Deep Learning for Metagenomic Data: using 2D Embeddings and Convolutional Neural Networks

Nguyen Thanh Hai
UPMC- Paris 6 University
nthai@cit.ctu.edu.vn

Yann Chevaleyre
Dauphine, PSL Research University, LAMSADE
yann.chevaleyre@dauphine.fr

Edi Prifti
Integromics, ICAN, Paris
e.prifti@ican-institute.org

Nataliya Sokolovska
UPMC- Paris 6 University
nataliya.sokolovska@upmc.fr

Jean-Daniel Zucker
IRD, UMMISCO, France
jean-daniel.zucker@ird.fr

Abstract

Deep learning (DL) techniques have had unprecedented success when applied to images, waveforms, and texts to cite a few. In general, when the sample size (N) is much greater than the number of features (d), DL outperforms previous machine learning (ML) techniques, often through the use of convolution neural networks (CNNs). However, in many bioinformatics ML tasks, we encounter the opposite situation where d is greater than N. In these situations, applying DL techniques (such as feed-forward networks) would lead to severe overfitting. Thus, sparse ML techniques (such as LASSO e.g.) usually yield the best results on these tasks. In this paper, we show how to apply CNNs on data which do not have originally an image structure (in particular on metagenomic data). Our first contribution is to show how to map metagenomic data in a meaningful way to 1D or 2D images. Based on this representation, we then apply a CNN, with the aim of predicting various diseases. The proposed approach is applied on six different datasets including in total over 1000 samples from various diseases. This approach could be a promising one for prediction tasks in the bioinformatics field.

1 Introduction

High throughput data acquisition in the biomedical field has revolutionized research and applications in medicine and biotechnology. Also known as omics data, they reflect different aspects of system's biology (genomics, transcriptomics, metabolomics and proteomics but also whole biological ecosystems acquired with the use of metagenomics). These datasets are increasingly available and numerous models use this information to make medical decisions [8] (personalizing health care [25], diagnosis and prognosis [6], pharmacogenomics [26], etc.). However, exploring omics data has met many challenges (large number of features - genes/species d, and few observations N). Up to now, the most successful techniques applied to omics datasets have been mainly Random Forests (RF), and sparse regression. In this paper, we applied DL directly on six metagenomic datasets which reflect bacterial species abundance and presence in the gut of diseased patients and healthy controls. Since this technology performs particularly well in image classification, we focus here in the use of CNNs. In this context, it is important to find representations of the data that would be biologically pertinent to apply CNN techniques in order to learn new representations that would be used for classification

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

purposes. Our objectives are to propose efficient representations which are compact in images, and to prove DL techniques as efficient tools for prediction tasks in the context of metagenomics.

There are numerous studies where authors have applied ML for analyzing large metagenomic datasets. Pasolli1 et al. [6] proposed a unified methodology to compare different state of the art methods (SVM, RF, etc) in various metagenomic datasets, which were processed with the same bioinformatics pipeline for comparative purposes. Authors in [4] introduced an overview of technologies in DL, which emerged and grew rapidly in recent years and have been applied to health informatics. Additionally, Zhao Y et al. in [27] leveraged ML methods to investigate prediction tasks for multiple sclerosis disease outcomes. Some studies applied ML techniques to predict dementia [18] and cancer [13]. Yoshua B. stated [2] that the performance of prediction tasks depends on selecting representations, numerous studies have attempted to propose efficient representations. Montavon et al [17] presented invariant representations of molecules to perform prediction tasks on atomization energy. The authors in [7] introduced Ph-CNN using CNNs applying to metagenomic data embed the phylogentic tree.

Image classification based on ML has achieved impressive results and has been performing better than human experts in numerous cases. Yann Lecun et al. [16] proposed LeNet-5 as one of the first standard architecture for CNNs, whereas AlexNet as one of the most remarkable CNN architectures, showed significant improvements upon previous architectures. Since AlexNet, numerous architectures were introduced to improve the performance. Some of the most famous architectures among them are ZFNet [28], GoogLeNet [23], VGGNet [22], ResNet [10], and WELDON [5]. A broad survey of CNNs was presented in [9] as proven that CNNs have gained a central position in image classification.

