Deep Learning for reliable detection of tandem repeats in 3D protein structures.*

Веселова Е.Р.[1]
veselova.er@phystech.edu
[1]Московский физико-технический институт (МФТИ)

В работе рассматривается не требующий предварительной обработки и аугментации входных данных алгоритм для выделения осей симметрии и повторяющихся элементов в трёхмерной структуре белка. Задачи анализа белков на наличие структурных особенностей с высокой эффективностью решаются с помощью свёрточных нейросетей. Относительно трансляции входных данных CNN обладают свойством устойчивости, или *эквивариантности* — коммутативности преобразования исходных данных и свёртки, однако для вращений данное свойство не сохраняется. Искусственно постоянство ответа нейросетей на одних и тех же данных, подвергнутых жестким (евклидовым) преобразованиям достигается аугментацией входных данных. Обеспечить эквивариантность всей нейросети без необходимости аугментации данных предлагается заменой свёрточных фильтров на линейную комбинацию трёхмерных сферических гармоник, эквивариантных относительно вращений. Решение данной задачи позволит увеличить скорость обработки белков и, гипотетически, точность выделения структурных повторений.

Ключевые слова: *CNN, свёрточные нейросети, сферические гармоники, структурные повторения, оси симметрии, 3D объект.*

1 Введение

Задачи структурной биологии . Большая часть получаемых на практике белков обладает повтояющимися элементами структуры или симметрией, которые влияют на функции белков и позволяют исследовать их эволюцию. Нахождение симметрий и повторов является важной задачей, во многих формулировках уже решённой с помощью классических методов машинного обучения в 2006 году [5] При наличии точечного представления плотности 3D объекта задача определения всех типов симметрий была решена аналитически [7, 9]. Поэтому особый интерес представляет применение методов глубокого обучения и свёрточных нейросетей (CNN — convolutional neural network), которые позволяют получать на нижних уровнях сети легко интерпретируемые характерные черты изучаемых объектов.

Предсказание трёхмерной структуры белка по его аминокислотному составу [1] и детектирование структурных повторений и внутренних симметрий с высокой точностью решается свёрточными нейросетями [8], однако существующие свёрточные нейросети не имеют возможности одинаково качественно обрабатывать входные данные при любых поворотах и сдвигах. Основная цель работы состоит в адаптации построенных нейросетей для выявления повторов и симметрий к различным преобразованиям входных данных.

Если трансляция входного объекта при обработке свёрточной нейросетью даёт пропорционально транслированную карту характеристик [4], то для вращений входного объекта подобное свойство не реализуется. Искомое свойство переноса преобразования входных данных на выходные называется эквивариантностью. Кроме того, сохраняющаяся во всех слоях нейросети эквивариантность позволяет отслеживать свойства исследуемых структур уже на нижних уровнях нейросети. 2D CNN, относительно которой данные были бы

*Задачу поставил: Grudinin S. Консультант: Pages G.

эквивариантны, была реализована заменой стандартных свёрточных фильтров на комплексные *круговые гармоники* (circular harmonics), обеспечивающие вращательную эквивариантность без необходимости использовать сильную аугментацию данных [12]. Трёхмерной интерпретацией данного подхода являются сферические гармоники. Первоначально идея сферических гармоник была развита в моделировании для эффективного представления и определения степени схожести поверхностей трёхмерных объектов [3, 6]. Далее идея была применена к анализу трёхмерных признаковых карт с помощью CNN. При замене стандартных трёхмерных свёрточных фильтров на линейную комбинацию аналитически определённого вращательного базиса из *сферических гармоник* CNN становится эквивариантна относительно любого преобразования из группы симметрий $SE(3)$ [11].

Любое движение $g \in SE(3)$ представимо как комбинация вращения $r \in SO(3)$ и трансляции $t \in \mathbb{R}^3$. При рассмотрении одного уровня свёрточной нейросети с K трёхмерных признаковых карт, соответствие между входом и выходом слоя может быть записано как $f : \mathbb{R}^3 \to \mathbb{R}^K$. Оператор трансляции выходного векторного поля легко описывается как $t : (x - t) \mapsto x$. Вращение описывается более сложным образом, так как при повороте всей каждый вектор меняет свою позицию и поворачивается с помощью матрицы $\rho(r)$. Поэтому оператор вращения $\pi(r)$ определяется как $[\pi(r)f](x) := \rho(r)f(r^{-1}x)$, где $r^{-1}x$ описывает перемещение векторов на новые позиции. Таким образом, $g = tr$ представимо как $[\pi(tr)f](x) := \rho(r)f(r^{-1}(x - t))$. Обобщая полученные выкладки на CNN, выражение свёрточного фильтра между n и $n + 1$ слоями нейросети через базис в пространстве эквивариантных преобразований между пространствами признаков $\mathcal{F}_n$ и $\mathcal{F}_{n+1}$ позволяет гарантировать, что любое преобразование входа слоя будет давать такое же преобразование выхода, т.е. обеспечивать эквивариантность. Кроме того, в силу одинакового изменения всех трёх RGB матриц изображения при рассматриваемых преобразованиях полученное представление может быть перенесено и на цветные изображения.

