

Оценка качества прогнозирования структуры белка с использованием графовых свёрточных нейронных сетей

Северилов Павел

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
2020 г.

Проблема

Последовательность аминокислот сворачивается в нативную структуру белка. Моделируется структура, в которую произойдет сворачивание. Вычислительно дорого определить качество смоделированной структуры по отношению к нативной.

Задача оценки качества структуры (Quality Assessment)

На основе данных о смоделированной структуре построить регрессию на значение качества структуры. Для решения задачи проводятся соревнования CASP.

Предлагается

Методами спектральной теории графов проанализировать спектр графовой свёртки. Применить графовые свёрточные нейронные сети к задаче Quality Assessment.

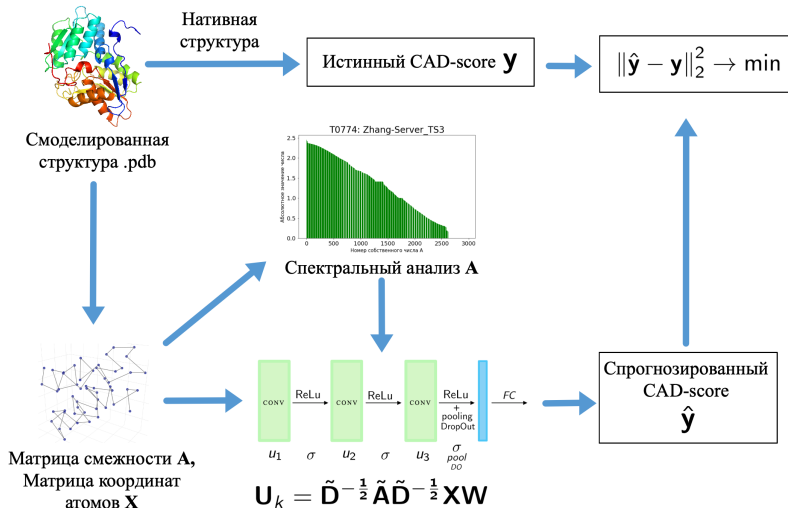
Работы по графовым свёрточным нейронным сетям

- Kipf T. N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks // Proceedings of the 5th International Conference on Learning Representations, 2017
- Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P. S. A Comprehensive Survey on Graph Neural Networks // IEEE Transactions on Neural Networks and Learning Systems, 2020

Работы по Quality Assessment

- Derevyanko G., Grudinin S., Bengio Y., Lamoureux G. Deep convolutional networks for quality assessment of protein folds // Bioinformatics (Oxford, England), 2018
- Pagès G., Charmettant B., Grudinin S. Protein model quality assessment using 3D oriented convolutional neural networks // Bioinformatics (Oxford, England), 2019

Оценка качества смоделированной структуры белка



Общая схема эксперимента

Постановка задачи регрессии CAD-score

- Дана выборка

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{R}^{n_i \times 3}$ – молекулы, каждая из которых описана множеством 3-мерных координат всех ее n_i атомов, $y_i \in \mathbb{R}$ – оценка близости смоделированной и нативной структуры белка $\text{CAD}_{\text{score}}$.

- Рассматривается множество графовых свёрточных нейронных сетей

$$\{\mathbf{f}_k: (\mathbf{w}, \mathbf{X}) \rightarrow \hat{\mathbf{y}} \mid k \in \mathcal{K}\},$$

где $\mathbf{w} \in \mathbb{W}$ – параметры модели, $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) \in \mathbb{R}^m$, $\mathbf{X} = \bigcup_{i=1}^m \mathbf{x}_i$.

- Функция ошибки

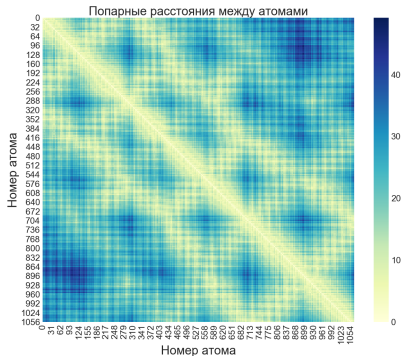
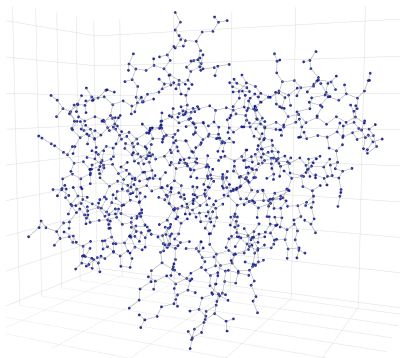
$$\mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2.$$

- Решается задача оптимизации:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{W}}{\operatorname{argmin}} (\mathcal{L}(\mathbf{w})).$$

Матрицы смежности графов молекул

Наличие связи между атомами молекулы вычисляется согласно химическим законам.



Трехмерное представление с помощью координат \mathbf{X} и полученной матрицы смежности \mathbf{A} и попарные расстояния между атомами модели BAKER-ROSETTASERVER_TS3 для нативной структуры T0870 из набора данных CASP12

Графовый Лапласиан

Матрица $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$, где \mathbf{A} – матрица смежности графа \mathbf{G} , \mathbf{D} – диагональная матрица степеней вершин, $\mathbf{D}_{ii} = \sum_j (\mathbf{A}_{ij})$.

