Глубокое обучение для обработки 3D структур белка

Веселова Е.Р.

Московский физико-технический институт

April 25, 2019

Обзор задачи

Цель работы

Обнаружение осей симметрии и повторяющихся элементов в белковых структурах с учетом сдвигов и вращений входных данных.

Проблематика задачи

1. Малые размеры датасетов исследованных белков
2. Неинвариантность имеющихся алгоритмов относительно вращения входных данных

Предложенное решение

Свёрточная нейросеть с применением сферических гармоник

Основная литература

- 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.
- Guillaume Pag‘es and Sergei Grudinin. DeepSymmetry: Using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. working paper or preprint, 2018.
- Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. ArXiv e-prints, 2018.

Входные данные

Данные

Для каждого белка имеются $K = 11$ карт атомных плотностей элементов. Карты плотностей переводятся в матрицу $x_i \in \mathbb{R}^{24\times24\times24} = 13824$ вокселей:

$$\rho(x, y, z) = \iiint_{C(x,y,z)} \sum_{j=1}^{11} \exp\left(\frac{-||p - a_i||^2}{\sigma}\right) dp$$

где $C(x, y, z)$ — координаты вокселея, a_i — позиция i атома, $\rho(x, y, z) \in [0, 255]$

Ответом на элементе выборки x_i является вектор

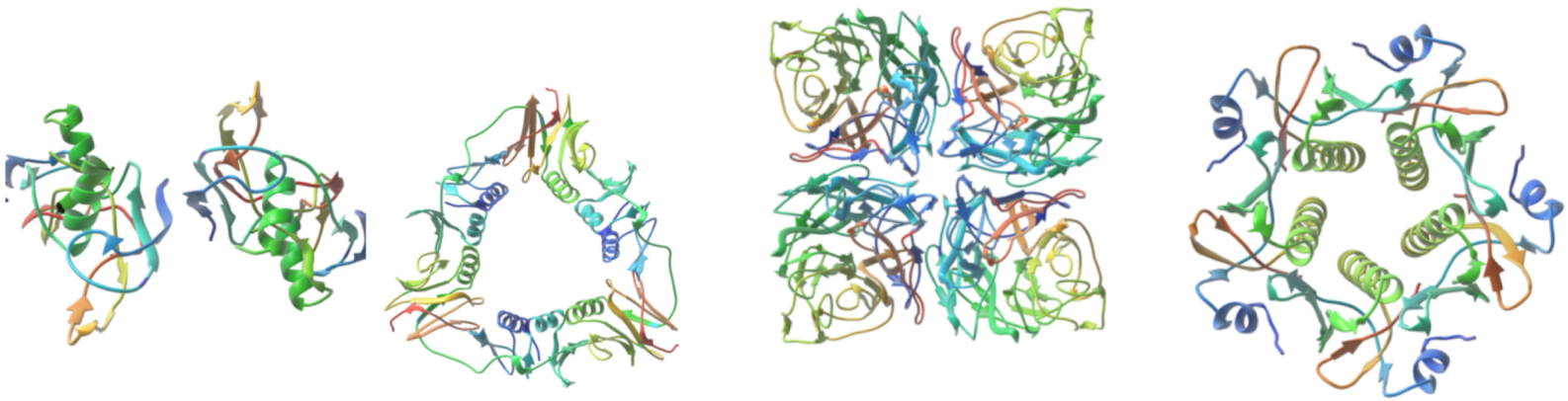
$$f(x_i) = f_i = (y_i, z_i) \in \mathbb{R}^7$$

где $y_i \in \mathbb{R}$ — порядок симметрии белка, $z_i \in \mathbb{R}^6$ — координаты оси симметрии белка в шестимерном представлении Веронезе

Данные

Обучающие данные

Обучающие данные получены искусственной симметризацией датасета Top8000 с порядком циклической симметрии от 1 до $N_{order} \in [10, 20]$.



