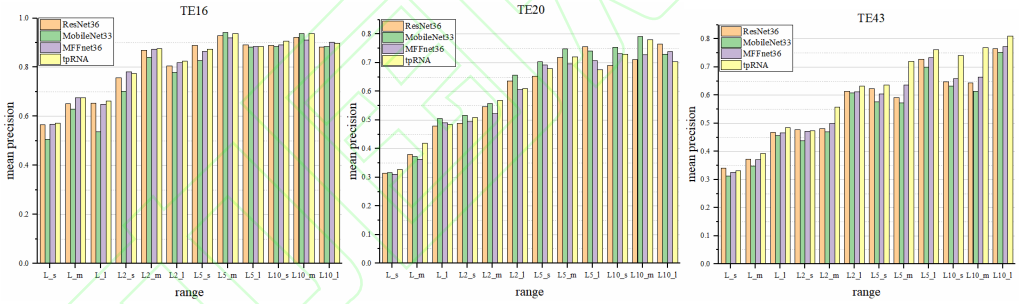

图 10 三个子集的模型测试结果

Fig. 10 Test results of model in three subsets

表 4 列出了在 TE79 的测试结果。与最先进的方法 RNAcontact 相比，在 $L/10$ 的长程相互作用预测上，tpRNA 已超过 RNAcontact。尽管 tpRNA 没有在整体范围上都优于 RNAcontact，但是 RNAcontact 采用的是 ResNet 结构，而且卷积神经网络的通道数很大，tpRNA 的训练集数量比 RNAcontact 少 11%。因此，RNAcontact 的模型复杂度要远大于 tpRNA，而且 tpRNA 可以在数据量少的情形下取得较好的精度。对比 ResNet36，MFFnet36 和 [2]MobileNet33，tpRNA 的预测精度具有全面的优势，更适用于此类数据特性的任务。

表 4 TE79 数据集的预测结果

Table 4 Prediction results of TE79 dataset

Model	*L*			*L*/2			*L*/5			*L*/10		
	short	mid	long	short	mid	long	short	mid	long	short	mid	long
ResNet36	0.379	0.431	0.508	0.536	0.577	0.658	**0.685**	0.692	0.768	0.707	0.716	0.789

1) https://github.com/xiaowang121/RNA

[2]MobileNet33	0.353	0.411	0.485	0.512	0.567	0.655	0.659	0.692	0.748	0.714	0.724	0.773
MFFnet36	0.37	0.431	0.51	0.54	0.582	0.652	0.68	0.709	0.757	0.725	0.731	0.791
RNAcontact	-	-	**0.59**	-	-	**0.73**	-	-	**0.8**	-	-	0.81
tpRNA	**0.385**	**0.452**	0.541	**0.548**	**0.612**	0.675	0.681	**0.731**	0.776	**0.736**	**0.794**	**0.818**

表 4 和图 10 表明，本文提出的方法 tpRNA，在 RNA 碱基相互作用预测任务上，通过处理数据量少和特异性数据所产生的问题，能进一步地提高模型的性能，具有计算成本低和精度高的优势。

结束语 RNA 碱基相互作用的准确预测对其结构预测具有至关重要的作用。本文提出利用迁移学习将类似特征分布的蛋白质残基接触模型运用于 RNA 碱基相互作用预测任务，改善了因 RNA 数据量少对模型学习造成的影响。此外，设计的一种不同尺度的特征融合模块 MFF，丰富了空间特征信息。同时，特制的损失函数可以优化梯度更新，规范了此类特性数据的模型训练。相比于目前最先进的方法，tpRNA 模型在 *L*/10 的长程预测精度上获得了最优的结果。tpRNA 虽然取得了较好的结果，但还有很多不足，如存在预训练模型的训练问题，协方差矩阵是否为最有效的输入特征等，因此下一步将对预训练模型进行优化，并考虑使用逆协方差矩阵作为输入特征。

参考文献

[1] ZHANG T, SINGH J, LITFIN T, et al. RNAcmap: A Fully Automatic Pipeline for Predicting Contact Maps of RNAs by Evolutionary Coupling Analysis [J]. Bioinformatics, 2021, DOI: 10.1093/bioinformatics/btab391.

