Science and Bio-engineering
Department Computer Science

Generalized deep learning cancer detection from epigenetic marks

Advanced Methods in BioInformatics report

Jari Beysen - 3rd Bachelor Physics and Astronomy

01/06/2022

Academic year: 2021 – 2022

Contents

1 Introduction II

2 Dimensionality reduction 1
2.1 Principal Component Analysis 1
2.2 Autoencoders and Variational Autoencoders 2

3 Classification and supervised learning 2

4 SVAE architecture 2
4.1 Data acquisition 3
4.2 Model 4

5 Generating synthetic data 6

6 Evaluation 6

7 Discussion 7

1 Introduction

Curing cancer is one of the big frontiers in the medical fields. The end is however not nigh, but major advances are being made year by year. In 2020 for example, the Pan-cancer project [1] was launched. This large scale international project aimed to collect and analyze large amounts of whole-cancer genomes. 2658 genomes and their corresponding normal tissues were made open access for anyone to use. Many papers regarding cancer detection have been published already. Specifically, attempting to accurately detect cancer from DNA Methylation data. DNA methylation is the most widely studied epigentic mark in mammals. It prohibits DNA transcription, this interaction is associated with gene expression. Keun H. and Erdenebileg [2] tackled the problem of cancer detection by applying a Supervised Variational AutoEncoder model (SVAE) to lung cancer DNA methylation data. In this report we will re-implement their deep learning approach and try to broaden and improve it.

2 Dimensionality reduction

Larger featurespaces are generally computationally ill-advised to work with. Sometimes noisy unfiltered data with a large amount of features can even be at a disadvantage because it can lead to overfitting. In the next two paragraphs two different dimesion reduction methods will be discussed.

2.1 Principal Component Analysis

The most frequently used method is PCA. This method tries to orient n orthogonal eigenvectors through a N dimensional (n$\ll$N) featurespace. This is done by finding a first axis on which the data retains the highest variance when projected. The following axis will be recursively created by finding an orthogonal axis which tries to retain the highest variance. This is repeated until a certain dimension is reached, or the last dimension cannot retain more than a percentage of the original data's variance.

In the field of DNA methylation analysis, most methods make use of Principal Component Analysis (PCA) to reduce the dimensionality of the data. This is very much needed since these dataset have either 27,578 or 450,000 features, which were respectively generated from the illumina infinium 27k and 450k beadchip. Each of these features represents a CpG island, this is a 200 base-pair region of DNA with more than 50% CG content and an observed/expected ratio larger than 0.6.

PCA falls short due to the fact that it is purely based in linear algebra. Since we cannot guarantee that Methylation pattern disruption can be linearly approximated, a more advanced dimensionality reduction method is needed. This is where AutoEncoders (AE) come into the picture.

2.2 Autoencoders and Variational Autoencoders

Autoencoders are a class of unsupervised neural networks. The encoder forces data through a ‘bottleneck’, from which the decoder tries recreate the input data. Doing this effectively forces the network to express the input data through a lower dimension. This space is formally called the ‘latent space’. Data encoded into this latent space does not intrinsically carry any meaning, it does however carry some sort of information linked to the input data. Notice that this bears a lot of similarity to PCA.

A problem with standard autoencoders is a lack of smoothness in the latent space. What we would like is that a small change in the input data would result in a small change in the latent space. This is however not the case for autoencoders. What Diederik P. K. and Max W. [3] figured out in 2014 is that this problem can be fixed by distribution matching the input data and the latent space. This is done by letting the encoder learn a mean and a variance and feed them into a multivariate normal distribution, which acts as a ‘sampler’. Distribution matching is done by introducing Kullback–Leibler divergence into the loss function. Jonathan S. [4] provides a short run through of the underlying mathematics. This model provides a distribution matching dimensionality reduction solution which is non-linear in nature.

3 Classification and supervised learning

In this report we will not attempt to differentiate between different cancer types. The goal here is to predict whether data is cancerous. This implies that we need a binary classification solution. A model which is frequently used in combination with PCA is logistic regression. We however would like a more robust and integrated solution. This requirement leads us to the concept of supervised varational autoencoders, as mentioned in section 1. The term ‘supervised’ purely refers to working with labeled data.

To integrate a classification solution into a variational autoencoder, we can add a simple branch of neurons, attached to the latent space, which will be able to classify data by adding a binary crossentropy term to the loss function.

4 SVAE architecture

In this section, data augmentation and the neural network architecture, which was used for testing will be discussed. Disagreements with the original paper and the resulting changes will also be mentioned.

