Сферические сверточные нейронные сети для прогнозирования химических или биологических свойств молекул

Наталия Викторовна Вареник

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 674, весна 2019

Цель исследования

Задача

Создать точный метод прогнозирования химических или биологических свойств молекулы по ее пространственной структуре.

Проблема

Существующие методы имеют недостаточно высокую точность.

Способ решения

Предлагается использовать модель сферических сверточных нейронных сетей, учитывающую пространственное представление исследуемого объекта.

Литература

Существующие методы:

- David S. Palmer and Noel M. O'Boyle and Robert C. Glen and John B. O. Mitchell. *Random Forest Models To Predict Aqueous Solubility*, 2007.
- George E. Dahl and Navdeep Jaitly and Ruslan Salakhutdinov. *Multi-task Neural Networks for QSAR Predictions*, 2014.

Модель SCNN:

- Taco S. Cohen, Mario Geiger, Jonas Koehler, Max Welling. *Spherical CNNs*, 2018.

Постановка задачи

Дано

Выборка $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$, где $\mathbf{X} \in \mathbb{R}^{m \times 3n}$ — матрица объект-признак, состоящий из m молекул, каждая из которых описана множеством 3D-координат всех ее атомов, а $\mathbf{y} \in \mathbb{R}^m$ — свойства молекул.

Модель

$f : (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}} \in \mathfrak{F}$, где $\mathbf{w} \in \mathbf{W}$ - параметры модели, а $\mathfrak{F}$ - семейство параметрических моделей из класса сферических сверточных нейронных сетей.

Постановка задачи

Ошибка

- для регрессии:

$$S(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|_2^2$$

- для классификации: ROC_AUC

Итоговая задача оптимизации

Параметры модели $\mathbf{w} \in \mathbf{W}$ подбираются в соответствии с минимизацией функции ошибки на обучении:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbf{W}}{\operatorname{argmin}} S(\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}}, \mathbf{w} \mid f)$$

Базовый эксперимент

Цель

Определить качество предсказания базовой модели, для дальнейшего сравнения результатов с исследуемой моделью.

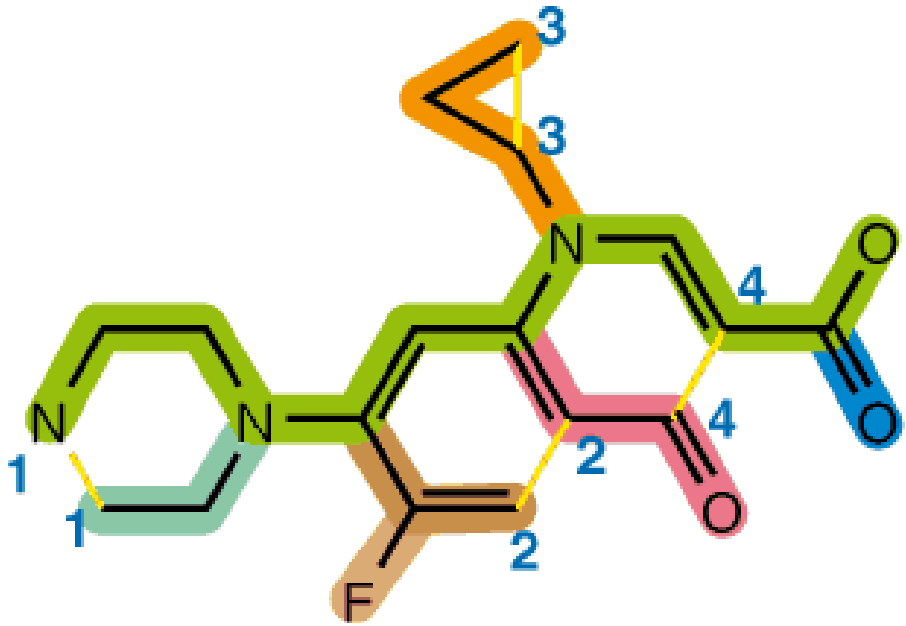
Данные

Формат представления молекулы - SMILES
Число молекул - 7165, число атомов - 23
Метки - энергия атомизации.

