doi:10.19638/j.issn1671-1114.20220304

利用自注意力机制优化费米网络的数值研究

王佳奇[1],高泽峰[2],李永峰[1],王 璐[1]
(1. 内蒙古科技大学 理学院,内蒙古 包头 014010;2. 中国人民大学 理学院,北京 100080)

摘 要:为了探索不使用特定形式的试探态研究多电子系统基态性质的方法,以至多约 10 个原子的小分子为例,利用神经网络的方法对多电子系统进行求解. 此外,利用包含自注意力机制的 Transformer 结构对费米网络(FermiNet)进行改进,并称之为 Transformer-FermiNet,同时对比了 Transformer 结构中不同数量的隐藏单元对网络表达能力的影响,并对 Transformer 结构取代线性连接结构的有效性进行分析. 结果表明:Transformer-FermiNet 能够在保证原费米网络结果精度的同时将网络参数的规模缩减为原来的 3/4.

关键词:多电子系统;基态;神经网络;Transformer 结构;自注意力机制
中图分类号:O641 文献标志码:A 文章编号:1671-1114(2022)03-0022-09

Numerical research of optimization FeimiNet using self-attention mechanism

WANG Jiaqi[1], GAO Zefeng[2], LI Yongfeng[1], WANG Lu[1]
(1. School of Science, Inner Mongolia University of Science and Technology, Baotou 014010, Inner Mongolia, China; 2. School of Science, Renmin University of China, Beijing 100080, China)

Abstract: To explore approaches which investigate properties of ground states of multi-electron systems, simulations were performed for the molecules including up to about 10 atoms as examples with the neural network methods. The Transformer structure with self-attention was employed to improve the efficiency of the FermiNet. The improved network is denoted as Transformer-FermiNet. In addition, the expressive power of the networks with different number of hidden units was compared, and the validity that substitutes the linear connection with the Transformer structure was studied. The results show that the number of parameters of the Transformer-FermiNet can be reduced to 3/4 without loss of accuracy compared to the FermiNet.

Keywords: multi-electron systems; ground states; neural network; Transformer structure; self-attention mechanism

多电子系统的理论计算一直是物理学基础研究的重要课题. 理论上,如果能够给出系统的波函数,则可由波函数得到系统的所有性质. 但多电子系统的薛定谔方程难以直接求解,为此研究人员提出了多种各有优劣的近似求解方法,如常用的密度泛函理论[1-3]受交换关联能项的近似限制,在计算相互作用复杂的多电子系统时,精度无法满足计算要求. 耦合簇方法中的 CCSD(T)(coupled cluster with single, double and perturbative triple excitation)方法[4]可以得到精确的结果,但所用计算资源与基函数数目的 7 次方成正比,造成 CCSD(T)几乎不可能计算大分子系统. 量子蒙特卡罗方法(quantum Monte Carlo method, QMC)[5]的计算精度很高,但结果收敛很慢,计算误差的衰减速度是计算时间平方根的倒数,要得到精确度足够高的结果所花费的时间代价无法接受. 这些传统近似求解方法的求解精度和计算速度虽然在缓慢进步,但远远无法满足人们的需求.

近几年,人工智能在图像识别和语音识别等多个领域取得了突破性进展,并广泛应用于物理和化学等学科的研究中[6-10],取得一定成果. 如人们使用受限玻尔兹曼机拟合晶格体系的基态及含时演化过程[11-12],其基态结果的精度可与耦合簇方法相比. 对于小分子系统,人们使用神经网络作为变分量子蒙特卡罗方法的波函数试探解,并通过优化神经网络得到系统的基态

收稿日期:2021-05-18
基金项目:国家自然科学基金资助项目(51961031);内蒙古自治区自然科学基金资助项目(2019MS01013).
第一作者:王佳奇(1995—),男,硕士研究生.
通信作者:王 璐(1982—),男,讲师,主要从事统计物理与计算物理方面的研究. E-mail:wanglustar@hotmail.com;
高泽峰(1995—),男,博士研究生. E-mail:zfgao@ruc.edu.cn.

能和基态波函数[13-15]，结果比传统变分蒙特卡罗方法更精确. 其中，费米网络[13]仅使用电子及原子核的位置关系作为神经网络的输入，实现了对费米子系统的计算，给出了具有相同自旋电子的交换反对称波函数，并在小分子系统的计算中达到了非常高的精度. 因此，研究费米网络有助于更有效地定量研究费米子系统的性质.

