Quality prediction of proteins models with spherical convolutions on three-dimensional graphs

Nikita Pavlichenko, Sergei Grudinin, Ilia Igashov.
pavlichenko.nv@phystech.ru, sergei.grudinin@inria.fr, igashov.i@yandex.ru
[1] Moscow Institute of Physics and Technolgy, Moscow, Russia

Convolutional neural networks have become very popular in recent years, and, in particular, have found widespread application in computer vision. Recently, active work has also begun on graph convolutional networks. In general, the graphs, unlike the pictures, are irregular structures, and in many tasks of learning on graphs sample objects also do not have unified topology. Therefore, the existing operations of convolution on the graphs are very much simplified, and the task of pulling on the graphs remain open in general. The purpose of this work is to research new operations of convolution on three-dimensional graphs within the framework of solving the problem of quality estimation of three-dimensional models of proteins (the problem of regression on the graph nodes).

Key words: *graph convolutional networks, spherical convlutions, three-dimensional graphs learning.*

1 Introduction

Protein molecules are an important part of any biological form. They determine cellular functions and behavior of various biological and chemical structures. It makes the discovery and prediction of proteins structure one of the most important points of medical, chemical and genetic science researches.

Molecules of proteins consist of smaller molecules called amino acids. These amino acids form a chain that is folded and placed in space. Thus, protein functions are determined by their positions in a 3D space. So, having this chain of amino acids we need to identify how they are located. There are ways to do this experimentally, but it can be time-consuming, expensive and not always possible. To solve these disadvantages, computational algorithms [1] [7] [12] were developed that generate different chain foldings. The problem is that no algorithm is the best one. Some of proteins are better modeled by one algorithm, others by others. Therefore, we are facing the problem of quality assessment (QA) of these protein models.

This problem has recently got attention from the machine learning community. Various artificial intelligence methods were applied such as neural networks [11] and support vector machines [9] [10]. More recent approaches mostly include deep learning methods [5] [4] [8] [3]. The newest approach is to use graph machine learning methods such as Graph Convolutional Networks (GCN) [2], where the protein is in some way represented as a graph. This work brings the new idea of capturing the 3D structure of this graph to improve the quality of GCN using convlutions based on spherical harmonics.

2 Problem statement

Consider a 3D model of a protein in space. The protein represents a chain of amino acids rolled up in space. Through dividing space around a protein into cells, for example, by the Voronoi method, we can get a 3D-graph, the vertices of which are amino acids of protein and edges are carried out between those amino acids that are in adjacent cells. Denote the resulting graph by $G = (V, E)$, where verticles $V = (v_1, \ldots, v_n)$ are a set of amino acids, E are edges.

For the i-th vertex we denote for $\mathcal{N}(v_i)$ the set of its neighbors in a graph G and for G an adjacency matrix $\boldsymbol{A}$:

$$A_{ij} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & \text{otherwise.} \end{cases}$$

Consider that each vertex v_i is described by some real d-dimensional vector of attributes $x(v_i) = \boldsymbol{x}_i$. In the simplest case, it can be a one-hot representation of an amino acid type. We can form feature matrix $\boldsymbol{X}$ from vectors $x(v_i)$ for every vertex x_i. Using these data we will solve a regression problem: to predict for each vertex v_i a real number - its "score". In other words, how correctly it is placed in the given 3D-model in comparison with the actual conformation of this protein.

3 Graph Convolutional Networks

We will develop approach for machine learning on graphs using Graph Convolutional Network (GCN) [6]. The idea of this approach is aggregation of neighbors features for every vertex and forming new feature vectors on the layer's output. So, for GCN with L layers, $L \geqslant 2$ we have:

$$A_{ij} = \begin{cases} \boldsymbol{H}^{(0)} = \boldsymbol{X} \\ \boldsymbol{H}^{(l)} = \sigma\left(\boldsymbol{A}\boldsymbol{H}^{l-1}\boldsymbol{W}^{(l-1)}\right) \quad \text{for } l \in \{1, \ldots, L-1\} \\ \boldsymbol{H}^{(L)} = \boldsymbol{A}\boldsymbol{H}^{L-1}\boldsymbol{W}^{(L-1)}. \end{cases}$$

In the above $\boldsymbol{H}^{(l)}$ denotes feature matrix at l-th layer and $\boldsymbol{W}^{(l)}$ denotes weight matrix. Common choise for activation function σ is a ReLU activation: $\mathrm{ReLU}(x) := \max(x, 0)$ or an ELU activation: $\mathrm{ELU}(x) := \begin{cases} x, & x \geqslant 0 \\ \alpha(e^x - 1), & x < 0 \end{cases}$. Wheights matrix $\boldsymbol{W}^{(l)}$ are trained by stochastic gradient descend or its modifications.