Модель RNN

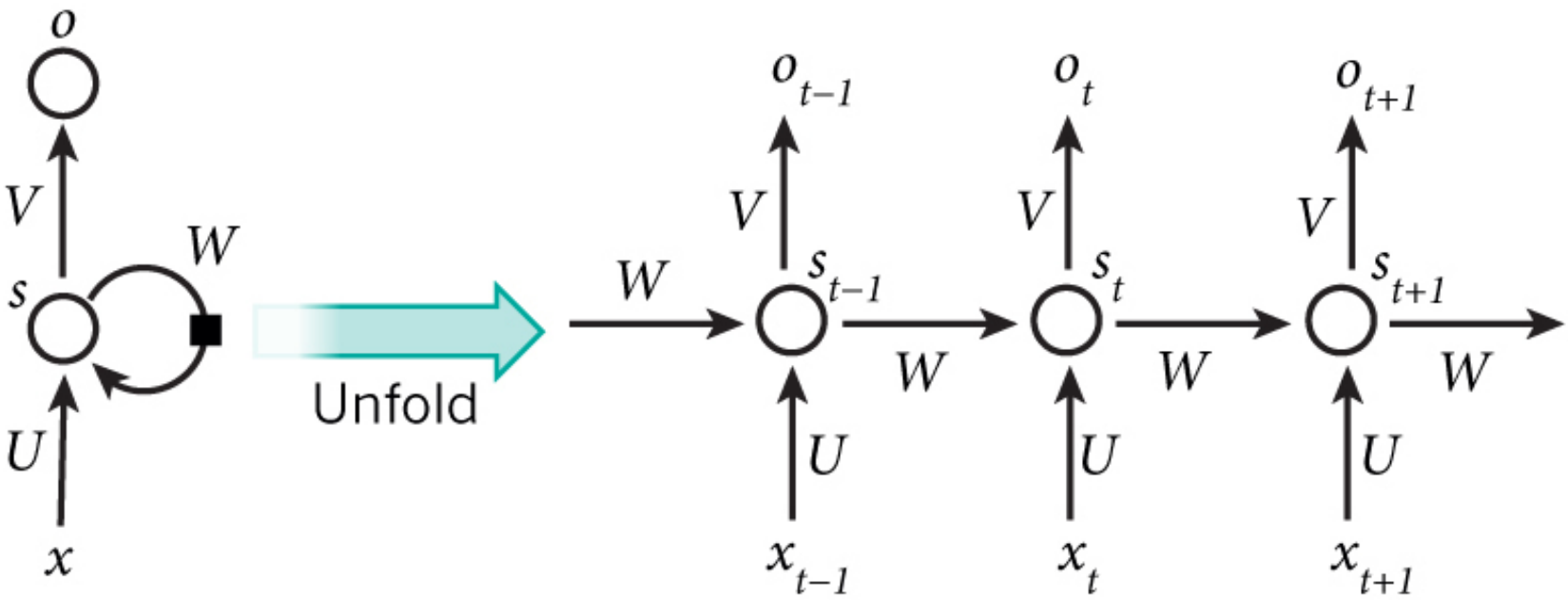
- Число слоев в энкодере - 2
- Размер скрытого слоя - 128
- Число эпох - 70
- Ошибка - MSELoss
- Оптимизатор - Adam

