Code and write-up

The goal of the assignment is to provide hands-on experience in implementing state-of-the-art algorithms. Students are encouraged to use the algorithms discussed in class or use ones that we have not discussed. Students are expected to submit a written report describing their implementation, results, and conclusions (up to 5 pages; can be a Jupyter notebook), as well as the code and data used (can be a link to the github repository).

Single-cell RNA sequencing and drug perturbations

Paper:

https://www.science.org/doi/10.1126/science.aax6234

Data:

https://cellxgene.cziscience.com/collections/00109df5-7810-4542-8db5-2288c46e0424
https://github.com/cole-trapnell-lab/sci-plex
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139944

Possible tasks:

- Unsupervised visualization of cells and their drug perturbation conditions
- Predicting drug perturbations from single cells and identify genes (features) that are predictive
- Identifying cell clusters and their associated genes
- Predicting the effect of drug perturbations on gene expression

Possible algorithms:

- https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language/
- UMAP/tSNE/PCA
- Graph Neural Networks
- https://arxiv.org/abs/2002.05709
- https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
- https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/
- https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Additional resources:

- https://scanpy.readthedocs.io/en/stable/
- https://cloud.r-project.org/web/packages/Seurat/index.html

Breast Histopathology

Paper:

https://pubmed.ncbi.nlm.nih.gov/27563488/
https://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872?SSO=1

Data:
https://www.kaggle.com/paultimothymooney/breast-histopathology-images

Possible tasks:

- Visualization of data features with UMAP, tSNE, PCA
- Use unsupervised learning to obtain an embedding for the image samples using a SOTA algorithm for image task (see possible algorithms list)
- Identify features in the data that are predictive of Invasive Ductal Carcinoma (IDC)
- Segmentation of malignant cells (semantic segmentation, saliency plot, attention visualization or similar)

Possible algorithms:

- DINO: https://arxiv.org/abs/2104.14294
- MAE: https://arxiv.org/abs/2111.06377
- ViT: https://arxiv.org/abs/2010.11929
- SimCLR: https://arxiv.org/abs/2002.05709
- NMCE: https://arxiv.org/abs/2201.10000

Crypto forecasting

Data and task:
https://www.kaggle.com/c/g-research-crypto-forecasting/
https://towardsdatascience.com/cryptocurrency-price-prediction-using-deep-learning-70cfca50dd3a
https://cryptodatum.io/
https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

Possible Algorithms:

- https://towardsdatascience.com/neural-odes-breakdown-of-another-deep-learning-breakthrough-3e78c7213795
- Transformers
- RNNs

Spotify

Data and tasks:
https://research.atspotify.com/datasets/
Possible tasks:

- Build a recommender system - predict new songs/playlists for listeners
- Unsupervised visualization of songs, playlists, or listeners

- Generate new playlists / predict optimal playlists

Possible algorithms:

- tSNE/PCA/UMAP
- Self supervised learning: DINO/SimCLR/ViT/data2vec
- Transformers/Bert
- http://www.recsyschallenge.com/2018/
- https://dl.acm.org/doi/proceedings/10.1145/3267471

Rubrics

The following rubrics will be used to grade your coding assignment:

1) How the question/problem is framed;
2) Choice and handling of the dataset to address the proposed problem;
3) Design of the method to analyze the data;
4) Quality of the report, which should consist of:
    a) Introduction
    b) Background / Related work
    c) Method description
    d) Results & Discussion
    e) Conclusion
    f) ~ 5 pages

Keep in mind that we will run your code, which implies you will need to provide the data you used, in case you decide to use a different dataset.

Feel free to reach out if you have questions.