Спектральное разложение

$\mathbf{L} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$, где $\mathbf{U} \in \mathbb{R}^{n \times n}$ – матрица собственных векторов, $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ – диагональная матрица собственных значений.

Графовое преобразование Фурье

для вектора признаков всех вершин $\mathbf{x} \in \mathbb{R}^n$ задается

$$\mathcal{F}(\mathbf{x}) = \mathbf{U}^T \mathbf{x} \equiv \hat{\mathbf{x}} \in \mathbb{R}^n,$$

обратное графовое преобразование Фурье: $\mathcal{F}^{-1}(\hat{\mathbf{x}}) = \mathbf{U} \hat{\mathbf{x}}$.

Теорема о свёртках

Преобразование Фурье свёртки двух сигналов является покомпонентным произведением их преобразований Фурье, т.е.

$$\mathcal{F}(\mathbf{f} * \mathbf{g}) = \mathcal{F}(\mathbf{f}) \odot \mathcal{F}(\mathbf{g}).$$

Применяя теорему, определяем спектральную свёртку на графах для сигнала \mathbf{x} и фильтра $\mathbf{g} \in \mathbb{R}^n$ как

$$\mathbf{x} * \mathbf{g} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{x}) \odot \mathcal{F}(\mathbf{g})) = \mathbf{U}(\mathbf{U}^T \mathbf{x} \odot \mathbf{U}^T \mathbf{g}) = \mathbf{U} \mathbf{g}_\theta \mathbf{U}^T \mathbf{x},$$

где $\mathbf{g}_\theta = \text{diag}(\mathbf{U}^T \mathbf{g})$ – спектральные коэффициенты фильтра.

Аппроксимируя \mathbf{g}_θ с помощью полиномов Чебышёва $\mathbf{T}_k(\mathbf{x})$, получаем

$$\mathbf{x} * \mathbf{g} = \sum_{k=0}^K \theta_k \mathbf{T}_k(\tilde{\mathbf{L}}) \mathbf{x},$$

где

$$\tilde{\mathbf{L}} = 2 \frac{\mathbf{L}}{\lambda_{\max}} - \mathbf{I}_n, \mathbf{T}_k(\mathbf{x}) = 2\mathbf{x}\mathbf{T}_{k-1}(\mathbf{x}) - \mathbf{T}_{k-2}(\mathbf{x}), \mathbf{T}_0(\mathbf{x}) = 1, \mathbf{T}_1(\mathbf{x}) = \mathbf{x}.$$

Свёрточный слой

Приняв $\lambda_{\max} \approx 2$, $K = 1$ и $\theta = \tilde{\theta}_0 = -\tilde{\theta}_1$, получаем

$$\mathbf{x} * \mathbf{g} \approx \tilde{\theta}_0 \mathbf{x} + \tilde{\theta}_1 (\mathbf{L} - \mathbf{I}_n) \mathbf{x} = \tilde{\theta}_0 \mathbf{x} - \tilde{\theta}_1 \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \mathbf{x} = \theta \left(\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \right) \mathbf{x}.$$

Трюк перенормировки:

$$\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \rightarrow \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}, \text{ где } \tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n, \quad \tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}.$$

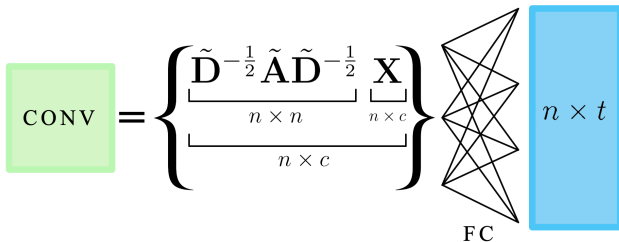
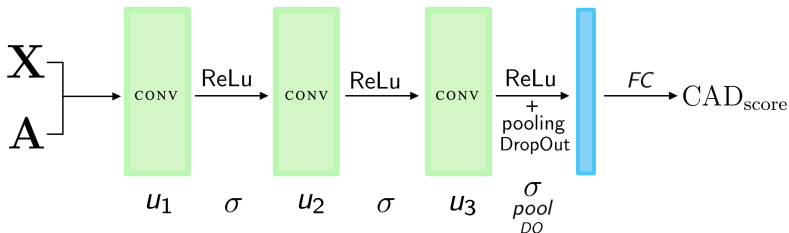


Схема свёртки графа с матрицей \mathbf{X} , t – число фильтров в свёртке, FC – полносвязный слой. Синий прямоугольник – выходная матрица

Модель нейронной сети



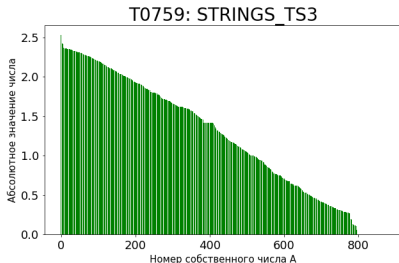
Схематическое представление архитектуры свёрточной нейронной сети, использованной в данной работе

$$f = FC \circ DO \circ pool \circ \sigma \circ u_3 \circ \sigma \circ u_2 \circ \sigma \circ u_1$$

