MGCN for Property Prediction in Large Molecules

Prof. Prabhkar B, **Anurag Mehta**, **Akshat Maheshwari** and **Sayak Kundu**

Abstract— **For chemical and pharmaceutical industries efficient prediction of molecular properties (for example, atomization energy) is a challenging problem. Solving such problems using quantum chemistry can speed up several researches, such as drug designing and substance discovery in a fast manner which can be very useful like in the current pandemic case. Traditional studies based on density functional theory (DFT) showed good results but are also time-consuming for predicting properties of large-sized molecules. Several machine learning methods have been developed in order to solve the problem but they miss one important factor, i.e., capturing the complex inherent quantum interactions. The complex inherent quantum interactions of molecules are largely under-explored by existing solutions. In this report, we present a generalizable (the model should be able to perform well even trained with a small amount of data) and transferable (irrespective of training data, small or large molecule, the model should be able to represent the whole demographic) Multilevel Graph Convolutional neural Network (MGCN) for molecular property prediction.**

KEYWORDS

Multilevel Graph Convolutional neural Network (MGCN), DFT, Molecular properties

I. LITERATURE REVIEW

- **DFT of electronic structure W Kohn 96**
  Density functional theory (DFT) is a theory of electronic structure, based on the electron density distribution over number of atoms participating, instead of the many-electron wave function ψ(r1,r2,r3,...).
  The theory has been in use for over 30 years by physicists working on the electronic structure of solids, surfaces, defects, etc..DFT is preferable when the demographic for the study is limited to 5-10 atomic structure and provides an approximation for the energy function given a sufficiently powerful computer. The drawback of DFT is that it is highly dependent on exchange-correlation energy functional, which can be unknown in several cases. Although several accurate studies are constantly evolving, there is no known systematic way to achieve an arbitrarily high level of accuracy without the exchange of high computation power.
- **Molecular graph convolutions Keranas 2016**
  The study defined molecules as undirected graphs and used molecular graph convolutions. Graph convolutions used simple encodings to represent a molecule as a graph with bonds, distances, etc represented mathematically. This allowed the model to take greater advantage of extra information in the graph structure. Although the presented graph convolutions didn't outperform all fingerprint-based methods, this study became a benchmark for further such approaches
- **Message passing for quantum Chemistry, Justin Gilmer 2017**
  The models were learnt for the graphs data structure, using message passing algorithm and aggregation procedure to compute a function of their entire input graph. The graph training was divided into two phases: message passing and readout. Message functions M are based on time step quantization and hidden states of the nodes, which are updated based on the message received after a cycle.The readout phase computes a feature vector for the whole graph using a readout function R. Appropriate message, update, and molecular properties output functions acted as good factors for predicting molecular properties, making it outperform several laid baselines methods.
- **MGCN: Semi-supervised Classification in Multi-layer Graphs with Graph Convolutional Networks Mahsa Ghorbani 19**
  The study conveyed the Multi-layer GCNs architecture that utilizes the GCN for multi-layer graphs. MGCN embeds nodes of multilayer graphs using both (within and between layers) relations and nodes attributes. The between-layer edges are not completely utilized in the GCN layers, so created a multilayer graph for which these layers were not independent, further introducing layer connections and a loss function that incorporates reconstruction of both the within-layer and between layer connections, also proved the superior-

