1 Interpretation of Laplacian based models as MPNNs

Another family of models defined in Defferrard et al. (2016), Bruna et al. (2013), Kipf & Welling (2016) can be interpreted as MPNNs. These models generalize the notion of convolutions a general graph G with N nodes. In the language of MPNNs, these models tend to have very simple message functions and are typically applied to much larger graphs such as social network data. We closely follow the notation defined in Bruna et al. (2013) equation (3.2). The model discussed in Defferrard et al. (2016) (equation 5) and Kipf & Welling (2016) can be viewed as special cases. Given an adjacency matrix $W \in \mathbb{R}^{N \times N}$ we define the graph Laplacian to be $L = I_n - D^{-1/2} W D^{-1/2}$ where D is the diagonal degree matrix with $D_{ii} = \deg(v_i)$. Let V denote the eigenvectors of L, ordered by eigenvalue. Let σ be a real valued nonlinearity (such as ReLU). We now define an operation which transforms an input vector x of size $N \times d_1$ to a vector y of size $N \times d_2$ (the full model can defined as stacking these operations).

$$y_j = \sigma \left(\sum_{i=1}^{d_1} V F_{i,j} V^T x_i \right) \quad (j = 1 \dots d_2) \tag{1}$$

Here y_j and x_i are all N dimensional vectors corresponding to a scalar feature at each node. The matrices $F_{i,j}$ are all diagonal $N \times N$ matrices and contain all of the learned parameters in the layer. We now expand equation 1 in terms of the full $N \times d_1$ vector x and $N \times d_2$ vector y, using v and w to index nodes in the graph G and i, j to index the dimensions of the node states. In this way $x_{w,i}$ denotes the i'th dimension of node w, and $y_{v,j}$ denotes the j'th dimension of node v, furthermore we use x_w to denote the d_1 dimensional vector for node state w, and y_v to denote the d_2 dimensional vector for node v. Define the rank 4 tensor $\tilde{L}$ of dimension $N \times N \times d_1 \times d_2$ where $\tilde{L}_{v,w,i,j} = (V F_{i,j} V^T)_{v,w}$. We will use $\tilde{L}_{i,j}$ to denote the $N \times N$ dimensional matrix where $(\tilde{L}_{i,j})_{v,w} = \tilde{L}_{v,w,i,j}$ and $\tilde{L}_{v,w}$ to denote the $d_1 \times d_2$ dimensional matrix where $(\tilde{L}_{v,w})_{i,j} = \tilde{L}_{v,w,i,j}$. Writing equation 1 in this notation we have

$$\begin{aligned}
y_j &= \sigma \left(\sum_{i=1}^{d_1} \tilde{L}_{i,j} x_i \right) \\
y_{v,j} &= \sigma \left(\sum_{i=1,w=1}^{d_1,N} \tilde{L}_{v,w,i,j} x_{w,i} \right) \\
y_v &= \sigma \left(\sum_{w=1}^{N} \tilde{L}_{v,w} x_w \right).
\end{aligned}$$

Relabelling y_v as h_v^{t+1} and x_w as h_w^t this last line can be interpreted as the message passing update in an MPNN where $M(h_v^t, h_w^t) = \tilde{L}_{v,w} h_w^t$ and $U(h_v^t, m_v^{t+1}) = \sigma(m_v^{t+1})$.

1.1 The special case of Kipf and Welling (2016)

Motivated as a first order approximation of the graph laplacian methods, Kipf & Welling (2016) propose the following layer-wise propagation rule:

$$H^{l+1} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^l W^l\right) \tag{2}$$

Here $\tilde{A} = A + I_N$ where A is the real valued adjacency matrix for an undirected graph G. Adding the identity matrix I_N corresponds to adding self loops to the graph. Also $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ denotes the degree matrix for the graph with self loops, $W^l \in \mathbb{R}^{D\times D}$ is a layer-specific trainable weight matrix, and $\sigma(\cdot)$ denotes a real valued nonlinearity. Each H^l is a $\mathbb{R}^{N\times D}$ dimensional matrix indicating the D dimensional node states for the N nodes in the graph.