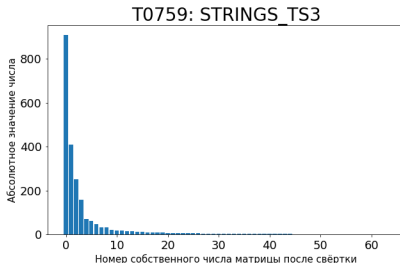
$$u_k = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} XW$$

Собственные числа матриц A , U_k

Для примера взята смоделированная структура STRINGS_TS3 для нативной T0759. Здесь U_k – матрица после прохождения свёртки.



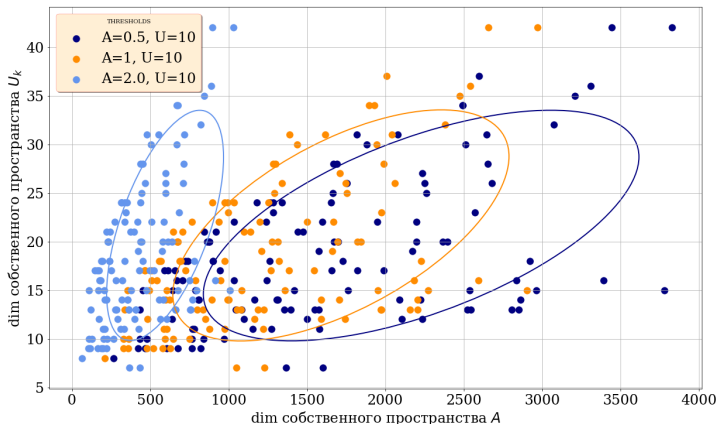
Собственные числа матрицы A



Собственные числа матрицы U_k

Собственное пространство матриц смежности \mathbf{A} , \mathbf{U}_k

Количество собственных чисел матриц \mathbf{A} , \mathbf{U}_k , превосходящих порог. Рассмотрены разные значения порогов A , U , соответствующих собственным числам до и после прохождения свёртки.



Наборы данных

Набор	Нативные структуры	Модели структур	Разбиение
CASP 9	117	35963	Train, Validation
CASP 10	103	15450	
CASP 11	84	12291	
CASP 12	37	5538	Test

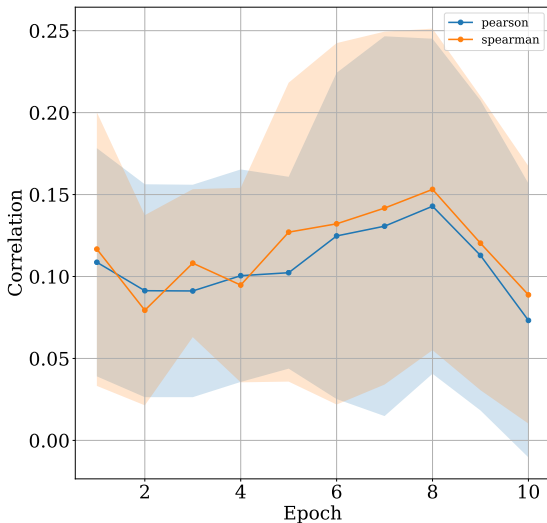
При обучении нейросети анализируются усредненные по T нативным структурам коэффициенты корреляции Пирсона и Спирмена

$$R = R(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T} \sum_{i=1}^T R_i^{\text{target}} = \frac{1}{T} \sum_{i=1}^T \text{PEARSON}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

$$\rho = \rho(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T} \sum_{i=1}^T \rho_i^{\text{target}} = \frac{1}{T} \sum_{i=1}^T \text{SPEARMAN}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Результаты обучения на 10 эпохах

Графики корреляций Пирсона и Спирмена стабилизируются



Сравнение с существующими методами Quality Assessment

Модель	Spearman ρ	Pearson R
ProQ3D	0.801	0.750
VoroMQA	0.803	0.766
SBROD	0.685	0.762
Ornate	0.828	0.781
SpectralQA (данная работа)	0.79	0.83

Сравнение корреляции Пирсона и Спирмена существующих современных алгоритмов с моделью SpectralQA на данных CASP12

Полученные результаты

- Предложено решение задачи Quality Assessment с использованием графовых сверток
- Проведен анализ графовых свёрток на задаче Quality Assessment
- Полученная модель дает качество, сравнимое с качеством альтернативных моделей

Дальнейшие исследования

- Использовать другие существующие улучшения спектральных свёрток (CayleyNet, Adaptive Graph Convolution Network)
- Учесть дополнительные химические свойства атомов
- Учесть в матрице смежности расстояния между атомами

К публикации

Севериков П.А., Стрижов В.В. Оценка качества прогнозирования структуры белка с использованием графовых свёрточных нейронных сетей // Computational and Applied Mathematics, 2020