Deep Learning For Lung Cancer Survival Prediction Based on Transcriptomic Data

Aymen MERROUCHE

Machine Learning for Medicine

March 23, 2021

Introduction

In this project we are interested in lung cancer survival prediction based on gene-expression data with deep learning approaches (binary classfication task).

The Data

- We use the TCGA Pan-Cancer data-set, containing $\approx 9k$ RNA-Seq gene-expression samples from 33 distinct tumor types. The data is separated into two subsets : **Lung Cancer samples** $\approx 1k$ and **Non-Lung Cancer samples** $\approx 8k$.
- The gene-expression data is exploited in two forms : **gene-expression vectors** with dimension 7509 and **gene expression images** with shape 175×175.

Gene-Expression Images 1/2

- The use of Convolutional Neural Networks on gene-expression vectors is not possible due to the unstructured nature of this data. CNNs are based on convolutional filters that take advantage of local information patterns. CNN based deep architectures have proven their efficiency on image data.
- In order to allow the use of CNNs on gene-expression data, [1] introduced a method to transform the gene-expression vectors into gene-expression images exploitable by CNNs.

Gene-Expression Images 2/2

The gene-expression vectors are transformed into images, by rearranging the gene-expression vectors using a biological criterion (KEGG BRITE hierarchical information). This transformation puts together genes that share domain-specific information thus providing them with some structure and allowing the use of CNNs.

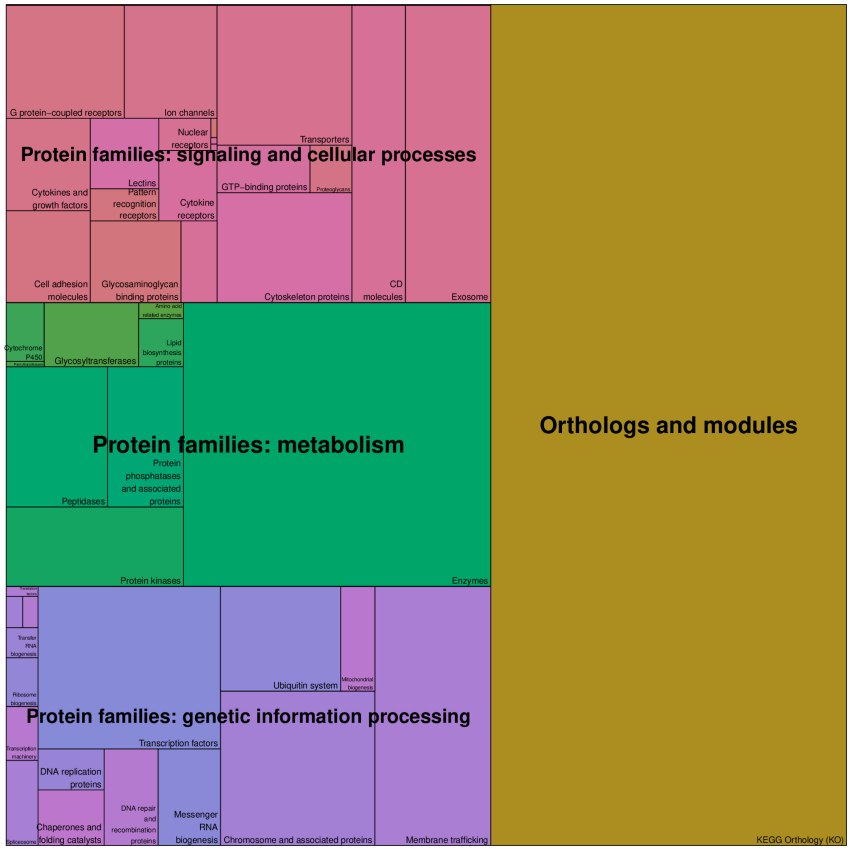


Figure: Gene-Expression vectors → Gene-Expression images

Dealing with Class Imbalance

- Our Dataset suffers from severe class imbalance with with a high number of negative examples ($\approx 85\%$ on Non Lung Cancer Samples and $\approx 90\%$ on Lung samples) and a low number of positive examples ($\approx 15\%$ on Non Lung Cancer Samples and $\approx 10\%$ on Lung samples).
- To deal with this caveat, we use two mechanisms when training our models : **Random Over Sampling** and the use of **the focal loss**.

