基于迁移学习的钙钛矿材料带隙预测*

孙涛[1)] 袁健美[1)2)†]

1) (湘潭大学数学与计算科学学院, 湘潭 411105)

2) (湘潭大学, 智能计算与信息处理教育部重点实验室, 湘潭 411105)

(2023 年 6 月 22 日收到; 2023 年 7 月 31 日收到修改稿)

针对快速获取钙钛矿材料带隙值的问题, 建立特征融合神经网络模型 (CGCrabNet), 利用迁移学习策略对钙钛矿材料的带隙进行预测. CGCrabNet 从材料的化学方程式和晶体结构两方面提取特征, 并拟合特征和带隙之间的映射, 是一个端到端的神经网络模型. 在开放量子材料数据库中数据 (OQMD 数据集) 预训练的基础上, 通过仅 175 条钙钛矿材料数据对 CGCrabNet 参数进行微调, 以提高模型的稳健性. 数值实验结果表明, CGCrabNet 在 OQMD 数据集上对带隙的预测误差比基于注意力的成分限制网络 (CrabNet) 降低 0.014 eV; 本文建立的模型对钙钛矿材料预测的平均绝对误差为 0.374 eV, 分别比随机森林回归、支持向量机回归和梯度提升回归的预测误差降低了 0.304 eV、0.441 eV 和 0.194 eV; 另外, 模型预测的 $SrHfO_3$ 和 $RbPaO_3$ 等钙钛矿材料的带隙与第一性原理计算的带隙相差小于 0.05 eV, 这说明 CGCrabNet 可以快速准确地预测钙钛矿材料的性质, 加速新材料的研发过程.

关键词: 特征融合神经网络, 回归模型, 带隙预测, 迁移学习

PACS: 89.90.+n **DOI:** 10.7498/aps.72.20231027

1 引 言

近年来, 随着密度泛函理论的不断完善和计算机计算能力的逐渐提升, 第一性原理高通量计算在预测新材料和优化材料性质等方面取得较大的发展[1]. 然而, 基于第一性原理计算在快速计算时的准确性是有限的, 在需要获取高精度的计算结果的时候, 利用该方法所需要的计算量是很大的, 巨大的计算量成为了高通量计算的瓶颈. 对此, 寻找一种更快、消耗计算量更少的方式来预测新材料的属性, 则可加快新材料的研发过程. 近年来, 随着机器学习技术的发展, 可以有效地对数据和目标标签之间复杂的非线性关系进行建模, 已经被广泛应用到材料信息学中[2].

2018 年, Xie 和 Grossman [3] 提出了一种晶体图卷积神经网络 (crystal graph convolutional neural networks, CGCNN), 预测了带隙等多种晶体材料的性质, 并且预测的平均绝对误差要低于或接近基于密度泛函理论 (Density functional theory, DFT) 的方法. 2019 年, Chen 等[4] 提出了一种通用的材料图网络 (MatErials graph network, MEGNet) 来预测分子和晶体的性质. 在 QM9 分子数据集上对 13 个分子属性进行测试, 提高了对大部分分子属性的预测精度. 2020 年, Karamad 等[5] 提出一种轨道图卷积神经网络 (orbital graph convolutional neural network, OGCNN), 在 CGCNN 模型的基础上加入了原子轨道相互作用特征, 在多个晶体材料数据库上验证模型的性能, 证明了模型具有高精度的晶体性质预测能力.

对于缺乏材料结构信息的数据集, 研究人员尝试不利用材料的结构信息, 通过其他的材料信息如

* 湖南省自然科学基金 (批准号: 2023JJ30567, 2021JJ30650) 资助的课题.

† 通信作者. E-mail: yuanjm@xtu.edu.cn

元素组成等对材料性质进行预测. 2018 年, Jha 等[6]提出一种仅通过元素组成来预测材料化学性质的深度学习模型 ElemNet. 在 OQMD 数据库构成的数据集上对每个材料化合物的最低形成焓进行预测, 能够以较快的速度和较高的精度预测材料性质. 2020 年, Goodall 和 Lee[7] 提出一种从化学计量比中深度表征学习方法 Roost, 它将化学计量式视为元素之间的密集加权图, 利用消息传递操作在组成材料的各元素之间传递信息, 从而更新元素的组成. 通过在 OQMD 数据库和 MP 数据库构成的数据集上进行实验测试, 发现 Roost 可以有效地对缺少晶体材料结构的数据进行性质预测任务. 2021 年, Wang 等[8] 提出一种基于注意力的成分限制网络 (compositionally restricted attention-based network, CrabNet), 在 28 个基准数据集上进行测试, 大多数的预测精度都与其他方法接近或更高.

钙钛矿材料的化学通式为 ABX_3, 是一类结构相似的化合物的总称. 钙钛矿材料具有合适的且可调的带隙宽度等优良的性质, 是太阳能电池、光催化材料和热电器件等领域的候选材料[9]. 带隙是一个重要的物理性质, 它一开始就划定了该材料的应用范围. Guo 和 Lin [10] 利用机器学习模型, 预测了无铅双钙钛矿卤化物的带隙性质. Gao 等[11] 利用梯度提升回归算法从 5796 种无机双钙钛矿材料中筛选带隙合适的材料, 得到 K_2NaInI_6 和 Na_2MgMnI_6 两种候选材料, 并且 DFT 计算证实了热稳定性和良好的光学性能. 对钙钛矿材料的带隙快速预测可以减小候选材料的搜索空间, 加快新材料的研发过程.

