基于级联卷积和递归神经网络的蛋白质二级结构预测

电子与信息工程学院　　计算机科学与技术
118532014013　　袁 超　　指导教师：游文杰

总览

蛋白质预测

什么是蛋白质的二级结构？

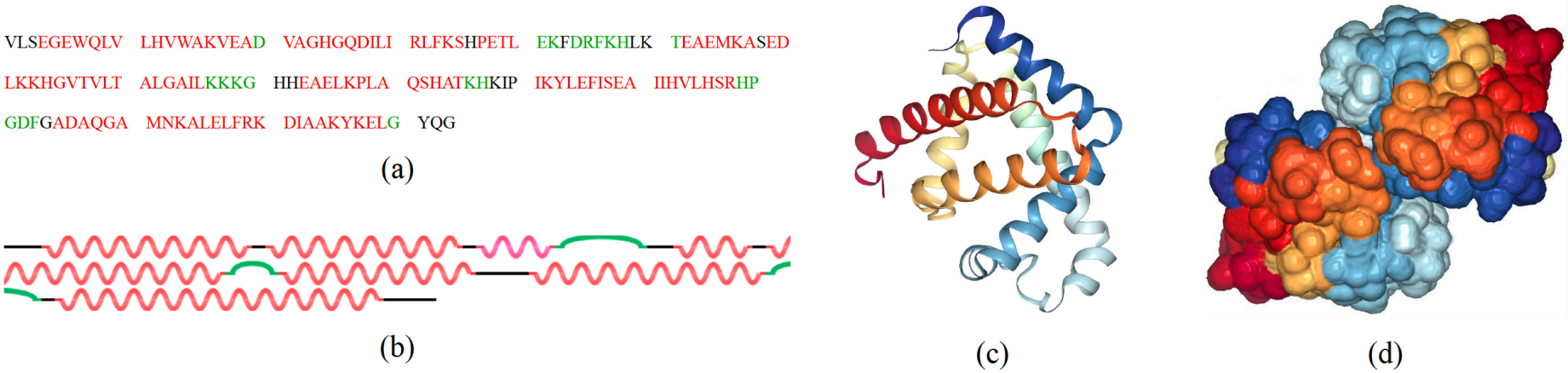


图: 数据来自 RCSB(https://www.rcsb.org/) 的蛋白质 PDB 105M

蛋白质预测

八类蛋白质二级结构预测的符号化定义：

给定一条蛋白质输入序列 $x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})^T$，存在模型 f，使得

$$y = f(x) = (y^{(1)}, y^{(2)}, \cdots, y^{(n)})^T \tag{1}$$

如果 $y^{(i)} \in \{H, G, I, E, B, T, S, C\}_8$，则称 y 为序列 x 的八类二级结构预测结果，$y^{(i)}$ 为第 i 个残基所对应的二级结构，共八种类别。

设第 i 条蛋白质序列 x_i 有 n_i 个残基，其中正确预测有 m_i 个残基，则

$$Q_8 = \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{N} n_i} \tag{2}$$

第 i 条蛋白质的 Q_8 准确率

$$q_i = \frac{m_i}{n_i} \tag{3}$$

也可以参考准确率

$$r = \frac{1}{N} \sum_{i=1}^{N} \frac{m_i}{n_i} \tag{4}$$

代价函数：

$$J_\theta = \frac{1}{N}\sum_{j=1}^{N} L(y_j, \hat{f}_\theta(x_j)) \tag{5}$$

梯度下降：

$$\theta \leftarrow \theta - \eta \nabla_\theta J \tag{6}$$

预备知识

位置特异得分矩阵（PSSM）

Article No. jmbi.1999.3091 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (1999) **292**, 195–202

JMB

COMMUNICATION

Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices

David T. Jones

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL United Kingdom

A two-stage neural network has been used to predict protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST. Despite the simplicity and convenience of the approach used, the results are found to be superior to those produced by other methods, including the popular PHD method according to our own benchmarking results and the results from the recent Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP3), where the method was evaluated by stringent blind testing. Using a new testing set based on a set of 187 unique folds, and three-way cross-validation based on structural similarity criteria rather than sequence similarity criteria used previously (no similar folds were present in both the testing and training sets) the method presented here (PSIPRED) achieved an average Q_3 score of between 76.5% to 78.3% depending on the precise definition of observed secondary structure used, which is the highest published score for any method to date. Given the success of the method in CASP3, it is reasonable to be confident that the evaluation presented here gives a fair indication of the performance of the method in general.

Keywords: protein structure prediction; secondary structure; protein folding; sequence analysis; neural network