近年来，相较于多层神经网络等以往常用的结构，在人工智能各个领域广泛应用的自注意力机制(self-attention)[16]展现出更强的表达能力. 基于自注意力机制的 Transformer 结构被应用于自然语言处理，如 BERT (bidirectional encoder representations from transformers)[17]和 GPT-2 (generative pre-training)[18]中，并取得了良好效果. Transformer 结构成为提高现有工作精度和准确性的首选工具之一. 本研究在保证系统费米子交换反对称的前提下，添加 Transformer 结构，改变费米网络仅有的单一线性连接结构，以期增强网络的表达能力，新网络结构被称为 Transformer-FermiNet (TFN)，并在此基础上分析整个神经网络的物理内涵，对比不同 Transformer 结构的表现力.

1　多电子系统

对于多电子系统，哈密顿量可以写为

$$\hat{H}=-\frac{1}{2}\sum_i\nabla^2{}_i-\sum_i\sum_k\frac{Z_k}{|\boldsymbol{r}_i-\boldsymbol{R}_k|}+\frac{1}{2}\sum_i\sum_{j\neq j}\frac{1}{|\boldsymbol{r}_i-\boldsymbol{r}_j|}\tag{1}$$

式(1)中：Z_k 和 $\boldsymbol{R}_k$ 分别表示编号为 k 的原子核的电量和位置；$\boldsymbol{r}_i$ 和 $\boldsymbol{r}_j$ 分别表示编号为 i 和 j 的电子的位置. 式(1)右侧第 1 项表示电子的动能，第 2 项表示电子与原子核之间的相互作用，第 3 项表示电子间的相互作用. 基于波恩奥本海默近似(Born-Oppenheimer approximation)，原子核的位置为给定的参量. 理论上，将哈密顿量代入薛定谔方程

$$\hat{H}\psi(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)=E\psi(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\tag{2}$$

并求解，即可得到系统的波函数. 式(2)中 $\boldsymbol{x}_i=(\boldsymbol{r}_i,\sigma)$ 表示编号为 i 的电子的坐标和自旋. 在理论计算中，常使用斯莱特(Slater)行列式

$$\sum_p\mathrm{sign}(\mathcal{P})\prod_i\phi^k{}_i(\mathcal{P}(\boldsymbol{x}))=\begin{vmatrix}\phi^k{}_1(\boldsymbol{x}_1)&\cdots&\phi^k{}_1(\boldsymbol{x}_n)\\ \vdots&\ddots&\vdots\\ \phi^k{}_n(\boldsymbol{x}_1)&\cdots&\phi^k{}_n(\boldsymbol{x}_n)\end{vmatrix}=\det[\phi^k{}_i(\boldsymbol{x}_j)]=\det[\boldsymbol{\Phi}^k]\tag{3}$$

$$\psi(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)=\sum_k\omega_k\det[\boldsymbol{\Phi}^k]\tag{4}$$

的线性组合给出变分法的试探波函数. 式(3)中：$\phi^k{}_m$ 为单电子轨道波函数，角标 $m=1,\cdots,n$ 为不同的轨道，n 为电子数，其中有 $n^{\uparrow}$ 个电子自旋向上，有 $n^{\downarrow}$ 个电子自旋向下，$n=n^{\uparrow}+n^{\downarrow}$；$\boldsymbol{\Phi}^k$ 为由元素 $\phi_i^k(\boldsymbol{x}_j)$ 组成的矩阵；ω_k 为电子波函数 ψ 用不同行列式 $\det|\boldsymbol{\Phi}^k|$ 展开的系数，$k=1,2,\cdots,K$ 为正整数；$\mathcal{P}$ 表示单电子波函数的一种排列；$\mathrm{sign}(\mathcal{P})$ 给出排列的奇偶性，$\mathcal{P}$ 为奇排列时，$\mathrm{sign}(\mathcal{P})$ 取负数，$\mathcal{P}$ 为偶排列时，$\mathrm{sign}(\mathcal{P})$ 取正数. $\mathcal{P}(\boldsymbol{x})$ 表示在某排列 $\mathcal{P}$ 下，$\boldsymbol{x}_i$ 变为 $\mathcal{P}(\boldsymbol{x})=\boldsymbol{x}_j$. 斯莱特行列式的维数等于电子的个数，其展开后的项数随电子数增加呈指数级增长. 对于多电子系统，求解这样的波函数极其困难，即使全结构相互作用求解方式在最近有了新的进展[19]，也只能计算为数不多的小分子.

选择 1 个优秀的试探波函数可以加快变分法迭代的速度，且更容易得到高精度的近似解. 在量子蒙特卡罗计算连续空间多体问题时，通常选用 Slater-Jastrow 作为试探解[5,20]. 其主要思想是将斯莱特行列式截断，并添加 Jastrow 系数项做近似. 近年来，研究人员选用神经网络作为变分蒙特卡罗的试探波函数. 由万能逼近定理[21]可知，多层神经网络能够以任意精度逼近目标函数，波函数也不例外，利用神经网络优秀的表达能力，良好地逼近复杂的波函数. 文献[13]提出的费米网络正是对上述想法的实现.