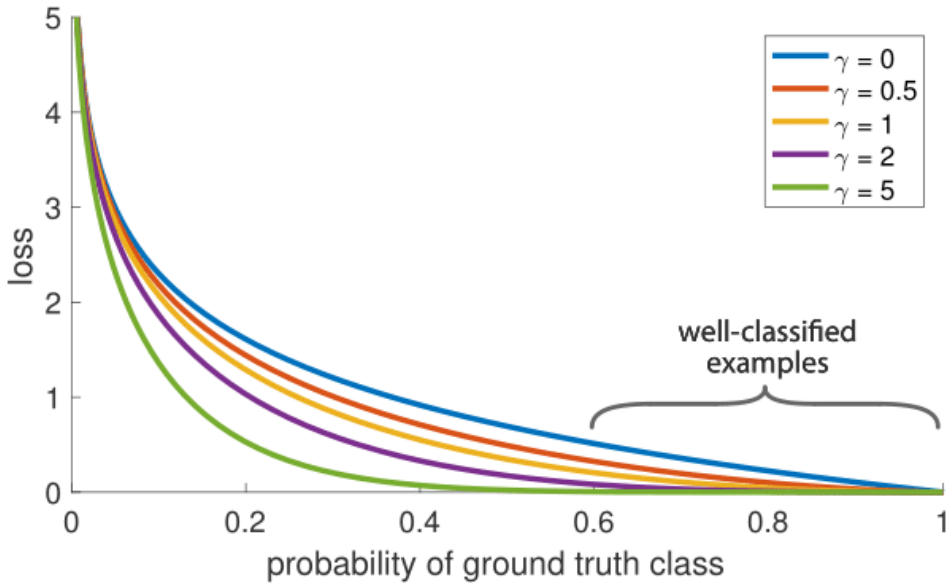
Random Over Sampling

- Random Over Sampling consists of supplementing the training data with copies of the positive class by randomly sampling, with replacement, from positive training examples. The resulting data-set contain roughly 50% of negative examples and 50% of negative examples.

Figure: Random Over Sampling

Focal Loss

- The focal loss introduced by [2] is used to train our models. It is given by the formula $\mathbf{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t)$ where $p_t = p^{1_{y=1}}(1-p)^{1_{y\neq 1}}$ and $\alpha_t = \alpha^{1_{y=1}}(1-\alpha)^{1_{y\neq 1}}$.
- It down-weighs the loss accorded to easy examples (negative examples) so that their contribution to the total loss is low, allowing the model to focus on hard examples (positive examples) during training.



Metrics

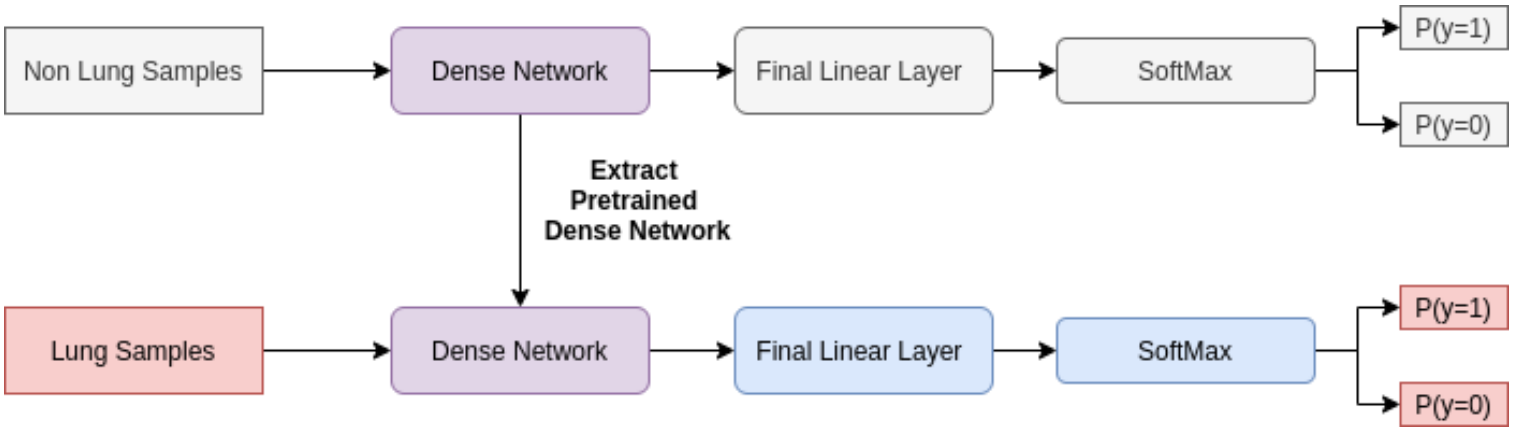
- Because of the imbalanced nature of our data-set, the accuracy metric is not informative : even a dummy model which always predicts 0, will achieve an accuracy score of 90%.
- In our experiments, We use the **AUC metric** as a criterion to judge the quality of a model. Therefore the best model is the one which achieves the best AUC score. Generally speaking, an AUC > 0.7 is considered acceptable.