预备知识

Fixed-size Ordinally-Forgetting Encoding

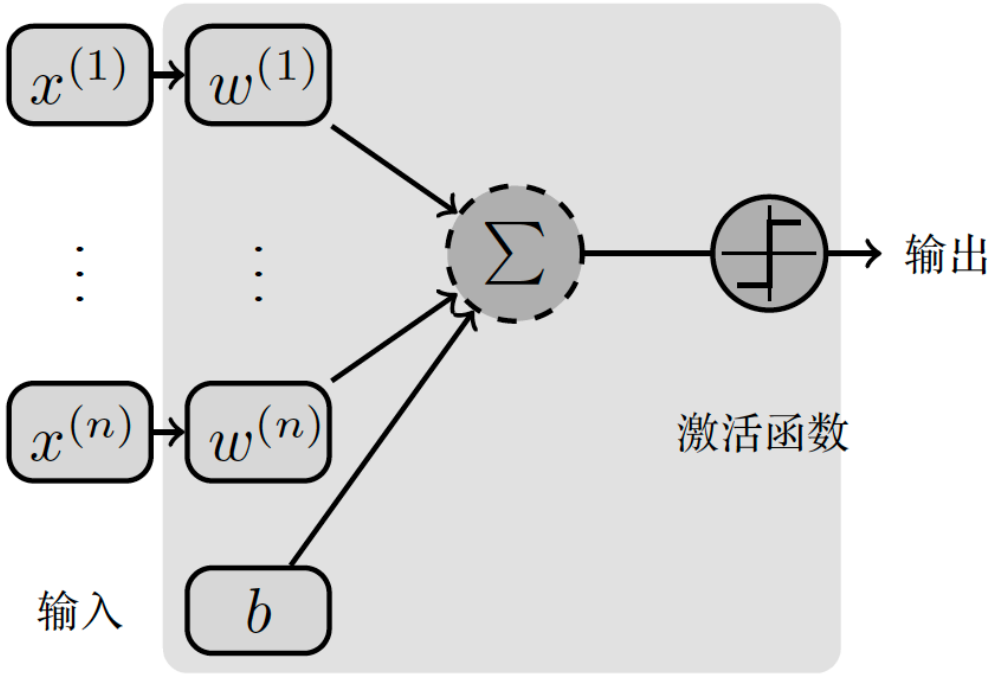
给定一条蛋白质序列 x 及其 one-hot 编码 x'，则 x 的 FOFE 编码矩阵 $A_{fofe} = (e^{(1)}, e^{(2)}, \cdots, e^{(t)})^T$

$$e^{(t)} = \alpha e^{(t-1)} + x'^{(t)} \tag{7}$$

PROTEIN SEQUENCE	FOFE
$w^{(6)},$	$0, 0, 0, 0, 0, 0, 1$
$w^{(6)}, w^{(4)},$	$0, 0, 0, 0, 1, 0, \alpha$
$w^{(6)}, w^{(4)}, w^{(5)},$	$0, 0, 0, 0, \alpha, 1, \alpha^2$
$w^{(6)}, w^{(4)}, w^{(5)}, w^{(0)},$	$1, 0, 0, 0, \alpha^2, \alpha, \alpha^3$
$w^{(6)}, w^{(4)}, w^{(5)}, w^{(0)}, w^{(5)},$	$\alpha, 0, 0, 0, \alpha^3, 1+\alpha^2, \alpha^4$
$w^{(6)}, w^{(4)}, w^{(5)}, w^{(0)}, w^{(5)}, w^{(4)}$	$\alpha^2, 0, 0, 0, 1+\alpha^4, \alpha+\alpha^3, \alpha^5$

