TP Introduction to Machine Learning – GP3 Porject

NOMAD 2018 Reproduction Challenge – Predicting Transparent Conductors

1 Problem Description

The Novel Materials Discovery (NOMAD) Center of Excellence (CoE) is a research lab dedicated to advance computational materials science to enable systematic studies and predictions of novel materials to solve urgent energy, environmental, and societal challenges. From the late 2017 to 2018 they launched a competition in Kaggle where the competitors proposed Machine Learning Solutions to a problem related to discovery of new materials. The competition was based on a dataset of 3,000 $\left(\mathrm{Al_xGa_yIn_{1-x-y}}\right)_2\mathrm{O_3}$ compounds where $(x + y + z = 1)$, Its aim was to identify the best machine learning (ML) model for the prediction of two key physical properties that are relevant for optoelectronic applications: the electronic bandgap energy and the crystalline formation energy.

For the discovery of materials for a targeted application, it is required to explore a big amount of compositional and configurational degrees of freedom. Density functional theory (DFT) is commonly used but a vast amount of computational power is required for exploring a big configurational space of alloys, for this reason machine learning (ML) promises to accelerate the discovery of new materials by evaluating candidate compounds at lower computational cost than electronic structure approaches.

The competition is centered around the discovery of a transparent conductor because they are an important class of compounds that are both electrically conductive and have a low absorption in the visible range, which are typically competing properties. A combination of both of these characteristics is key for the operation of a variety of technological devices such as photovoltaic cells, light-emitting diodes for flat-panel displays, transistors, sensors, touch screens, and lasers. However, only a small number of compounds are currently known to display both transparency and conductivity suitable enough to be used as transparent conducting materials. Aluminum, gallium, indium sesquioxides are some of the most promising transparent conductors because of a combination of both large bandgap energies, which leads to optical transparency over the visible range, and high conductivities. These materials are also chemically stable and relatively inexpensive to produce. Alloying of these binary compounds in ternary or quaternary mixtures could enable the design of a new material at a specific composition with improved properties over what is current possible.

The training set is 2400 compounds, while the test set is 600 compounds. For each line of the CSV file, the corresponding spatial positions of all of the atoms in the unit cell expressed in Cartesian coordinates are provided as a separate file in the subfolders train/ and test/. files with spatial information about the material are provided according to the id in the respective csv files as follows: **{train/test}/{id}/geometry.xyz**.

The python script "**read_xyz.py**" uses the id entry in the csv file to print the structure object from ASE (https://wiki.fysik.dtu.dk/ase/about.html).

The following information has been included for each compound in the training and test sets:

- Spacegroup label identifying the symmetry of the material
- Total number of Al, Ga, In and O atoms in the unit cell (Ntotal)
- Relative compositions of Al, Ga, and In (x, y, z)
- Lattice vectors and angles: lv1, lv2, lv3 which are lengths given in units of angstroms (i.e., 10^−10 meters) and α, β, γ, which are angles in degrees between 0° and 360°.

A domain expert will understand the physical meaning of the above information but those with a data mining background may simply use the data as input for their models.

The task for this competition is to predict two target properties:

1. Formation energy (an important indicator of the stability of a material)
2. Bandgap energy (an important property for optoelectronic applications)

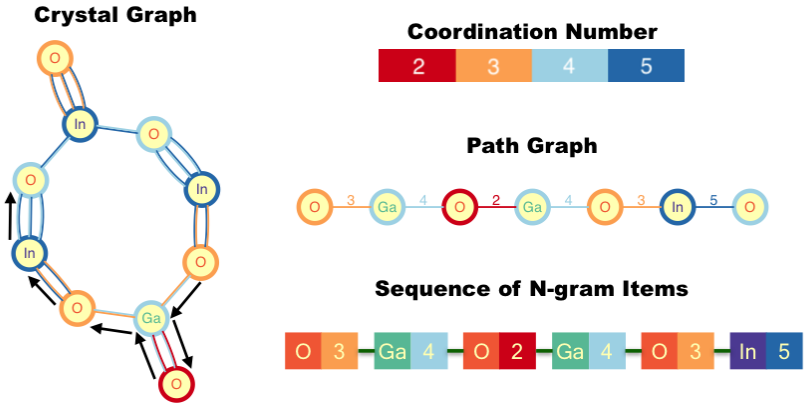
2 METHODOLOGY

2.1 FIRST STEP: PRE-PROCESS THE DATA

The first part of the problem consists in pre-processing the data. From the features already provided by the dataset it is useful to do some kind of encoding applying domain knowledge in order to have features that facilitate the regression problem. Two of the proposed feature processing strategies are:

2.1.1 Crystal Graph n-grams

The 1st place winning solution was obtained using the metal-oxygen coordination number derived from the number of bonds that are within the sum of the ionic Shannon experimental radii (which were enlarged by 30-50% depending on the crystal structure type). These ionic bonds are then used for building a crystal graph, where each atom is a node in the graph and the corresponding edges between nodes are defined by the ionic bond, which are shown as coordination numbers for each atom for a sequence of 6 atoms.



