PES Challenge

September 8, 2019

1 Description

The goal of the challenge is to model the inter-atomic potential energy surface (PES) for seven small molecules: aspirin, ethanol, malonaldehyde, naphthalene, sallcylic acid, toluene, and uracil. We denote by $r = (r_1, r_2, \ldots, r_N)^T$ the position of atoms in a molecule and denote by $E(r)$ the PES. One should define a unified model for all the molecules. This amounts to a high dimensional regression problem.

The most challenging component of this problem is to respect the symmetry constraint on the PES. In other words, we define the translational, rotational, and permutational operations as

$$T_b E(r) = E(r + b),\ R_U E(r) = E(rU),\ P_\sigma E(r) = E(r_{\sigma(1)}, r_{\sigma(2)}, \ldots, r_{\sigma(N)})$$

respectively. Here $b \in \mathcal{R}^3$ is an arbitrary 3-dimensional translation vector, $U \in \mathcal{R}^{3\times 3}$ is an orthogonal rotation matrix, and σ denotes a permutation of the set of indices. The permutation can only be performed on indices corresponding to the same type of atoms. $E(r)$ should be invariant under the operation T_b, R_U, and P_σ.

One more concept that might be helpful is the atomic force, namely the negative gradient of energy $f_i = -\bigtriangledown_{r_i} E(r)$. We provide the atomic force in the training data, which might help to regularize the fitting process.

2 Dataset

The data source contains 7 folders: asp_data, eth_data, mal_data, nap_data, sal_data, tol_data, ura_data, for 7 different molecules. The data was originally obtained at `http://quantum-machine.org/datasets/`.

In each folder, there are 4 different files written in a plain text format: coord.dat, energy.dat, force.dat, type.dat, denoting the atomic coordinates, the energy, the atomic force, and the atomic type.

We use the following convention of units

Property	Unit
Length	A
Energy	eV
Force	eV/A

Suppose the molecule that we consider has N atoms. The type.dat file contains one line with the type of the N atoms written one by one. The atomic types are integers. Here we use 0, 1, 2, 3 for C, H, O, N, respectively. The coord.dat file contains 10000 lines. Each line provides all the 3 coordinate components of N atoms in 1 frame. The first three numbers are the 3 coordinate components of the first atom, while the second three numbers are the 3 coordinate components of the second atom. The force file force.dat is organized similarly to the coordinate file. The energy file energy.dat contains 10000 lines. Each line provides a number, namely the energy of the corresponding structure.

For each molecule, there are 1000 testing structures in the test folder. The format is the same with the training structures, but the energies and forces are not given.

General introduction to the PES problem and benchmark studies involving this dataset can be found in [5] [4] [1] [2] [3]

3 Evaluation

We adopt RMSE metric, for prediction $\hat{E}_i$ and actual energy E_i, the loss function is

$$L = \sqrt{mean_i (\hat{E}_i - E_i)^2}$$

References

[1] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

[2] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pages 991–1001, 2017.

[3] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8:13890, 2017.

[4] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14):143001, 2018.

[5] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, and E Weinan. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In *Advances in Neural Information Processing Systems*, pages 4436–4446, 2018.