Summer Student Project: Crystal Structure Map

By: Lauri Himanen
Date: Wednesday 20th June, 2018

Introduction

When working with large datasets of heterogeneous crystal structures, understanding how the different structures are distributed in the structural and chemical space is an important task. With the advent of large material repositories and datasets, this task is becoming vital in order to effectively visualize structures in large repositories e.g. for search purposes or in order to analyze the composition of a dataset used in machine learning.

Crystal structures are often described through their symmetry properties, such as space group number, bravais lattice, crystal lattice or Wyckoff positions. Such a mapping helps, but can be misleading because the symmetry properties are discontinuous with respect to small changes in the coordinates of atoms in the unit cell, and the results are often difficult to interpret for the non-expert.

In this project we will investigate an alternative approach, where the structural and chemical features are represented as a feature vector that is projected to an easily visualizable lower dimensional space by using data clustering or projection algorithms which are a part of unsupervised machine learning methods.

Objectives and methods

The final objective of the project is to create a method for mapping and visualizing large datasets of crystal structures. In order to test the ideas we have prepared a set of roughly 30 000 crystal structures from the AFLOW database[**aflow1**]. The project is divided into three key topics: choosing the descriptor, clustering and projection, and presentation. Depending on the time and your interests, you can choose how you want to divide your time on these tasks.

Choosing the descriptor

In order for clustering methods to work, we need to define how crystal structures are presented to the algorithm. Typically such structures are represented by using a unit cell, atom positions and chemical elements of atoms. Such an input feature is however not directly suitable for clustering or machine learning in general. There are multiple reasons for this, including the fact than an XYZ-coordinate representation is not invariant to rotations or translations. In order to address this problem, multiple feature transformation have been proposed[**cm**, **mbtr**, **soap**, **voronoi**, **acsf**]. These transformations are here rererred to as descriptors. We will start by using the many-body tensor representation (MBTR) [**mbtr**], but can investigate other models as needed. MBTR captures the crystal structure as a "spectrum" of atomic distances and angles, as illustrated in image 1.

Clustering and projection

The chosen descriptor effectively captures the structural features of a crystal in a n-dimensional space, which cannot be visualized or understood directly. In order to make sense of these feature vectors, we will use an unsupervised machine learning method that can be used to transform the data into a 2D or 3D.

A simple approach is to define an euclidean metric in the descriptor space, and calculate distances between crystal as the euclidean distance between two descriptor vectors. This approach can be used together with clustering algorithms such as DBSCAN[**dbscan**] and graph layout algorithms to produce a map of the crystal space.

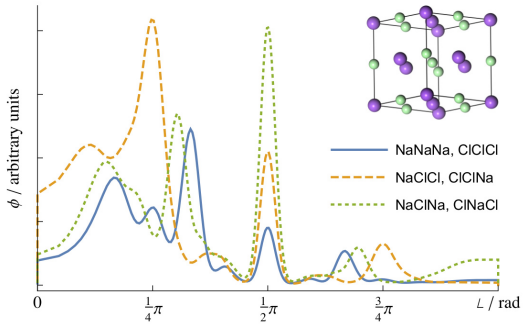

Figure 1: The MBTR spectrum for angles in a sodium-chloride crystal.

An alternative approach is to perform a projection of the n-dimensional space into a lower dimensional one that can be more easily visualized. There are multiple approaches for doing such tasks, including PCA[**pca**], t-SNE[**tsne**] and UMAP[**umap**]. An example of t-SNE clustering of hand-writtend digits is given in figure 2.

There is a very nice introduction to different visualization methods given here: `http://colah.github.io/posts/2014-10-Visualizing-MNIST/`.

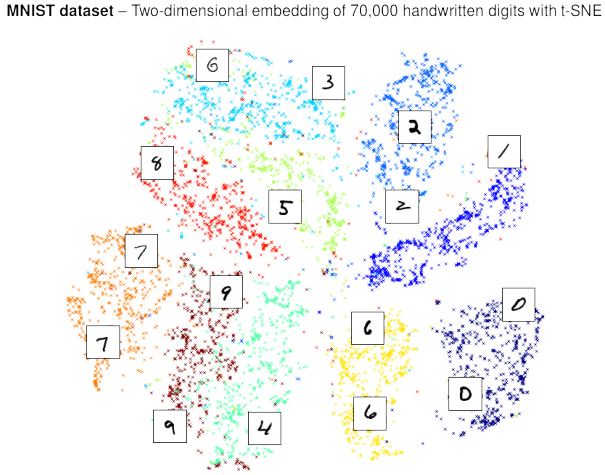

Figure 2: Illustration of t-SNEs capability to reduce multidimensional input into meaningful categories in a lower-dimensional space. The input data is an MNIST dataset of hand-written digits as 28x28 grayscale images, which has been clustered into groups representing different digits. Original image from `http://alexanderfabisch.github.io/t-sne-in-scikit-learn.html`

Presentation
In order to make sense of the dataset we need to present the results in 2D or 3D space. For this purpose we could create an interactive web-based presentation that allows one to visualize the data in a browser. To make the interpretation easy, we can provide a visualization of the individual crystals directly in this tool, and encode other crystal properties (energy, band gap, symmetry properties, etc.) as colours or sizes of the nodes in the visualization.

Getting started

- Understanding what unsupervised machine learning is, and how it differs from supervised learning. Depending on your interest also getting to know other machine learning techniques.

  - Machine learning course in Coursera, by Andrew Ng: https://www.coursera.org/learn/machine-learning.
  - https://en.wikipedia.org/wiki/Machine_learning.

- Reading about dimensionality reduction techniques for visualization:
  - Visualizing MNIST: An Exploration of Dimensionality Reduction: http://colah.github.io/posts/2014-10-Visualizing-MNIST/

- Reviewing basic theory behind crystal structures:
  - E.g. Girolami, Gregory S. (2016). X-Ray Crystallography. University Science Books, Section 1 https://app.knovel.com/web/toc.v/cid:kpXRC00001/viewerType:toc/root_slug:x-ray-crystallography/url_slug:x-ray-crystallography (or any basic book on crystallography).

- Reading about the rational and techniques behing descriptors:
  - E.g. Many-Body Tensor Representation for Machine Learning of Atomistic Systems https://arxiv.org/abs/1704.06439