Именно поэтому в качестве решения поставленной задачи в статье предложена эффективная имплементация сферических гармоник в существующую CNN модель выделения тандемных повторов и симметрий в белках для получения идентичных результатов при любых вращениях исходных карт атомных плотностей белковых 3D моделей [8]. В качестве входных данных выступает синтетический датасет, полученный «симметризацией» белковых структур датасета Top8000*, состоящий из карт плотностей размеров $24 \times 24 \times 24$.

2 Постановка задачи

Данными в задаче являются $K = 11$ карт атомных плотностей элементов белка. Каждая карта $\mathbf{x}_i \in \mathbb{R}^{24\times24\times24}$ обучающей выборки $\mathbf{X}$ получена искусственной симметризацией карт датасета Top8000 с порядком циклической симметрии от 1 до $N_{order} \in [10, 20]$. Ответом на элементе выборки $\mathbf{f}(x_i) = \mathbf{f}_i \in \mathbb{R}^{N_{order}+6}$ является композиция двух векторов $\mathbf{y}_i \in \mathbb{R}^{N_{order}}$ и $\mathbf{z}_i \in \mathbb{R}^6$, где $\mathbf{y}_i$ определяет вероятность каждого порядка симметрии, $\mathbf{z}_i$ определяет положение оси симметрии. Размерность 6 для задания оси симметрии выбрана неслучайно: модель переводит ось в трёхмерном пространстве в её представление в шестимерном пространстве посредством *отображения Веронезе* $V(x, y, z) = (x^2, y^2, z^2, \sqrt{2}yz, \sqrt{2}zx, \sqrt{2}xy)$.

Исходная модель $\mathbf{f}(\mathbf{X})$ — свёрточная нейросеть, состоящая из двух свёрточных слоёв, имеющих по 4 фильтра каждый, и трёх полносвязных слоёв, входом для которых является конкатенация столбцов выхода второго свёрточного слоя. Зададим $\mathbf{W} \in \mathbb{R}^{n\times4\times2}$ — матрицу

*http://kinemage.biochem.duke.edu/databases/top8000.php

коэффициентов разложения каждого из 4 фильтров двух свёрточных слоёв нейросети на n базисных сферических гармоник.

Функция потерь при определении порядка симметрии объекта $\mathbf{x}_i$ при известном истинном порядке k_t с помощью вероятности каждого порядка $P(k) = \frac{\exp(p_k)}{\sum_{j=1}^{N_{order}} \exp(p_j)}$ для вектора $\mathbf{y}_i = [p_1, \ldots, p_{N_{order}}]^\intercal$ может быть записана как

$$L_c(\mathbf{W}) = -\log(P(k_t)) = \log\left(\sum_{j=1}^{N_{order}} \exp(p_j)\right) - p_{k_t}.$$

Функция потерь при определении оси симметрии объекта $\mathbf{x}_i$ по известным координатам одной из точек (x_t, y_t, z_t) истинной оси симметрии для вектора $\mathbf{z}_i$ может быть записана как

$$L_a(\mathbf{W}) = \|\mathbf{z}_i - V(x_t, y_t, z_t)\|_2.$$

Получаем решаемую в процессе обучения задачу оптимизации

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \big(L_c(\mathbf{W}) + L_a(\mathbf{W})\big)$$

3 Особенности задачи

Требуется адаптировать модель $\mathbf{f}(\mathbf{X}, \mathbf{W})$ так, что при воздействии на исходные данные любым оператором $\pi(tr)$ трансляции t и вращения r реализуется свойство эквивариантности, т.е.

$$\mathbf{f}(\pi(tr)\mathbf{X}, \mathbf{W}) = [\pi(tr)\mathbf{f}](\mathbf{X}, \mathbf{W}).$$

Большинство функций активации, обрабатывающих выход свёрточного слоя, заранее не обладают эквивариантностью.

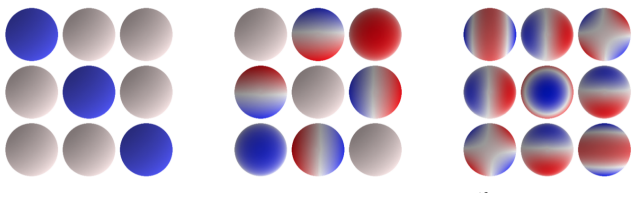

Рис. 1 сферические гармоники 0, 1 и 2 порядков

4 Базовый эксперимент

С целью проверки наличия эквивариантности у нейросети, построенной из сферических свёрточных фильтров с учетом ограничений на нелинейную активацию, был проведён базовый эксперимент по реализации структуры базовой нейросети из работы [8] с применением свёрточных слоёв из модуля se3cnn [11]. Базовая архитектура была переработана с учётом требований модуля и необходимого порядка сферических гармоник. Согласно эмпирическим данным, достаточно второго–третьего порядков сферических для проявления эквивариантности в структуре, поэтому свёрточный фильтр в конволюционном слое представляет из себя комбинацию трёх наборов сферических функций порядков 0, 1 и 2 с двумя наборами коэффициентов разложения для каждого из наборов. Таким образом,

выход конволюционного слоя имеет $2 * \big((2*0+1)+(2*1+1)+(2*2+1)\big) = 18$ карт. Синтетические данные для обучения тестовой сети представляют из себя 400 симметричных с порядком 3 относительно любой из главных диагоналей матриц. Выходом для тестовой сети является заданная тремя координатами ось симметрии.

