Transformer-Based Interatomic Potentials for Efficient Materials Modeling

Your Name
Your Affiliation

September 6, 2025

Abstract

Machine learning interatomic potentials (MLIPs) have emerged as powerful tools for materials modeling, offering near-quantum accuracy at significantly reduced computational cost. However, the computational expense associated with large MLIP models remains a bottleneck, limiting their applicability to complex systems and long timescales. In this work, we explore the use of transformer neural networks as MLIPs, leveraging their ability to capture long-range interactions and handle complex data representations. We further investigate the application of low-rank approximation techniques to the attention mechanism within the transformer architecture to improve computational efficiency. Our results demonstrate that transformer-based MLIPs can achieve competitive accuracy compared to existing MLIPs, while offering the potential for significant speedups through compression.

1 Introduction

The accurate and efficient simulation of materials at the atomic scale is crucial for a wide range of scientific and engineering applications, including materials discovery, design, and optimization. First-principles methods, such as density functional theory (DFT), provide a high level of accuracy but are computationally expensive, limiting their application to relatively small systems and short timescales. Machine learning interatomic potentials (MLIPs) have emerged as a promising alternative, offering near-DFT accuracy at a fraction of the computational cost [**?**, **?**, **?**].

Existing MLIPs, such as neural network potentials (NNPs) [**?**], Gaussian approximation potentials (GAPs) [**?**], and moment tensor potentials (MTPs) [**?**], have shown remarkable success in modeling a variety of materials and properties. However, these methods still face challenges in terms of computational cost, transferability, and the ability to accurately capture long-range interactions. The computational cost of evaluating MLIPs can become significant for

large systems or long molecular dynamics simulations. Furthermore, the transferability of MLIPs, i.e., their ability to accurately predict properties of materials that are different from those used in the training data, remains a concern.

In recent years, transformer neural networks have revolutionized the field of natural language processing and have shown great promise in other domains, including computer vision and scientific computing [**?**]. Transformers excel at capturing long-range dependencies and handling complex data representations through the use of self-attention mechanisms. These properties make transformers a promising candidate for developing MLIPs that can overcome some of the limitations of existing methods.

In this work, we explore the use of transformer neural networks as MLIPs. We propose a novel approach for representing atomic environments as input to a transformer architecture and train the transformer to predict the potential energy of the system. Furthermore, we investigate the application of low-rank approximation techniques to the attention mechanism within the transformer architecture to improve computational efficiency, building upon the work of Vorotnikov et al. [**?**].

The main contributions of this work are:

- A novel transformer-based architecture for learning interatomic potentials.
- A method for representing atomic environments as input to a transformer.
- An investigation of low-rank approximation techniques for compressing the attention mechanism in transformer-based MLIPs.
- A demonstration of the accuracy and efficiency of our approach on benchmark materials systems.

2 Related Work

Machine learning interatomic potentials (MLIPs) have become a central focus in computational materials science, offering a pathway to overcome the computational limitations of traditional first-principles methods. Several approaches have been developed, each with its strengths and weaknesses.

Neural Network Potentials (NNPs) pioneered by Behler and Parrinello [**?**], utilize artificial neural networks to map atomic environments to potential energy. NNPs have demonstrated impressive accuracy in various systems, but their computational cost can be substantial, especially for large networks.

Gaussian Approximation Potentials (GAPs) introduced by Bart'ok et al. [**?**], employ Gaussian process regression to construct potential energy surfaces. GAPs offer a good balance between accuracy and computational efficiency, but their performance can be sensitive to the choice of kernel function and hyperparameters.

Moment Tensor Potentials (MTPs) developed by Shapeev [**?**], utilize a basis of moment tensors to represent atomic environments. MTPs are sys-

tematically improvable and have been successfully applied to a wide range of materials, but their computational cost can be high for complex systems.

More recently, there has been growing interest in applying deep learning techniques to materials modeling. For example, graph neural networks (GNNs) have been used to learn interatomic potentials by representing atomic environments as graphs. However, GNNs typically focus on local interactions and may struggle to capture long-range dependencies.