We rely on this approach because it has already showed outperformance in protein quality assessment problen [2].

4 Spherical convolution

Consider a vertex v_i. Since all the amino acids in the protein are connected in a peptide chain, it is easy to construct its local coordinate system for the amino acid v_i under consideration — on two dihedral corners, which are unambiguously determined from the geometry of the peptide chain. Write the coordinates of all the neighbors of $v_j \in \mathcal{N}(v_i)$ of the amino acid in the obtained coordinate system. Then proceed to spherical coordinates, and project all vertices onto a unit sphere with the center in v_i. Now, each vertex $v_j \in \mathcal{N}(v_i)$ can be matched with a pair of angles $\Omega_i^j = (\varphi_i^j, \psi_i^j)$ that specify the angular position of the projection of the vertex v_j onto a unit sphere in the local coordinate system of v_i.

Now, having an unambiguous orientation for each vertex of the graph, we can introduce the convolution operation. Let us consider a certain matrix function $f(\Omega) : [0, \pi] \times [0, 2\pi) \to \mathbb{R}^{d \times d'}$ on a single sphere. We can expance it into a series on the basis of spherical functions $\{Y_l^m\}_{l,m}$ and leave the first few components:

$$f(\Omega) \approx f_W(\Omega) = \sum_{l=0}^{L} \sum_{m} \boldsymbol{W}_l^m Y_l^m(\Omega),$$

where $\boldsymbol{W}_l^m$ denotes coefficient matrix in the expancion of matrix function f on the basis of $\{Y_l^m\}_{l,m}$. Then we can introduce the spherical convolution operation for the vertex v_i in the following way:

$$f_W \circ v_i = \sum_{v_j \in \mathcal{N}(v_i)} f_W(\Omega_i^j) x(v_i).$$

Considering $\boldsymbol{W}_l^m$ matrices to be optimized parameters, we will thus train spherical filters.

4.1 Spherical convolution layer

Let us have a 3D model of a protein consisting of N amino acids, the i-th amino acid is described by the trait vector $\boldsymbol{x}_i \in \mathbb{R}^d$. Let us denote all the vertices through the $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ feature matrix. Then one layer of spherical convolution is written down as follows:

$$\boldsymbol{X} \longrightarrow \boldsymbol{X}' = \sigma(f_W \circ \boldsymbol{X}) = \sigma\left(\sum_{l,m} Y_l^m(\boldsymbol{A}_\Omega)\boldsymbol{X}\boldsymbol{W}_l^m\right),$$

where σ is an activation function, Y_l^m are spherical functions, $\boldsymbol{W}_l^m$ are optimized parameters and $\boldsymbol{A}_\Omega$ is the adjacency matrix of graph G, which cells contains the spherical coordinates of the vertices in the corresponding local coordinate system:

$$[\boldsymbol{A}_\Omega]_{i,j} = \begin{cases} \Omega_i^j, & (v_i, v_j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

References

[1] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, nov 2005.

[2] Federico Baldassarre, Hossein Azizpour, David Menéndez Hurtado, and Arne Elofsson. Graphqa: Protein model quality assessment using graph convolutional network. 2020.

[3] Matthew Conover, Max Staples, Dong Si, Miao Sun, and Renzhi Cao. AngularQA: Protein model quality assessment with LSTM networks. feb 2019.

[4] Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, jun 2018.

[5] David Menéndez Hurtado, Karolis Uziela, and Arne Elofsson. Deep transfer learning in the assessment of the quality of protein models.

[6] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks.

[7] Jesper Lundström, Leszek Rychlewski, Janusz Bujnicki, and Arne Elofsson. Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Science*, 10(11):2354–2362, dec 2008.

[8] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 3d oriented convolutional neural networks. *Bioinformatics*, 35(18):3313–3319, feb 2019.

[9] Arjun Ray, Erik Lindahl, and Björn Wallner. Improved model quality assessment using ProQ2. *BMC Bioinformatics*, 13(1):224, 2012.

[10] Karolis Uziela, Nanjiang Shu, Björn Wallner, and Arne Elofsson. ProQ3: Improved model quality assessments using rosetta energy terms. *Scientific Reports*, 6(1), oct 2016.

[11] Björn Wallner and Arne Elofsson. Can correct protein models be identified? *Protein Science*, 12(5):1073–1086, may 2003.

[12] Jinbo Xu. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34):16856–16865, aug 2019.