67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF)

21.08. - 25.08.2022, online

Meeting Abstract

Potential Applications of Transfer Learning in Limited Biomedical Data

Youngjun Park - Universitätsmedizin Göttingen, Georg-August-Universität, Institut für Medizinische Informatik, Göttingen, Germany

Anne-Christin Hauschild - Universitätsmedizin Göttingen, Georg-August-Universität, Institut für Medizinische Informatik, Göttingen, Germany

Dominik Heider - Department of mathematics and computer science at the Philipps University Marburg, Marburg, Germany

Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF). sine loco [digital], 21.-25.08.2022. Düsseldorf: German Medical Science GMS Publishing House; 2022. DocAbstr. 75

doi: 10.3205/22gmds112 , urn:nbn:de:0183-22gmds1121

Published: August 19, 2022

Text

Introduction: Tremendous advances in next-generation sequencing technology have enabled the accumulation of large amounts of omics data and various applications in medicine. However, study limitations due to small sample sizes, especially in rare diseases, technological heterogeneity, and batch effects still limit the applicability of traditional statistics and machine learning analysis. Here, we present a transfer-learning-based approach to transfer knowledge from big data and reduce the search space in data with small sample sizes.

Methods: Transfer learning is conducted with a few-shot learning model [1]. The model was pre-trained with a large-scale dataset. After that, the trained network is transferred to a new task and fine-tuned with the small dataset for the new task. We investigated three public datasets, TCGA (The Cancer Genome Atlas), GTEx dataset, and five different human pancreas single-cell datasets. The used TCGA cancer dataset consists of 33 different cancer types. The used GTEx dataset consists of 48 tissue types. First, we investigated the transfer of knowledge between the GTEx and TCGA datasets focussing on tissue classification. The heterogeneous human pancreas single-cell datasets consist of several pancreas-specific cell types which only partially overlap amongst datasets. Therefore, we transferred a model trained on the GTEx dataset to improve cell-type classification in the human pancreas single-cell data.

Results: First, we pretrained a model on the GTEx dataset and transferred it to the TCGA dataset. We evaluated the performance of the model classifying all 33 TCGA cancer types showing a 78.91% ± 0.76% accuracy. The model pre-trained on TCGA and transferred to GTEx to distinguish the 48 tissue types, shows a 84.57% ± 0.66% accuracy. Moreover, when we used 5% of the original TCGA data for fine-tuning, the accuracy reached over 94%, which is close to the accuracy of the state-of-the-art method (95.7%) that was trained using 80% of the data. Afterward, we evaluated the performance of the cell-type classification model on the human pancreas single-cell dataset. The model is pretrained with the GTEx dataset and fine-tuned with a small subset of the target single-cell sequencing dataset. We observed that, when the fine-tuning dataset is limited, the pre-training with bulk-cell sequencing significantly improved the cell-type classification accuracy on the single-cell sequencing data.

Discussion: Our analysis of the TCGA and GTEx datasets confirmed that transfer learning can offer benefits in training time cost and robustness, in particular, if the fine-tuning dataset is very limited. In the single-cell data analysis, we also find that transfer learning can reduce training time cost and improve accuracy when the target dataset is limited. This finding is concordant with transfer learning in other fields, computer vision or bio-imaging [2]. This demonstrates the potential of transfer learning on large-public data to improve analysis in small data scenarios, such as rare diseases.

Conclusion: We investigated the potential value of transfer learning using different types of transcriptomics data. The application of transfer learning can be beneficial for circumstances of where data is limited frequently, such as the medical domain. In particular, it can mitigate noise from batch effects and overcome technological heterogeneity by leveraging existing public data resources. This approach could be especially useful for low data abundance of rare diseases and to overcome systematic biases amongst different medical institutions.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. This work reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

The authors declare that they have no competing interests.

The authors declare that an ethics committee vote is not required.

References

1. Sung F, Yang Y, Zhang L, Xiang T, Torr PHS, Hospedales TM. Learning to Compare: Relation Network for Few-Shot Learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. p. 1199-1208. DOI: 10.1109/CVPR.2018.00131
2. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019).
3. Park Y, Hauschild AC, Heider D. Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. NAR Genom Bioinform. 2021 Nov 12;3(4):lqab104. DOI: 10.1093/nargab/lqab104