LDEncoder: Reference deep learning-based feature detector for transfer learning in the field of epigenomics

Gun Woo (Warren) Park[1] and Kevin Bryson[1,2]

1. Department of Computer Science, University College London, 2. School of Computing Science, University of Glasgow

Introduction

Motivation

- Many impactful machine learning approaches on epigenomics are not suitable to solve different, but relatively similar problems [1-3]
- If the approach can generalise well, then the model developed is limited to be used as a translational model [4]
- This creates inefficient workflow when researchers first encounter novel epigenomics data since they need to devise a bespoke solution approach for the problem that they want to solve

Goal of the project

1. Development of generalisable model trained using a pan-cancer dataset
2. Formalising the research problem

Research Questions

RQ1: Can there be a transferable, pre-trained feature detector for the DNA methylation data?
RQ2: Is it possible to encode the input features from DNA methylation data into a significantly lower dimension?
RQ3: Is it possible to perform common machine learning tasks using the pre-trained feature detector? (e.g., transfer learning and meta-learning)

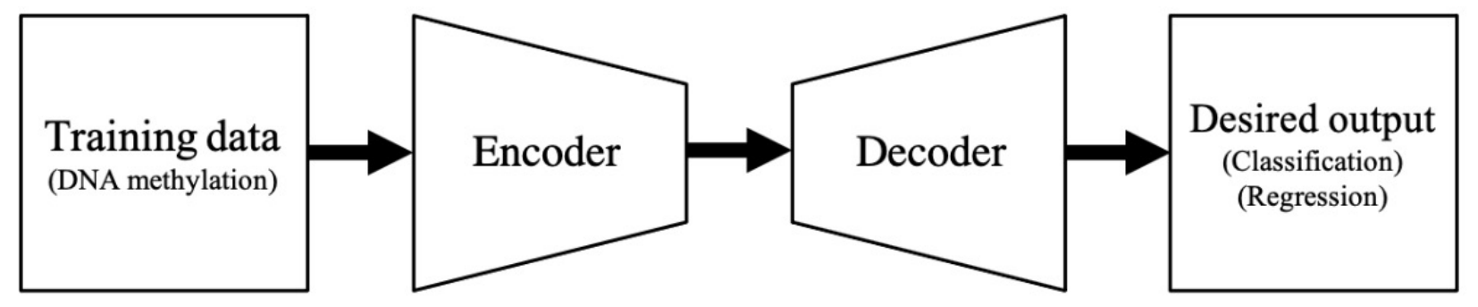
Methods

Dataset

The Cancer Genome Atlas (Used the same dataset used for [4], and preprocessing method)

- Methylation beta values
- RNA-seq gene expression values
- PAM 50 subtypes
- For the classification problem, the Synthetic Minority Oversampling Technique [7] was used to oversample the training set in order to solve the class imbalance problem

Models



- Encoder-decoder models used
- Inspiration from VGG [6], ResNet[5], and TDImpute [4]
- Encoding dimension of 512 was enforced for novel models

Sorting

- Illumina's CG reference number are not sorted in terms of the coordinates
- Unsorted dataset does not matter for fully connected networks but becomes important when we try to utilise locality features
- Sorting ensures locality features could be exploited, which means convolutional neural networks can be used for solving problems.

Final model

- LDEncoder: Enhanced model based on the model used for the TDImpute [4]
- Trained for 600 epochs, which other models were trained for 300 epochs (if early stopping is not used). This was practical as the number of parameters in LDEncoder was fewer

Transfer learning and meta learning

- The final regressor was substituted by classifiers to solve downstream tasks while utilising the trained parameter which contains knowledge from pretrained pan-cancer dataset
- Meta leaning was tried to see whether the trained model can obtain problem-specific knowledge

