ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Multi-omics data integration using cross-modal neural networks

Ioana Bica, Petar Veličković, Hui Xiao and Pietro Liò

Department of Computer Science and Technology, University of Cambridge
Cambridge CB3 0FD - United Kingdom

Abstract. Successful integration of multi-omics data for prediction tasks can bring significant advantages to precision medicine and to understanding molecular systems. This paper introduces a novel neural network architecture for exploring and integrating modalities in omics datasets, especially in scenarios with a limited number of training examples available. The proposed cross-modal neural network achieves up to 99% accuracy on omics datasets. Moreover, we show how analysis of the weights and activations in the network can give us biological insights into understanding which genes are most relevant for the decision process and how different types of omics influence each other.

1 Introduction

Multi-omics data integration is paramount for medical research and subsequent diagnoses since the preliminary signs of a disease are noticeable first in the omics. [1]. Moreover, in developmental biology, multi-omic changes are critical factors in the way stem cells differentiate and gain a certain functionality [2].

The Cancer Genome Atlas (TCGA) [3] consists of a comprehensive collection of multi-omics data. However, the move from heterogeneous data produced by genome projects to final diagnosis in medical research, requires a powerful machine learning tool capable of approximating the underlying correlated relationships between different types of omics data through supervised learning. Neural networks achieve state-of-the-art performance on many classification tasks [4] and they are also well-suited for analysing complex multi-omics data [5]. Nevertheless, due to the high costs involved in obtaining omics data, finding large datasets in order to reliably train neural network models remains a challenge.

Therefore, we propose a cross-modal superlayered neural network architecture (SNN) that is capable of extracting cross-correlations present in multimodal datasets, thus achieving good performance, particularly on datasets with a limited number of training examples. A similar approach has been used by Veličković *et al.* to improve the performance of neural networks on small image datasets [6]. Modalities are inherent to multi-omics datasets which contain information from different data sources including epigenome, transcriptome and proteome. When this is not the case, unsupervised learning can be employed to identify implicit modalities present in the omics data, such as groups of genes that are co-expressing under a specific condition.

The SNN achieves up to 99% accuracy on the omics datasets analysed. In addition, through t-SNE visualizations, we show how this cross-modal method could provide a way to study how different omics influence each other.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

2 Methods

2.1 Datasets and preprocessing

The datasets were chosen to illustrate that the superlayered architecture achieves outstanding results on both binary and multi-class classification. Moreover, since for prognosis or diagnosis in personal medicine there is usually a limited amount of data available, the datasets used in this paper are representative of this fact.

Therefore, we will use a dataset from TCGA that involves binary classification of patients for breast cancer [1]. We focused on the activity of genes in the tumour necrosis factor receptor superfamily (TNFRS) which has been proven to play important roles during tumorigenesis and could be the potential targets for cancer therapy [7]. The dataset consists of 528 positive and 62 negative training examples, where each example has 26 gene expression measurements (transcriptomics) and 26 DNA methylation measurements (epigenomics).

The second dataset used includes transcriptomes of 90 human preimplantation embryos corresponding to seven embryonic developmental stages [8]. A training example consists of more than 20,000 gene expression levels, which requires an initial pre-processing step to select up to 200 genes that have the highest entropy across different classes. Then, k-means clustering was used to find groups of genes that are co-expressed across the seven developmental stages and therefore to explore the potential mechanisms underlying the transcriptome data. The two largest gene clusters are used as input modalities to the SNN.

2.2 Cross-modal neural network

The SNN consists of two superlayers, each of which receives a data modality. A superlayer represents a feedforward neural network that learns to analyse its input modality. The separation into data modalities through domain-specific knowledge (gene expression and DNA methylation) or unsupervised learning (clustering) leverages the width of the data, which is essential for overcoming the problem of having sparse datasets for training.

Cross-connections are added between the superlayers in order to allow the information to flow freely between the different modalities, but also to explore the interactions between them. The features learnt by the superlayers are eventually concatenated, and passed into two fully connected (FC) layers before the output layer computes the network's predictions. Figure 1 illustrates the SNN architecture and the data flow through this network.

The positioning of cross-connections required careful consideration due to its influence on feature extraction. If we do not use cross-connections and just concatenate the superlayers at the end, then the correlations between the modalities would not be exploited to their full potential. Conversely, if the cross-connections are incorporated too early, the different modalities would have excessive influence on each other.