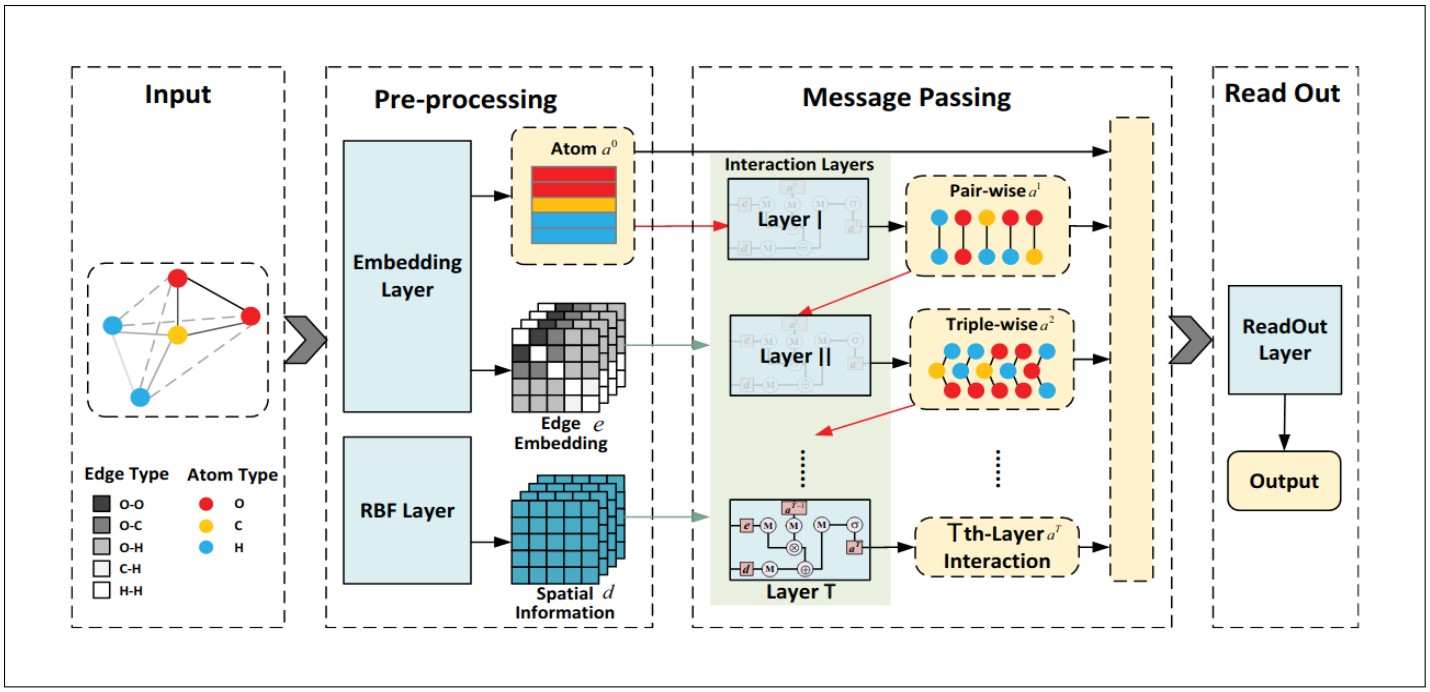

Fig. 1. The Complete Model Architecture, cited from Original Research Paper

ity of MGCN in considering between-layer edges to some single-layer graph embedding methods.

II. Original Model and Changes in it

A. Model Architecture

The overall architecture of the model (the original as developed by Chengqiang Lu et al. 2019 and the one implemented by us) can be broken down into 4 parts.

1) Embedding Layer

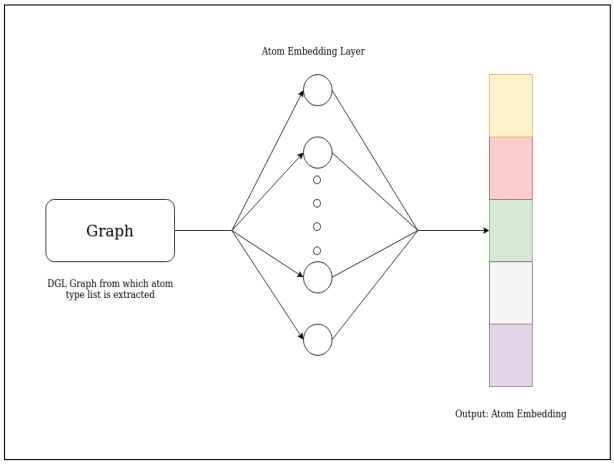

Fig. 2. Atom Embedding Layer

To model the interactions between atoms and bonds, the molecules first need to be embedded into a tensor/vector, making it a valid input argument for convolutional layers. The Embedding layer can be further decomposed into 2 components: *AtomEmbedding Layer* and *EdgeEmbedding Layer*. The atom embeddings of all the molecules are generated randomly before training, and the embeddings corresponding to an atom or edge will have the same value throughout the molecule. As of the edge embeddings, a mathematical formula is used to generate a unique value from the integers assigned to the atoms at the edge of that edge.