预备知识

神经元、多输入感知机

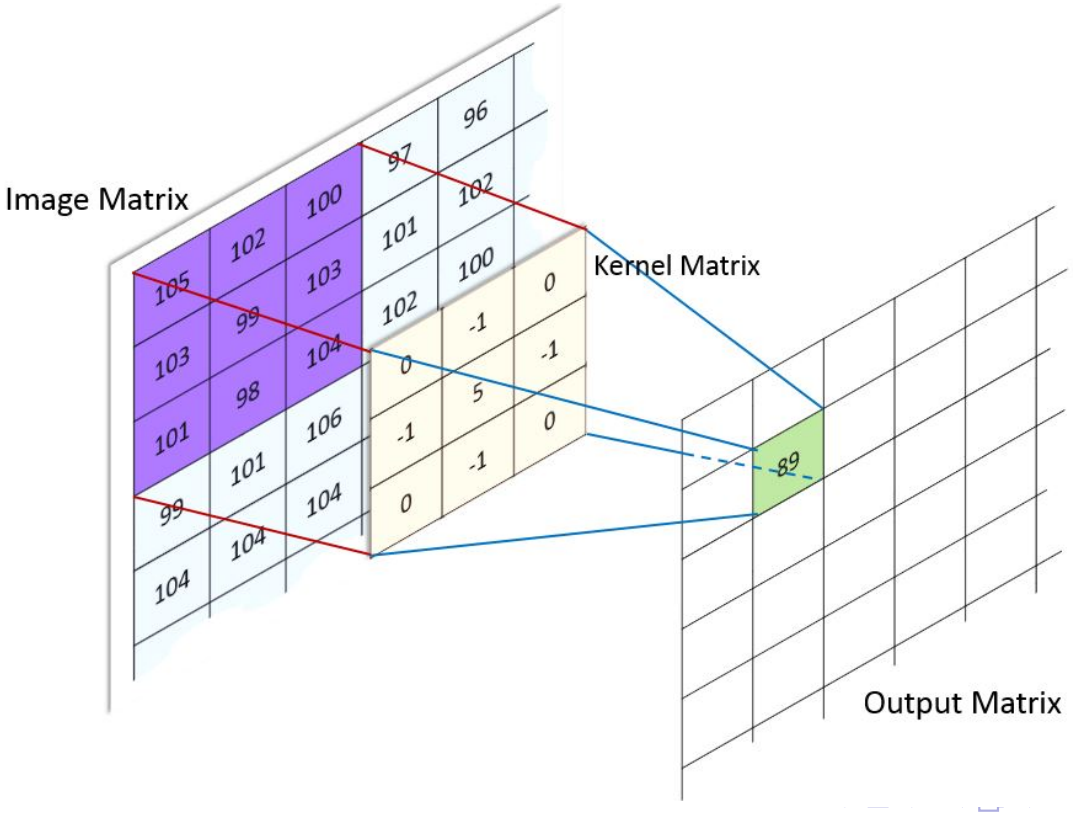


连接权重

$$f_\theta = \sigma(\sum_{i=1}^{n} w^{(i)} x^{(i)} + b) \tag{8}$$

设卷积核 $w \in K$，w 与矩阵 I 进行一次卷积运算，则矩阵 I 第 i 行第 j 列的运算结果为：

$$I'_{i,j} = (I * w)_{i,j} = \sum_m \sum_n I_{i+m,j+n} w_{m,n} \tag{9}$$