本文构建了一个深度学习模型 CGCrabNet 用于研究钙钛矿材料的带隙性质. 该模型通过输入化学式和结构信息, 使材料的带隙值作为模型输出, 通过反向传播算法更新模型参数. 首先在 OQMD 数据集中预训练了 CGCrabNet 模型, 并且与其他深度学习带隙预测模型进行比较. 其次将预训练的 CGCrabNet 模型在钙钛矿材料数据集上进行微调. 最后, 将本文建立的深度学习模型用于对钙钛矿材料的带隙进行预测, 并与随机森林回归、支持向量回归和梯度提升回归的预测结果进行对比分析.

2 特征融合神经网络模型

2.1 模型架构介绍

本文建立了一种基于化学式和晶体结构作为原始输入的特征融合神经网络模型, 简称为 CGCrabNet 模型. 模型的数学表达式为

$$z = f(\boldsymbol{x}_1, \boldsymbol{x}_2, y_1, y_2, \theta), \tag{1}$$

其中 z 为带隙值, f 表示神经网络模型, $\boldsymbol{x}_1 \in \mathbb{R}^{N_f}$ 表示元素信息矩阵, $\boldsymbol{x}_2 \in \mathbb{R}^{N_f}$ 表示原子占比矩阵, N_f 表示化学式中元素种类数, $y_1 \in \mathbb{R}^{p\times I}$ 表示节点特征, 每个节点的特征由 One-Hot 编码表示, y_2 表示相邻节点信息, p 表示晶体图中节点总数, I 表示节点特征维度, θ 表示模型参数.

CGCrabNet 模型对化学式信息的特征提取模块参考 CrabNet 模型[8] 构建. 模型输入的化学式用于构造元素信息矩阵 $\boldsymbol{x}_1$ 和原子占比矩阵 $\boldsymbol{x}_2$, 其中元素信息矩阵是由组成化学式的所有元素的原子序数构成的, 原子占比矩阵是由化学式中每个元素对应原子数占晶体总原子数的比例分数构成的. 例如 SiO_2 的元素信息矩阵为 $[8\ \ 14]^{\mathrm{T}}$, 原子占比矩阵为 $[0.67\ \ 0.33]^{\mathrm{T}}$.

对于元素信息矩阵, 使用元素嵌入将原子序数表示成固定维度的向量, 向量维度取决于通过测试确定的元素嵌入方法. 对于原子占比矩阵, 通过分数编码将其表示为一个 d_m 维的向量. 分数编码灵感来源于 Vaswani 等[12] 提出的位置编码器, 是将一个 0 到 1 之间的分数表示成一个固定维度向量的过程. 元素特征构造中的全连接层的作用是进行维度变换, 使其可以和分数编码后的向量相加, 得到化学式中每个元素的元素特征向量. 然后将元素特征向量经过 N 个多头注意力机制, 多头注意力机制的头数用 n 表示, 这样可以使化学式的每个元素特征向量都包含化学式中其他元素信息. 最后在多头注意力机制后连接一个全连接残差网络, 用来得到化学式中所有元素的元素原贡献向量 $\boldsymbol{p}$、元素不确定性向量 $\boldsymbol{u}$ 和元素对数向量 $\boldsymbol{l}$, 通过 (2) 式和 (3) 式来计算 z_1 和偶然不确定性 S:

$$z_1 = \mathrm{mean}\left(\boldsymbol{p} \otimes \sigma\left(\boldsymbol{l}\right)\right), \tag{2}$$

$$S = \mathrm{mean}\left(\boldsymbol{u}\right), \tag{3}$$

其中 $\sigma\left(\boldsymbol{l}\right) = \dfrac{1}{1+\mathrm{e}^{-l}}$, $\mathrm{mean}(\cdot)$ 表示求向量各分量均值.

对于输入的晶体结构, 将其转化成晶体图数据, 包括节点特征 y_1 和相邻节点信息 y_2. 首先由节点嵌入层将每个节点 i 的 One-Hot 特征变成一个连续、密集的分布式特征向量 $\boldsymbol{v}_i$. 然后考虑相邻节

点信息 y_2, 利用图卷积层通过非线性图卷积函数迭代更新原子特征向量 $\boldsymbol{v}_i$, 更新公式为

$$\boldsymbol{v}_i^{(t+1)} = \boldsymbol{v}_i^{(t)} + \sum_{j,k} \sigma\left(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_f^{(t)} + \boldsymbol{b}_f^{(t)}\right) \odot \left(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_s^{(t)} + \boldsymbol{b}_s^{(t)}\right), \tag{4}$$

其中 $\boldsymbol{z}_{(i,j)_k}^{(t)} = \boldsymbol{v}_i^{(t)} \oplus \boldsymbol{v}_j^{(t)}$, 由节点 i 和节点 j 的特征向量拼接得到. 对 $\boldsymbol{z}_{(i,j)_k}^{(t)}$ 分别通过两个全连接神经网络, σ 表示 sigmoid 函数, $\odot$ 表示两个张量对应元素进行乘积, $\boldsymbol{W}_f^{(t)}$, $\boldsymbol{W}_s^{(t)}$, $\boldsymbol{b}_f^{(t)}$, $\boldsymbol{b}_s^{(t)}$ 分别为第 t 层图卷积层中的两个全连接神经网络的权重矩阵和偏置向量. 晶体图数据经过 T 层图卷积后, 网络可以通过迭代, 自动学习每个原子包含其周围的环境的特征向量 $\boldsymbol{v}_i^{(T)}$.