Примеры симметризованных белков с порядками симметрии 2, 3, 4 и 5 соответственно и направленными к наблюдателю осями симметрии

Генерация данных

1. Для каждого белка обучающей выборки случайно выбирается порядок и ось симметрии, после чего создается симметричный объект
2. Тестовая выборка генерируется независимо тем же способом из других белков
3. Для увеличения объема обучающей выборки алгоритм генерации постоянно создаёт новые данные из тех же белков с другими порядками и осями симметрии, что позволяет увеличить размер обучающей выборки в 1000 раз

Модель задачи

Исходная модель $f(X)$ — свёрточная нейросеть, $W \in \mathbb{R}^{n\times 4\times 2}$ — матрица коэффициентов разложения каждого из 4 фильтров двух свёрточных слоёв нейросети на n базисных сферических гармоник.

Выход нейросети

$$f(x_i) = f_i = (y_i, z_i) \in \mathbb{R}^{N_{order}+6}$$

где $y_i \in \mathbb{R}^{N_{order}}$ — вектор оценки наличия симметрии порядка от 1 до N_{order} во входной структуре

Выход модели получается применением softmax-активации к y_i и получением вектора вероятностей наличия порядка симметрии

$$p_k = \frac{e^{y_k}}{\sum_{j=1}^{N_{order}} e^{y_j}}$$

Постановка задачи

Функции качества

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

$$L_a(W) = \|z_i - V(x_t, y_t, z_t)\|_2.$$

Функция потерь

$$W^* = \min_W \left(L_c(W) + L_a(W)\right)$$

Постановка задачи

Особенность задачи

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

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

В нашем случае эквивариантность означает сохранение ответа при преобразованиях из группы SE3 входных данных, так как трансляция и вращение скаляра оставляют его неизменным.

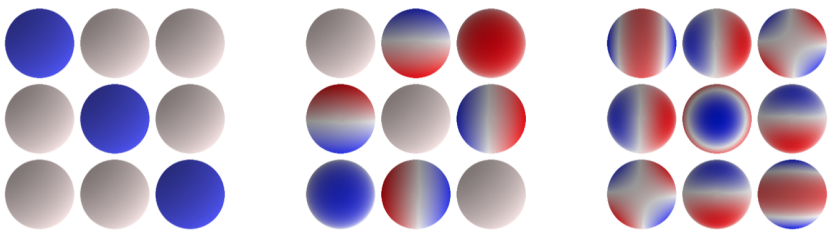
Сферические гармоники

Устойчивость к трансляции входных данных заложена в структуре свёрточной нейросети. Устойчивость к вращениям реализуется при помощи разложения свёрточных фильтров в линейные комбинации сферических функций.

Сферические гармоники

собственные функции оператора Лапласа в сферической системе координат, образующие ортонормированную систему в пространстве функций на сфере.

$$Y_{lm} = \frac{1}{2\pi} e^{im\varphi} \Theta_{lm}(\theta)$$



Сферические гармоники 0, 1 и 2 порядков

Архитектура нейросети

Использование сферических гармоник накладывает ограничения на применяемые в архитектуре модули:

- Max Pooling не является инвариантным относительно вращений, поэтому применяется Average Pooling[1]
- Нелинейные функции активации не являются инвариантными относительно трансляций, поэтому в работе используется ReLU
- Батч-нормализация реализуется по радиальным слоям для сохранения эквивариантности

$$\mathrm{f}(\mathrm{x_i}) \mapsto \mathrm{f_i}(\mathrm{x})\Big(\frac{1}{|\mathcal{B}|}\sum_{\mathrm{j}\in\mathcal{B}}\frac{1}{\mathrm{V}}\int ||\mathrm{f_i}(\mathrm{x})||^2\mathrm{dx} + \epsilon\Big)^{-1/2}$$