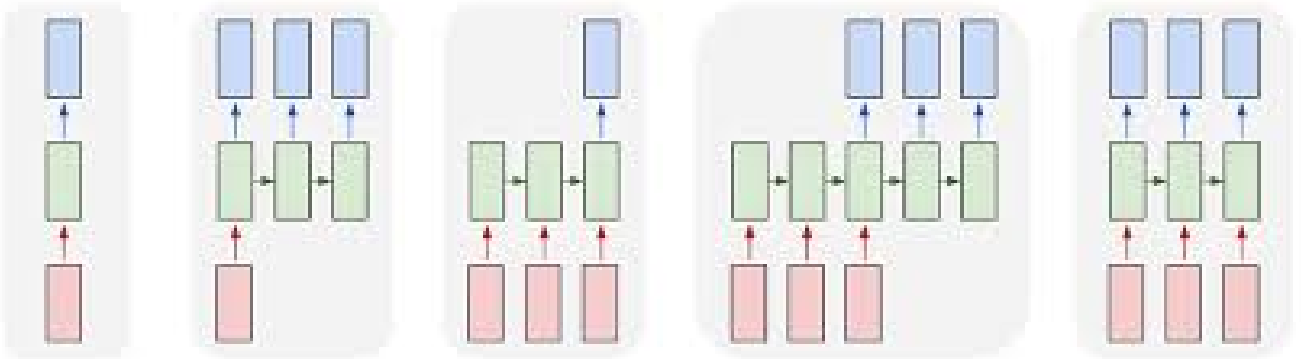


循环神经网络的子结构是一个动态系统

$$h^{(t)} = f_\theta(h^{(t-1)}, x^{(t)}) \tag{10}$$

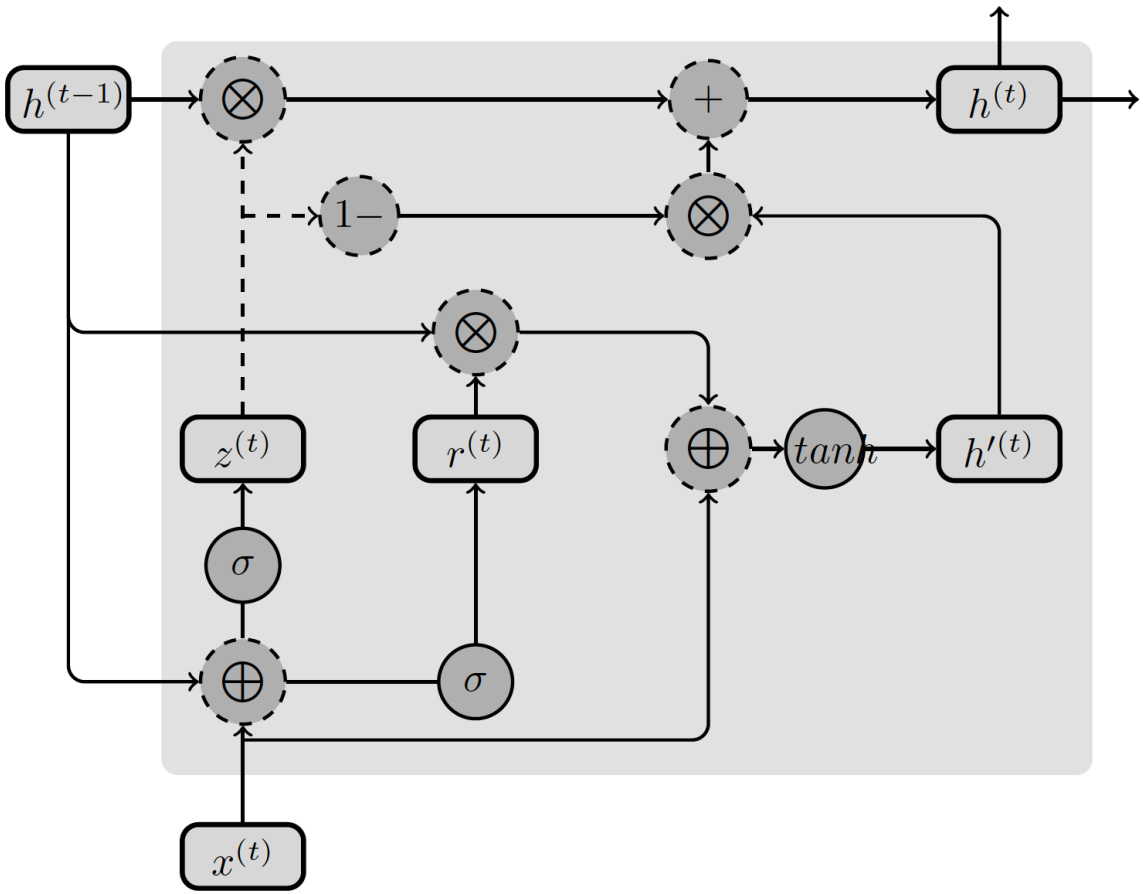
若保存所有迭代的输出，循环神经网络模型可以构建出如下的映射关系：

$$f_\theta : \{x^{(1)}, x^{(2)}, \cdots, x^{(t)}\} \rightarrow \{h^{(1)}, h^{(2)}, \cdots, h^{(t)}\} \tag{11}$$