Therefore, in order to strike a balance, we decided to add the cross-connections in the middle of the superlayers. This way, the data is allowed to flow freely

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

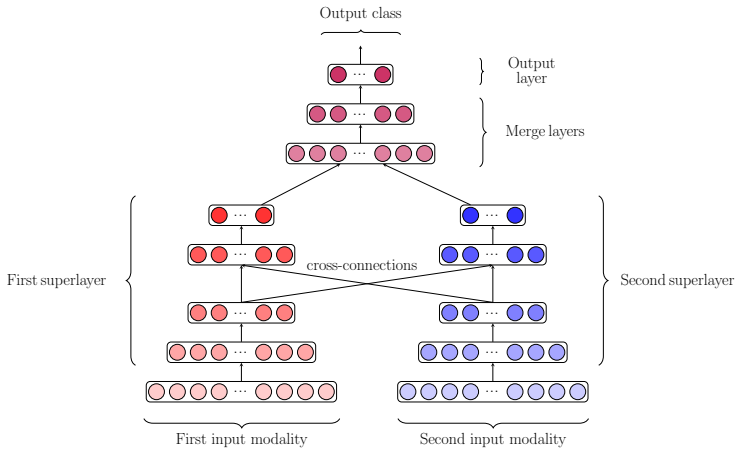

Fig. 1: Data flow through the SNN. Each superlayer consists of four FC layers where each neuron is connected to all of the neurons in the layer indicated by the arrow. The added cross-connections are also FC layers. The neurons in the later layers learn to extract more and more complex features, concept illustrated by the gradient in colour.

between the superlayers, but also each individual modality is capable of playing a significant role in the prediction.

2.3 Comparison with baseline models

The SNN is compared against a multilayer perceptron (MLP) and a recurrent neural network (RNN), which both concatenate the modalities in the input data. The MLP consists of several FC layers extracting features from the input data. On the other hand, the RNN uses Long Short-Term Memory (LSTM) units in order to learn long-term dependencies from the input sequence [9]. The architecture used for each model is described in Table 1.

The neurons in the FC layers use the ReLU activation function [10]. He initialization is used for the weights, while the biases are initialised to zeroes. Dropout [11], batch normalization [12] and weight decay are used for regularization. All of the neural network models were trained for 100 epochs using Adam SGD optimizer [13] with batch size of 64 and hyperparameters selected through cross-validation.

The weights in the LSTM units use Xavier initialization, while the forget gate biases are initialised to a vector of ones to establish gradient flow and to encourage long-term dependencies at the onset of training [14]; other biases in the LSTM unit are initialised to zero. Dropout is used as a form of regularization on the hidden-to-hidden connections in the LSTM layer.

In order to show that the SNN is suitable for performing inference on multi-omics datasets, we used as additional baselines Support Vector Machines (SVMs) with RBF kernel, Random Forests (RFs) and k-Nearest Neighbours (kNNs).

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

MLP	**RNN**	**SNN**	
~ 57000 params	~ 27000 params	~ 22000 params	
FC 256-D	LSTM 128 features	FC 128-D	FC 128-D
FC 128-D	LSTM 32 features	FC 64-D ↘	↙ FC 64-D
FC 64-D	FC 64-D	FC 32-D	FC 32-D
FC 32-D	FC 32-D	FC 16-D	FC 16-D
		FC 64-D	
		FC 16-D	

FC number of output classes-D

Table 1: Architectures for the neural network models compared. ↘ and ↙ denote the cross-connections between the superlayers. The dimension (D) of a FC layer is given by the number of neurons it contains.

3 Results

Stratified nested k-fold cross-validation was used in order to determine the best hyperparameters for the models (inner cross-validation) and to evaluate their performance (outer cross-validation).

Due to the class imbalance in the multi-omics datasets, the accuracy alone is insufficient in assessing the performance of the models. Additional evaluation metrics were chosen to obtain a better indication of the discriminative ability of the models. The evaluation metrics were extended to multiple classes by using micro-averaging and macro-averaging. Micro-averaging assigns equal weights to each test example, while macro-averaging gives the same weight to each class. [15]

Metric	**kNN**	**RF**	**SVM**	**MLP**	**RNN**	**SNN**
Accuracy	0.97 ± 0.02	0.96 ± 0.02	0.98 ± 0.01	0.93 ± 0.03	0.95 ± 0.02	$\mathbf{0.99} \pm 0.01$
Precision	0.98 ± 0.01	0.96 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	0.96 ± 0.02	$\mathbf{0.99} \pm 0.01$
Sensitivity	0.97 ± 0.01	0.97 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.98 ± 0.01
F1 - score	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.96 ± 0.02	0.97 ± 0.01	$\mathbf{0.99} \pm 0.01$
MCC	0.85 ± 0.04	0.79 ± 0.05	0.88 ± 0.02	0.61 ± 0.09	0.72 ± 0.04	$\mathbf{0.91} \pm 0.02$
ROC AUC	0.97 ± 0.02	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	0.93 ± 0.03	$\mathbf{0.99} \pm 0.01$

Table 2: Mean results obtained for each model after 10-fold outer cross-validation for the breast cancer patients dataset and the standard error in the results.