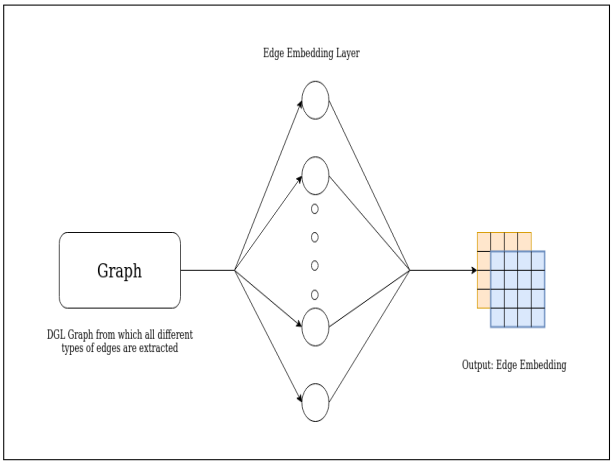

Fig. 3. Edge Embedding Layer

PS- Find images of other section of the model in the layers_images folder.

2) **Radial Basis Function** A radial basis function (RBF) is a real-valued function whose value depends only on the distance between the input and some fixed point, either the origin, so that $\psi(x) = \psi(\|x\|)$, or some other fixed point c, called a centre so that $\psi(x) = \psi(\|x - c\|)$ (definition: Wikipedia). Since the interactions are heavily affected by complex inherent quantum interactions between atoms of a molecule, hence this RBF layer is used to convert this spatial information to usable tensors/vectors. A Gaussian RBF Kernel is being used with Euclidean distance between the atoms to convert the distance matrix obtained from the graph to a usable tensor.
3) **Interaction Layer** To model the multilevel molecular structure with all the conformation and spatial information, the interactions between the atoms are recorded in a level-wise manner (atom-wise, atom-pair, atom-triplets, etc.), and these levels define the graph convolution layer in the network. The groups of atoms in each layer that are taken into account are the groups of neighbour atoms, which are found by the edges of the graph. In each layer, one more neighbouring atom is added to the existing group of interacting atoms, and hence, with each progressing layer, the interacting groups' size increases by 1. The ith level values will be used to generate the interaction terms in the i+1th layer.
   The first level here represents the atom embeddings, i.e., the Embedding layer. The layer, being hierarchical in nature, has a message-passing phenomenon, which means that the lower layer representations are fed into the network to create the next higher-level representations.
   The forward inference involves the first-order neighbour node and spatial information with the message passed by the embedding layer and so on the cycle continues. This layer, hence, is a multi-level interaction layer, where the lowest layer consists of individual atoms, and the highest layer consists of the whole molecule.
4) **Readout Layer** The final layer or the output layer in the network. The layer predicts molecular properties. The layer consists of 2 dense sub-layers that get the input from the interaction layer and using that, the final output of the predictions of properties is being generated. The final values of the properties are generated by summing up the values of these properties of individual nodes in the entire graph.
5) **Activation Function**: Throughout the network, the *SoftPlus* activation function is used, as the author in discussion revealed a differentiable function with finite bounds was required to limit the weights, hence this acts as a better alternative to the traditional activation functions (sigmoid and tanh).

B. *Few changes introduced*

- A discussion with the author suggested adding some linear and dropout layers can affect the training, both in result and time, so a dropout layer was introduced after the RBF Kernel, to avoid overfitting.
- Further variations of the dropout layer were also tried, like introducing extra layers in the network after the edge weights/features generation from the DGL Graph object.
  Adding dropout layers certainly made the model converge at early epochs however substantially increased the MAE hence not suggested.
- Few variations were also introduced in the CNN layers, however, the results showed an even greater loss, so we decided to go with the addition of only 1 more linear layer for updating the edge feature values.

III. Performance

Models	Loss	MAE
Author Model		0.0229
Model with only 1 Dropout layer	42.387	0.0517
Model with 1 Dropout and 1 extra Linear layer	58.08	0.0595
Model with Dropout after Convolution	43.04	0.0520

As we can see in the table, the Mean Absolute Error (MAE) with only 1 dropout layer is less compared to adding an extra linear layer or adding the dropout layer after the convolution layer. In all the cases, it was observed that the Loss decreases at a very fast rate. It comes to a gradual and slow decrease even before the 25th epoch mark.