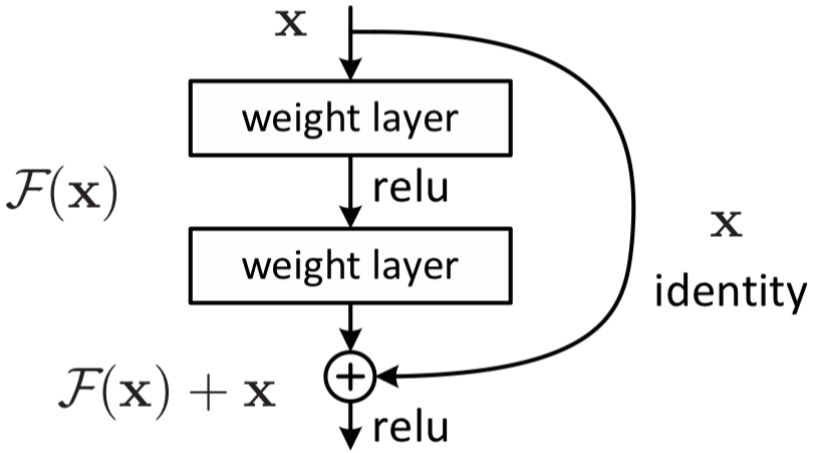
[1] D.E. Worrall and G.J. Brostow. CubeNet: Equivariance to 3D Rotation and Translation.

Архитектура нейросети

Была реализована ResNet (Residual Network) архитектура.

ResNet

Особенностью данной архитектуры является добавление промежуточных связей между слоями — shortcut connections. Выход последовательности слоёв, соединённых shortcut connection, представим в виде $x \mapsto \mathcal{F}(x) + x$ в отличие от стандартной последовательной архитектуры $x \mapsto \mathcal{F}(x)$.



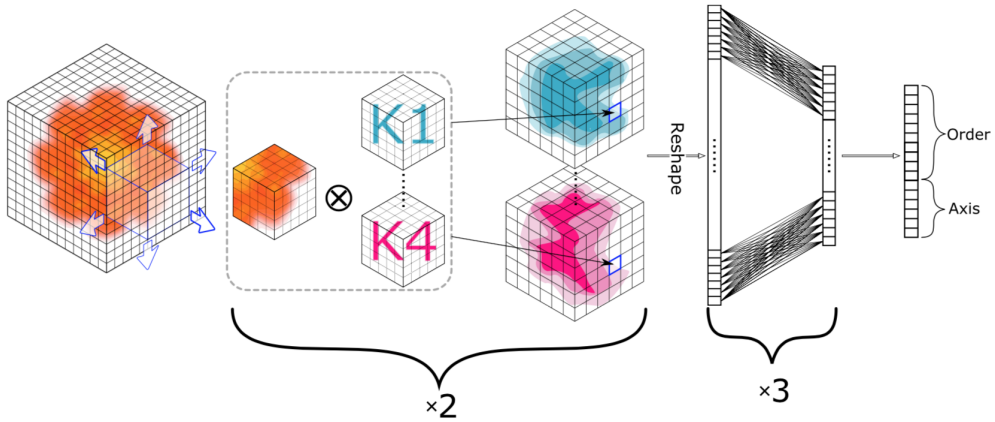
Структура одного слоя ResNet

Реализация нейросети

- Реализация разложения свёрточного слоя взята из библиотеки se3cnn[2].
- Для улучшения точности модели использовалась взвешенная функция ошибки:

$$L(W) = -\lambda \cdot \log\left(\sum_{j=1}^{N_{order}} p_{t_k} \exp(p_j)\right) + \|z - V(x_t, y_t, z_t)\|_2$$

$\lambda \in [0, 1)$ для повышения точности определения оси симметрии.