2.1.2 SOAP-based Descriptor

In the 3rd place winning solution, the smooth overlap of atomic positions (SOAP) kernel developed by Bartók et al. that incorporates information on the local atomic environment through a rotationally

integrated overlap of neighbor densities. The SOAP kernel describes the local environment for a given atom (i) through the sum of Gaussians centered on each of the atomic neighbors (j) within a specific cutoff radius (rij):

$$\rho_i(r) = \sum_j \exp\left(\frac{-(r - r_{ij})^2}{2\sigma_{\text{atom}}^2}\right) f_{\text{cut}}(r_{ij})$$

The functions for getting the **n-gram** and **SOAP** descriptors will be provided.

2.2 Second Step: Use a NN for regression

There are many methods for performing regression, for the project and because this is what we studied in the course, you will have to build a neural network for predicting the band gap energy and the formation energy given the processed features, from this you will be able to predict these properties for the materials in the test set and determine the best transparent conductors. The winner of the third place of the competition employed SOAP + NN so you are already in a good path, you can also use n-gram + NN and get similar results if not better.

For building the Neural Network model you have to use pytorch, the architecture used for this problem are simple MLPs, you can find details about the hyperparameters of the model in the additional documentation provided. The metric used for evaluation of the performance of the model is the root mean squared logarithmic error:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log\left(\frac{\hat{y}_i + 1}{y_i + 1}\right)\right)^2}$$

Where N is the total number of samples. The error is calculated as the log ratio of the predicted target property $\hat{y}_i$ and the corresponding reference value y_i. The error for both the formation energy and bandgap energy is then averaged for a final assessment of the model performance. The log ratio of the errors is a convenient choice because it prevents the bandgap, which is an order of magnitude larger than the formation energy, from dominating an analysis of the predictive capability of each model.

Note: It is recommended to train 2 NNs for predicting the Band Gap and formation energy separately, these networks will be identical and use the same inputs but trained with their respective objectives. Also it is sufficient to use the MSE as training target, but remember to calculate de RMSLE for evaluation.

3 Instructions for doing the practice

You should do everything in a jupyter notebook which should be reproductible for evaluation, you can use custom .py modules but they should be callable in the notebook and remember to include this in your submission. Also, at the end you should write a report about your implementation with the following sections:

- Introduction: describe in your words what the problem consists of.
- Methodology: describe the methods you used, details of the neural network you used for the problem, hyperparameters (learning rate, number of epochs, batch size).

- Results: Here you need to present the training loss curve of the model. Also, you now have to use the trained NN for determining the formation and band gap energies for the materials in the test set, do a table indicating these properties for the best 5 materials.
- Conclusions: Discuss the results obtained, what advantage does the use of ML have for this problem? What effects does the parameters, feature processing method, size of the neural network have over the performance of the model? In what ways you think these results may be improved? What where the difficulties you encountered doing this project and following the course material? What have you learned by doing this and the course? Finally, state how you imagine Machine Learning will be useful for you in the future, in your discipline, whether you become a practitioner or someone else do it in your workplace, give a concrete example.

The report can be either in English or French.

4 Implementation Tips

- You can use pytorch-lightning, a higher level API that reduces the need to write boilerplate and have cleaner code, you can see the docs here https://pytorch-lightning.readthedocs.io/en/latest/ , also you can ask me on how to use it.
- Use google colab since most of the libraries you need will be already installed and can use the free gpu.
- If you will stop working on the project in colab and want to continue later, remember to save anything you have accomplished. If you are logged in in your google account the notebook will be saved in drive, if not you can download it before closing it.
- After you have trained the model, you can save the parameters. Pytorch-Lightning can do this automatically or you can do it manually also with pytorch.
    - https://pytorch-lightning.readthedocs.io/en/latest/common/weights_loading.html (for lightning)
    - https://pytorch.org/tutorials/beginner/basics/saveloadrun_tutorial.html