接下来, 图池化层用于生成晶体的整体特征向量 $\boldsymbol{v}_{\mathrm{c}}$, 如下所示:

$$\boldsymbol{v}_{\mathrm{c}} = \kappa\left(\mathrm{mean}\left(\boldsymbol{v}_0^{\mathrm{T}}, \boldsymbol{v}_1^{\mathrm{T}}, \cdots, \boldsymbol{v}_M^{\mathrm{T}}\right)\right), \tag{5}$$

其中 $\kappa(x) = \ln(1+\mathrm{e}^x)$, M 表示晶体图中节点个数. 最后将其输入到由全连接层组成的全连接网络中, 就可以得到 z_2.

CGCrabNet 模型架构如下图 1 所示. CGCrabNet 模型将基于成分网络得到的 z_1 和基于结构网络得到的 z_2 加权求和, 如 (6) 式所示, 即为目标晶体性质:

$$z = w_1 z_1 + w_2 z_2, \tag{6}$$

其中 $w_1 + w_2 = 1$.

2.2 训练方式

在 CGCrabNet 模型训练过程中, 需要一个合适的损失函数来衡量模型输出值和真实值之间的差异. 在建立损失函数的过程中, 这里主要考虑偶然不确定性, 并且假设偶然不确定性是异方差的, 即不确定性取决于模型的输入, 不同的输入对应的噪声输出是不同的[13]. 假设材料数据不确定性的概率分布为拉普拉斯分布, 拉普拉斯分布如下:

$$P\left(\hat{z}_i | X_i, \theta\right) = \frac{1}{2\rho\left(X_i\right)} \exp\left(-\frac{\left|\hat{z}_i - z\left(X_i\right)\right|}{\rho\left(X_i\right)}\right), \tag{7}$$

其中 X_i 为模型第 i 个输入材料数据, θ 为 CGCrabNet 模型参数, $\rho\left(X_i\right)$ 为 X_i 的偶然不确定性, $z\left(X_i\right)$ 为材料带隙性质的计算值, 从数据库中直接获得. 对 (7) 式取对数如下:

$$\begin{aligned} &\ln P\left(\hat{z}_i | X_i, \theta\right) \\ &= -\ln(2\rho\left(X_i\right)) - \frac{1}{\rho\left(X_i\right)}\left|\hat{z}_i - z\left(X_i\right)\right|. \end{aligned} \tag{8}$$

因为希望似然估计越大越好, 也就是负对数似然值越小越好, 所以用等式 (8) 右边构造一个损失函数:

$$L_1 = \sum_i \frac{1}{\rho\left(X_i\right)}\left|\hat{z}_i - z(X_i)\right| + \ln\left(2\rho(X_i)\right), \tag{9}$$

式中, X_i 的偶然不确定性 $\rho\left(X_i\right)$ 可以看成关于 X_i 的一个未知函数, 也可以通过模型学习得到, 所以模型同时预测材料的性质 $\hat{z}_i$ 和方差 $\rho\left(X_i\right)$. 可以看到, 在模型学习偶然不确定性时, 是不需要真实样本值的. 因为如果一个样本 i 很难预测, 为了最小

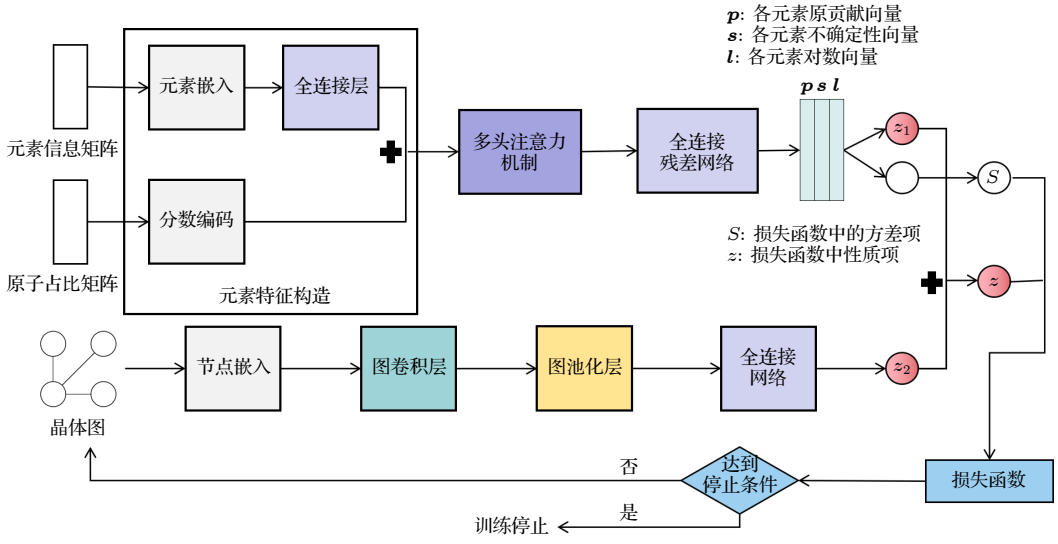

图 1 CGCrabNet 模型算法

Fig. 1. CGCrabNet model algorithm.