Models

Different deep learning approaches were tested :

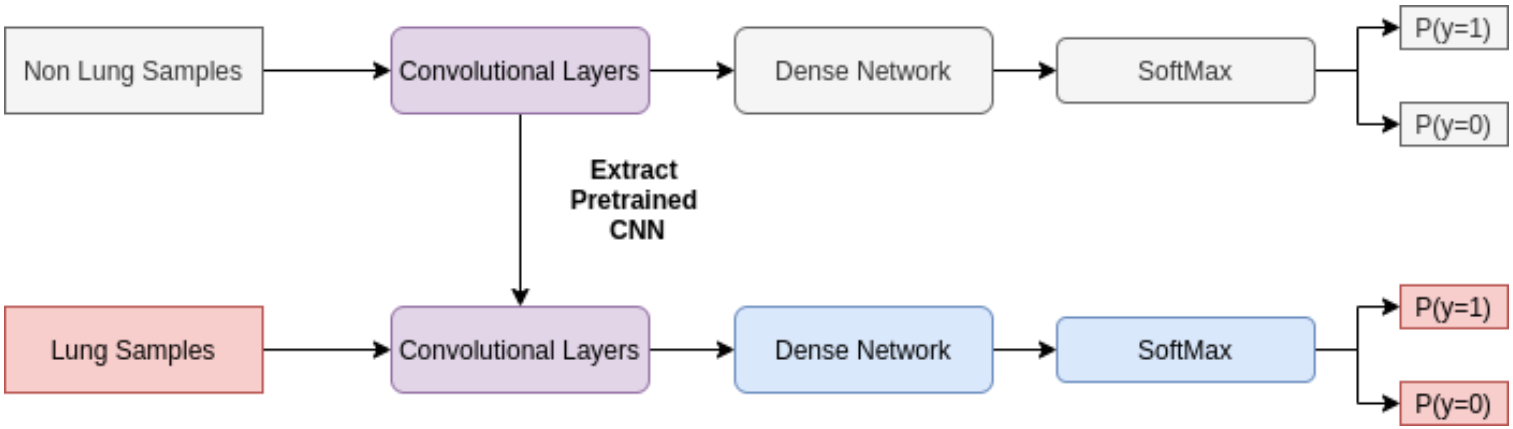
- Multi layer perceptron, with or without transfer learning.
- CNN with transfer learning, (convolution and deformable convolution).
- MLP with VAE embeddings, with or without transfer learning.

Multi Layer Perceptron



- With Pretraining On Non Lung Samples : **AUC = 0.70**.
- Without Pretraining On Non Lung Samples : **AUC = 0.67**.

