Neural Message Passing for Quantum Chemistry

Kalashnikov Vadim
vadym.kalashnykov@gmail.com

Core concepts

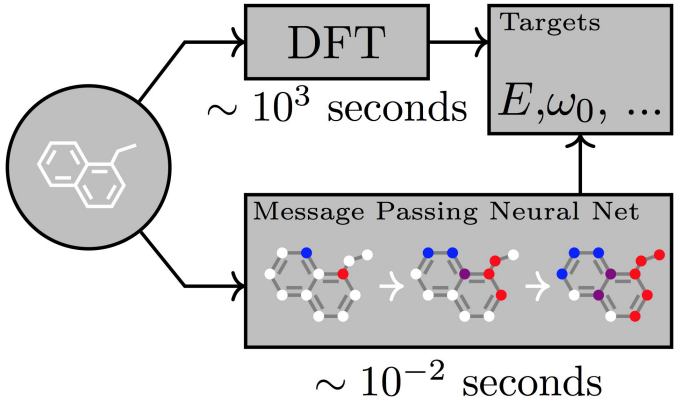
MPNN - general RNN architecture for graph-structured data

DFT(Density functional theory) - slow but precise way to simulate chemical properties

QM9 Dataset - dataset with both DFT-computed targets and ‘true’ targets

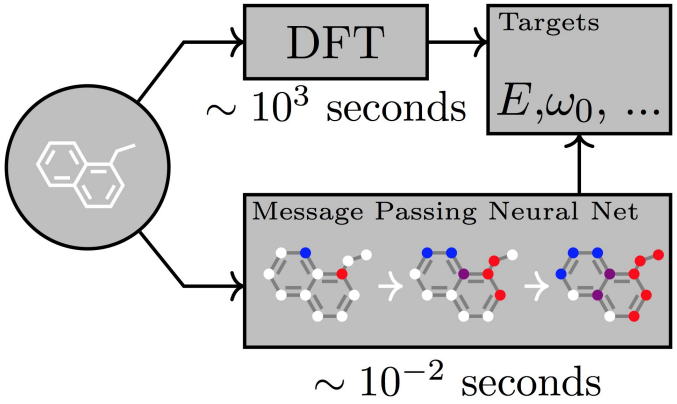
Chemical accuracy - target error that has been established by the chemistry community

Chemical bond



Claimed results

- We develop an MPNN which achieves state of the art results on all 13 targets and predicts DFT to within chemical accuracy on 11 out of 13 targets.
- We develop several different MPNNs which predict DFT to within chemical accuracy on 5 out of 13 targets while operating on the topology of the molecule alone (with no spatial information as input).
- We develop a general method to train MPNNs with larger node representations without a corresponding increase in computation time or memory, yielding a substantial savings over previous MPNNs for high dimensional node representations.



QM9 Dataset

- Hydrogen (H), Carbon (C), Oxygen (O), Nitrogen (N), and Flourine (F) atoms
- Up to 9 heavy (non Hydrogen) atoms.
- 134k organic molecules
- 13 targets

two settings:

- the complete molecular geometry is known (atomic distances, bond angles, etc.)
- only the atom and bond information (i.e. graph) is available as input.

QM9 Dataset

Table 1. Atom Features

Feature	Description
Atom type	H, C, N, O, F (one-hot)
Atomic number	Number of protons (integer)
Acceptor	Accepts electrons (binary)
Donor	Donates electrons (binary)
Aromatic	In an aromatic system (binary)
Hybridization	sp, sp2, sp3 (one-hot or null)
Number of Hydrogens	(integer)

QM9 Dataset

- 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.

Target
mu
alpha
HOMO
LUMO
gap
R2
ZPVE
U0
U
H
G
Cv
Omega

Message Passing Neural Networks

Undirected graph G
Node (vertex) features
Edge features
Node (vertex) hidden states
Edge hidden states (for some models)

Two phases:
1. Message passing phase runs T times and updated hidden states with information from neighbour-nodes.
2. Readout phase to compute graph-level features vector

MPNN is invariant to graph isomorphisms.