化损失函数, $\rho(X_i)$ 会适当变大, 而损失函数中的 $\ln(2\rho(X_i))$ 防止了网络预测样本数据不确定性为无穷大的情况.

在实际训练中, 我们训练网络来预测对数方差 $S_i=\ln(2\sigma(x_i))$, 从而损失函数为

$$L=\sum_i 2\exp(-S_i)|\hat{z}_i-z(X_i)|+S_i. \tag{10}$$

这样做是因为在数值上对数方差比方差更稳定, 而且可以避免学习到方差为 0 从而导致无法计算损失函数的情况. 另外, 通过学习对数方差 S_i, 使得方差 $\rho(X_i)$ 被解析到正值, 给出有效的方差值.

本文使用 Lamb 优化器[14] 更新模型权重, 在进行大批量数据训练时, 可以保持梯度更新的精度. 在训练过程中, 学习率是一个很关键的超参数. 学习率设置的过大, 模型可能很难收敛, 设置的过小, 则参数更新过于缓慢. 这里使用周期性学习率 (CLR)[15], 学习率每 2 个 epoch 在 1×10^{-4}—1×10^{-2} 之间循环, 以实现一致的模型收敛.

3 数值实验分析

3.1 实验训练细节

本文使用了 OQMD (the open quantum materials database)[16] 和 Materials Project[17] 数据库中的材料结构和性质数据. 针对 OQMD 数据库, 参考 Yamamoto[18] 的数据获取方法, 得到包含材料化学式、晶体图结构和带隙的样本数据, 这里获取的带隙是由 GGA+U 计算得到的[19]. 由于获取的 OQMD 数据库中大部分材料带隙值为 0, 为了样本的数据均衡性, 考虑将原数据集中带隙值为 0 的样本移除数据集, 最终得到 30368 个样本构成本文使用的 OQMD 数据集. 针对 Materials Project 数据库, 按照钙钛矿结构的一些性质例如化学通式 (ABX_3)、空间群数 (221) 和空间群名称 ($Pm\bar{3}m$) 等, 使用 API 接口筛选数据, 然后清除一些非钙钛矿结构数据, 得到 175 条包含 CIF 文件、化学式和带隙值的钙钛矿数据, 参考 CGCNN[3], 由 CIF 文件得到相应的晶体图信息, 构成本文钙钛矿材料数据集①. 本文 CGCrabNet 模型先在 OQMD 数据集中和文献中提出的模型进行对比, 验证其对带隙的预测能力. 然后将 CGCrabNet 应用到钙钛矿材料带隙预测任务中, 两个数据集的训练集、验证集和测试集的划分比例均为 6:2:2. 为了防止过拟合, 训练中加入提前停止策略, 超参数取值如表 1 所列.

表 1 超参数取值

Table 1. Hyperparameter value.

超参数名称	含义	值
$d_{\rm m}$	元素特征构造得到的向量维度	512
N_f	化学式中最大元素种类	7
N	注意力机制层数	3
n	注意力机制头数	4
I	参与训练的元素种类和	89
T	图卷积层数	3
$V_{\rm cg}$	节点嵌入后元素向量维度	16
$w_1:w_2$	权重比参数	7:3
Epochs	最大迭代次数	300
batch_size	批处理大小	256

3.2 实验结果

首先, 在 OQMD 数据集上对 One-Hot[20], Magpie[21] 和 Mat2vec[22] 三种元素嵌入方法进行测试. Magpie 是一系列原子特征的集合, 包括化学计量特征、元素属性特征、电子结构特征和离子化合物属性. Mat2vec 是将物理、化学、材料科学方向的一些词汇表示成 200 维词嵌入向量的方法, 包含了对每个元素的向量表示. 三种方法表示的元素向量均可以在开源项目 CrabNet[8] 中获得, 测试结果如表 2 所列. 由测试结果, 确定模型中元素嵌入方法使用 Mat2vec 方法.

表 2 元素嵌入法测试结果 (单位: eV)

Table 2. Elemental embedding method test results (in eV).