预备知识

Gate Recurrent Unit



向量拼接

向量元素相乘

向量和

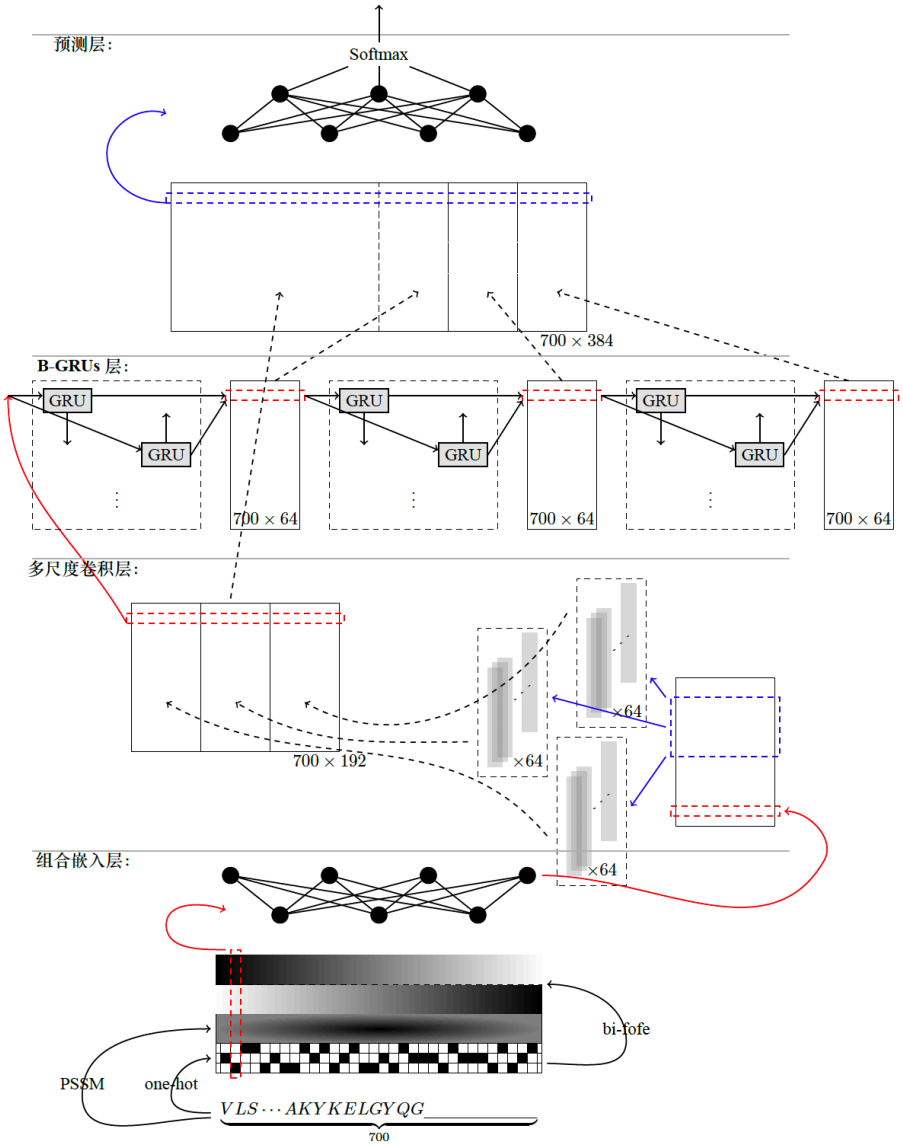
$$z^{(t)} = \sigma(W_z x^{(t)} \oplus U_z h^{(t-1)} + b_z) \tag{12}$$

$$r^{(t)} = \sigma(W_r x^{(t)} \oplus U_r h^{(t-1)} + b_r) \tag{13}$$

$$h'^{(t)} = tanh(W_h x^{(t)} \oplus U_h(r^{(t)} \otimes h^{(t-1)}) + b_h) \tag{14}$$

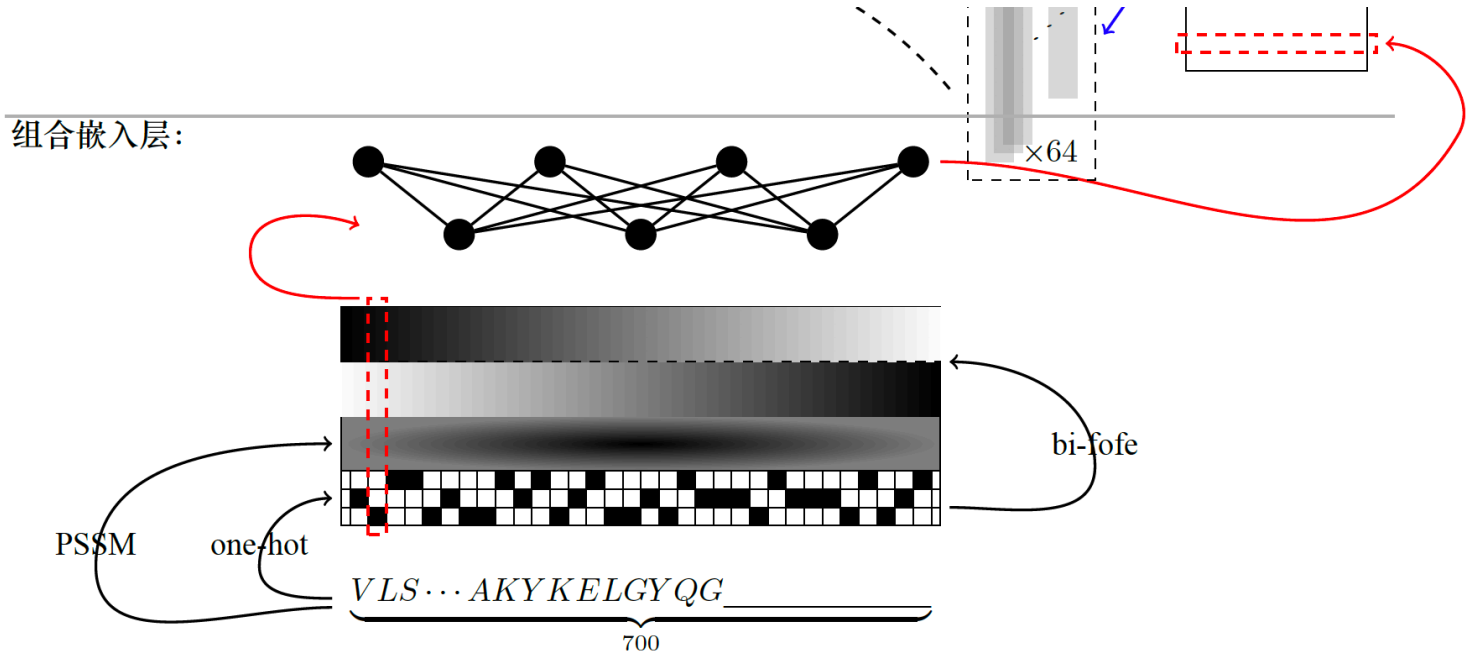
$$h^{(t)} = z^{(t)} \otimes h^{(t-1)} + (1 - z^{(t)}) \otimes h'^{(t)} \tag{15}$$