Message Passing Neural Networks

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw})$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1})$$

$$\hat{y} = R(\{h_v^T \mid v \in G\})$$

Where M, U, R are all learned differentiable functions (nn, linear projections, concatenation, etc.)

Convolutional Networks for Learning Molecular Fingerprints, Duvenaud et al. (2015)

$$M(h_v, h_w, e_{vw}) = (h_w, e_{vw})$$

$$U_t(h_v^t, m_v^{t+1}) = \sigma(H_t^{\deg(v)} m_v^{t+1})$$

$$R = f\left(\sum_{v,t} \text{softmax}(W_t h_v^t)\right)$$

separately sums over connected nodes and connected edges
unable to identify correlations between edge states and node states

Gated Graph Neural Networks (GG-NN), Li et al. (2016)

$$M_t(h_v^t, h_w^t, e_{vw}) = A_{e_{vw}} h_w^t$$

$$U_t = \mathbf{GRU}(h_v^t, m_v^{t+1})$$

$$R = \sum_{v \in V} \sigma\left(i(h_v^{(T)}, h_v^0)\right) \odot \left(j(h_v^{(T)})\right)$$

A is a learned matrix (d x d), one for each edge label e (discrete edge types).

i and j are both nns.

Molecular Graph Convolutions, Kearnes et al. (2016)

$$M(h_v^t, h_w^t, e_{vw}^t) = e_{vw}^t$$

$$U_t(h_v^t, m_v^{t+1}) = \alpha(W_1(\alpha(W_0 h_v^t), m_v^{t+1}))$$

$$e_{vw}^{t+1} = \alpha(W_4(\alpha(W_2, e_{vw}^t), \alpha(W_3(h_v^t, h_w^t))))$$

edge representations are updated during the message passing phase (!)

alpha is ReLU

W are learned matrices

Deep Tensor Neural Networks, Schutt et al. (2017)

$$M_t = \tanh\left(W^{fc}((W^{cf}h_w^t + b_1) \odot (W^{df}e_{vw} + b_2))\right)$$

$$U_t(h_v^t, m_v^{t+1}) = h_v^t + m_v^{t+1}$$

$$R = \sum_v \mathrm{NN}(h_v^T)$$

MPNN Variants

GG-NN model is a strong baseline.
d - dimension of the hidden representation
n - number of nodes in the graph
G considered to be directed graph (ever undirected edge is splitted)

Input:
set of feature vectors for every vertex
adjacency matrix A:
1. Chemical graph (bond type: single, double, triple, etc.)
2. Distance bins (one hot for bond type + one hot for distance (for non-bonded pairs))
3. Raw distance features (distance + one-hot for bond type)

MPNN Variants

Matrix Multiplication: We started with the message function used in GG-NN which is defined by the equation $M(h_v, h_w, e_{vw}) = A_{e_{vw}} h_w$.

Edge Network: To allow vector valued edge features we propose the message function $M(h_v, h_w, e_{vw}) = A(e_{vw}) h_w$ where $A(e_{vw})$ is a neural network which maps the edge vector e_{vw} to a $d \times d$ matrix.

Virtual Graph Elements

1. “virtual” edge type for pairs of nodes that are not connected

2. “master” node with separate dimension, which is connected to every input node

Readout Functions

1. $$R = \sum_{v \in V} \sigma\left(i(h_v^{(T)}, h_v^0)\right) \odot \left(j(h_v^{(T)})\right)$$

2. Use set2set model from [arXiv preprint arXiv:1511.06391, 2015.]
   takes set of (hidden node state, node feature vector) and constructs graph-level embedding invariant to the order of input tuples (vertex information).
   This embedding is also transformed by nn to get prediction.

