Costrita Alexandru – GP3 – ESITech

Report

Machine Learning Project

Introduction/Problematics

This project represents the use of machine learning into the subject of identification of a model for the prediction of two physical properties: the electronic bandgap energy and the crystalline formation energy. These two properties are essential and key for optoelectronic applications. We will do that by finding the best compound of $(Al_xGa_yIn_{1-x-y})_2 O_3$ with (x+y+z = 1) which is characterized by its conductivity and transparency.

Method used

The first part of the problem consists in pre-processing the data. We need this step to find an encoding to simplify the data and further facilitate the regression problem. We will use a specific type of Networks which is NN (Neural Networks). We will use as well the SOAP-based descriptor to pre-process the data. The SOAP descriptor characterizes the local environment for a given atom influenced by the other atoms of the molecule.

Firstly, we use pytorch to build the NN with a simple Multi-Layers Perception (MLP).

We have to mention that in this project we took advantage and inspired from the 3rd place winner in the competition that the Novel Materials Discovery (NOMAD) Center of Excellence (CoE) launched from 2017 to 2018 on the online platform Kaggle. They used the SOAP descriptor and the Neural Networks for this problem.

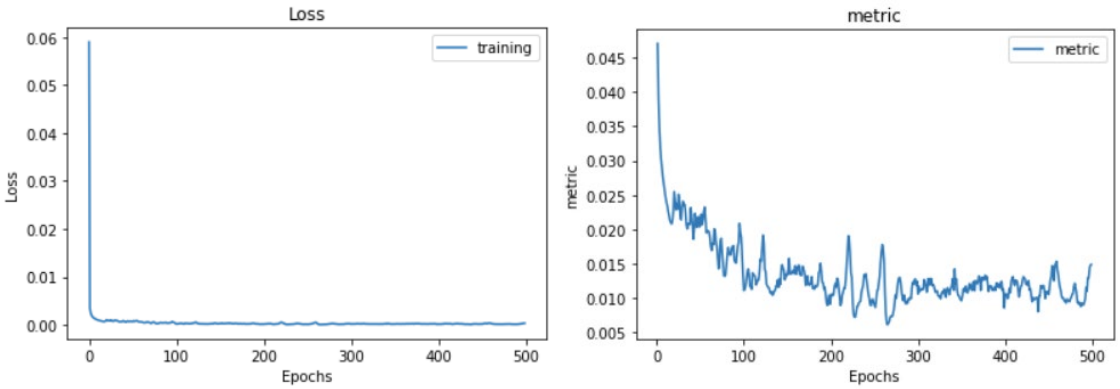
From the Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition article: “The average SOAP features were used in a three-layer feed-forward NN by using PyTorch99 with batch normalization and 20% dropout in each layer. For predicting the bandgap energies and the formation energies, the initial layer had 1024 and 512 neurons, respectively. In both cases, the remaining two layers had 256 neurons each. The neural networks were trained for 200 or 250 epochs, and the final predictions were based on 200 independently trained NNs by using the same architecture but with different initial weights.”.

We used a set of predetermined hyperparameters which are presented below:

- 500 epochs for scientific and machine learning behavior interest
- ReLU (Rectified Linear Unit) function
- Tanh (hyperbolic tangent) function
- MSE (Mean Squared Error) as a loss function
- RMSLE (Root Mean Squared Logarithmic Error) as the metric function.
- Adam optimizer (learning rate set as 0,005)

Obtained results/Discussion

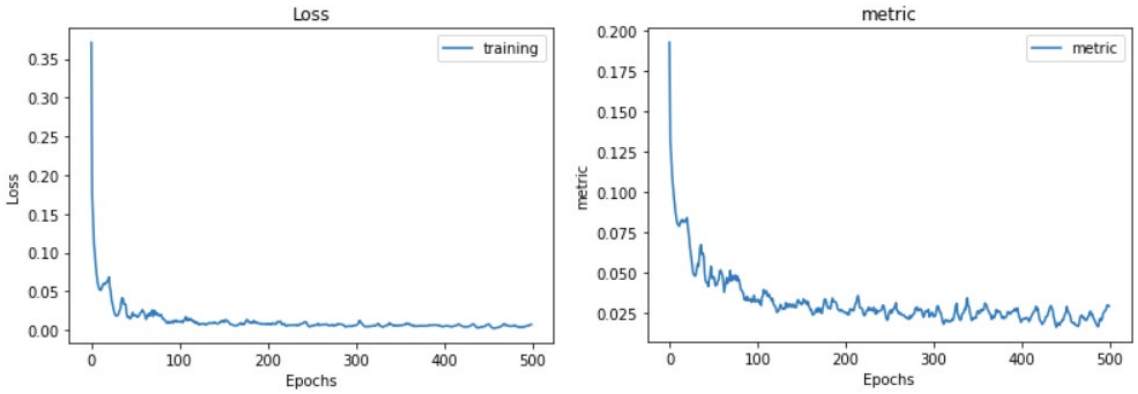
We obtain the following graphs for the formation energy problem:



The trained NN gives has an accuracy of 93.5% with a safety interval of ±0,065.

The smallest given Eform values are: ['0.0158'; '0.0231'; '0.0235'; '0.0235'; '0.0241'] with their indexes as: [532; 236; 586; 267; 585]

We obtain the following graphs of the energy gap problem:



And once againt the NN gives has an accuracy of 95.33% with a safety interval of ±0,65.

The smallest given Gap values are: ['0.1609'; '0.1708'; '0.1829'; '0.1922'; '0.2040'] with their indexes as: [397; 438; 164; 576; 441].

Conclusion

The present work can be considered as a quick overview on what compounds fall into the classifications of good materials with a decent conductivity and transparency. This work is just a brief touch onto the difficult and complex subject of the electronic bandgap energy and the crystalline formation energy in particular $(Al_xGa_yIn_{1-x-y})_2 O_3$ materials.

We can certainly say that Neural Networks are just perfect for trained datasets. However, they tend to have some issues with new datasets which we think is ok since the dataset presents a huge amount of data that needs to be processed and to be trained on before manipulating with it.

We obtained good results which gave us approximate and a good look at what small values are alike and whether or not they are coherent with the context of the given problem.

It was a very informative course and I enjoyed listening and understanding the proposed problems. I tried my best into solving these problems and sometimes with the help of my classmates.

It seemed as an innovative engineer-problem-solving environment and I felt as a real engineer working in a group of other intelligent and very capable people over a difficult problem using methods we never used before.

As of the industry and how Neural Networks can be applied – a good example can be weather predictions. Why weather predictions? - well it's one the most difficult and chaotic domains there is at the moment and I would really love to see how such a complex and complicated system of Networks can solve such unreal and absolute chaotic problems as the prediction of the weather – there are so many hyperparameters to take into consideration, it seems almost unreal.