深度模型架构


预测层:
Softmax
700 × 384
B-GRUs 层:
GRU
GRU
700 × 64
700 × 64
700 × 64
多尺度卷积层:
700 × 192
×64
×64
×64
组合嵌入层:
bi-fofe
PSSM
one-hot
VLS···AKYKELGYQG
700

深度模型架构

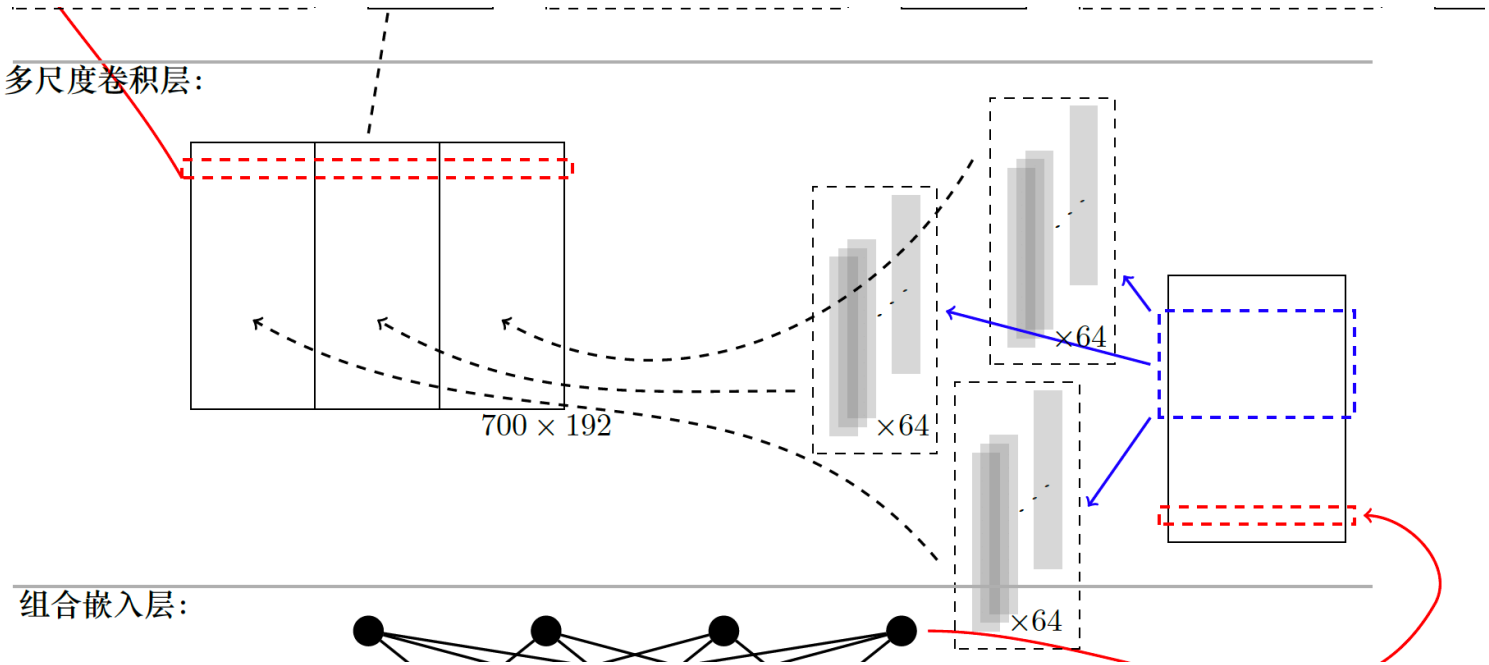
组合嵌入层



$$A = concatenate\{A_{0-1}, sigmoid(A_{pssm}), A_{bifofe}\} \tag{16}$$

深度模型架构

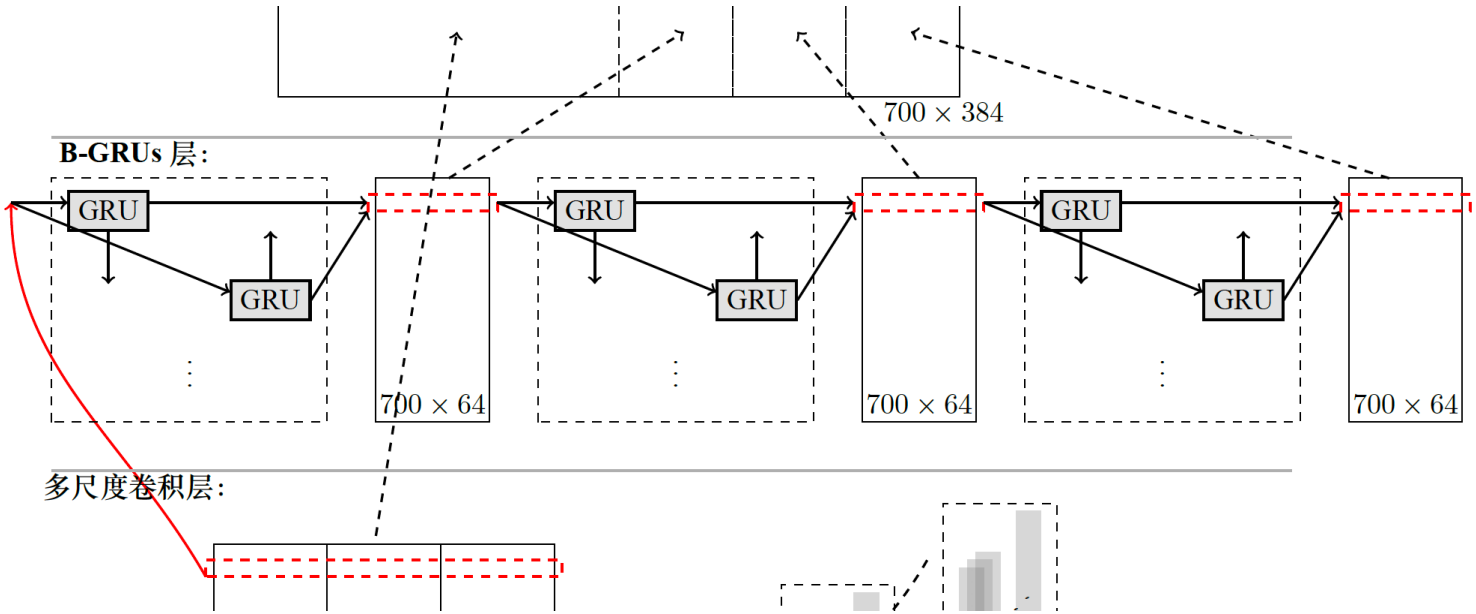
多尺度卷积层



$$C^{(i)} = ReLU((A * K^{(i)}) + b^{(i)}) \tag{17}$$

深度模型架构

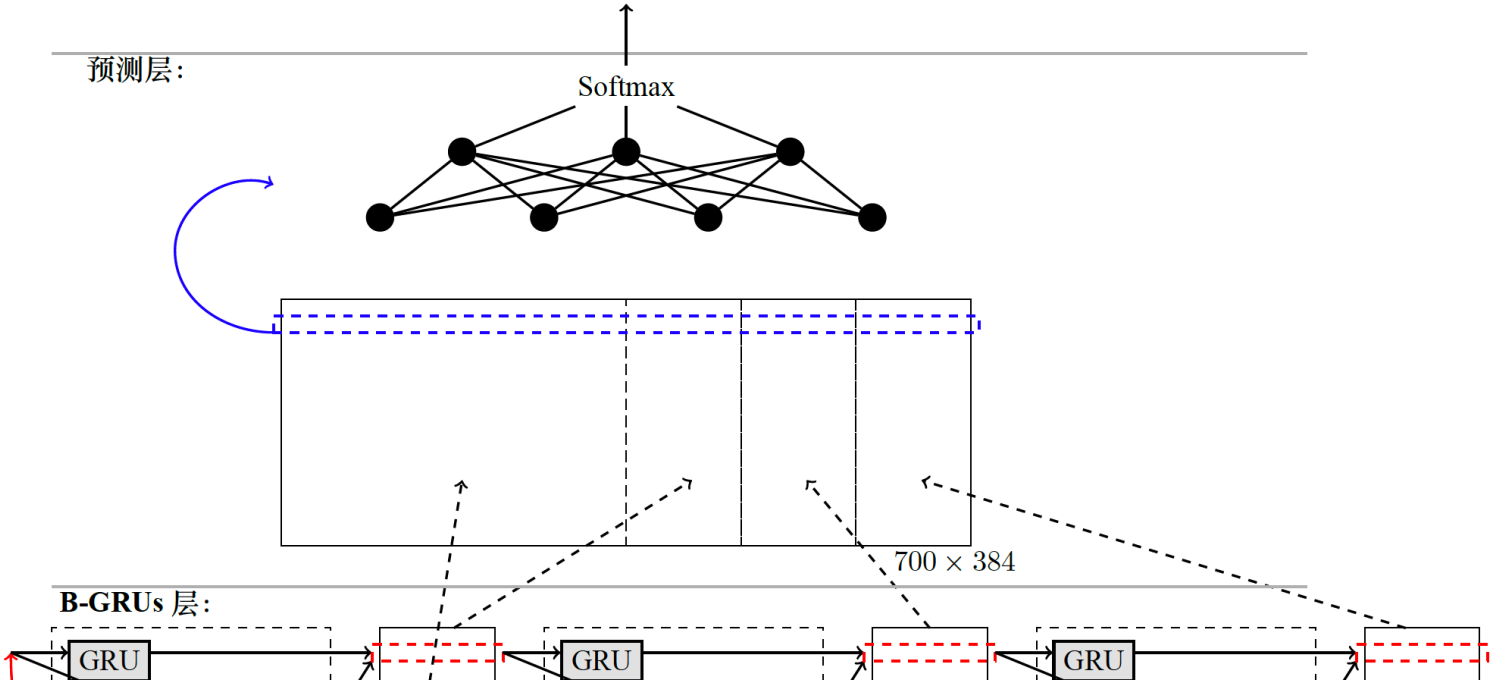
