

Применение эволюционных аугментаций для работы с геномными последовательностями

Пензин Н.¹, Пензар Д.^{1, 2*}

¹ Московский государственный университет им. М.В. Ломоносова, Москва Россия

² Институт белка РАН, Пущино, Россия

* dmitrypenzar1996@gmail.com

Ключевые слова: аугментации; EvoAug; LegNet

Мотивация и цель: В прошлом году американскими исследователями из Центра количественной биологии Симонса (Simons Center for Quantitative Biology) был опубликован фреймворк EvoAug на PyTorch с открытым исходным кодом [1]. В биологии частой проблемой является недостаточность данных для обучения глубокой нейронной сети (DNN), вследствие чего нейронная сеть склонна переобучаться и показывать плохую способность к генерализации на новых данных. Авторы показывают, что обучение нейронных сетей с помощью фреймворка EvoAug, применяющего в ходе обучения набор аугментаций, основанных на псевдоэволюционных процессах, приводит к улучшению генерализации моделей и повышает эффективность стандартных методов апостериорного сравнения, например, интерпретируемости фильтров и анализа параметров, в решении важных задач регуляторной геномики.

Целью данной работы является проверка эффективности и апробация фреймворка для различных задач регуляторной геномики.

Методы и алгоритмы:

- Basset – конволюционная нейронная сеть, используемая в качестве многозадачной бинарной классификации сайтов доступности хроматина в 161 типе клеток [2].
- DeepSTARR – конволюционная нейронная сеть, представляющая собой многозадачную количественную регрессию для прогнозирования активности энхансера на основе данных секвенирования самотранскрибирующейся активной регуляторной области (STARR-seq) [3].
- LegNet – конволюционная нейронная сеть для предсказания экспрессии генов из массового параллельного репортерного анализа [4].
- Данные массового параллельного репортерного анализа для трех основных типов клеток ENCODE: гепатоциты человека (HepG2), лимфоциты (K562) и индуцированные плюрипотентные стволовые клетки (WTC11) [5].

Результаты: Авторы данного фреймворка показывают улучшение генерализации моделей Basset и DeepSTARR с использованием набора аугментаций, предоставляемых их алгоритмом. Нами были воспроизведены результаты, представленные в статье, и сделаны выводы о действительном улучшении качества предсказаний для задач классификации сайтов доступности хроматина и прогнозирования активности энхансера.

Рассматриваемый фреймворк позволил также повысить качество модели, направленной на предсказание эффективности трансляции у прокариот. Также EvoAug был испробован для задачи предсказания экспрессии эукариотических

генов для человека из массового параллельного репортерного анализа и для дрожжей из гигантского параллельного репортерного анализа. Были подобраны оптимальные гиперпараметры для каждого типа аугментации, поддерживаемого алгоритмом. Обучение модели с набором аугментаций, предоставляемых фреймворком, не позволило улучшить результаты исходной нейронной сети для обеих задач.

Выводы: Для части задач регуляторной геномики фреймворк EvoAug действительно позволяет получать модель с лучшим качеством, нежели исходные. Однако для части датасетов может вообще не приводить к улучшению качества предсказания, а дообучение может не позволить избавиться от искажений в данных, вносимых аугментациями.

Application of evolutionary augmentations to work with genomic sequences

Penzin N.¹, Penzar D.^{1, 2*}

¹ Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow, Russia

² Institute of Protein Research, RAS, Pushchino, Russia

* dmitrypenzar1996@gmail.com

Key words: augmentation; EvoAug; LegNet

Motivation and Aim: Last year the EvoAug, an open-source PyTorch package, was published by American researchers from the Simons Center for Quantitative Biology [1]. A common biology problem is insufficient data to train a deep neural network (DNN), causing the neural network to overtrain and exhibit poor generalization ability on new data. The authors of the original paper show that training neural networks using the EvoAug framework, that provides a suite of evolution-inspired data augmentations, leads to better generalization performance and improves efficacy with standard post hoc explanation methods, including filter interpretability and attribution analysis, across prominent regulatory genomics prediction tasks for well-established DNNs.

The purpose of this work is to test the effectiveness and test the framework for various tasks of regulatory genomics.

Methods and Algorithms:

- Basset is a convolutional neural network used as a multi-task binary classification of chromatin accessibility sites in 161 cell types [2].
- DeepSTARR is a convolutional neural network that is a multi-task quantitative regression for predicting enhancer activity based on self-transcribing active regulatory region sequencing (STARR-seq) data [3].
- LegNet is a convolutional neural network for predicting gene expression from yeast gigantic parallel reporter assays [4].
- Massive parallel reporter assay data for three major ENCODE cell types: human hepatocytes (HepG2), lymphoblasts (K562) and induced pluripotent stem cells (WTC11) [5].

Results: The authors of this framework show improved generalization of the Basset and DeepSTARR models using the set of augmentations provided by their algorithm. We reproduced the results presented in the article, additionally tested them against more

strict baseline methods and drew conclusions about a real improvement in the quality of predictions for the tasks of classifying chromatin accessibility sites and predicting enhancer activity. Next we tested EvoAug on a task of prediction translation efficiency in prokaryotes that resulted in improved model performance.

After that we decided to apply EvoAug for eukaryotic gene expression prediction using data from a massively parallel reporter assay. Optimal hyperparameters were selected for each type of augmentation provided by the algorithm. Training the model with a set of augmentations provided by the framework did not improve the results of the original neural network.

As the last step we applied the framework for the yeast GPRA promoter dataset and showed that using the framework for this task results in suboptimal results.

Conclusion: For some regulatory genomic problems the EvoAug framework actually allows one to obtain a model with better quality than the original ones. However, for some datasets it may not lead to an improvement in the quality of prediction at all or even result in suboptimal results. Moreover, further model fine-tuning suggested by authors may not allow it to get rid of bias in the data introduced by augmentations and in general doesn't result in performance gains.

Список литературы/References

1. Lee N.K., Tang Z., Toneyan S. et al. EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biol.* 2023;24(1):105. doi 10.1186/s13059-023-02941-w
2. Penzar D., Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics.* 2023;39(8):btad457. doi 10.1093/bioinformatics/btad457
3. Kelley D.R., Snoek J., Rinn J.L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 2016;26(7):990-999. doi 10.1101/gr.200535.115
4. de Almeida B.P., Reiter F., Pagani M., Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet.* 2022;54(5):613-624. doi 10.1038/s41588-022-01048-5
5. Agarwal V., Inoue F., Schubach M. et al. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv.* 2023. doi 10.1101/2023.03.05.531189