BRAC UNIVERSITY

Inspiring Excellence

CSE-474

Assignment 3

Literature Review

Fall 2022

Samiha Afaf Neha
ID: 20101266

Research Title: Protein Sequence Classification using 1D-CNN

Advanced technologies have developed tools to extract the raw protein sequence, which has given researchers the opportunity to analyze these sequences in order to discern the corresponding proteins' functions and biological families. The cellular mechanism of a living thing depends heavily on proteins. When performing any proteomic research on a specific protein, it is crucial to understand the function of the protein. However, as of 2013, more than 40% of the protein sequences in the NCBI database had no known function[1]. This may be due to the fact that it is costly, time-consuming, and difficult to ascertain a protein's function utilizing functional assays. This has brought about the need for computational methods to analyze the raw sequences in order to decipher the hidden amino acid features or patterns. Taken from **NLP** concepts, protein sequences are considered to be a language of their own, where the amino acids are the words and the motifs are phrases. Earlier research in this domain has used clustering algorithms and other machine learning algorithms such as the **Random Forest, SVM, and Naive Bayes** for the classification task, but the accuracy or speed of computation had not been satisfactory[2]. Unlike genomic medicine and the image research field, less work has been done on the sequence analysis problem. The main issue with the sequencing problem is feature extraction, as there is no ready to use feature in a raw protein sequence. Deep learning approaches have been applied and met with state-of-the art results where the models learn features by themselves in each epoch of the training process. Analysis among deep learning models such as the **RNN**, **LSTM**, and **1D-CNN** further proved the superiority of the 1D CNN for the specific task[3]. The CNN model can be parallelized, unlike the sequential RNN and LSTM. Stacking convolutional layers creates a more receptive field for classification, and the backpropagation computation of the model only depends on the depth of the model instead of the sequence itself, which is not the case for RNNs and LSTMs.

The article "Protein classification from scratch using 1D-CNN" deals with the classification task by employing a customized 1D CNN model that runs on sequences of length 50 to 1200 amino acids taken from the UniProt dataset. It used encoding, embedding, CNN, and a fully connected module. Their model has reached an accuracy of 97%, which is significantly greater than other machine learning models. The model incorporates max pooling, dropout, and activation functions such as ReLu and SoftMax for the task. Along with the training, hyperparameters such as batch size, number of epochs, dropout, etc. were tuned while checking accuracy metrics such as F1, recall, and precision scores.

One similar sequencing problem was also researched in the paper **"Deep Learning Architectures for DNA Sequence Classification,"** where DNA sequences were up for classification from the 16S data-set. The model has the similar encoding, embedding steps and uses tanh, softmax activation functions along with cross-validation for the training. It compares the LSTM and CNN for a total of 5 bacteria classification tasks, where CNN outperforms the LSTM 4 out of 5 times [9]. It also concludes that multi-tasking approach on a model works better for LSTM but worsens CNN performance A review of the deep learning approach for protein classification stipulates that CNN algorithms are capable of extracting distinctive geographical information[4] from the sequence data while reducing data size by amplification and representing local features in the final output. Modern sequence matching tools such as **BLAST** and **FASTA** use heuristic algorithms for sequence matching against already identified protein sequences and give results based on sequence or subsequence similarity scores. Studies have shown that sequence matching is not the most accurate matrix for prediction, as highly identical protein sequences are not always expected to have the same gene ontology. Comparison between RNN, Multimodal architectures and CNN shows that 45% of the time CNN alone is used for such sequence classification tasks. Although protein sequences might have potential attention components just like a text, not much research has been done on the transformer[5] (e.g., the use of ProtBERT) models for sequence classification so far.

Some non-biological sequences related studies have also been conducted on time series data, which can be expanded to include the protein sequences as well. Time series data has a natural temporal ordering, as does the amino acid sequence of a protein[6]. Thus, the protein sequence can also be treated as a TSC (time series classification) problem. In the paper **"Classification of Time-Series Images Using Deep Convolutional Neural Networks,"** a unified network of CNN and Recurrence Plot (RP) is used to boost the general time series classification. The data for this task is used from the UCR time-series classification archive, which is first converted into a 2D image based on the recuurecce plot of the values, and then a 2D CNN model is applied to it, which then automatically learns the data map features and classification simultaneously in a supervised method. Like the majority of the sequence classification CNN models, this architecture also used the "categorical cross entropy" as the loss function. Other hyperparameters, like batch size and epochs, were tuned using the SGD algorithm. The model was trained using TensorFlow.