2　费米网络

2.1　费米网络的基本结构

费米网络的细节非常丰富，为方便讲解 TFN，在此对费米网络进行简要介绍，详细讲解可见文献[13]. 费米网络的结构如图 1 所示，图 1 上半部分表示神经网络的输入，下半部分展示了神经网络中单电子流与斯莱特行列式的连接及网络的最终输出.

图 1 中，神经网络的输入分为 2 个部分，上半部分的输入为电子-原子核之间的位置关系，称为单电子流. $h^0{}_n$ 层中，d_{nN_x}、d_{nN_y} 和 d_{nN_z} 分别表示第 n 个电子与第 N 个原子核间位矢在三维空间的 3 个投影，$|\boldsymbol{d}_{nN}|$ 则为这个距离的绝对值. 下半部分的输入为电子-电子之间的位置关系，称为双电子流. 与单电子流同理，电子 i 和电子 j 间的位置关系均用二者距离在三维空间中的 3 个投影 r_{ijx}、r_{ijy} 和 r_{ijz} 及其绝对值 $|\boldsymbol{r}_{ij}|$ 共 4 个量来表示. 文献[13]强调，添加距离的绝对值能够大幅提高网络的学习精度及训练效率. 这是因为三维分量与其所对应的位置关系是一体相关的，而神经网络无法准确地自我优化学习到这种关联，所以人为添加距离的

绝对值是必要的. 在单电子流和双电子流中，人为将自旋向上 $r^{\uparrow}$ 和自旋向下 $r^{\downarrow}$ 的电子分离输入. 在最后 1 个隐藏层 h^{L}_{n} 的输出中，自旋向上和向下的数据流分别对应网络后边的 2 个行列式. 费米网络在每一层中不断将双电子流信息混入单电子流中，最后将单电子流的输出组合成行列式 $\det^{\uparrow}$ 和 $\det^{\downarrow}$ 的形式，对多个行列式求行列式值并线性组合，得到神经网络最终的输出，即系统的波函数 ψ.