[2] DE L E, LUTZ B, RATZ S, et al. Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction [J]. Nucleic Acids Res, 2015, 43(21):10444-10455.

[3] SUN S, WANG W, PENG Z, et al. RNA inter-nucleotide 3D closeness prediction by deep residual neural networks [J]. Bioinformatics, 2021, 37(8): 1093-1098.

[4] LIU W Y, GUO Y B, LI W H. Identifying Essential Proteins by Hybrid Deep Learning Model [J]. Computer Science, 2021, 48(8): 240-245.

[5] WU Q, PENG Z, ANISHCHENKO I, et al. Protein contact prediction using metagenome sequence data and residual neural networks [J]. Bioinformatics,2020,36(1):41-48.

[6] XIE L X, LI F, XIE J P, et al. Predicting Drug Molecular Properties Based on Ensembling Neural Networks Models[J]. Computer Science, 2021, 48(9): 251-256.

[7] PAN S, YANG Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016:770-778.

[9] MORCOS F, PAGNANI A, LUNT B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families [J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(49):E1293-E1301.

[10] MARKS D, COLWELL L, SHERIDAN R, et al. Protein 3D structure computed from evolutionary sequence variation [J].PLoS One, 2011,6(12):e28766.

[11] EKBERG M, LOVKVIST C, LAN Y, et al. Improved contact prediction in proteins: using pseudo likelihoods to infer Potts models [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2013,87(1):012707.

[12] JIAN Y, WANG X, QIU J, et al. DIRECT: RNA contact predictions by integrating structural patterns [J]. BMC Bioinformatics,2019,20(1):497.

[13] LI Y, HU J, ZHANG C, et al. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks [J]. Bioinformatics, 2019,35(22):4647-4655.

[14] YU F, KOLTUN V, FUNKHOUSER T. Dilated Residual Networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017:636-644.

[15] YANI I, DUNCAN R, ROBERTO C, et al. Deep Roots: Improving CNN Efficiency with Hierarchical Filter Groups [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017:5977-5986.

[16] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press,2018:7132-7141.

[17] OLAF R, PHILIPP F, THOMAS B. U-Net: Convolutional Networks for Biomedical Image Segmentation [J]. Medical Image Computing and Computer-Assisted Intervention, 2015, 9351:234-241.

[18] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018:4510-4520.

[19] LIN T, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection [C]// Proceedings of IEEE International Conference on Computer Vision. Venice: IEEE Press, 2017:2999-3007.

[20] BERMAN HM, WESTBROOK J, FENG Z, et al. The Protein Data Bank [J]. Nucleic Acids Res, 2000,28(1):235-242.

[21] REMMERT M, BIEGERTA, HAUSER A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment [J]. Nature Methods, 2011,9(2):173-175.

[22] The UniProt Consortium. UniProt: the universal protein knowledgebase [J]. Nucleic Acids Res, 2017,45(D1):D158-D169.

[23] JONES DT, SINGH T, KOSCIOLEK T, et al. Meta PSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins [J]. Bioinformatics, 2015,31(7):999-1006.

[24] JONES D T, KANDATHIL S M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features [J]. Bioinformatics,2018, 34(19): 3308-3315.

[25] ZHANG C X, ZHENG W, MORTUZA S M, et al. Deep MSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins [J]. Bioinformatics,2020,36(7):2105-2112.

[26] CHEN M C, LI Y, ZHU Y H, et al. SSCpred: Single-Sequence-Based Protein Contact Prediction Using Deep Fully Convolutional Network [J]. Journal of Chemical Information and Modeling, 2020, 60(6):3295-3303.

WANG Xiaofei, born in 1995, postgraduate. His main research interests include computer vision and bioinformatics.

LI Zhangwei, born in1967, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include intelligent information processing and so on.

（责编：柯颖）

1) https://github.com/xiaowang121/RNA