元素嵌入方法	Train MAE	Val MAE	Test MAE
One-Hot	0.185	0.423	0.433
Magpie	0.428	0.546	0.566
Mat2vec	0.203	0.408	0.420

为了验证 CGCrabNet 模型预测能力, 在 OQMD 数据集上, 按照同样的训练集、验证集、测试集划分方式, 复现了 CGCNN 模型[3]、Roost 模型[7]、CrabNet 模型[8] 和 HotCrab 模型[8](即 CrabNet 中元素嵌入方法改为 One-Hot 得到的模型) 进行对照实验, 实验结果如表 3 所列, 可以看出 CGCrabNet

① https://github.com/STsuntao/CGCrabNet/tree/master

模型在验证集和测试集上的平均绝对误差都要低于其他模型, 在测试集上平均绝对误差比 CrabNet 模型降低了 0.014 eV. 图 2 是 CGCrabNet 模型训练过程训练集和验证集损失值变换图, 可以看到验证集损失值趋于稳定, 并且没有出现过拟合的情况.

表 3 深度学习模型测试结果 (单位: eV)
Table 3. Deep learning model test results (in eV).

	Train MAE	Val MAE	Test MAE
CGCNN	0.502	0.605	0.601
Roost	0.178	0.447	0.455
CrabNet	0.226	0.422	0.427
HotCrab	0.177	0.422	0.440
CGCrabNet	0.187	0.408	0.413

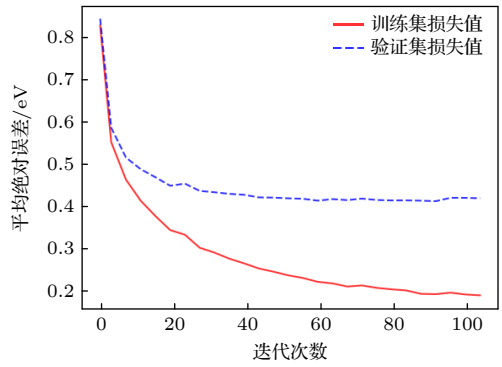

图 2 CGCrabNet 预训练损失值变化
Fig. 2. CGCrabNet pre-training loss value change.

图 3 和图 4 分别是 CGCrabNet 模型在验证集和测试集上预测带隙值和计算带隙值的散点图, 图中黑色虚线是由坐标轴上的点线性拟合得到的一次函数, 该拟合函数越接近函数 $y=x$, 说明预测效果越好, 进一步验证该模型的预测能力.

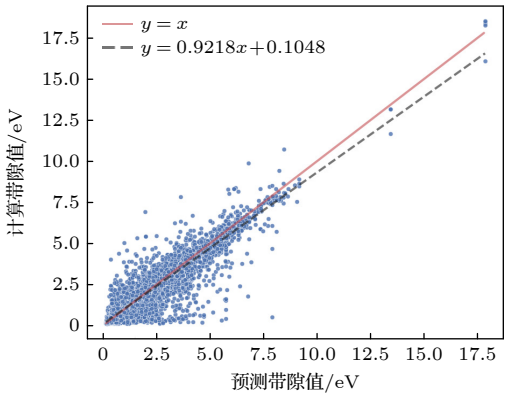

图 3 验证集预测带隙值
Fig. 3. Predicted band gap values on the validation set.

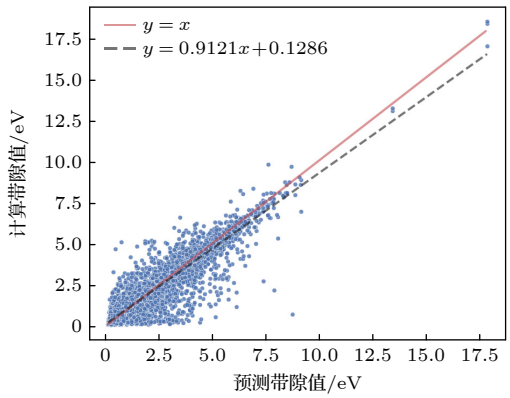

图 4 测试集预测带隙值
Fig. 4. Predicted band gap values on the test set.

3.3 模型应用

为了研究钙钛矿材料的带隙性质, 将 CGCrabNet 模型用于钙钛矿材料的带隙预测. 由于本文使用的钙钛矿数据集较小, 所以训练模型时加入迁移学习策略. 首先对 OQMD 数据集进行检验, 将其中包含的钙钛矿结构数据移除数据集, 最终得到 29827 条样本; 接下来在调整后的 OQMD 数据集上预训练 CGCrabNet 模型, 并保存模型参数; 最后, 将预训练的 CGCrabNet 模型在钙钛矿材料数据集上进行微调. 模型微调时 batch_size 变为 16, 其他训练细节和预训练时保持不变, 训练 39 代时模型训练提前终止. 可以看到, 在预训练的基础上微调模型, 加上设置的提前终止条件, 可以在快速收敛的同时有效防止过拟合.

此外, 还用随机森林回归模型 (RF)[23]、支持向量回归模型 (SVR)[24]、梯度提升回归模型 (GBR) 对钙钛矿材料的带隙性质进行预测, 在建立钙钛矿材料特征时, 使用 Atom_DL 原子分布式特征 [25] 拼接得到. 为得到稳定的预测模型, 对 3 种回归模型的参数进行网格搜索, 利用 5 折交叉验证对模型进行评估. 5 折交叉验证将数据集平均划分成 5 份, 依次用其中的一份作为测试集, 其他数据作为训练集来得到误差. 最后, 计算 5 个误差的平均值作为模型最终的误差. 所有的机器学习算法模型都是使用开源库 Scikit-learn[26] 实现的, 各回归模型的超参数如表 4 所列.