Evaluation

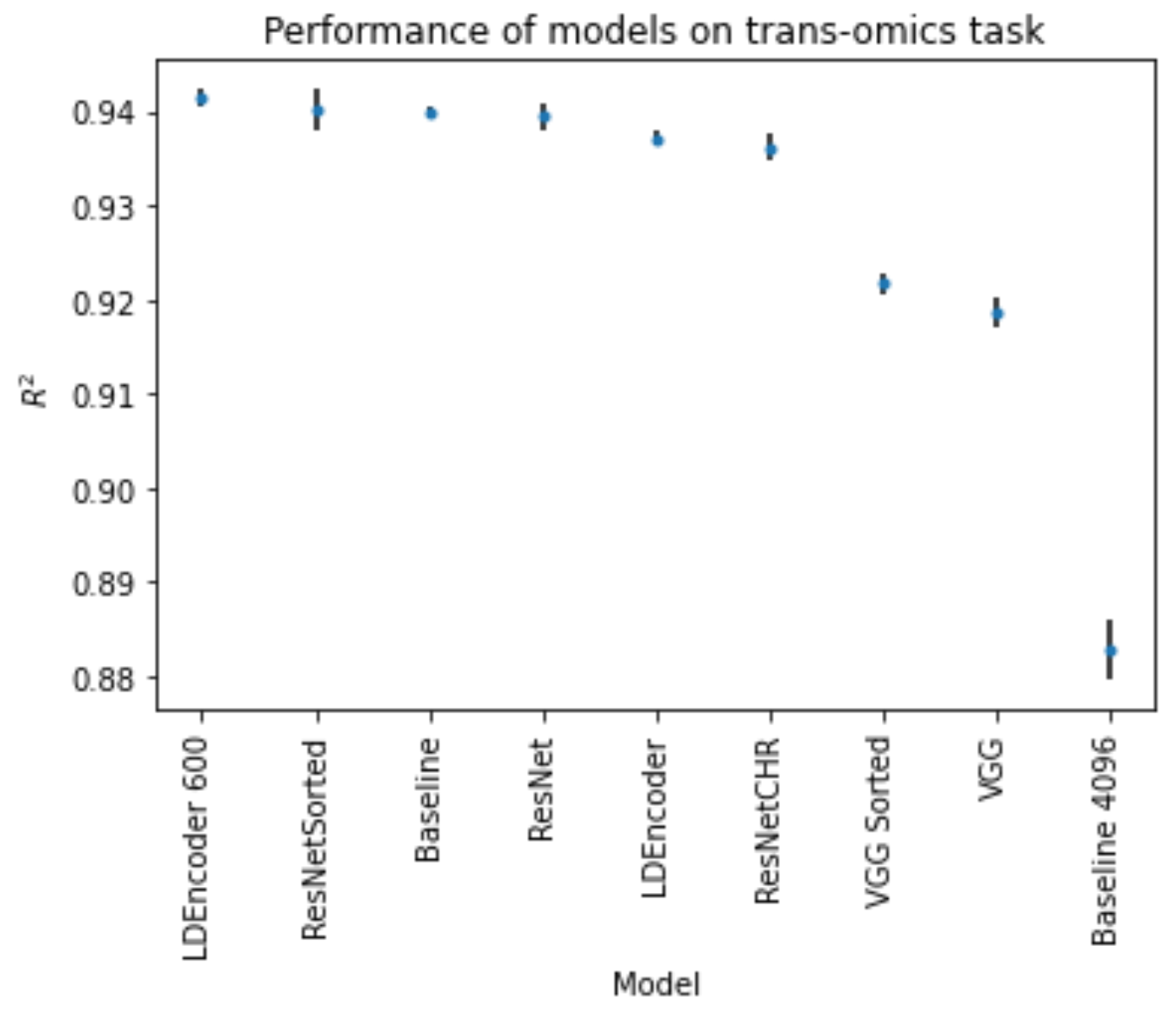
- Regression: 5 repeats - randomly selected held-out set for each repeat
- Classification: 15 repeats - randomly selected held-out set for each repeat

Result 1: LDEncoder performs better

Models including the state-of-the-art model (Baseline) were trained with the TCGA (The Cancer Genome Atlas) pan-cancer dataset and tested on the held-out set.

- Prediction of RNA-seq gene expression value from DNA methylation beta values (**Regression** – trans-omics)
- 5 repeats of the experiment occurred
- For each repeat, training set, testing set, and validation set were randomly generated from the pool of all the sample entries
- Set split ratio: 81% training set, 10% testing set, and 9% validation set
- Early stopping used for models except for the baseline and the LDEncoder
- Root mean squared error and squared Pearson's correlation coefficient were used for metric (R^2). The R^2 was used for ranking the performance.

	RMSE	$\mathbf{R^2}$
LDEncoder 600[1]	0.6496± 0.0048	**0.9414± 0.0009**
ResNetSorted[2]	0.6568±0.1240	**0.9402±0.0022**
Baseline	0.6566±0.0030	**0.9400±0.0007**
ResNet	0.6608±0.0085	**0.9395±0.0014**
LDEncoder	0.6725±0.0050	**0.9371±0.0008**
ResNetCHR[3]	0.6778±0.0083	**0.9362±0.0014**
VGGSorted[4]	0.7549±0.0054	**0.9217±0.0011**
VGG	0.7696±0.0071	**0.9186±0.0015**
Baseline 4096[5]	0.9201±0.0117	**0.8828±0.0031**
PCA	4.1664±0.0248	**0.2750±0.0047**



[1] LDEncoder model with 600 epochs
[2] ResNet [5] inspired model with a sorted dataset
[3] ResNet [5] inspired model with a sorted dataset with chromosomal separation of the data
[4] VGG [6] inspired model with a sorted dataset
[5] Baseline experiment with a batch size of 4096

*PCA model was excluded because the R^2 value for the PCA model was significantly different