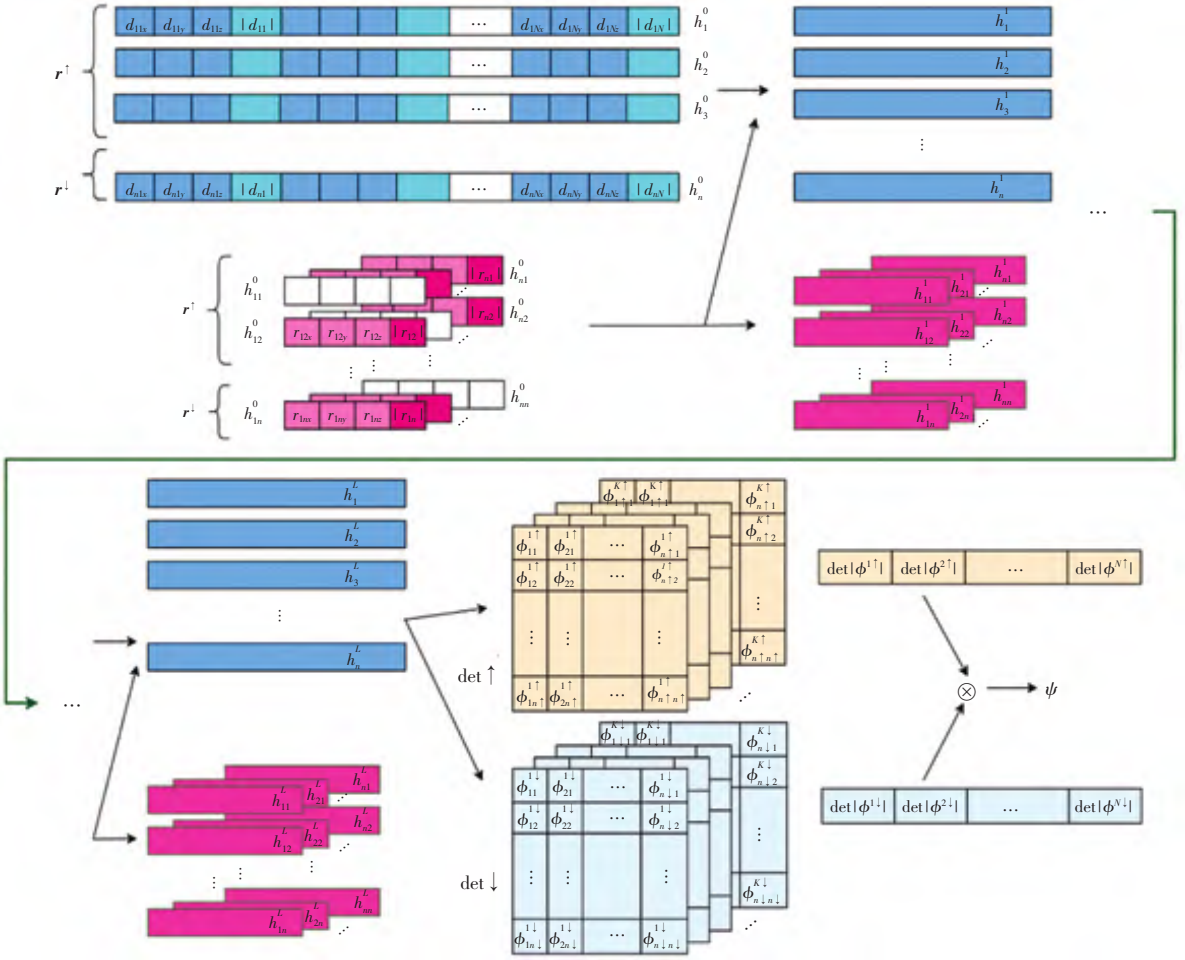

图 1 费米网络结构

Fig. 1 Structure of the FermiNet

费米网络的训练分为 2 个步骤：①使用 Hartree Fock 理论计算结果作为标签值，使用 Adam 算法[22]进行有监督学习的网络训练，称为预训练过程. 预训练可以将网络参数优化到真实波函数的附近，减少网络陷入局部最优的可能性. 预训练损失函数为

$$L^{\mathrm{pre}}(\theta)=\int\left\{\sum_{a}\sum_{ijk}[\phi_{i}^{ka}(r_{j}^{a})_{\mathrm{Net}}-\phi_{ia}^{\mathrm{HF}}(r_{j}^{a})]^{2}\right\}\cdot p^{\mathrm{pre}}(X)\,\mathrm{d}X \tag{5}$$

式 (5) 中：θ 为神经网络的可训练参数；$\phi_{i}^{ka}(r_{j}^{a})_{\mathrm{Net}}$ 和 $\phi_{ia}^{\mathrm{HF}}(r_{j}^{a})$ 分别表示由神经网络和 Hartree Fock 给出的单电子波函数；蒙特卡罗抽样中构型 X 的概率

$$p^{\mathrm{pre}}(X)=\frac{1}{2}\left\{\prod_{a\in\{\uparrow,\downarrow\}}\prod_{i}[\phi_{ia}^{\mathrm{HF}}(r_{i}^{a})]^{2}+\psi^{2}(X)_{\mathrm{Net}}\right\} \tag{6}$$

②将系统能量的期望值视为损失函数，使用经过优化的 KFAC (Kronecker-factored approximate curvature) 算法[23]最小化系统能量. 电子的初始位置正态分布在原子核周围随机生成，在随后的迭代过程中，电子位置使用蒙特卡罗方法根据神经网络波函数给出.

2.2 费米网络的基准测试

为加深对费米网络的理解，本文首先计算一些已有结果. 根据电子数由少到多，选择计算 LiH、NH_3、CH_3NH_2 和 C_2H_5OH 共 4 种分子，并比较计算结果. 预训练的结果如图 2 所示，其纵坐标用 $\log_{10}$ 标度.

由图 2 可以看出，只有 4 个电子的 LiH 的预训练损失函数收敛得非常快，在 1 000 步时，其损失函数已经降低至 0.001 以下. 而电子数为 26 的 C_2H_5OH 在 1 000 步时，其损失函数稳定性较差，且下降速度缓慢. 造成二者收敛速度不同的原因是电子数越多，其在空

间中的态密度分布越复杂. 因此，在预训练过程中，当系统的电子数较多时，应使用足够多的预训练步数使损失函数降到一个合理的范围，以保证预训练效果.