为说明迁移学习的重要性, 在仅用钙钛矿数据的情况下使用 CGCrabNet 模型对其带隙进行了预测, 测试集的平均绝对误差为 0.536 eV; 加入迁移学习策略之后, 测试集的平均绝对误差为

0.374 eV, 预测精度有明显提高. 同时, 在钙钛矿数据集上复现了 MEGNet 和 CGCNN 模型, 用于对比预测效果. 各模型预测效果的平均绝对误差值如图 5 所示, 可以看到, 随机森林回归模型和梯度提升回归模型的平均绝对误差 (MAE) 要低于支持向量回归模型, 这是因为 RF 和 GBR 都是集成学习模型, 考虑了多个基本学习器; MEGNet 和 CGCNN 模型在测试集上的 MAE 较高, 这是由于钙钛矿数据集较小, 在小数据集上训练 MEGNet 和 CGCNN 模型容易出现过拟合现象.

分别利用 CGCrabNet 模型、随机森林回归、支持向量回归和梯度提升回归预测带隙值与计算带隙值绘制成散点图如图 6 所示, 两条黑色实线内

表 4 回归模型参数

Table 4. Regression model parameters.

机器学习方法	超参数名称	取值
RF	子学习器数量	90
SVR	核函数	多项式核
	多项式核次数	3
	正则化强度	2
	伽马参数	2
	零系数	1.5
GBR	子学习器数量	500
	学习率	0.2
	最大深度	4
	损失函数	绝对误差函数

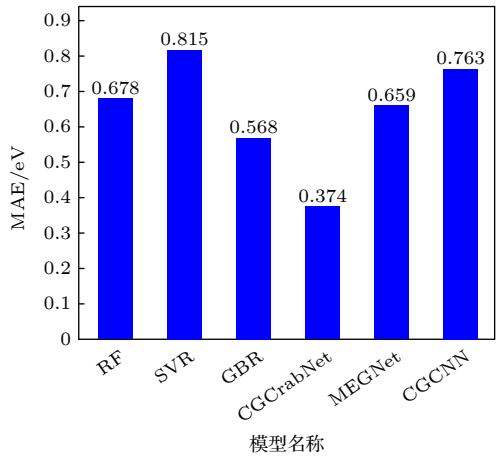

图 5 钙钛矿材料带隙预测的平均绝对误差对比

Fig. 5. MAE of band gap prediction for perovskite materials.

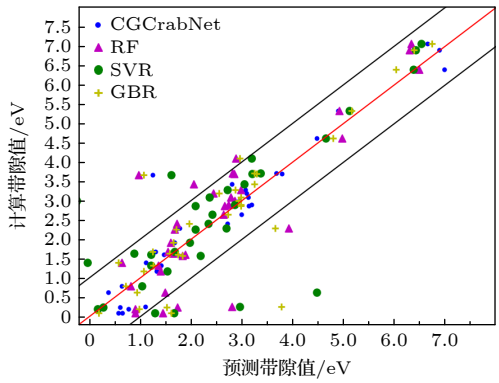

图 6 预测带隙与计算带隙散点图

Fig. 6. Predicting and calculating band gap scatter maps.

表 5 钙钛矿材料预测值和计算值对比 (单位: eV)

Table 5. Comparison of predicted and calculated values for perovskite materials (in eV).

化学式	带隙计算值	CGCrabNet	RF	SVR	GBR
$NbTlO_3$	0.112	0.658	1.458	1.296	1.614
$ZnAgF_3$	1.585	1.776	1.836	2.194	1.840
$AcAlO_3$	4.102	3.212	2.881	3.197	2.963
$BeSiO_3$	0.269	1.116	2.813	2.963	3.777
$TmCrO_3$	1.929	1.682	1.612	1.987	1.668
$SmCoO_3$	0.804	0.644	0.821	1.043	0.724
$CdGeO_3$	0.102	0.586	0.911	1.675	0.196
$CsCaCl_3$	5.333	4.891	4.918	5.116	5.157
$HfPbO_3$	2.415	2.724	1.733	2.346	1.967
$SiPbO_3$	1.185	1.327	1.407	1.543	1.079
$SrHfO_3$	3.723	3.683	2.821	3.370	3.253
$PrAlO_3$	2.879	3.139	2.665	2.091	2.984
$BSbO_3$	1.405	1.123	0.653	-0.025	0.579
$CsEuCl_3$	0.637	0.388	1.500	4.477	0.949
$LiPaO_3$	3.195	3.100	2.443	-0.306	2.553
$PmErO_3$	1.696	1.309	1.550	1.682	1.252
$TlNiF_3$	3.435	2.806	2.063	3.049	3.255
$MgGeO_3$	3.677	1.256	0.979	1.623	1.073
$NaVO_3$	0.217	0.785	0.911	0.180	0.989
$RbVO_3$	0.250	0.616	1.736	0.290	1.534
$KZnF_3$	3.695	3.785	2.853	3.203	3.295
$NdInO_3$	1.647	1.587	1.653	0.889	1.590
$RbCaF_3$	6.397	6.974	6.482	6.372	6.028
$RbPaO_3$	3.001	2.952	2.864	-0.234	2.937
$PmInO_3$	1.618	1.480	1.896	1.222	1.754
$KMnF_3$	2.656	2.991	2.647	2.428	2.730
$NbAgO_3$	1.334	1.419	1.369	1.227	1.265
$CsCdF_3$	3.286	3.078	2.990	2.724	2.879
$KCdF_3$	3.101	3.125	2.789	2.365	2.990
$CsYbF_3$	7.060	6.641	6.325	6.523	6.736
$NaTaO_3$	2.260	1.714	1.680	2.093	1.715
$CsCaF_3$	6.900	6.874	6.291	6.416	6.379
$RbSrCl_3$	4.626	4.470	4.966	4.647	4.795
$AcGaO_3$	2.896	3.199	2.740	2.869	2.981
$BaCeO_3$	2.299	1.789	3.918	2.696	3.655