Transformer neural networks, originally developed for natural language processing, have shown remarkable capabilities in capturing long-range dependencies and handling complex data representations [**?**]. Transformers have been applied to various scientific domains, including protein folding and drug discovery. However, to the best of our knowledge, there has been limited work on using transformers directly as MLIPs.

Vorotnikov et al. [**?**] recently explored the use of low-rank matrix and tensor approximations to compress existing MLIPs, such as MTPs. Their work demonstrated that significant compression can be achieved without sacrificing accuracy. We build upon this work by investigating the application of low-rank approximation techniques to the attention mechanism within transformer-based MLIPs.

Our work differs from previous approaches in that we propose a novel transformer-based architecture for learning interatomic potentials and investigate the application of compression techniques to improve its efficiency. This approach has the potential to combine the strengths of transformers in capturing long-range interactions with the efficiency of compressed MLIPs, leading to a powerful new tool for materials modeling.

3 Methodology

Our approach involves training a transformer neural network to predict the potential energy of a materials system, given the atomic coordinates as input. We represent the local environment of each atom using the coordinates of its neighbors within a certain cutoff radius. We then apply low-rank approximation techniques to the attention mechanism within the transformer to improve computational efficiency.

3.1 Representation of Atomic Environments

For each atom i in the system, we consider all neighboring atoms j within a cutoff radius R_c. We represent the local environment of atom i as a set of vectors $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$, where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the coordinates of atoms i and j, respectively. The number of neighbors within the cutoff radius will vary for each atom. To handle this variable number of neighbors, we pad the set of vectors with zero vectors up to a maximum number of neighbors N_{max}. This results in a fixed-size input for the transformer. Each vector $\mathbf{r}_{ij}$ is then embedded into a higher-dimensional space using a linear embedding layer.

3.2 Transformer Architecture

We employ a standard transformer encoder architecture [**?**]. The encoder consists of multiple layers, each containing a multi-head self-attention mechanism and a feed-forward network. The multi-head self-attention mechanism allows the transformer to capture long-range dependencies between atoms in the system. The feed-forward network further processes the information and provides non-linearity. The output of the transformer encoder is a set of hidden state vectors, one for each atom in the system. We then sum these hidden state vectors and pass the result through a linear layer to predict the total potential energy of the system.

3.3 Low-Rank Approximation of Attention Matrices

The self-attention mechanism in the transformer involves computing an attention matrix for each layer. The attention matrix captures the relationships between all pairs of atoms in the system. Computing the attention matrix requires $O(N^2)$ operations, where N is the number of atoms. To reduce this computational cost, we apply low-rank approximation techniques to the attention matrices. Specifically, we approximate each attention matrix $\mathbf{A}$ with a low-rank matrix $\mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are matrices with r columns, and $r << N$ is the rank of the approximation. This reduces the computational cost of computing the attention matrix from $O(N^2)$ to $O(Nr)$. We use singular value decomposition (SVD) to compute the low-rank approximation of the attention matrices.

3.4 Training Procedure

We train the transformer-based MLIP on a dataset of atomic configurations and their corresponding potential energies and forces, obtained from DFT calculations. The loss function is a combined loss, consisting of the mean squared error (MSE) on the energies and the MSE on the forces:

$$L = \frac{1}{N_{configs}} \sum_{i=1}^{N_{configs}} \left[(E_i^{pred} - E_i^{DFT})^2 + \frac{1}{N_{atoms}} \sum_{j=1}^{N_{atoms}} ||\mathbf{F}_{ij}^{pred} - \mathbf{F}_{ij}^{DFT}||^2 \right] \tag{1}$$

where E_i^{pred} and E_i^{DFT} are the predicted and DFT energies of configuration i, respectively, $\mathbf{F}_{ij}^{pred}$ and $\mathbf{F}_{ij}^{DFT}$ are the predicted and DFT forces on atom j in configuration i, respectively, $N_{configs}$ is the number of configurations in the training set, and N_{atoms} is the number of atoms in the system. We use the Adam optimizer to train the transformer.