Also, the convergence of the model with only 1 dropout layer was faster than the other models. This model gave us the ability to balance speed with

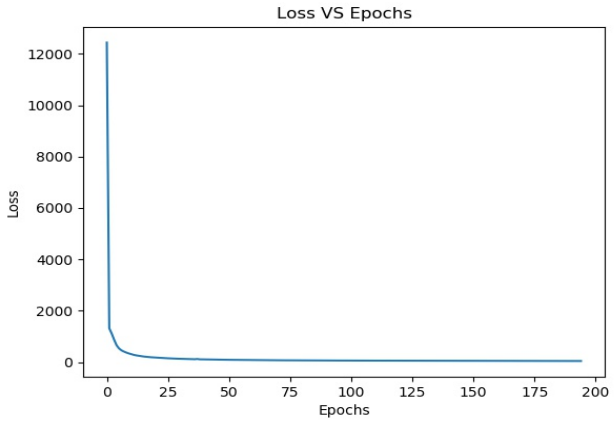

Fig. 4. Loss

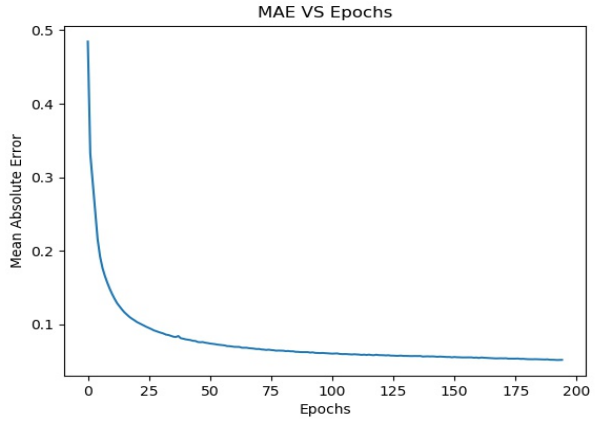

Fig. 5. MAE

accuracy. Fig. 4 and 5 show the curves for the loss (Root Mean Squared Error) and Mean Absolute Error (MAE) for the single dropout layer model respectively. We can see that the implementations cannot cross the SOTA benchmark, which was published in June 2019. We can further try tweaking the convolution layer and try experimenting with it so that we end up with a better result than the SOTA.

IV. Future Work

The meeting with the author revealed that the process of training the model can be improved. To speed up the convergence and check for better predictions, training process can be modified in a way that first the model is trained on small molecules, and the training then gradually shifts to larger molecules.
Improvements can also be done to improve the computational-efficiency of the model.

V. Discussions in Meetings

- An interesting question that arose during the discussion with Prof. Prabhakar was that the initializations of atom embeddings are random so this can lead to non-valance atoms. A discussion with authors suggested also that this can be possible as the bond types are not labeled molecules, but no such observation was noticed in any of the results.
- Another question asked by the professor was about the conformers' energy differentiation since both molecules share the same smile structures. Since the distance matrix is used as the parameter for a sample, the parameter output can be considered as a result of the pseudo function of distances, helping in differentiating the conformers.
- How is the spatial information conserved when the edges in graph representation are not even initialized? The graphs maintain the structure of the molecule i.e. the neighbouring atoms, in the interacting layer every nth depth neighbour is considered and optimized over the edges leading to spatial information conservation.

ACKNOWLEDGMENT

- We would like to thank all the Professors of the course ML in Natural Sciences for providing us this opportunity to work.

References

[1] Chengqiang Lu , Qi Liu , Chao Wang , Zhenya Huang , Peize Lin , Lixin He (2019), "Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective".
[2] W. Kohn, A. D. Becke, R. G. Parr (1996), "Density Functional Theory of Electronic Structure".
[3] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, Patrick Riley (2016), "Molecular Graph Convolutions: Moving Beyond Fingerprints".
[4] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, George E. Dahl(2017), "Neural Message Passing for Quantum Chemistry".
[5] Mahsa Ghorbani, Mahdieh Soleymani Baghshah, Hamid R. Rabiee(2019), "MGCN: Semi-supervised Classification inMulti-layer Graphs with Graph ConvolutionalNetworks".
[6] Yann leCun, Bell labs (1996), "Convolutional Networks for Images Speech and Time Series".
[7] Guangyong Chen, Pengfei Chen, Chang-Yu Hsieh, Chee-Kong Lee, Benben Liao, Renjie Liao, Weiwen Liu, Jiezhong Qiu, Qiming Sun, Jie Tang, Richard Zemel, Shengyu Zhang (2019) , "Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models"