In what follows, given a matrix M we use $M_{(v)}$ to denote the row in M indexed by v (v will always correspond to a node in the graph G). Let $L = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$. To get the updated node state for node v we have

$$\begin{aligned} H^{l+1}_{(v)} &= \sigma\left(L_{(v)}H^l W^l\right) \\ &= \sigma\left(\sum_w L_{vw}H^l_{(w)}W^l\right) \end{aligned}$$

Relabelling the row vector $H^{l+1}_{(v)}$ as an N dimensional column vector h^{t+1}_v the above equation is equivalent to

$$h^{t+1}_v = \sigma\left((W^l)^T\sum_w L_{vw}h^t_w\right) \tag{3}$$

which is equivalent to a message function

$$M_t(h^t_v, h^t_w) = L_{vw}h^t_w = \tilde{A}_{vw}(\text{deg}(v)\text{deg}(w))^{-1/2}h^t_w,$$

and update function

$$U_t(h^t_v, m^{t+1}_v) = \sigma((W^t)^T m^{t+1}).$$

Note that the L_{vw} are all scalar valued, so this model corresponds to taking a certain weighted average of neighboring nodes at each time step.

2 A More Detailed Description of the Quantum Properties

First there the four atomization energies.

- Atomization energy at $0K$ U_0 (eV): This is the energy required to break up the molecule into all of its constituent atoms if the molecule is at absolute zero. This calculation assumes that the molecules are held at fixed volume.

- Atomization energy at room temperature U (eV): Like U_0, this is the energy required to break up the molecule if it is at room temperature.
- Enthalpy of atomization at room temperature H (eV): The enthalpy of atomization is similar in spirit to the energy of atomization, U. However, unlike the energy this calculation assumes that the constituent molecules are held at fixed pressure.
- Free energy of atomization G (eV): Once again this is similar to U and H, but assumes that the system is held at fixed temperature and pressure during the dissociation.

Next there are properties related to fundamental vibrations of the molecule:

- Highest fundamental vibrational frequency ω_1 (cm^{-1}): Every molecule has fundamental vibrational modes that it can naturally oscillate at. ω_1 is the mode that requires the most energy.
- Zero Point Vibrational Energy (ZPVE) (eV): Even at zero temperature quantum mechanical uncertainty implies that atoms vibrate. This is known as the zero point vibrational energy and can be calculated once the allowed vibrational modes of a molecule are known.

Additionally, there are a number of properties that concern the states of the electrons in the molecule:

- Highest Occupied Molecular Orbital (HOMO) $\varepsilon_{\text{HOMO}}$ (eV): Quantum mechanics dictates that the allowed states that electrons can occupy in a molecule are discrete. The Pauli exclusion principle states that no two electrons may occupy the same state. At zero temperature, therefore, electrons stack in states from lowest energy to highest energy. HOMO is the energy of the highest occupied electronic state.
- Lowest Unoccupied Molecular Orbital (LUMO) $\varepsilon_{\text{LUMO}}$ (eV): Like HOMO, LUMO is the lowest energy electronic state that is unoccupied.
- Electron energy gap $\Delta\varepsilon$ (eV): This is the difference in energy between LUMO and HOMO. It is the lowest energy transition that can occur when an electron is excited from an occupied state to an unoccupied state. $\Delta\varepsilon$ also dictates the longest wavelength of light that the molecule can absorb.

