Convolutional Neural Network Architectures for CAFA4

Jari Björne
Department of Future Technologies, University of Turku
20014 University of Turku, Finland
firstname.lastname@utu.fi

1 Introduction

The Critical Assessment of protein Function Annotation (CAFA) challenge concerns the prediction of annotated ontology terms for given protein sequences [Radivojac et al., 2013]. In the current CAFA 4 challenge, the goal is to predict terms from the Gene Ontology (GO), the Human Phenotype Ontology (HPO) and the Disorder Ontology (DO) ontology for a set of 100k sequences provided as prediction targets.

In this work, different convolutional neural network architectures were examined for the task of protein function prediction. UniProt protein sequence data was enriched with additional information from protein taxonomical hierarchies and InterProScan sequence analyses before being processed with the networks, predicting the different ontologies' terms in a multi-label classification task.

2 Data Preprocessing

The same data preprocessing approach was applied with all ontologies. The used protein sequences consisted of the 560k union of the UniProt XML release and the set of sequences provided as CAFA4 prediction targets [UniProt Consortium, 2019]. GO annotations and protein lineage information were imported from the UniProt release. HPO and DO annotations were extracted from their respective databases and added to the sequence information.

All annotated terms were propagated using the GOA Tools package, so that all of their ancestor terms were included as individual labels [Klopfenstein et al., 2018]. InterProScan sequence analyses were imported from the dataset release and added to the protein sequence information [Jones et al., 2014].

3 Neural Architectures

All neural network methods were implemented with the Keras package and were based on a one-dimensional, convolutional approach [Chollet et al., 2015, LeCun et al., 1995]. Both parallel and nested convolutional architectures were tested. In the parallel convolution approach, the outputs of the convolutional layers were concatenated together, whereas in the nested approach each convolutional layer produced the output of the next layer. In addition, the suitability of a one-dimensional variant of the MobileNet V2 image recognition network was tested, as well as a network based on the DeepGOPlus architecture [Kulmanov and Hoehndorf, 2020, Sandler et al., 2018].

All networks received as inputs sequences of a maximum of 1000 elements, where each element corresponded to a single amino acid. Each element contained an 8-dimensional embedding of an amino acid, as well as optionally 8-dimensional embeddings for InteProScan *domain*, *family* and *homologous superfamily* features, with a maximum of 1000 most common unique features included in each category. These sequences were processed by the convolutional layers and merged together with 4-dimensional embeddings representing five levels of the protein's organism's taxonomy lineage, before being finally processed with a dense layer followed by a multi-label prediction layer, which generated individual predictions for up to 1000 of the most common annotated terms.

A dropout of 0.1 was used after the inputs and the convolutional layers to improve regularization. Several optimizers were tested, with the Adam and SGD ones achieving the most promising results.

The models were trained with a learning rate of 1×10^{-5} and with early stopping, for 10–100 epochs depending on the computational complexity of the model, the size of the dataset and the available resources.

4 Experimental Setup

The entire dataset was randomly divided into six approximately equal subsets so that all homologs of a protein were in the same subset. Training was done by cross-validation, so that four subsets at a time were used for training, one for parameter optimization and one for prediction. The final predictions were acquired by joining the six predicted subsets together, thus covering the whole dataset. Finally, the CAFA4 targets were separated from the whole set of predicted sequences, and converted to the submission format.

Experiments were performed with using either all known annotations as training labels, or using only those annotations which were not marked as IEA (inferred from electronic annotation). Experiments were also performed using either all available protein sequences, or only those sequences which had at least one annotated term, with the aim of reducing the presence of yet unknown terms as false negatives in the training data.

5 Results

System performance was evaluated using micro- and macro-averaged F1 and AUC metrics, evaluated for the multilabel predictions as well as for each predicted term separately. Submitted models were chosen based on these metrics.

References

F. Chollet et al. Keras. `https://keras.io`, 2015.

P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9): 1236–1240, 2014.

D. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. W. Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, et al. Goatools: A python library for gene ontology analyses. *Scientific reports*, 8(1):1–17, 2018.

M. Kulmanov and R. Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019.