4.1 Data acquisition

In the original paper, the 5000 highest variance features from data acquired by the 450k beadchip were used as training data. In this report however, all the features from the 27k beadchip were used. They reasoned that the 5000 highest variance features would reduce features that do not carry information. Variational autoencoders however do not take features into account that do not carry information inherently. The 27k beadchip was chosen because the system which the network ran on did not have enough memory to store 450,000 input nodes and the resulting network.

The data consisted of 2520 total cancerous and normal samples of the following types: Lung Squamous Cell Carcinoma (LUSC), Lung Adenocarcinoma (LUAD), Rectum adenocarcinoma (READ), Glioblastoma (GBM), Uterine Corpus Endometrial Carcinoma (UCEC), Colon adenocarcinoma (COAD), Ovarian serous cystadenocarcinoma (OV), Acute Myeloid Leukemia (LAML), Breast invasive carcinoma (BRCA), and Kidney renal clear cell carcinoma (KIRC). Kein et al [2] however only used samples of two different lung cancer types, namely LUAD and LUSC.

4.2 Model

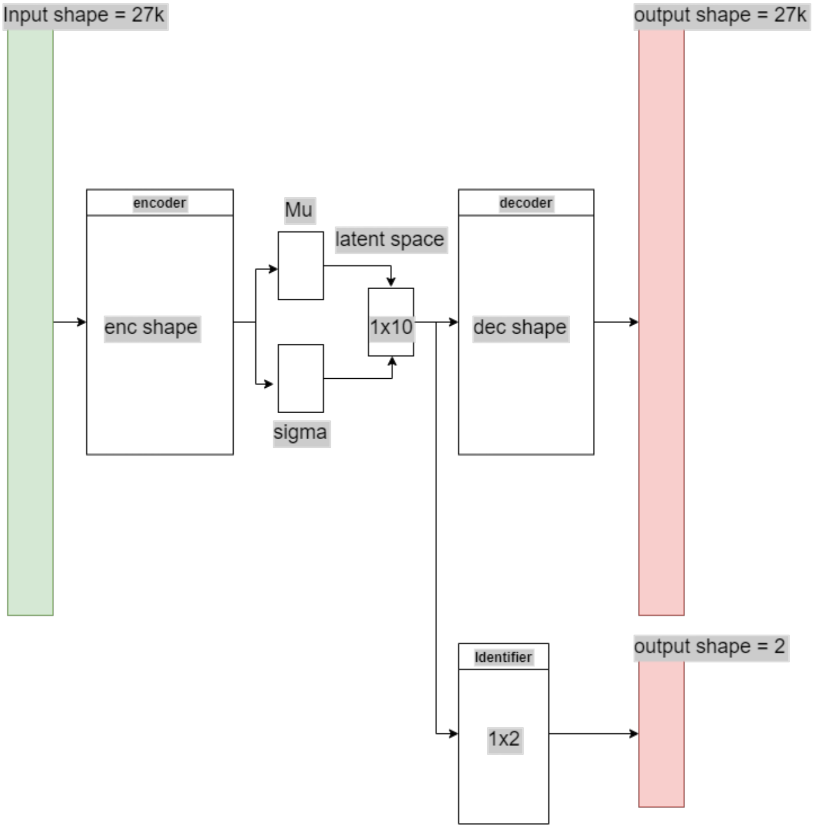

Figure 1: SVAE architecture

The final model used 1 is a pretty standard SVAE. Its loss function takes the following form:

$$L = \frac{1}{N}\sum_i (x_i - \hat{x}_i)^2 + \frac{1}{N}\sum_i \left[-y_i log(\hat{y}_i) - (1-y_i) log(1-\hat{y}_i)\right] + 0.01 \cdot D_{KL}(q(z|x), \mathcal{N}(\mu,\sigma))) \tag{1}$$

In this equation the first part is the Mean Squared Error (MSE) of the output, the second part is Binary CrossEntropy (BCE) of the label prediction, and D_{KL} is the Kullback–Leibler divergence of the encoded data. In table 1 the model's hyperparamaters are shown together with some specific architecture dimensions.

Input & output shape	27,578
latent space size	10
encoder & decoder layers	3
encoder & decoder nodes/layer	0.005*input shape * [(N-n)/N & n/N]
activation function	gaussian error function
optimizer	Nadam
learning rate	1e-4
epochs	30
batch size	20

Table 1: Hyperparameters & model details

This set of hyperparameters are not per se the most optimal, but displayed very nice behaviour. The model, in comparison with larger models which were tested, did not take long to converge on the training data. On an Intel i7-1165G7 it took about 14 minutes for a training set of 2200 samples to finish.

