Computer Vision and Pattern Recognition

Project report

Predicting the secondary structure of proteins using Neural Networks

02.07.2019

Leading teacher: dr inż. Piotr Fabian

Section:
Patryk Cieślak
Szymon Dyrała
Piotr Gazda
Mateusz Koźlik
Marta Miler

1. Training and testing data

We've conducted two experiments. In the first one, for training data we used CB6133 dataset and for testing we used CB513 dataset. In the second one, for training and testing our neural network we used prepared dataset of proteins CB513 and cross-validation technique.

Datasets:

a) CB513

CB513 dataset contains 513 non redundant sequences, that can be used to test new secondary structure prediction methods. The format is as simple comma separated variable file e.g.: DSSP:-,-,-,G,G,G,-,-,-,E,E,E,E,E,-,-,-,-,H,H,H,H,H,-,

b) CB6133

CB6133 was produced by PISCES CullPDB and is a larger non-homologous protein dataset with known secondary structure for every protein. It contains 6128 proteins, in which 5600 proteins are training samples, 256 proteins are validation samples and 272 proteins are testing samples.

2. Progress of work

After analyzing the problem in the article, we selected the appropriate tools and libraries - Python language and, among others, tensorflow and numpy libraries, we found and retrieved data with the protein structure. Then we installed the environment for Python, configured the project in which we read the downloaded data.

Later we created a six-layer convolutional neural network that predicts a single amin in Q8 accuracy. Network consists of 3 convolutional layers and 3 fully connected layers. Each layer uses relu activation function except the last one, where it uses softmax function. Loss function used was cross entropy. During training we used adam optimizer. We used a 15-element window to analyze individual proteins. Missing elements in the window at the beginning and end of proteins were replaced by NOSEQ symbol. To train the network we used a set of 6133 proteins and to test it we used a set of CB513 proteins, but the probability that the letter predicted by our network is correct was too high, so we suspect an error that was probably caused by poor reading of the data from the file. After a more detailed analysis of the problem, we noticed that the problem was in the structure describing the protein. Many lines that were mostly filled up with NOSEQ symbol, the neural network predicted their occurrence and therefore the probability of predicting a good value was high. After removing these values, the probability level equalled that of typical cases - around 55%

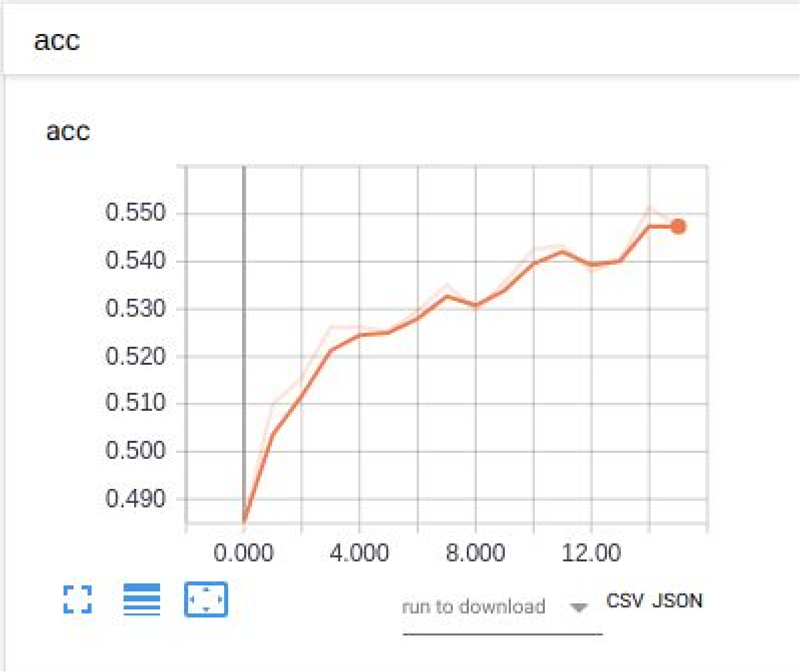
After receiving the results for the Q8 accuracy, we mapped the proteins to the Q3 accuracy to compare the results with our colleges SVM's protein prediction. We have reduced the Q8 accuracy to Q3 accuracy by mapping the values to other models. There are 3 categories - H (Helix), E (Elipse) and C (Coil) - which represent proteins that are not included in any of the previously mentioned categories. We then used CB513 to teach and test the neural network, but split it into four subsets by cross-validation technique. We teach neural network on 3/4 of the CB513 set, and on the last part we test its accuracy. We get 4 results and the final result is the average. The obtained results are presented in the third point - the results.

3. Results

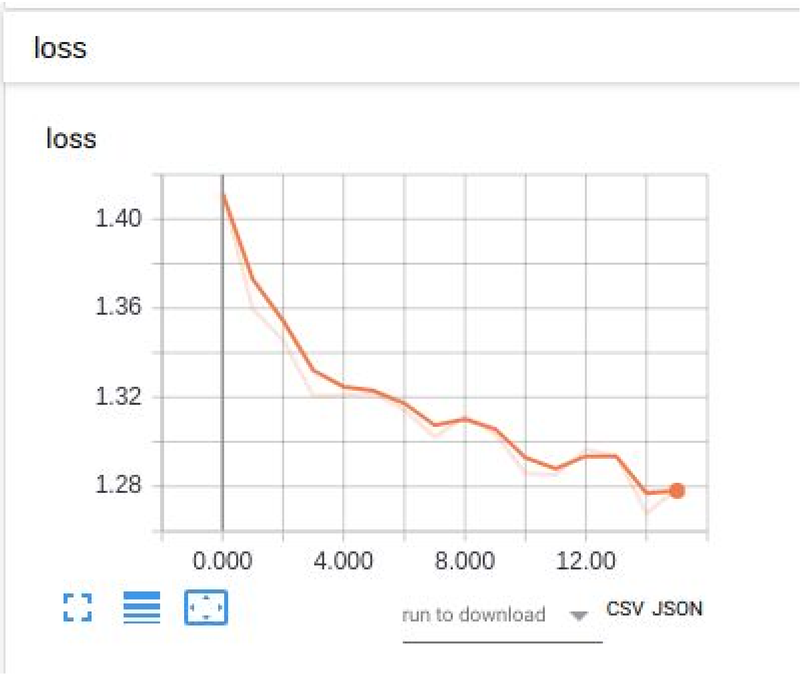
In the first, erroneous approach we received results of 80%, but we quickly found their incorrectness and excluded them from the study. After the error had been corrected, the average results were 49,11% for the Q8 protein accuracy. For Q3 protein accuracy average results were 66,33%.

(Q8,Q3)
(0,5060,0,6653)
(0.4884,0.7028)
(0.4852,0.6368)
(0,4849,0,6482)

Table 1 - The results of subsequent validations in this cross validation



Img. 1 - Acc of neural network training



Img. 2 - Loss of neural network training

4. Conclusions

While creating a Computer vision and Pattern Recognition project, we learned how convolutional neural networks work and how to create them. We were also able to face the challenge of creating code in the Python language and we familiarizes ourselves with the construction of proteins. Unfortunately, we were not able to beat the best result in protein prediction so far, but we received satisfactory results. Working on this project allowed us to develop programming skills and logical thinking.