Layer	Type	Input dimensions	Output dimensions	Parameters
1	SE3Convolution	$24 \times 24 \times 24 \times 1$	$21 \times 21 \times 21 \times 18$	Size 4, stride 1
2	Average pooling	$21 \times 21 \times 21 \times 18$	$8 \times 8 \times 8 \times 7$	Size 2, stride 2
5	Reshape	$8 \times 8 \times 8 \times 7$	1512	
6	Linear	1512	800	
7	LeakyReLU	800	800	
8	Linear	800	3	

По результатам эксперимента наблюдается постоянство ответов при вращениях на 90^{deg} и 120^{deg}. Для повышения выделяемых порядков симметрии требуется применение и комбинация слоёв сферических гармоник более высоких порядков вплоть до порядка 20, максимального в нашей постановке задачи.

5 Вычислительный эксперимент

Для проведения основного вычислительного эксперимента была построена ResNet (Residual Network) архитектура [2], эффективно используемая в обработке изображений (3.08% ошибка в Imagenet Recognition challenge) [13].Особенностью данной архитектуры является добавление промежуточных связей между слоями — shortcut connections, передающих неизменное входное значение слоя. Выход последовательности слоёв, соединённых shortcut connection, представим в виде $\mathbf{X} \mapsto \mathbf{f}(\mathbf{X}) + \mathbf{X}$ в отличие от стандартной последовательной архитектуры $\mathbf{X} \mapsto \mathbf{f}(\mathbf{X})$.

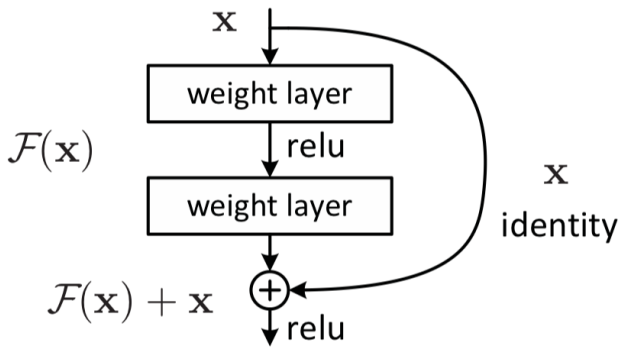

Рис. 2 структура ResNet

Выбор данной архитектуры в первую очередь обусловлен лучшей устойчивостью к переобучению по сравнению с последовательной нейронной сетью. Согласно математическому представлению архитектуры, каждая последовательность слоёв, соединенная shortcut connection, учится предсказывать разность между входом и выходом. Обычная архитектура не позволяет весам слоя полностью обнулиться без потери возможности обратного распространения ошибки в сети. В ResNet–архитектуре допустима последовательность слоёв с нулевыми коэффициентами — данная последовательность является тождественным преобразованием $\mathbf{X} \mapsto \mathbf{X}$. Аналогично первым статьям, исследующим residual–метод, было решено использовать связи через один свёрточный слой [2, 10].

Литература

[1] Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *ArXiv e-prints*, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[3] Michael Kazhdan and Thomas A. Funkhouser. Harmonic 3d shape matching. In *SIGGRAPH Abstracts and Applications*, page 191. ACM, 2002.

[4] Karel Lenc and Andrea Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. *International Journal of Computer Vision*, 2018.

[5] Niloy J. Mitra, Leonidas Guibas, and Mark Pauly. Partial and Approximate Symmetry Detection for 3D Geometry. *ACMTG: ACM Transactions on Graphics*, 25, 2006.

[6] Mohamed-Hamed Mousa, Raphaëlle Chaine, Samir Akkouche, and Eric Galin. Toward an efficient triangle-based spherical harmonics representation of 3D objects. *Computer Aided Geometric Design*, 25(8):561–575, 2008.

[7] Guillaume Pagès and Sergei Grudinin. Analytical symmetry detection in protein assemblies. II. Dihedral and Cubic symmetries. *Journal of Structural Biology*, 203(3):185–194, September 2018.

[8] Guillaume Pagès and Sergei Grudinin. DeepSymmetry: Using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. working paper or preprint, 2018.

[9] Guillaume Pagès, Elvira Kinzina, and Sergei Grudinin. Analytical symmetry detection in protein assemblies. I. Cyclic symmetries. *Journal of Structural Biology*, 203(2):142–148, August 2018.

[10] Andreas Veit, Michael J. Wilber, and Serge J. Belongie. Residual networks are exponential ensembles of relatively shallow networks. *CoRR*, abs/1605.06431, 2016.

[11] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *CoRR*, abs/1807.02547, 2018.

[12] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic Networks: Deep Translation and Rotation Equivariance. In *CVPR*, pages 7168–7177. IEEE Computer Society, 2017.

[13] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.