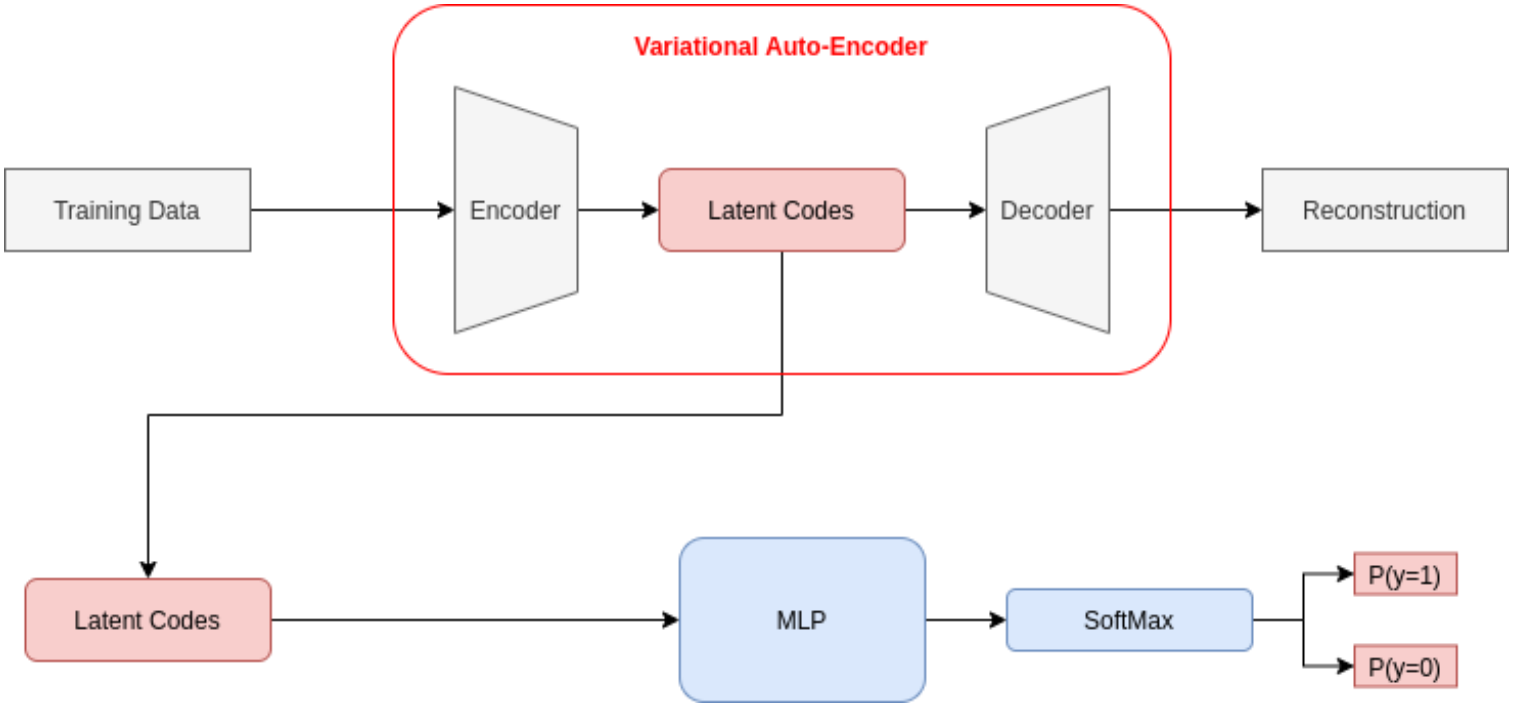
Convolutional Neural Network 1/2



- With Pretraining On Non Lung Samples : **AUC = 0.70**.

- We can say that transforming the gene-expression vectors into gene-expression images did allow CNNs to extract high-level features from the data and achieve an AUC of 0.7. However this approach is still limited when compared to what can be done with real image data.
- We tried to enhance the transformation modeling capacity of CNNs by using deformable convolution introduced in [3] which augments the spatial sampling locations of CNNs by learning the offsets. This methods, although it has shown it's effectiveness on traditional computer vision tasks, achieved poor performances on our gene-expression images with no generalization ability.

Learning Embeddings with a VAE



- With Pretraining On Non Lung Samples : **AUC = 0.66**.
- Without Pretraining On Non Lung Samples : **AUC = 0.73**.

Final Results

Table: Deep Architectures

Approach	Model	AUC
Transfer Learning	MLP	0.70
	CNN	0.70
	VAE+MLP	0.66
No Transfer Learning	MLP	0.67
	VAE+MLP	**0.73**

Table: ML Algorithms

Approach	Model	AUC
No Transfer Learning	PCA + SVM	**0.68**
	PCA + Random Forests	0.53
	PCA + Gradient Bossting	0.64

Conclusion

- Random Over Sampling as well as the use of the focal loss allows to surpass the class imbalance problem.
- Gene-Expression images as defined in [1] allows the use of CNNs with limitations inherited from the unstructured nature of gene-expression data.
- Transfer Learning, when possible, allows to prevent over-fitting when dealing with limited data-sets (1000 non lung cancer samples only).
- Variational Autoencoders, allowed us to surpass the curse of dimensionality (reduce the dimension from 7509 to 300 and achieve a good AUC score of 0.73).

References

Franco L Veredas FJ López-García G, Jerez JM.

Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data, 2020.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár.

Focal loss for dense object detection.

CoRR, abs/1708.02002, 2017.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei.

Deformable convolutional networks, 2017.