Общая структура нейросети

Результаты

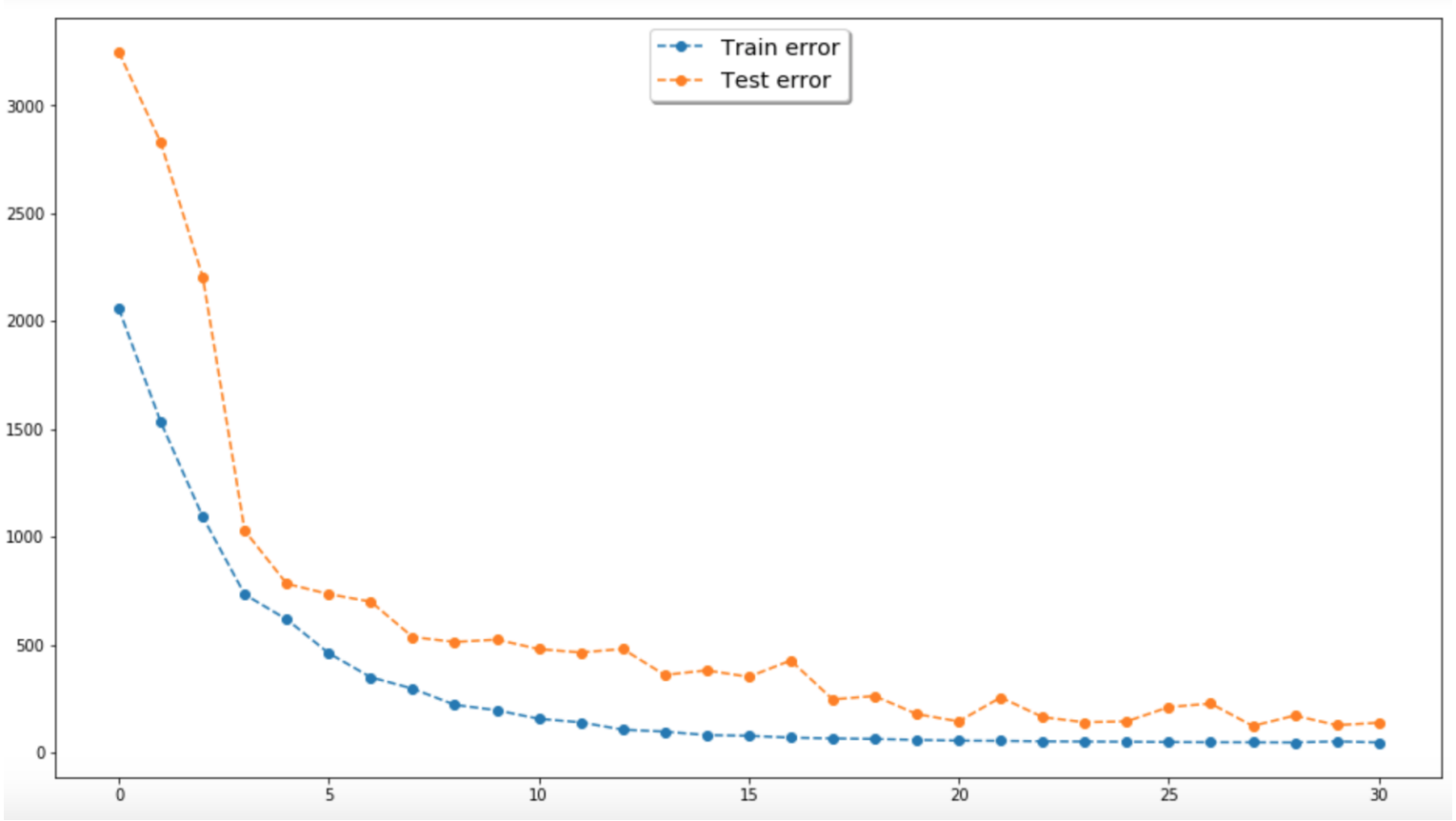


График обучения

Результаты

Данные	порядок симметрии	ось симметрии
Train	0.91	0.82
Test	0.78	0.59

Ошибка выделения осей симметрии

Дальнейшие исследования

- Поиск оптимальных порядков сферических функций в свёрточных слоях для точного выделения высоких порядков симметрии
- Увеличение количества слоёв нейросети для повышения точности выделения осей симметрии