2 Methodology

We evaluated our method with each of the different representations in the six datasets, which include bacterial species related to various disease, namely: liver cirrhosis (CIR), colorectal cancer (COL), obesity (OBE), inflammatory bowel diseases (IBD) and Type 2 Diabetes (T2D) [6, 21, 29, 15, 20, 19, 11], with CIR (n=232 samples with 118 patients), COL (n=48 patients and n=73 healthy individuals), OBE (n=89 non-obese and n=164 obese individuals), IBD (n=110 samples of which 25 were affected by the disease) and T2D (n=344 individuals of which n=170 are T2D patients). In addition, WT2 (n=96 European women with n=53 T2D patients and n=43 healthy individuals). The abundance datasets (ABD) are transformed to obtain another representation based on feature presence (PRE) when the abundance is greater than zero. These data were obtained using the default parameters of MetaPhlAn2 [24] as detailed in Pasolli1 et al. [6]. For each sample, species abundance is represented as a real number and the total abundance of all species sums to 1. Our approach consists of the

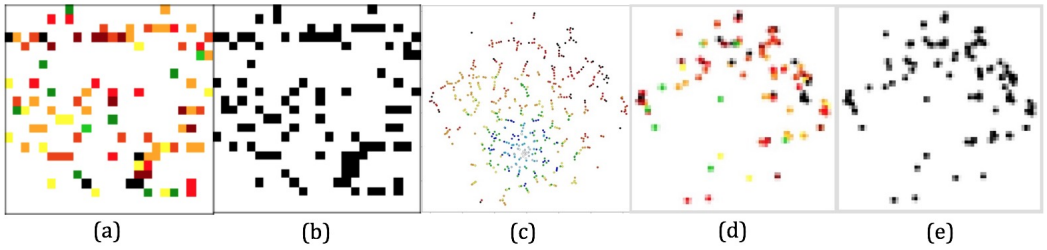

Figure 1: Visualization of image-based representations. (a) square filling up from left to right/top to bottom with species abundances and a phylogenetic ordering; (b) same as a) but with presence instead of abundance; (c) a global t-SNE map was generated from a training dataset (Cirrhosis dataset), color represents mean abundance; (d) t-SNE image of a particular sample based on the same map, color represents abundance levels (e) same as d) but using presence instead of abundance.

following steps: First, a set of colors is chosen and applied to different bins based on the abundance distribution. The binning can be performed on a linear or logarithmic scale. Here we present results from the latter. Then, the features are visualized into images by one of two different ways (**phylogenetic-sorting** (PLG) or visualized based on t-Distributed Stochastic Neighbor Embedding (**t-SNE**) [14]). t-SNE technique is so useful to find faithful representations for high-dimensional points visualized in a compact space, typically the 2D plane. For phylogenetic-sorting, the features which are bacterial species are arranged based on their taxonomical annotation ordered increasingly

by the concatenated strings of their taxonomy (i.e. phylum, class, order, family, genus and species). This ordering of the variables integrates within the data an external biological knowledge, which reflects the evolutionary relationship between these bacterial species. Each visualization method will be used to either represent abundance or presence data. A fifth representation, which serves as control is the 1D of the raw data (with the species also sorted phylogenetically). For the t-SNE approach, we use only training sets to generate global t-SNE maps, images of training and test set are created from these global maps. The representations are evaluated in a framework with 100 CNN architectures and a network with a fully connected layer without convolutions. We used accuracy (ACC) to measure model performances. Results represent average accuracy values through a 10-fold cross-validation. All the above mentioned architectures are implemented in Torch7 [3].

2.1 Approaches to generate images: Fill up and t-SNE

"Fill up" images are created by arranging abundance/presence values into a matrix in a left-to-right order by row. The image is square and empty bottom-left of the image are set to zero (white). As an example for a dataset containing 542 features (i.e. bacterial species) in the cirrhosis dataset, we need a matrix of 24x24 to fill up 542 values of species into this square. The first row of pixel is arranged from the 1st species to the 24th species, the second row includes from the 25th to the 48th and so on till the end. We use distinct colors in a logarithmic binning scale to illustrate abundance values of species and black and white for presence/absence, where white represents absent values.

T-SNE maps: are built based on training sets from raw data with perplexity = 10 after 500 epochs. These maps are used to generate images for training and testing sets. Each species is considered as a point in the map and only species that are present are showed either in abundance or presence using the same color schemes as above. Figure 1 illustrates images generated by "fill-up" and t-SNE.