Some features which are not standard are the amount of nodes present in the encoder and decoder and the activation function used. The nodes per layer have been laid out to quickly slope towards the latent space.

A lot of activation functions were not able to achieve high levels of accuracy on this smaller neural network, except for the GELU activation function (Gaussian Error Linear Unit), and the standard Gaussian error function. The Gaussian error function is not typically used as an activation function, but it does however work fine since it is non linear. The Gaussian error function converged faster consistently, so it was used.

5 Generating synthetic data

A commonly used trope in variational autoencoders is the fact that they can generate synthetic data. Since their latent space is smooth, it is quite easy to sample from it. This can be done in many ways. In this report, data was uniformly sampled from a hypercuboid bound by the 90th percentile of the data fed through the encoder mean. This biases that generated data towards values already passed through the latent space. It is however not a guarantee that the generated methylation data is plausible nor of any use. It is mostly an interesting concept to explore.

6 Evaluation

To test the SVAE, 10 fold K-Crossfold was executed on the entire dataset. An important statistic the original paper, and other papers seem to miss or neglect is the false negative rate. The least desirable error a model can make in case of medicine is a false negative. This can easily be calculated with Bayes theorem. The percentage of tumor data is 87.4%.

$$P(p|n_g) = \frac{P(n_g|p)P(p)}{P(n_g|p)P(p) + P(n_g|n)P(n)} \tag{2}$$

Accuracy	99.7 ±0.2%
normal accuracy	98.8 ±1.9%
tumor accuracy	99.9 ±0.2%
false negative rate	1.2 ±1.5%
normal gen accuracy	99.9 ±0.2%
tumor gen accuracy	100.0 ±0%

Table 2: Relevant statistics

7 Discussion

The goal of this report was to test the plausibility of the paper presented by Keun H. et al [2], and maybe improve it. In their paper an accuracy of 96.2% was achieved, whilst the model in this report achieved a final accuracy of 99.7%. This result is also 8% higher than the results they achieved for a logistic regression applied to PCA. This makes the model in this report 200 times less likely to wrongly classify data than the standard procedure. We also calculated the false negative rate which was about 1%. This is quite good when taking into consideration that a PCR test for SARS-CoV-2 has an observed false negative rate of 10% to 14% [5].

References

[1] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, “Pan-cancer analysis of whole genomes,” *Nature*, vol. 578, no. 7793, pp. 82–93, 6th Feb. 2020, ISSN: 0028-0836, 1476-4687. DOI: `10.1038/s41586-020-1969-6`. [Online]. Available: `http://www.nature.com/articles/s41586-020-1969-6` (visited on 05/06/2022).

[2] K. H. Ryu and E. Batbaatar, “Improved cancer classification with supervised variational autoencoder on DNA methylation data,” in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, J.-S. Pan, J. Li, K. H. Ryu, Z. Meng and A. Klasnja-Milicevic, Eds., vol. 212, Series Title: Smart Innovation, Systems and Technologies, Singapore: Springer Singapore, 2021, pp. 36–43, ISBN: 978-981-336-756-2 978-981-336-757-9. DOI: `10.1007/978-981-33-6757-9_5`. [Online]. Available: `https://link.springer.com/10.1007/978-981-33-6757-9_5` (visited on 05/06/2022).

[3] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2013, Publisher: arXiv Version Number: 10. DOI: `10.48550/ARXIV.1312.6114`. [Online]. Available: `https://arxiv.org/abs/1312.6114` (visited on 06/06/2022).

[4] J. Shlens, “Notes on kullback-leibler divergence and likelihood,” 2014, Publisher: arXiv Version Number: 1. DOI: `10.48550/ARXIV.1404.2000`. [Online]. Available: `https://arxiv.org/abs/1404.2000` (visited on 06/06/2022).

[5] V. Pecoraro, A. Negro, T. Pirotti and T. Trenti, “Estimate false-negative RT-PCR rates for SARS-CoV-2. a systematic review and meta-analysis,” *European Journal of Clinical Investigation*, vol. 52, no. 2, Feb. 2022, ISSN: 0014-2972, 1365-2362. DOI: `10.1111/eci.13706`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/10.1111/eci.13706` (visited on 06/06/2022).