Another research work[7] deals with a similar kind of sequential data analysis using 1D-CNN with the title **"A few filters are enough: Convolutional Neural Network for P300 Detection."** This paper deals with EEG signals from brains to detect the p300 component, which is an event related potential used in **Brain - Computer Interfacing (BCI)**. The main objective of the paper was to reduce the parameters of the previously employed models as the detection rate comparatively was not impressive despite the huge parameters needed . They eventually devised a novel model called the Sep1D along with a Fully-connected neural network to beat the state-of-the-art models (DeepConvNet, ShallowConvNet, CNN1, CNNR) for p300 detection but with fewer parameters for both within subject and cross subject training. The network architecture incorporates a separable depthwise 1D convolutional layer followed by a classification sigmoid neuron. The model has six layers, including the input-output layer and five filters, along with the tanh and sigmoid activation functions. The paper used data with EEG signals from BCI competitions as well as from the BNCI Horizon and P300 -LINI datasets. The architectures were implemented using Keras with Tensorflow as the backend.

In protein domain classification from third generation sequencing reads, deep learning techniques, especially CNN, are used, as the convolutional filters are capable of representing motifs that are essential for sequence classification. The paper[8], **"ProDOMA: improving PROtein DOMAin classification for third-**

generation sequencing reads using deep learning," studies this concept and designs a CNN model named **"ProDOMA"** that outperforms the state-of-the-art HMMER and DeepFam models for domain classification. It also works well for long, noisy reads without depending on error correction. 3-frame encoding is used to transform DNA readings into a 3-channel tensor for neural network input. ProDOMA uses a convolutional layer and a max-over-time pooling layer to automatically extract features from the input. The probability of the sequence against all of the input protein domains was calculated using a classifier with two fully connected layers. The CNN was trained with a modified loss function to make out-of-distribution samples more likely to have a uniform distribution on softmax values, thereby excluding unrelated coding or non-coding DNA sequences. Simulated G-protein coupled receptor coding sequences from **PacBio** and human genome from **Oxford Nanopore Human Reference** Datasets Rel6 were used for training the model. Precision was evaluated based on F1 scores and run time was measured by averaging 5 independent trials with 10,000 random sequences. All the code implementation was done using Tensorflow.
For our research purpose the dataset chosen for the sequence classification problem is primary taken from the Protein Data Bank(PDB) which is an online database for biological molecules. Our dataset has DNA/RNA and protein sequences along with their corresponding class labelling. We discard the DNA/RNA sequences and delete data rows with either missing label or missing sequences using the PANDAS library. For better training of our CNN model we choose the top 10 frequently occurring classes of protein for classification and discard the rest. The sequences are then tokenized and sent as the model input using which the deep learning model learns the hidden sequence feature in order to finally classify a given unknown sequence into its' corresponding protein family.

References

[1,2,3] Zhang, D., Kabuka, M. R. (2020). Protein family classification from scratch: A cnn based deep learning approach. IEEE/ACM transactions on computational biology and bioinformatics, 18(5), 1996-2007.

[4,5] Aggarwal, D., Hasija, Y. (2022). A Review of Deep Learning Techniques for Protein Function Prediction. arXiv preprint arXiv:2211.09705.

[6] Hatami, N., Gavet, Y., Debayle, J. (2018, April). Classification of time-series images using deep convolutional neural networks. In Tenth international conference on machine vision (ICMV 2017) (Vol. 10696, pp. 242-249). SPIE.

[7] Montserrat Alvarado-Gonzalez, A., Fuentes-Pineda, G., Cervantes-Ojeda, J. (2019). A few filters are enough: Convolutional Neural Network for P300 Detection. arXiv e-prints, arXiv-1909.

[8] Du, N., Shang, J., Sun, Y. (2021). Improving protein domain classification for third-generation sequencing reads using deep learning. BMC genomics, 22(1), 1-13.

[9] Lo Bosco, G., Di Gangi, M. A. (2016, December). Deep learning architectures for DNA sequence classification. In International Workshop on Fuzzy Logic and Applications (pp. 162-171). Springer, Cham.