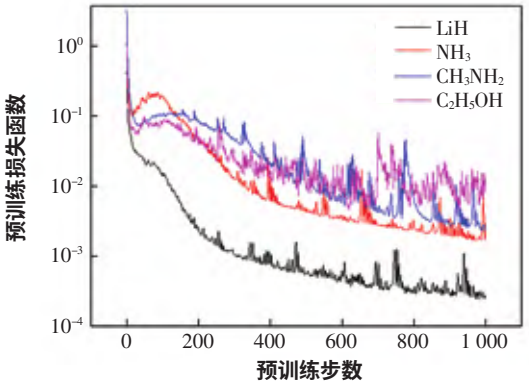

图 2　不同分子预训练损失函数的对比

Fig. 2　Comparisons of the loss functions from pretraining for different molecules

图 3 为不同分子基态能随正式训练步数的下降趋势. 由图 3 可以看出，4 条能量下降曲线在形状上相似，但前 3 条曲线下降趋势不明显，尤其是 LiH 对应的基态能下降曲线几乎为一条直线，而电子数为 26 的 C_2H_5OH 系统的基态能表现出明显的不稳定. 由此可知，系统的电子数越少，神经网络给出的系统基态能收敛越快，且数值相对稳定. 随着电子数的增加，需要拟合的目标波函数变得复杂. 神经网络需要使用更多的训练步数使能量逐渐收敛，并改善网络输出的数值稳定性.

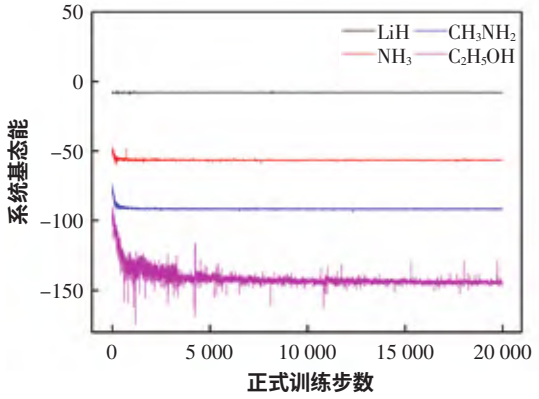

图 3　不同分子正式训练所得系统基态能

Fig. 3　Ground energies of different molecules from training

为精确对比本文与文献[13]中费米网络的计算结果，给出正式训练 20 000 步时，二者平均能量的具体数值列，结果如表 1 所示. 表 1 中的 CCSD (T) 使用了 CBS (complete basis set，CBS) 基底集，能量单位为 Hartree.

表 1　本研究与文献[13]计算结果的对比

Tab. 1　Comparisons of the results of this study and reference [13]

分子	原子核数/电子数	文献[13]的基态能 E_A	CCSD (T)	本研究的基态能 E_B	$\lvert\frac{E_B-E_A}{E_A}\rvert$/%
LiH	2/4	−8.070 50 (1)	−8.070 7	−8.065 (2)	6.814×10^{-2}
NH_3	4/10	−56.562 95 (8)	−56.564 4	−56.528 (1)	6.170×10^{-2}
CH_3NH_2	7/18	−95.855 40 (2)	−95.865 3	−91.527 (9)	4.520
C_2H_5OH	9/26	−155.030 80 (3)	−155.054 5	−144.570 (5)	6.750

在计算如 LiH 和 NH_3 等电子数较少的小分子时，费米网络对计算资源的要求不高，可达到很好的精度. 但当计算如 CH_3NH_2 和 C_2H_5OH 等电子数较多的分子时，利用现有计算资源所得结果与准确值存在一定差距. 原因主要有两点：①有限计算资源下，较大 batch size 的并行训练无法实现，但选用大 batch size 可以有效降低能量值的波动性并加快训练速度. ②对于较大分子，并未使用足够长的时间等待能量的进一步下降.

文献[13]中，费米网络在计算多电子系统，如甲烷、乙烯、甚至 30 个电子的双环丁烷时，多数情况精度可以比肩甚至超越变分量子蒙特卡罗方法和耦合簇方法等传统方法，并在描述氢分子链及氮气的解离能方面表现良好. 但费米网络的缺点也很明显，在现有技术手段下，费米网络的误差收敛到化学精度以内所花费的计算资源随电子数的增加呈指数级增长. 如 LiH 等小分子的计算仅需使用一张 Nvidia 1650 GPU，其计算结果在几小时内即可收敛；但在计算双环丁烷时，需要使用 16 张 V100 GPU 进行计算，并花费约 1个月的时间，才得到 97% 准确率的结果[13]. 目前，多数研究组难以获得这样规模的计算资源.

3　TFN 网络

3.1　TFN 网络结构

本研究使用的 Transformer 结构包含自注意力机制和线性连接结构 2 个部分，结构如图 4 所示. 图 4 中红色和绿色部分均代表可训练的参数，其中，自注意力参数的下标 q、k 和 v 分别源自自然语言处理问题中的 query、key 和 value，线性连接参数的下标 lc 表示 linear connection. 网络中可训练的参数包括自注意力机制所用到的权重 W_q、W_k 和 W_v 以及线性连接部分的权重 W_{lc}. 本研究在网络中不同位置添加不同的Transformer 结构，但均基于图 4 结构的微调.

