Получение информации о ко-эволюции белков с помощью языковых моделей

Зверев Егор

Московский физико-технический институт

Курс: Автоматизация научных исследований (практика, В. В. Стрижов)/Группа 821

Эксперт: Сергей Грудинин

Консультант: Илья Игашов

2021

Проблема классификации белков

Цель

Исследовать возможности генерации дескриптор с помощью языковых моделей вместо MSA (множественного выравнивания последовательностей).

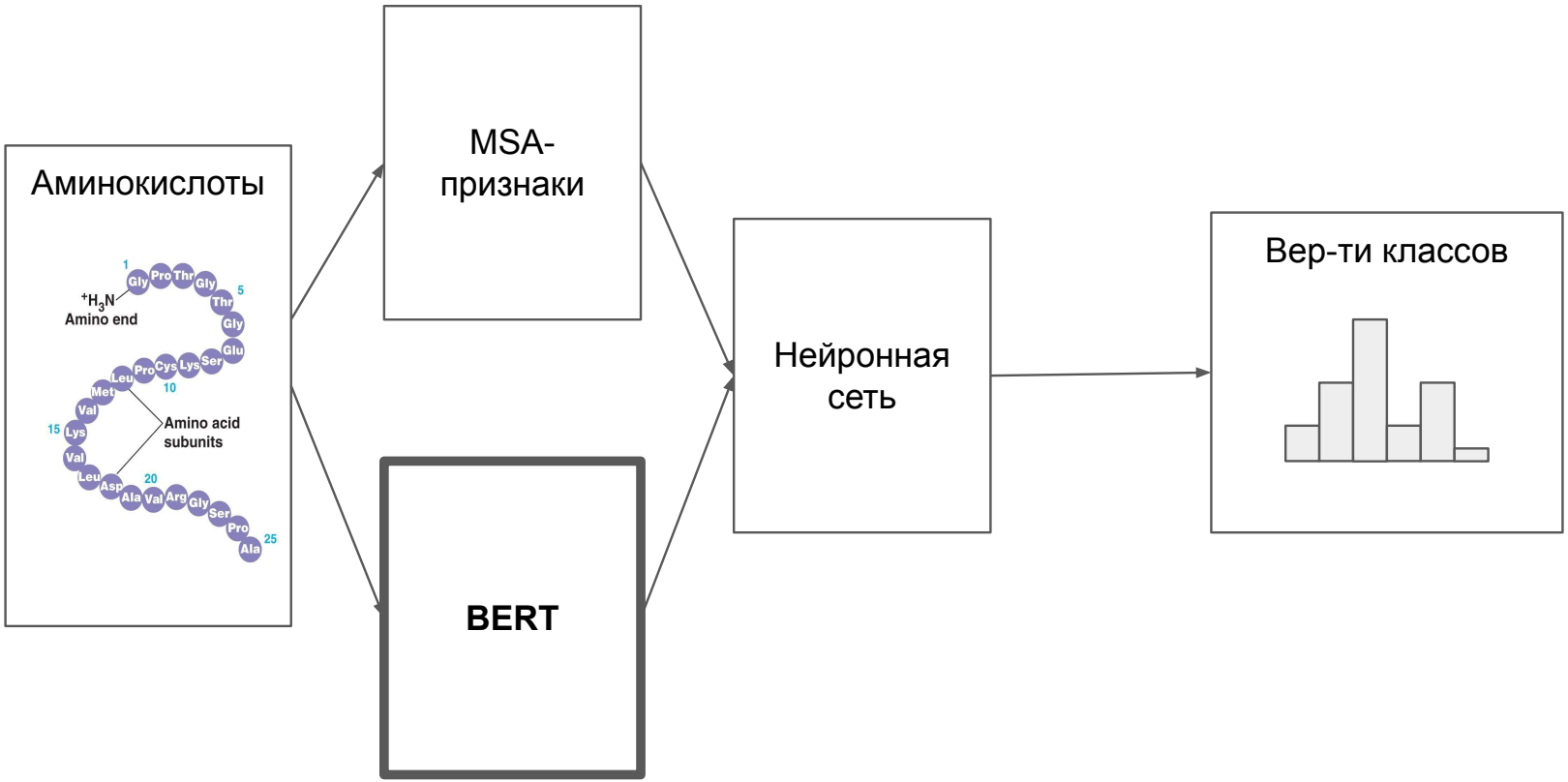
Задача

Решается задача классификаций белков (по свёрткам).

Решение

Использовать предобученный BERT для последовательностей аминокислот и применить свёрточные нейронные сети к выходу BERT-а.

Идея: заменить MSA генерацию признаков на BERT



Публикации по теме


S. de Oliveira and C. Deane.
Co-evolution techniques are reshaping the way we do structural bioinformatics.
F1000Res, 6:1224, 2017.


Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.


Elnaggar et al.
Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing.
bioRxiv, 2020.


Jie Hou, Badri Adhikari, and Jianlin Cheng.
DeepSF: deep convolutional neural network for mapping protein sequences to folds.
Bioinformatics, 34(8):1295–1303, 12 2017.

Задача классификации белков

Дано: $A = \{a_i\}_{i=1}^n$ - последовательности аминокислот.

$$A = A_{\text{train}} \cup A_{\text{val}}$$

$\mathbb{Y} = \{y_i\}_{i=1}^n$ - истинные классы.

Требуется: используя A_{train}, построить модель, предсказывающую вероятности принадлежности белка к классам.

Внешняя метрика качества: Accuracy Score на A_{val}

$$acc = \frac{1}{l} \sum_{i=0}^{l} I_{\hat{Y}_i = Y_i}$$

Данные: SCOP2, 10 классов, 1000 последовательностей.

Решение: BERT + CNN

Применить BERT

Сгенерировать дескрипторы для белков применением BERT-а.
На выходе имеем $S = \{(x_i, Y_i)\}_{i=1}^{n}$, где $x_i \in \mathbb{R}^D$ - дескрипторы.

Обучение нейронной сети

Решить задачу максимизации функции правдоподобия:

$$L_K(w) = \prod_{i=1}^{m} p_{w,x_i}(Y_i) \text{ для } K = \{(x_i, Y_i)\}_{i=1}^{m}$$

$$L_{S_{\text{train}}}(w) \to \max_{w \in W}$$

Данная работа предлагает новый способ генерировать дескрипторы. На шаге 2 берётся модификация существующей архитектуры DeepSF.

Эксперимент

- Упростили DeepSF, избавившись от ненужных слоёв.
- Взменили параметры k-max-pooling слоя (k = 60 вместо k = 30), что привело к увеличению Accuracy Score на валидации с 87% до 97%.
- Accuracy Score = 97.4% на валидации после 30 эпох обучения.
- Accuracy Score = 93% на тестовой выборке.

Задача решена успешно

Вывод:

задача классификации белков решена успешно. BERT извлёк необходимую информацию из последовательностей аминокислот.

Дальнейшая работа

Решить задачу для большего числа классов. Подробнее исследовать механизмы внимания, основываясь на работе Voita et al., 2019.