Metric	**kNN**	**RF**	**SVM**	**MLP**	**RNN**	**SNN**
Micro F1-score	0.90 ± 0.05	0.91 ± 0.02	0.96 ± 0.01	0.93 ± 0.02	0.90 ± 0.03	$\mathbf{0.98} \pm 0.01$
Micro MCC	0.88 ± 0.06	0.89 ± 0.03	0.95 ± 0.02	0.92 ± 0.03	0.87 ± 0.01	$\mathbf{0.97} \pm 0.01$
Macro F1-score	0.85 ± 0.04	0.91 ± 0.01	0.96 ± 0.01	0.85 ± 0.02	0.90 ± 0.02	$\mathbf{0.96} \pm 0.01$
Macro MCC	0.86 ± 0.04	0.84 ± 0.02	0.92 ± 0.01	0.85 ± 0.03	0.86 ± 0.02	$\mathbf{0.95} \pm 0.01$
Accuracy	0.90 ± 0.03	0.91 ± 0.02	0.96 ± 0.01	0.93 ± 0.01	0.89 ± 0.02	$\mathbf{0.98} \pm 0.01$

Table 3: Mean results obtained for each model after 6-fold outer cross-validation for embryo development dataset and their standard error in the results.

The SNN most explicitly exploits the interactions between the different modalities present in the dataset, thus obtaining higher metric averages, as it can be

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

noticed in Table 2 and Table 3.

After studying the weights of the input modalities in the SNN, for the cancer patients dataset, we have identified that the TNFRSF13C, TNFRSF13B, TNFRSF14 genes have the highest impact in the gene expression modality, while TNFRSF13C, TNFRSF13B and NGFR influence the most the DNA methylation modality. Moreover, gene ontology analysis of the genes involved in the cancer patients dataset has shown that the following pathways are enriched: Cytokine-cytokine receptor interaction, Apoptosis and NF-kappa B signalling pathway. These results illustrate how the SNN can be integrated into workflows with other clinical bioinformatics services to support medical decisions.

4 Analysis of cross-connections

In order to understand whether the cross-connections aid with the classification tasks, we have utilised a t-SNE visualization of the neurons' activations in specific layers of the SNN during testing on the cancer patients dataset. Figure 2 indicates that adding the cross-connections between the superlayers—thus allowing for the different modalities to interact during independent feature extraction—significantly helps in discriminating the patients with cancer from healthy ones.

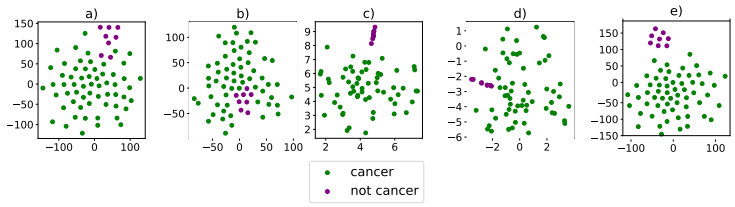

Fig. 2: t-SNE plots for the activations of SNN layers on the test set for the breast cancer dataset. a) and b) illustrate the t-SNE embeddings for the first and second superlayer, just before the cross-connections. c) and d) illustrate the respective embeddings just after the cross-connections e) represents the activations after merge layer.

These results show that integrating transcriptomics and epigenomics data is beneficial for medical diagnosis. Therefore, by analysing the cross-connections, biologists could determine the mutual influence of various omics measurements, particularly as these can vary for different regulatory circuits or genome regions.

5 Conclusion

This paper introduces a cross-modal neural network architecture capable of integrating modalities in multi-omics data thus achieving good performance in situations where only small training datasets are available. The model achieves up to 99% accuracy on classification problems with both multi-omic data (transcriptomics and epigenomics) and single-omic data (where k-means clustering was used to identify modalities).

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

The SNN model can be easily scaled up: more superlayers can be added, thus allowing more types of omic data to be involved, such as the genome copy number variation and the somatic mutations. Moreover, the SNN could be applied on regression problems such as survival time analysis of cancer patients or on predicting the expression levels of genes from their corresponding epigenetic modifications (e.g. different histone modification markers) based on data from the Roadmap Epigenomics Project.

Therefore, the proposed cross-modal neural architecture represents a powerful tool that can be used for integrating and understanding the role of multi-omics data in biomedical decisions.

References

[1] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[2] T. Nafee, W. Farrell, W. Carroll, A. Fryer, and K. Ismail, "Review article: Epigenetic control of fetal gene expression," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 115, no. 2, pp. 158–168, 2008.

[3] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[5] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, no. 12, 2016.

[6] P. Veličković, D. Wang, N. D. Lane, and P. Liò, "X-cnn: Cross-modal convolutional neural networks for sparse datasets," *IEEE Symposium Series on Computational Intelligence*, 2016.

[7] D. A. Schaer, D. Hirschhorn-Cymerman, and J. D. Wolchok, "Targeting tumor-necrosis factor receptor pathways for tumor immunotherapy," *Journal for immunotherapy of cancer*, vol. 2, no. 1, p. 7, 2014.

[8] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, *et al.*, "Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural & Molecular Biology*, vol. 20, no. 9, pp. 1131–1139, 2013.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, 1997.

[10] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, 2014.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *International Conference on Machine Learning*, 2015.

[15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, pp. 427–437, 2009.