Result 2: LDEncoder generalises well

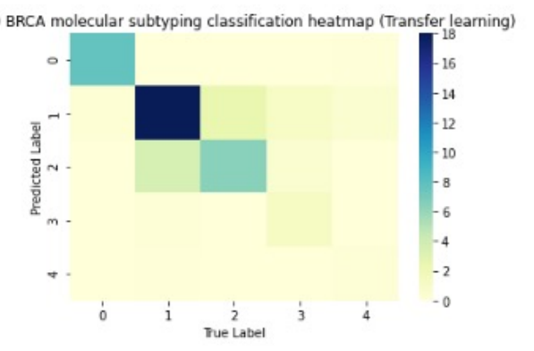
Two conditions were used for testing the **classification** capability of the model, as well as the usefulness of training with a pan-cancer dataset

- LDEncoder trained with a pan-cancer dataset without BRCA samples was used for the transfer learning
- LDEncoder without pre-training was used for the random initialisation (control condition)

PAM50 Molecular subtype Classification

	Test loss	Test accuracy/%	Test (micro) F1-score
Transfer learning	0.6495±0.0965	78.14±5.19	0.7835±0.0093
Random initialisation	0.6689±0.1137	77.98±6.02	0.7743±0.0137

0: Basal-like, 1: Luminal A. 2: Luminal B
3: HER2-enriched, 4: Normal-like

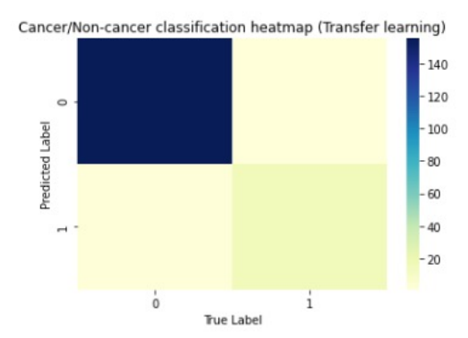


(a) Transfer learning

Cancer/Non-cancer Classification

	Test loss	Test accuracy/%	Test (micro) F1-score
Transfer learning	0.0719±0.0461	98.35±0.81	0.9815±0.0032
Random initialisation	0.0959±0.0862	97.78±2.72	0.9759±0.0046

0: Cancer, 1: Non-cancer



(a) Transfer learning

Conclusion

- LDEncoder was shown to be outperforming the state-of-the-art model in a trans-omics task
- Complicated models, such as ResNet inspired model have shown some potential of performing well in the trans-omics task
- The suggested reference model does not only solve the training goal of inferring the gene expression data from the DNA methylation data but also can be used with transfer learning that either link epigenomics with other neighbouring -omics domains or solves other problems in epigenomics.
- LDEncoder architecture outperforms other epigenomic state-of-the-art models in terms of BRCA PAM50 molecular subtype prediction, while the result is not fully conclusive within this research due to the restrictions on the use of datasets posed by the ethics approval status and resource availabilities
- We can reduce the number of parameters in the state-of-the-art model by introducing a low dimensional latent feature representation layer

Future works

- Experiments with more repeats to more accurately demonstrate the performance benefits of the novel model
- More thorough hyperparameter search to improve the performance of more complicated models e.g., ResNet inspired models
- Experiments with more samples so that Y chromosome could be included in the analysis and more conclusive results could be given

References

[1] Karim Malki, E Koritskaya, F Harris, K Bryson, M Herbster, and MG Tosto. Epige-netic differences in monozygotic twins discordant for major depressive disorder.*Translational psychiatry*, 6(6):e839–e839, 2016.
[2] Xuesi Dong, Lijuan Lin, Ruyang Zhang, Yang Zhao, David C. Christiani, YongyueWei, and Feng Chen. ToBMI: Trans-omics block missing data imputation using ak-nearest neighbor weighted approach.*Bioinformatics*, 35(8):1278–1283, 2019.
[3] Joshua J. Levy, Alexander J. Titus, Curtis L. Petersen, Youdinghuan Chen, Lucas A.Salas, and Brock C. Christensen. MethylNet: An automated and modular deeplearning approach for DNA methylation analysis.*BMC Bioinformatics*, 21(1):1–15,2020.
[4] Xiang Zhou, Hua Chai, Huiying Zhao, Ching Hsing Luo, and Yuedong Yang.Imputing missing RNA-sequencing data from DNA methylation by using a transferlearning-based neural network.*GigaScience*, 9(7):1–10, 2020.
[5] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition 2015.
[6] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique *Journal of articial intelligence research*, vol. 16, pp. 321-357, 2002.

Acknowledgement

We thank TCGA Research Network for providing the necessary datasets required for the completion of this research work.

We also thank University College London for providing the necessary support for the completion of this project.

warrenjamespark@gmail.com
Kevin.Bryson@glasgow.ac.uk