Пример построения последовательности SMILES

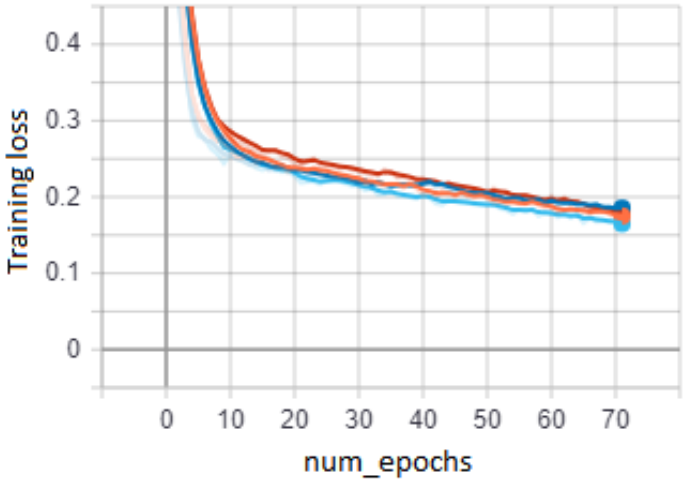


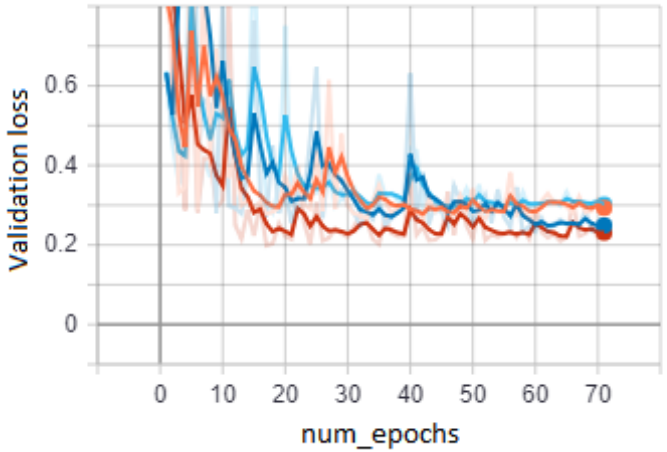
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

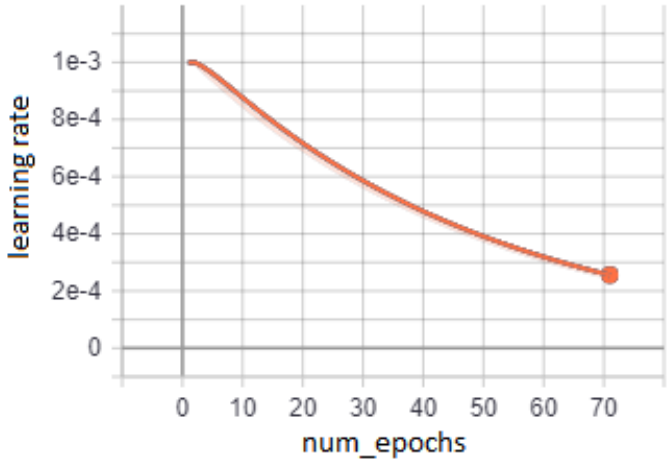
Идея RNN

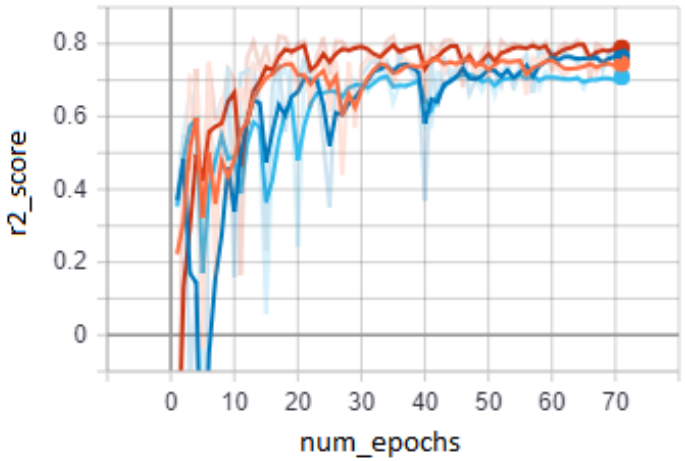


Результаты RNN









Сравнение с результатами Random Forest

Еще одна базовая модель - Random Forest с параметрами n_estimators = 500, max_depth = 21

	RMSE	R2_score
RNN	0.522	0.747
Random Forest	0.647	0.518

Выводы

- Была поставлена задача
- Проведен базовый эксперимент
- Получены результаты, которые в дальнейшем будут сравниваться с результатами исследуемой модели.