Multiple Towers

- break the d dimensional node embeddings h into k different d/k dimensional embeddings and run a propagation step on each of the k copies separately to get temporary embeddings, using separate message and update functions for each copy.
- mix together using nn: $\left(h_v^{t,1}, h_v^{t,2}, \ldots, h_v^{t,k}\right) = g\left(\tilde{h}_v^{t,1}, \tilde{h}_v^{t,2}, \ldots, \tilde{h}_v^{t,k}\right)$

k times faster in theory (with some overhead due to mixing network)
faster in practice (2 times for k = 8 vs k = 1 if n = 9; d = 200)

Training

Each model and target combination was trained using a uniform random hyper parameter search with 50 trials.

- $3 \leq T \leq 8$
- The number of set2set computations $1 \leq M \leq 12$.
- All models were trained using SGD with the ADAM optimizerwith batch size 20 for 3 million steps (540 epochs).
- The initial learning rate was chosen uniformly between $1e-5$ and $5e-4$.
- The QM-9 dataset has 130k molecules in it. 10000 samples for validation, 10000 samples for testing.
- We use the validation set to do early stopping and model selection and we report scores on the test set.
- We minimize the mean squared error between the model output and the target, although we evaluate mean absolute error.

Results

In all of our tables we report the ratio of the mean absolute error (MAE) of our models with the provided estimate of chemical accuracy for that target. Thus any model with error ratio less than 1 has achieved chemical accuracy for that target.

Note, unless otherwise indicated, all tables display result of models trained individually on each target (as opposed to training one model to predict all 13)

Found some evidence that the multi-tower structure improves generalization performance

Results

Table 2. Comparison of Previous Approaches (left) with MPNN baselines (middle) and our methods (right)

Target	BAML	BOB	CM	ECFP4	HDAD	GC	GG-NN	DTNN	enn-s2s	enn-s2s-ens5
mu	4.34	4.23	4.49	4.82	3.34	0.70	1.22	-	**0.30**	0.20
alpha	3.01	2.98	4.33	34.54	1.75	2.27	1.55	-	**0.92**	0.68
HOMO	2.20	2.20	3.09	2.89	1.54	1.18	1.17	-	**0.99**	0.74
LUMO	2.76	2.74	4.26	3.10	1.96	1.10	1.08	-	**0.87**	0.65
gap	3.28	3.41	5.32	3.86	2.49	1.78	1.70	-	**1.60**	1.23
R2	3.25	0.80	2.83	90.68	1.35	4.73	3.99	-	**0.15**	0.14
ZPVE	3.31	3.40	4.80	241.58	1.91	9.75	2.52	-	**1.27**	1.10
U0	1.21	1.43	2.98	85.01	0.58	3.02	0.83	-	**0.45**	0.33
U	1.22	1.44	2.99	85.59	0.59	3.16	0.86	-	**0.45**	0.34
H	1.22	1.44	2.99	86.21	0.59	3.19	0.81	-	**0.39**	0.30
G	1.20	1.42	2.97	78.36	0.59	2.95	0.78	.84[2]	**0.44**	0.34
Cv	1.64	1.83	2.36	30.29	0.88	1.45	1.19	-	**0.80**	0.62
Omega	0.27	0.35	1.32	1.47	0.34	0.32	0.53	-	**0.19**	0.15
Average	2.17	2.08	3.37	53.97	1.35	2.59	1.36	-	**0.68**	0.52

Results

Table 3. Models Trained Without Spatial Information

Model	Average Error Ratio
GG-NN	3.47
GG-NN + Virtual Edge	2.90
GG-NN + Master Node	2.62
GG-NN + set2set	**2.57**

Table 4. Towers vs Vanilla GG-NN (no explicit hydrogen)

Model	Average Error Ratio
GG-NN + joint training	1.92
towers8 + joint training	**1.75**
GG-NN + individual training	1.53
towers8 + individual training	**1.37**

Arxiv

Neural Message Passing for Quantum Chemistry :
https://arxiv.org/pdf/1704.01212.pdf

Order matters: sequence to sequence for sets:
https://arxiv.org/pdf/1511.06391.pdf