B-GRUs 层



$$B = concatenate\{C, B^{(1)}, B^{(2)}, B^{(3)})\} \tag{18}$$

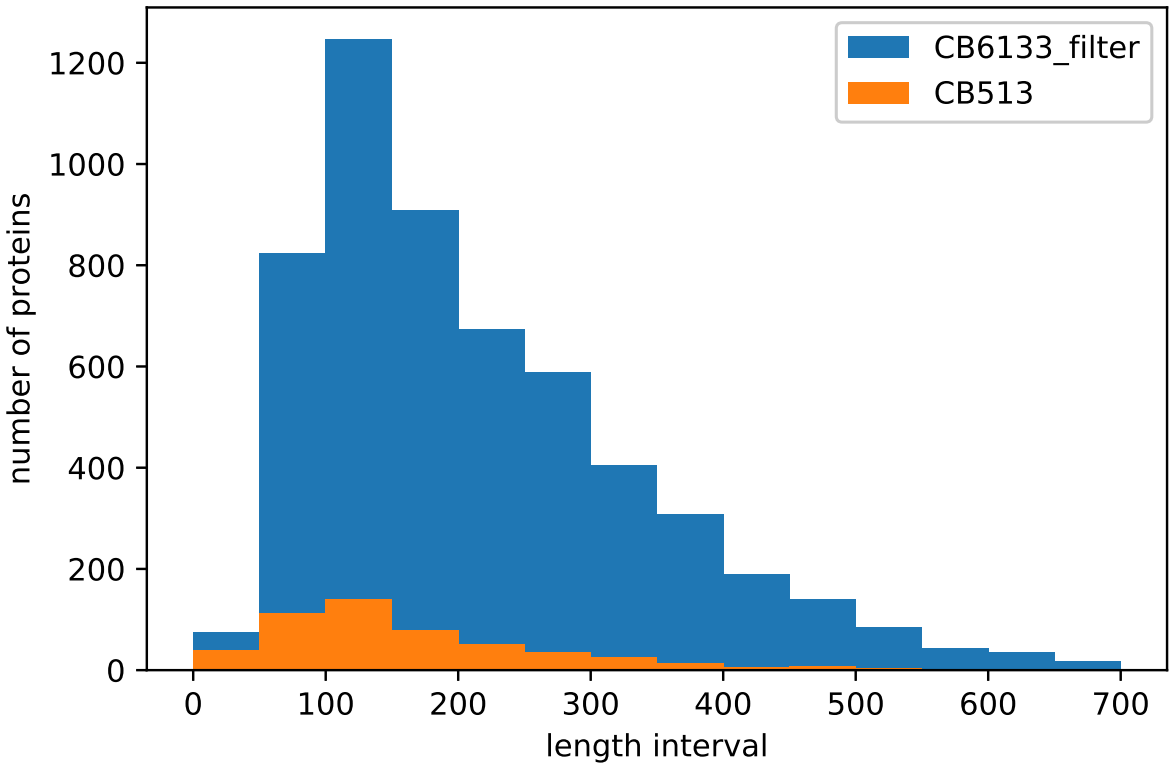
深度模型架构

预测层



实验和总结

数据集：蛋白质序列的长度分布图：



图：数据源自 PSICES CullPDB（http://dunbrack.fccc.edu/PISCES.php）

实验和总结

开发及测试环境

具体的，深度学习模型用 Keras(https://keras.io/) 库实现，该库基于深度学习开源框架 TensorFlow(https://www.tensorflow.org/) 的高级的 API。
整个神经网络都在一台带有 32GB 内存，NVIDIA GTX 1080 Ti GPU，Intel Core i7-8700k CPU 的服务器上训练。同时，使用提前停止来防止模型发生过拟合，训练大约需要半天时间。