2.2 Architecture configurations of CNNs

Our proposed architectures are mainly inspired by the philosophy of VGG nets [22], including 3x3 filters with stride 1, and max pooling of 2x2 with stride 2, using ReLU after each convolution. In the experiments, image dimensions range from 32x32 for fill up and 64x64 for t-SNE. These images are passed through a stack of convolutional layers (the depth ranging from one to five convolutional layers, and the width of each convolutional layer increasing from one, up to twenty filters), followed by one max pooling, and one Fully-Connected layer (**FC**). For the first convolutional layer, wide

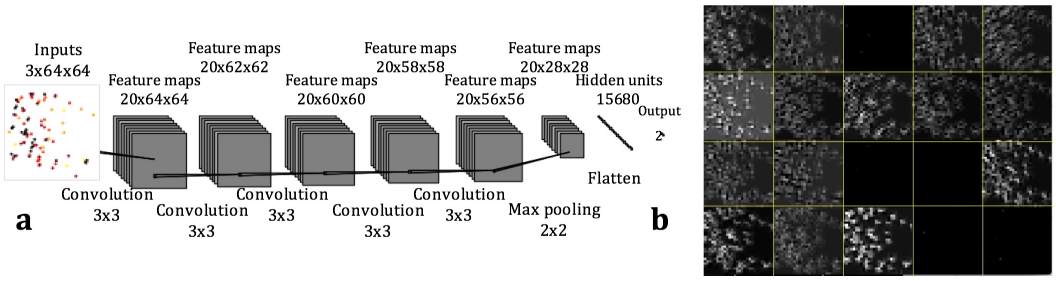

Figure 2: An illustration of the CNN architecture for representations based on synthetic images. **(a)** The architecture receives a three-channel color image as input passing a stack of five convolutional layers (with 20 kernels of 3x3 (stride 1) for each), followed by a max pooling 2x2 (stride 2). The last layer is fully connected, taking feature maps from a stack of convolutional and max pooling layers as input in a vector form, and computes the scores for the output by LogSoftMax function. **(b)** Visualization of 20 feature maps generated after a stack of convolutional layers and max pooling of a fully trained model (the architecture (a)) with the input image at (a).

.

convolution [1] is implemented to preserve the dimension of input. Besides 2D convolutions for images, we also apply a 1D convolution to raw data with kernel 3x1 (step 1), max pooling with size 2, stride 2. There are two approaches for binary classification including using two output neurons (two-node technique) or one output neuron (one-node technique). For the two-node approach, the final layer is the LogSoftMax layer with 2 outputs, one indicating patient was affected by the disease,

Table 1: Prediction performance (in ACC) of representations after 200 epochs. The results that are better the state of the art ML approaches are formatted in **bold**. PLG, t-SNE are ways to arrange species. ABD, PRE are two types of features (ABD: species abundance, and PRE: marker presence). (1) are results of the neural network with a fully connected layer without convolutions (FC), (2) indicates performance of CNN architectures (Con1D: 1D convolutions, Con2D: 2D convolutions). MetAML [6] is a computational tool to perform prediction tasks on metagenomic data using machine learning classifiers such as support vector machines (SVM), RF, Lasso, and Elastic Net. We selected the best classifiers of MetAML including RF, and SVM for comparison.

Datasets	PLG Raw	PLG ABD	PLG PRE	t-SNE ABD	t-SNE PRE	MetAML	
	Con1D	Con2D	Con2D	Con2D	Con2D	RF	SVM
CIR(1)	0.852	**0.878**	0.865	0.743	0.817	0.877	0.834
CIR(2)	0.870	**0.891**	0.874	0.865	0.852		
COL(1)	0.675	0.767	0.775	0.650	0.608	0.805	0.743
COL(2)	0.717	0.742	0.725	0.783	0.708		
IBD(1)	**0.818**	**0.855**	**0.827**	0.682	0.755	0.809	0.809
IBD(2)	0.763	**0.836**	0.764	0.809	**0.818**		
OBE(1)	0.588	0.636	**0.656**	0.628	0.596	0.644	0.636
OBE(2)	0.616	**0.660**	0.592	0.628	0.628		
T2D(1)	0.635	0.662	0.638	0.565	0.588	0.664	0.613
T2D(2)	0.639	0.626	0.656	0.635	0.647		
WT2(1)	0.656	0.678	**0.733**	0.611	0.622	0.703	0.596
WT2(2)	0.556	0.589	0.633	**0.711**	**0.711**		

the other shows the patient was unaffected (see an example in Figure 2). For the one-node approach, we used a sigmoid activation function at the final layer and binary cross-entropy for Sigmoid. In our experiments, the two-node approach with LogSoftMax gave better results, so the architectures in Table 1 used this method. The networks are trained using stochastic gradient descent for the two-node approach or Adam [12] for the one-node approach with a mini-batch size of 16, momentum of 0.1, and weight decay of 10^{-5}. Constant learning rate is set at 0.0005.