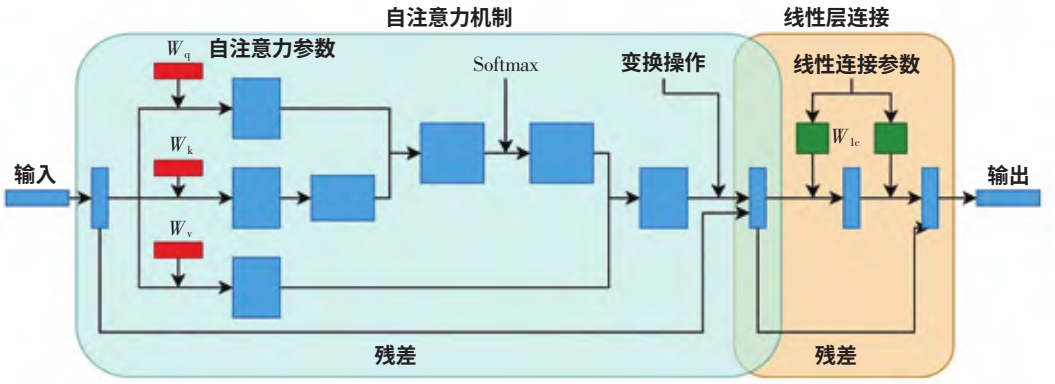

图 4 Transformer 结构图

Fig. 4 Transformer structure

本研究在费米网络中使用了 3 种 Transformer 结构，添加位置分别在单电子流的输入层和中间层以及双电子流的中间层，每种添加方式均有其对应的物理考虑.

(1) 选择对单电子流输入层中的电子–原子核距离的绝对值添加 Transformer 处理，相当于额外分配参数用于专门学习电子与原子核间距的绝对值对波函数的贡献，这种添加了 Transformer 的网络结构称为HT (head transformer)，结构如图 5 所示，图 5 中深蓝色和蓝绿色部分分别代表电子–原子核位置关系的三维投影和绝对值. 在同一个分子中，以甲烷 (CH_4) 为例，其原子核中 C 和 H 所带电量不同，与电子的相互作用强度也不同，在输入层添加额外的 Transformer 结构有助于学习到这种差异，因此，可将 HT 视为一种额外的数据处理.

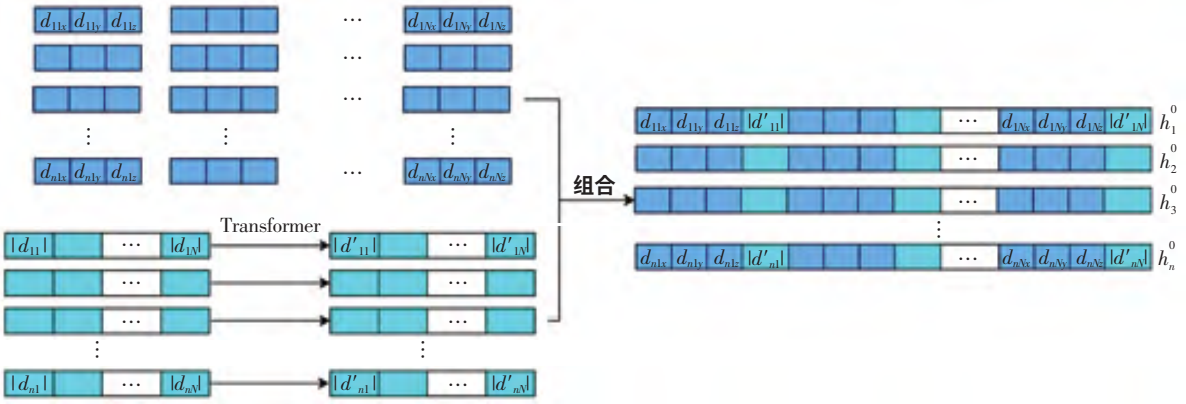

图 5 HT 结构图

Fig. 5 Structure of the HT

(2) 将单电子流中间层的连接方式完全改为 Transformer，从而改变原有费米网络单一的线性连接结构，称为 ST (single stream transformer). 图 6 分别展示了单电子流和双电子流的 Transformer.

(3) 将双电子流中间层的连接方式完全用 Transformer 替换则为 DT (double stream transformer) (图 6(b))，这种替换方式与单电子流完全一致. 由于许多新奇复杂的物理性质与电子–电子间相互作用有关，因此，使用更强表达能力的网络学习双电子流的信息是必要的. 需要注意的是，DT 中的 Transformer 结构并没有用来强化学习不同电子与同一电子间的关联，这是因为不同于电子与原子核间的作用，由于电子所带电量相同，所有电子–电子相互作用均可视为使用同一种相互作用规则，没有必要额外添加 Transformer 结构进行学习.