实验和总结

模型训练和验证集损失变化

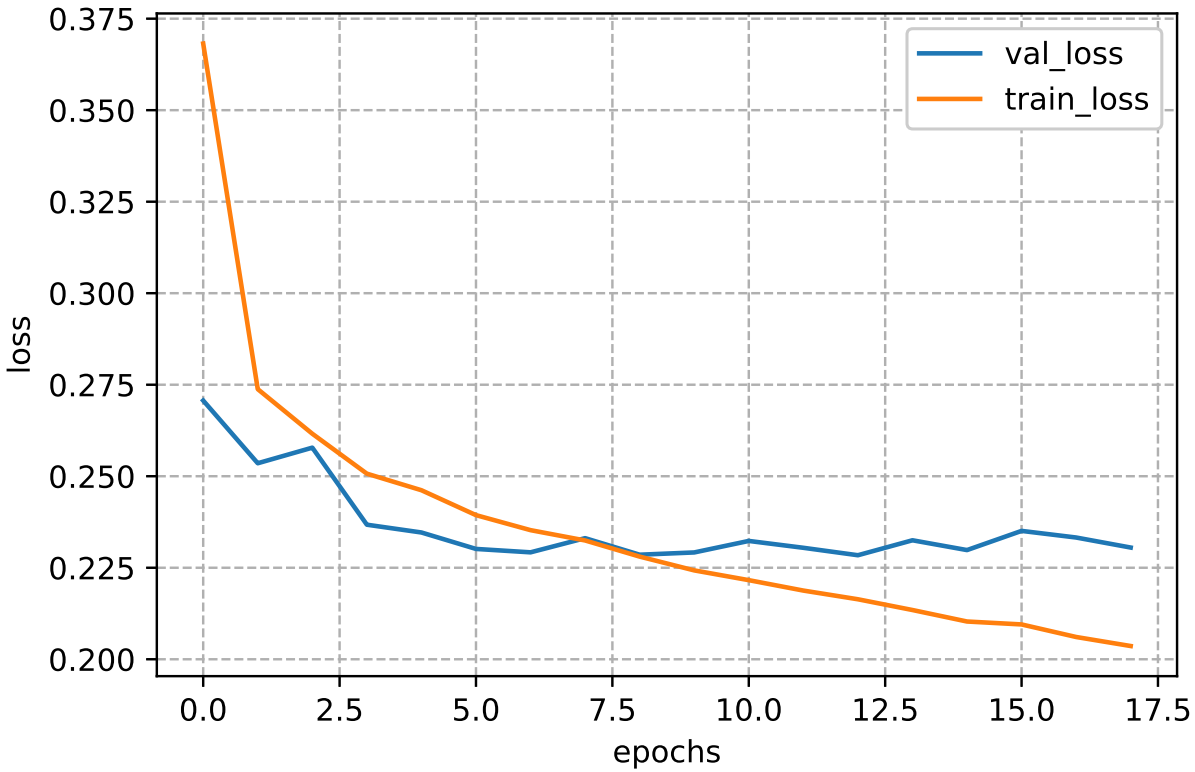


图: 使用 CB6133 的 99.6% 做训练数据，剩余 0.4% 做验证数据

实验和总结

Q_8 准确率随迭代次数的变化

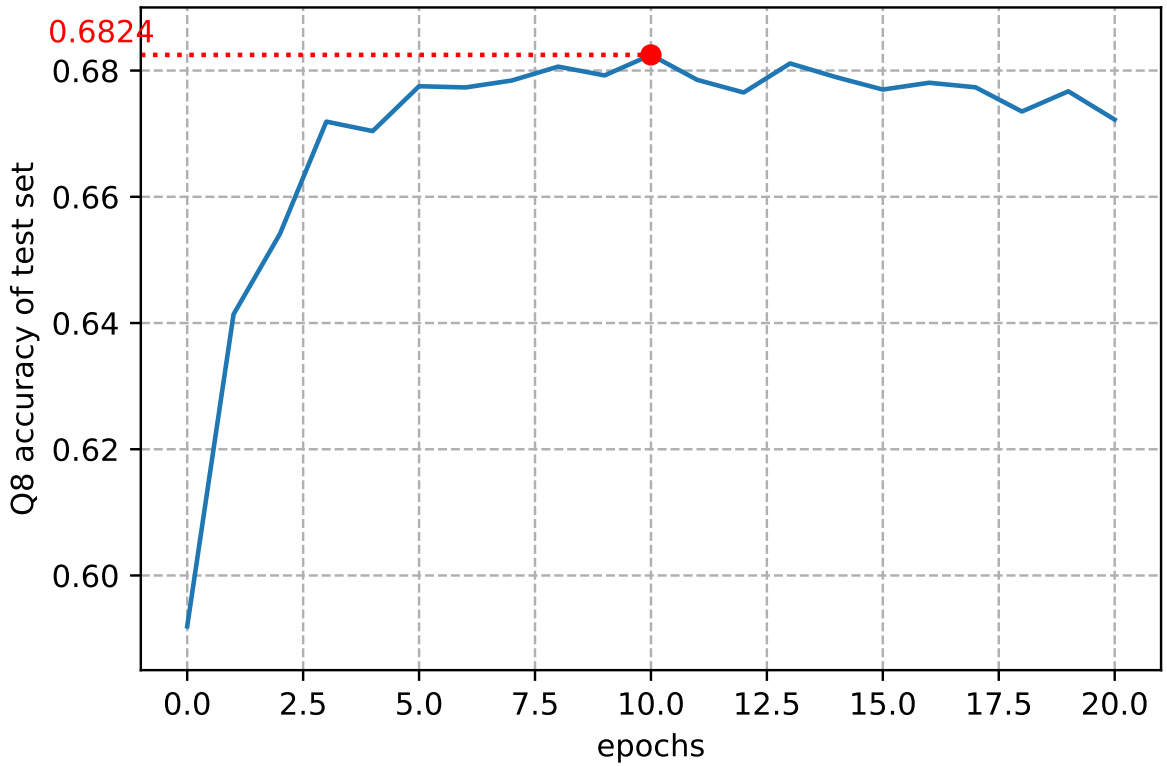


图: 过滤 CB6133 的全部数据做训练，用 CB513 来测试

实验和总结

算法在 CB513 测试集上的混淆矩阵

	H_{pred}	B_{pred}	E_{pred}	G_{pred}	I_{pred}	T_{pred}	S_{pred}	L_{pred}
H	**92.62**%	0.0%	0.97%	1.1%	0.0%	2.46%	0.3%	2.55%
B	9.14%	**1.27**%	27.18%	0.93%	0.0%	10.58%	5.59%	45.3%
E	2.2%	0.05%	**82.08**%	0.46%	0.0%	2.38%	1.53%	11.31%
G	27.14%	0.0%	6.1%	**23.63**%	0.0%	23.21%	2.68%	17.24%
I	66.67%	0.0%	3.33%	0.0%	**0.0**%	13.33%	3.33%	13.33%
T	18.37%	0.0%	5.38%	3.49%	0.0%	**53.07**%	4.49%	15.2%
S	8.35%	0.02%	11.1%	1.74%	0.0%	21.37%	**21.38**%	36.04%
L	6.17%	0.06%	17.7%	1.03%	0.0%	8.97%	5.48%	**60.59**%

实验和总结

算法在 CB513 测试集上的分类性能比较

算法	$Q_8(\%)$
CNF	64.9%
SC-GSN	66.4%
LSTM large	67.4%
本文算法	**68.2%**

演示

总结

Thank you!