注: CGCrabNet, RF, SVR和GBR分别代表特征融合神经网络、随机森林回归、支持向量回归和梯度提升回归模型计算得到的带隙值.

部的点表示预测值与计算值相差小于 1 eV, 数据点越靠近红色实线表示模型的预测效果越佳. 可以看出微调之后的 CGCrabNet 模型在钙钛矿材料数据集上预测效果良好.

表 5 列出了钙钛矿材料数据集测试集中的 35 个样本的带隙计算值和带隙预测值, 带隙预测值包括特征融合神经网络模型 (CGCrabNet)、随机森林回归 (RF)、支持向量回归 (SVR) 和梯度提升回归 (GBR) 预测的带隙值. 从表 5 可以看出, CGCrabNet 模型对 $SrHfO_3$ 和 $RbPaO_3$ 等钙钛矿材料预测的带隙值与计算值的平均绝对误差小于 0.05 eV, 并且对于大多数预测的钙钛矿材料, CG CrabNet 预测的带隙比其他 3 个回归模型预测的带隙更接近带隙计算值.

4 结 论

本文针对钙钛矿材料带隙预测任务, 构建了一个深度学习带隙预测模型. 比较了 Roost 模型、Crab Net 模型、HotCrab 模型、CGCNN 模型和 CGCrab Net 模型在 OQMD 数据集上对材料化合物的带隙预测效果. 利用迁移学习将在 OQMD 数据集上预训练的 CGCrabNet 模型, 用于对钙钛矿材料的带隙性质进行预测研究. 同时, 本文在钙钛矿材料数据集上建立了随机森林回归模型、支持向量回归模型和梯度提升回归模型, 用于和本文建立的模型进行比较. 实验结果表明, 在 OQMD 数据集上, CGCrabNet 对带隙的预测精度比 CrabNet 等模型有一定提升; 在钙钛矿材料数据集上, 本文建立的模型利用迁移学习策略对钙钛矿材料预测的平均绝对误差为 0.374 eV, 分别比随机森林回归、支持向量机回归和梯度提升回归的预测误差降低了 0.304 eV, 0.441 eV 和 0.194 eV; 另外, 模型预测的 $SrHfO_3$ 和 $RbPaO_3$ 等钙钛矿材料的带隙与第一性原理计算的带隙的误差小于 0.05 eV, 这说明 CGCrabNet 模型可以快速准确地预测新材料的性质, 加速新材料的研发过程.

此外, 本文的 CGCrabNet 模型考虑了材料的化学式和晶体图结构两种特征作为输入, 未来可以探索更多的材料特征输入模型, 通过多种特征精准学习出材料特征和性质的映射关系; 另一方面, 本文将特征融合神经网络模型应用到钙钛矿材料的带隙预测任务中, 进一步可以将其应用到其他材料的物理化学性质的研究中.

参考文献

[1] Fan X L 2015 *Mater. China* **34** 689 (in Chinese) [范晓丽 2015 中国材料进展 **34** 689]

[2] Wan X Y, Zhang Y H, Lu S H, Wu Y L, Zhou Q H, Wang J L 2022 *Acta Phys. Sin.* **71** 177101 (in Chinese) [万新阳, 章烨辉, 陆帅华, 吴艺蕾, 周登桦, 王金兰 2022 物理学报 **71** 177101]

[3] Xie T, Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301

[4] Chen C, Ye W K, Zuo Y X, Zheng C, Ong S P 2019 *Chem. Mater.* **31** 3564

[5] Karamad M, Magar R, Shi Y T, Siahrostami S, Gates L D, Farimani A B 2020 *Phys. Rev. Materials* **4** 093801

[6] Jha D, Ward L, Paul A, Liao W K, Choudhary A, Wolverton C, Agrawal A 2018 *Sci. Rep.* **8** 17593

[7] Goodall R E A, Lee A A 2020 *Nat. Commun.* **11** 6280

[8] Wang A Y T, Kauwe S K, Murdock R J, Sparks T D 2021 *NPJ Comput. Mater.* **7** 77

[9] Hu Y, Zhang S L, Zhou W H, Liu G Y, Xu L L, Yin W J, Zeng H B 2023 *J. Chin. Chem. Soc.* **51** 452 (in Chinese) [胡扬, 张胜利, 周文瀚, 刘高豫, 徐丽丽, 尹万健, 曾海波 2023 硅酸盐学报 **51** 452]