此外，本文在处理 HT 中自注意力机制的结果时并未使用图 4 中的求平均操作，而是使用打平操作 (将结果矩阵按选定维度首尾相接成低维矩阵，称为 flatten). 这样做的目的是为了更多地保留自注意力机制中的信息，满足 HT 中对电子–原子核距离绝对值的强调. 对 ST 和 DT 中自注意力机制结果的处理使用了图 4 中的求平均操作，这样做可以减少线性连接部分的参数数量. 修改后网络的优化过程依然符合预训练和正式训练梯度下降的算法要求 (Adam 和 KFAC)，因此，TFN 所使用的优化方式与原费米神经网络相同.

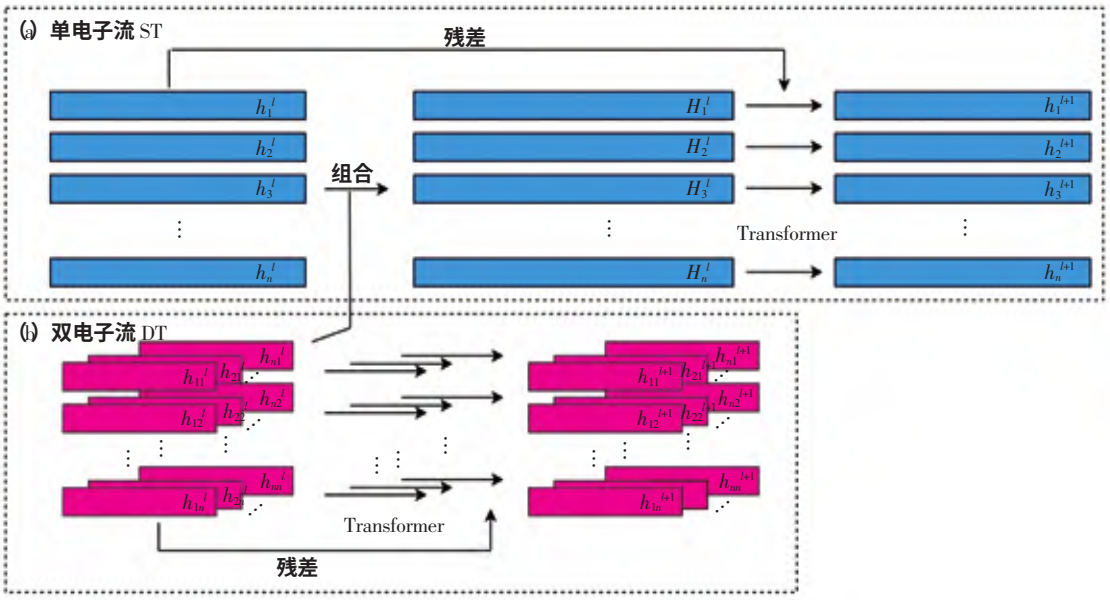

图 6　ST 和 DT 结构图

Fig. 6　Structure of the ST and DT

3.2　TFN 的自对比

为研究 Transformer 结构包含的自注意力机制和线性连接这 2 种不同结构的参数对网络拟合能力的影响，以 HT 计算甲烷（CH_4）基态能为例，分别对比了 Transformer 结构中参数的影响. 为突出 Transformer 结构本身的表达能力，网络的中间连接仅使用了 2 层线性连接的隐藏层. 每个隐藏层中单电子流使用 32 个隐藏单元，双电子流使用 8 个隐藏单元.

首先，使用固定数量的线性连接隐藏单元，并逐级增加自注意力机制参数的数量. 图 7 为 TFN 拟合所得体系基态能与自注意力机制参数数量的关系. 图 7 中，绿点代表在网络计算的基态能优化到稳定值后，分段截取数据并取平均，红点代表对整体的平均（对绿点求平均）. 第 1 组数据的自注意力机制隐藏单元数为 4，线性连接隐藏单元数为 8. 后续计算中，线性连接隐藏单元数量保持不变，自注意力机制参数数量逐步由 4 增加到 256.

