# DEEP GRAPH TRANSFORMER WITH POINTWISE DENSE POOLING

A PREPRINT

**Jiwei Liu**
NVIDIA
Princeton, NJ 08540
jiweil@nvidia.com

October 3, 2022

**ABSTRACT**

In this technical report, we introduce a novel pointwise dense pooling instead of the conventional global mean pooling for graph transformers. We train the model using different seeds and average the predictions. The approach achieved a $0.089$ mean absolute error on the $valid$ dataset of NeurIPS 2022 OGB-LSC PCQM4Mv2 dataset.

***Keywords*** GNN · Transformer · Pooling · Molecule DFT

## 1 Dataset

In this challenge, we are asked to predict a quantum property, Density Functional Theory (DFT), of molecular graphs. DFT calculations are known to be time-consuming and could take up hours even for small molecules. With the rapid advancement of machine learning (ML) technology, it is promising to use fast and accurate ML to replace the expensive DFT calculations.

The PCQM4Mv2 dataset, based on the PubChemQC project Nakata and Shimazaki [2017], provides us with a well-defined ML task of predicting DFT of molecules given their 2D molecular graphs. Additional 3D structures are also provided for training molecules but our approach only used the 2D molecular graphs.

## 2 Model Architecture

Our model uses the given $PygPCQM4Mv2Dataset$ which converts $SMILEs$ strings to $pytorch$ tensors. Each node and each edge are represented by a vector of dimensions of 9 and 3, respectively. We use embeddings to encode node features. Our model is implemented with 'pytorch geometric' Fey and Lenssen [2019]. The architecture of our model is shown in Figure 1. The core building block is $DeepGCNLayer$ Li et al. [2019] wrapper of $TransformerConv$ Shi et al. [2020] with skip connections. $DeepGCNLayer$ enables the net to grow deeper. There are 15 $DeepGCNLayer$s in our net.

The $TransformerConv$ implements the following equation, where $\mathbf{x}_i$ and $\mathbf{x}'_i$ are the input and output embeddings of node $i$ of a $TransformerConv$ layer, respectively. $\mathcal{N}(i)$ represent the neighbors of node $i$.

$$\mathbf{x}'_i = \mathbf{W}_1\mathbf{x}_i + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}\mathbf{W}_2\mathbf{x}_j \tag{1}$$

where the attention coefficients $\alpha_{i,j}$ are computed via multi-head dot product attention:

$$\alpha_{i,j} = \text{softmax}\left(\frac{(\mathbf{W}_3\mathbf{x}_i)^\top(\mathbf{W}_4\mathbf{x}_j)}{\sqrt{d}}\right) \tag{2}$$

At the end of the last $TransformerConv$, we have transformed embeddings for each node. To convert node embeddings to graph embeddings $\mathbf{G}_i$, global mean pooling is often used as shown in equation 4.

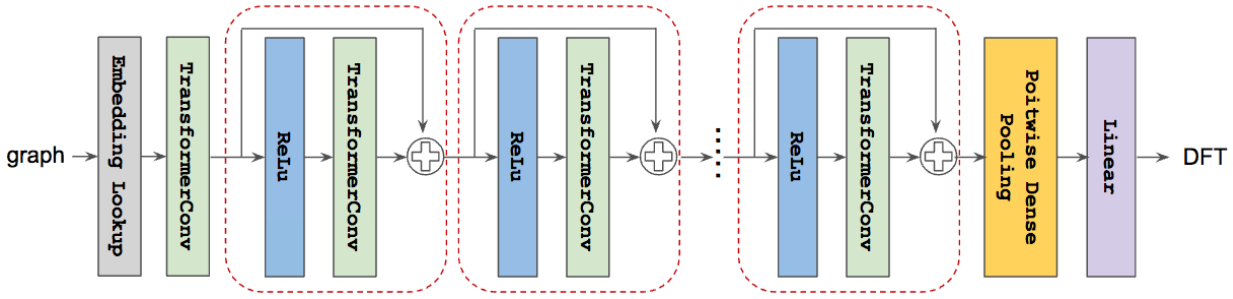


Figure 1: Architecture of our deep graph transformer. The boxes with red dashed lines are $DeepGCNLayer$ with skip connections.

Table 1: Experiment results

| model | Description | # of params | valid MAE (single) | valid MAE (ensemble) |
|---|---|---|---|---|
| Deep graph transformers | global mean pooling | 63.6M | 0.096 | 0.090 |
| Deep graph transformers | pointwise dense pooling | 63.6M | 0.094 | 0.089 |

$$\mathbf{G}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{x}_n. \tag{3}$$

With global mean pooling, each node contributes equally to the graph embedding, which could be suboptimal. An intuitive idea to improve global mean pooling is to use the weighted sum of the node embeddings to compute graph embedding where weights are learned from node embeddings. We propose a novel pointwise dense pooling layer:

$$\mathbf{G}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \alpha_n \mathbf{x}_n. \tag{4}$$

where $\alpha_n$ is learned from $\mathbf{x}_n$:

$$\alpha_n = \text{softmax}\left(\mathbf{W}\mathbf{x}_n\right) \tag{5}$$

## 3 Experiment Results

We train our models with 4 NVIDIA V100 GPUs. The model is trained with the given $train$ subset and validated with the given $valid$ subset. The model is trained with 4 different seeds, $[0, 1, 2, 3]$. Training one model with one seed on a single GPU takes 10 hours. Inference with one model and one seed on the $testdev$ takes 50 seconds. Experiment results are shown in table 1.

## 4 Conclusion

In this tech report, we present a novel pointwise dense pooling for deep graph transformers, which outperforms conventional global mean pooling.

## References

Maho Nakata and Tomomi Shimazaki. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of chemical information and modeling*, 57(6):1300–1308, 2017.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9267–9276, 2019.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.