[10] Guo Z, Lin B 2021 *Sol. Energy* **228** 689

[11] Gao Z Y, Zhang H W, Mao G Y, Ren J N, Chen Z H, Wu C C, Gates I D, Yang W J, Ding X L, Yao J X 2021 *Appl. Surf. Sci.* **568** 150916

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I 2017 arXiv: 1706.03762v5 [cs. CL]

[13] Nix D A, Weigend A S 1994 *Proceedings of 1994 Ieee International Conference on Neural Networks (ICNN'94)* Orlando, FL, USA, 28 June—02 July, 1994 p55

[14] You Y, Li J, Reddi S, et al. 2020 arXiv: 1904.00962v5 [cs. LG]

[15] Smith L N 2017 arXiv: 1506.01186v6 [cs. CV]

[16] Saal J E, Kirklin S, Aykol M, Meredig B, Wolverton C 2013 *JOM* **65** 1501

[17] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson K A 2013 *APL Mater.* **1** 011002

[18] Yamamoto T 2019 *Crystal Graph Neural Networks for Data Mining in Materials Science* (Yokohama: Research Institute for Mathematical and Computational Sciences, LLC)

[19] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S, Wolverton C 2015 *NPJ Comput. Mater.* **1** 15

[20] Calfa B A, Kitchin J R 2016 *AIChE J.* **62** 2605

[21] Ward L, Agrawal A, Choudhary A, Wolverton C 2016 *NPJ Comput. Mater.* **2** 16028

[22] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z Q, Kononova O, Persson K A, Ceder G, Jain A 2019 *Nature* **571** 95

[23] Breiman L 2001 *Mach. Learn.* **45** 5

[24] Wu Y R, Li H P, Gan X S 2013 *Adv. Mater. Res.* **848** 122

[25] Sun T, Yuan J M 2023 *Acta Phys. Sin.* **72** 028901 (in Chinese) [孙涛, 袁健美 2023 物理学报 **72** 028901]

[26] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É 2011 *J. Mach. Learn. Res.* **12** 2825

Band gap prediction of perovskite materials based on transfer learning*

Sun Tao [1)] Yuan Jian-Mei [1)2)†]

1) (*School of Mathematics and Computational Science, Xiangtan University, Xiangtan* 411105, *China*)

2) (*Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan* 411105, *China*)

(Received 22 June 2023; revised manuscript received 31 July 2023)

Abstract

The band gap is a key physical quantity in material design. First-principles calculations based on density functional theory can approximately predict the band gap, which often requires significant computational resources and time. Deep learning models have the advantages of good fitting capability and automatic feature extraction from the data, and are gradually used to predict the band gap. In this paper, aiming at the problem of quickly obtaining the band gap value of perovskite material, a feature fusion neural network model, named CGCrabNet, is established, and the transfer learning strategy is used to predict the band gap of perovskite material. The CGCrabNet extracts features from both chemical equation and crystal structure of materials, and fits the mapping between feature and band gap. It is an end-to-end neural network model. Based on the pre-training data obtained from the Open Quantum Materials Database (OQMD dataset), the CGCrabNet parameters can be fine-tuned by using only 175 perovskite material data to improve the robustness of the model.

The numerical and experimental results show that the prediction error of the CGCrabNet model for band gap prediciton based on the OQMD dataset is 0.014 eV, which is lower than that obtained from the prediction based on compositionally restricted attention-based network (CrabNet). The mean absolute error of the model developed in this paper for predicting perovskite materials is 0.374 eV, which is 0.304 eV, 0.441 eV and 0.194 eV lower than that obtained from random forest regression, support vector machine regression and gradient boosting regression, respectively. The mean absolute error of the test set of CGCrabNet trained only by using perovskite data is 0.536 eV, and the mean absolute error of the pre-trained CGCrabNet decreases by 0.162 eV, which indicates that the transfer learning strategy plays a significant role in improving the prediction accuracy of small data sets (perovskite material data sets). The difference between the predicted band gap of some perovskite materials such as $SrHfO_3$ and $RbPaO_3$ by the model and the band gap calculated by first-principles is less than 0.05 eV, which indicates that the CGCrabNet can quickly and accurately predict the properties of new materials and accelerate the development process of new materials.

Keywords: feature fusion neural network, regression model, band gap evaluation, transfer learning

PACS: 89.90.+n **DOI:** 10.7498/aps.72.20231027

* Project supported by the Natural Science Foundation of Hunan Province, China (Grant Nos. 2023JJ30567, 2021JJ30650).

† Corresponding author. E-mail: yuanjm@xtu.edu.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于迁移学习的钙钛矿材料带隙预测

孙涛 袁健美

Band gap prediction of perovskite materials based on transfer learning

Sun Tao Yuan Jian-Mei

引用信息 Citation: Acta Physica Sinica, 72, 218901 (2023) DOI: 10.7498/aps.72.20231027
在线阅读 View online: https://doi.org/10.7498/aps.72.20231027
当期内容 View table of contents: http://wulixb.iphy.ac.cn