Finally, there are several measures of the spatial distribution of electrons in the molecule:

- Electronic Spatial Extent $\langle R^2 \rangle$ (Bohr^2): The electronic spatial extent is the second moment of the charge distribution, $\rho(r)$, or in other words $\langle R^2 \rangle = \int dr r^2 \rho(r)$.
- Norm of the dipole moment μ (Debye): The dipole moment, $p(r) = \int dr' p(r')(r - r')$, approximates the electric field far from a molecule. The norm of the dipole moment is related to how anisotropically the charge is distributed (and hence the strength of the field far from the molecule). This degree of anisotropy is in turn

Table 1: Chemical Accuracy For Each Target

Target	DFT Error	Chemical Accuracy
mu	.1	.1
alpha	.4	.1
HOMO	2.0	.043
LUMO	2.6	.043
gap	1.2	.043
R2	-	1.2
ZPVE	.0097	.0012
U0	.1	.043
U	.1	.043
H	.1	.043
G	.1	.043
Cv	.34	.050
Omega	28	10.0

related to a number of material properties (for example hydrogen bonding in water causes the dipole moment to be large which has a large effect on dynamics and surface tension).

- Norm of the static polarizability α (Bohr^3): α measures the extent to which a molecule can spontaneously incur a dipole moment in response to an external field. This is in turn related to the degree to which i.e. Van der Waals interactions play a role in the dynamics of the medium.

2.1 Chemical Accuracy and DFT Error

In Table 1 we list the mean absolute error numbers for chemical accuracy. These are the numbers used to compute the error ratios of all models in the tables. The mean absolute errors of our models can thus be calculated as $(\text{Error Ratio}) \times (\text{Chemical Accuracy})$. We also include some estimates of the mean absolute error for the DFT calculation to the ground truth. Both the estimates of chemical accruacy and DFT error were provided in Faber et al. (2017).

3 Additional Results

In Table 2 we compare the performance of the best architecture (edge network + set2set output) on different sized training sets. It is surprising how data efficient this model is on some targets. For example, on both R2 and Omega our models are equal or better with 11k samples than the best baseline is with 110k samples.

In Table 3 we compare the performance of several models trained without spatial information. The left 4 columns show the results of 4 experiments, one where we train the GG-NN model on the sparse graph, one where we add virtual edges (**ve**), one where we add a master node (**mn**), and one where we change the graph level output

Table 2: Results from training the edge network + set2set model on different sized training sets (N denotes the number of training samples)

Target	N=11k	N=35k	N=58k	N=82k	N=110k
mu	1.28	0.55	0.44	0.32	0.30
alpha	2.76	1.59	1.26	1.09	0.92
HOMO	2.33	1.50	1.34	1.19	0.99
LUMO	2.18	1.47	1.19	1.10	0.87
gap	3.53	2.34	2.07	1.84	1.60
R2	0.28	0.22	0.21	0.21	0.15
ZPVE	2.52	1.78	1.69	1.68	1.27
U0	1.24	0.69	0.58	0.62	0.45
U	1.05	0.69	0.60	0.52	0.45
H	1.14	0.64	0.65	0.53	0.39
G	1.23	0.62	0.64	0.49	0.44
Cv	1.99	1.24	0.93	0.87	0.80
Omega	0.28	0.25	0.24	0.15	0.19

to a set2set output (**s2s**). In general, we find that it's important to allow the model to capture long range interactions in these graphs.

In Table 4 we compare GG-NN + towers + set2set output (**tow8**) vs a baseline GG-NN + set2set output (**GG-NN**) when distance bins are used. We do this comparison in both the joint training regime (**j**) and when training one model per target (**i**). For joint training of the baseline we used 100 trials with $d = 200$ as well as 200 trials where d was chosen randomly in the set $\{43, 73, 113, 153\}$, we report the minimum test error across all 300 trials. For individual training of the baseline we used 100 trials where d was chosen uniformly in the range $[43, 200]$. The towers model was always trained with $d = 200$ and $k = 8$, with 100 tuning trials for joint training and 50 trials for individual training. The towers model outperforms the baseline model on 12 out of 13 targets in both individual and joint target training.

In Table 5 right 2 columns compare the edge network (**enn**) with the pair message network (**pm**) in the joint training regime (**j**). The edge network consistently outperforms the pair message function across most targets.