3 Results

Table 1 illustrates the results of the CNN architecture with 5 convolutional (consisting of 2D convolutions) layers and 20 filters per layer to images-based representations and the convolutional architecture including 2 convolutional layers (1D convolutions) and 20 filters per layer for the raw data. IBD dataset illustrated the greater improvement with most representations performing better than state of the art results. Images-based representations are more efficient than raw data, while 2D convolutions also surpasses 1D convolutions. For the t-SNE representation, the CNN architecture outperforms FC.

4 Conclusion

We proposed the MET2IMG approach to predict patients' diseases using metagenomic data. Our method is very promising in the context of prediction using metagenomic data. We used two main approaches to build "synthetic images" including the "fill-up" and the t-SNE embedding. For the "fill-up" approach, we are able to use smaller and simple images. For the t-SNE representations, features are embedded in a two-dimensional space using a classic embedding approach in ML. The dimension of t-SNE images is required to be higher to get a higher performance because some data points might be overlapped with a small scaled dimension, while each feature in the "fill-up" approach is visible. Hence, learning with t-SNE images may be more complicated as well as require more time and more memory to process compared to "fill-up". We only evaluated small images with dimensions of 32x32 and 64x64 due to limitations of computational resources, so the deeper architectures of CNNs should be investigated to evaluate on larger-scaled images. Besides, further research should compare the performance of image-based representations against tree-based representations.

Bibliography

[1] *A convolutional neural network for modelling sentences*, 2014. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL http://dx.doi.org/10.1109/TPAMI.2013.50.

[3] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[4] C. W. Daniele RavA and F. Deligianni. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21, 1 2017.

[5] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4743–4752. URL http://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Durand_WELDON_Weakly_Supervised_CVPR_2016_paper.html.

[6] D. T. T. Edoardo Pasolli, F. Malik, L. Waldron, and N. Segata. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology, Edited by Jonathan A. Eisen*, 12, 7 2016.

[7] D. Fioravanti, Y. Giarratano, V. Maggio, C. Agostinelli, M. Chierici, G. Jurman, and C. Furlanello. Phylogenetic convolutional neural networks in metagenomics. URL http://arxiv.org/abs/1709.02268.

[8] G. GS and W. HF. Genomic and personalized medicine: foundations and applications. *Translational Research*, 154:277–287, 2009.

[9] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang. Recent advances in convolutional neural networks. *CoRR*, abs/1512.07108, 2015. URL http://arxiv.org/abs/1512.07108.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[11] T. V. Karlsson FH, N. I, B. G, B. CJ, and e. a. Fagerberg B. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 2013.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

[13] E. T. Kourou K., E. KP., K. MV., and F. DI. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[14] G. H. Laurens van der Maaten. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 8 2008.

[15] N. T. Le Chatelier E, Q. J, P. E, H. F, and e. a. Falony G. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 2013.

[16] Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551, 1989. ISSN 0899-7667.

[17] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, and K.-R. Muller. Learning invariant representations of molecules for atomization energy prediction. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 440–448, USA, 2012. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999134.2999184.

[18] A. L. Moraes, S. Eivazzadeh, E. Mendes, J. Berglund, and P. Anderberg. Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. 12: e0179804, 06 2017.

[19] L. Y. Qin J, C. Z, L. S, Z. J, and e. a. Zhang F. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 2012.

[20] R. J. Qin J, Li R, A. M, B. KS, and e. a. Manichanh C. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 2010.

[21] L. A. Qin N, Yang F, P. E, C. Y, and e. a. Shao L. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 2014.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL `http://arxiv.org/abs/1409.1556`.

[23] C. Szegedy, W. Liu, Y. J. andVGGNet Pierre Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL `http://arxiv.org/abs/1409.4842`.

[24] D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. 12(10):902–903. ISSN 1548-7091. doi: 10.1038/nmeth.3589. URL `http://www.nature.com/nmeth/journal/v12/n10/full/nmeth.3589.html`.

[25] H. W. Virgin and J. A. Todd. Metagenomics and personalized medicine. *Cell*, 147, 2011.

[26] M. H. Voora D., E. C., and G. BF. The pharmacogenetics of coumarin therapy. *Pharmacogenomics*, 205.

[27] B. C. H. Yijun Zhao, D. Rotstein, C. R. G. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, and T. Chitnis. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLOS Computational Biology, Edited by Jonathan A. Eisen*, 4 2017.

[28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL `http://arxiv.org/abs/1311.2901`.

[29] T. J. Zeller G, V. AY, S. S, K. JR, and e. a. Costea PI. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Nature*, 2014.