In Table 6 we compare our MPNNs with different input featurizations. All models use the Set2Set output and GRU update functions. The no distance model uses the matrix multiply message function, the distance models use the edge neural network message function.

References

Bruna, Joan, Zaremba, Wojciech, Szlam, Arthur, and LeCun, Yann. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

Defferrard, Michaël, Bresson, Xavier, and Vandergheynst, Pierre. Convolutional neural

Table 3: Comparison of models when distance information is excluded

Target	GG-NN	ve	mn	s2s
mu	3.94	**3.76**	4.02	3.81
alpha	2.43	2.07	**2.01**	2.04
HOMO	1.80	**1.60**	1.67	1.71
LUMO	1.73	1.48	1.48	**1.41**
gap	2.48	2.33	**2.23**	2.26
R2	14.74	17.11	**13.16**	13.67
ZPVE	5.93	3.21	3.26	**3.02**
U0	1.98	0.89	0.90	**0.72**
U	2.48	0.93	0.99	**0.79**
H	2.19	0.86	0.95	**0.80**
G	2.13	0.85	1.02	**0.74**
Cv	1.96	**1.61**	1.63	1.71
Omega	1.28	1.05	0.78	**0.78**
Average	3.47	2.90	2.62	**2.57**

networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pp. 3837–3845, 2016.

Faber, Felix, Hutchison, Luke, Huang, Bing, Gilmer, Justin, Schoenholz, Samuel S., Dahl, George E., Vinyals, Oriol, Kearnes, Steven, Riley, Patrick F., and von Lilienfeld, O. Anatole. Fast machine learning models of electronic and energetic properties consistently reach approximation errors better than dft accuracy. *https://arxiv.org/abs/1702.05532*, 2017.

Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv e-prints*, September 2016.

Table 4: Towers vs Vanilla Model (no explicit hydrogen)

Target	GG-NN-j	tow8-j	GG-NN-i	tow8-i
mu	2.73	**2.47**	**2.16**	2.23
alpha	1.66	**1.50**	1.47	**1.34**
HOMO	1.33	**1.19**	1.27	**1.20**
LUMO	1.28	**1.12**	1.24	**1.11**
gap	1.73	**1.55**	1.78	**1.68**
R2	**6.07**	6.16	4.78	**4.11**
ZPVE	4.07	**3.43**	2.70	**2.54**
U0	1.00	**0.86**	0.71	**0.55**
U	1.01	**0.88**	0.65	**0.52**
H	1.01	**0.88**	0.68	**0.50**
G	0.99	**0.85**	0.66	**0.50**
Cv	1.40	**1.27**	1.27	**1.09**
Omega	0.66	**0.62**	0.57	**0.50**
Average	1.92	**1.75**	1.53	**1.37**

Table 5: Pair Message vs Edge Network in joint training

Target	pm-j	enn-j
mu	**2.10**	2.24
alpha	2.30	**1.48**
HOMO	**1.20**	1.30
LUMO	1.46	**1.20**
gap	1.80	**1.75**
R2	10.87	**2.41**
ZPVE	16.53	**3.85**
U0	3.16	**0.92**
U	3.18	**0.93**
H	3.20	**0.93**
G	2.97	**0.92**
Cv	2.17	**1.28**
Omega	0.83	**0.74**
Average	3.98	**1.53**

Table 6: Performance With Different Input Information

Target	no distance	distance	dist + exp hydrogen
mu	3.81	0.95	**0.30**
alpha	2.04	1.18	**0.92**
HOMO	1.71	1.10	**0.99**
LUMO	1.41	1.06	**0.87**
gap	2.26	1.74	**1.60**
R2	13.67	0.57	**0.15**
ZPVE	3.02	2.57	**1.27**
U0	0.72	0.55	**0.45**
U	0.79	0.55	**0.45**
H	0.80	0.59	**0.39**
G	0.74	0.55	**0.44**
Cv	1.71	0.99	**0.80**
Omega	0.78	0.41	**0.19**
Average	2.57	0.98	**0.68**