由图 7 可以看出，当自注意力机制的参数数量较少时，网络的拟合能力较差，计算结果非常不稳定，表现为多个高能量点. 随着自注意力机制参数的增加，网络的稳定性和精度均得到明显改善. 但不断增加自注意力机制参数的数量并不能持续改进结果. 在自注意力机制参数数量达到 256 个时，虽然计算结果的精度依然小幅提升，但计算的收敛速度明显减慢. 此外，图 7 中常出现一些偏离整体的值，如自注意力机制参数为 64 个时，所得极个别能量值甚至高于参数为 32 个时的能量值，这些值导致整体平均能量有所上升，但依然低于参数为 32 个时的平均能量. 本研究仅对单层 Transformer 结构内部参数进行了对比，而实际计算中则需使用多层网络，因此在计算资源有限的情况下，使用 32~128 个隐藏单元即可得到较理想的训练结果.

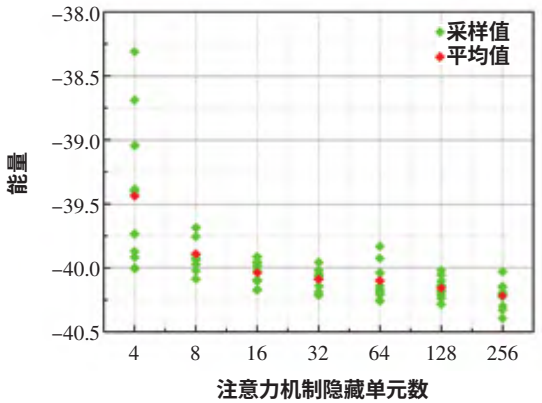

图 7　自注意力机制隐藏单元数量对 TFN 表达能力的影响

Fig. 7　Influence of the number of self-attention hidden units on the expressive power of the TFN

图 8 为 TFN 拟合的体系基态能与线性连接隐藏单元数量的关系，其中绿点、红点以及第 1 组数据的选取方式同图 7. 后续计算中，自注意力机制参数数量保持不变，线性连接隐藏单元数逐步由 8 增加到 256. 此外，图 8 中隐藏单元数由 64 增加到 256 的过程中，网络的精度依然小幅提升.

由图 8 可以看出，随着线性连接隐藏单元数量的增加，网络的精度和稳定性均逐步提升. 但相较于 128个隐藏单元，当隐藏单元数量增加到 256 个后，所带来的精度提升已经很小. 因此，在线性连接中使用 64 或

128 个隐藏单元是对计算速度和计算精度平衡后不错的选择.

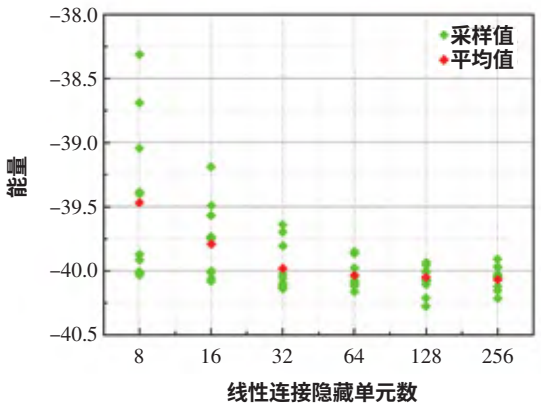

图 8 线性连接隐藏单元数量对 TFN 表达能力的影响

Fig. 8 Influence of the number of linear connection hidden units on the expressive power of the TFN

需要注意的是，线性连接的参数是 1 个二维矩阵，而自注意力机制的参数是 3 个一维矩阵. 显然随隐藏单元数量的增加，线性连接参数数量的增长速度是自注意力机制参数数量增长速度的平方量级. 即对比线性连接结构，自注意力机制的参数越少，表达能力越强.

为了统一对比，在对比 Transformer 结构内部参数的所有计算中，使用的并行训练个数 batch size = 20，预训练次数 = 1 000，正式训练次数 = 40 000，学习率 lr = 0.000 1，随着训练 (step) 的进行，学习率以 lr (1/(1.0 + (step/1 000.0))) 的方式衰减. 图 7 和图 8 中，计算数据在均值附近表现出显著偏离的现象主要是因为网络的训练过程使用了较小的batch size. 在计算资源充足时，使用较大的 batch size 可以使数据偏离的程度降低. 原则上，使用大的batch size 能够得到精度和稳定性更好的结果.

3.3 TFN 的计算和参数讨论

为了展示添加 Transformer 结构对神经网络表达能力的影响，以氨气 (NH_3) 的计算为例，对比 TFN 和标准费米网络的优化过程，结果如图 9 所示. 图 9 中，纵坐标 ΔE 为神经网络计算所得能量与 CCSD (T) /CBS 方法给出的能量之差，并用 $\log_{10}$ 标度；横坐标中能量为分阶段取平均的结果，横轴上 1~1 000 为每 10 步对 ΔE 取平均，1 001~10 000 为每 90 步取平均，10 000 后是每 100 步取平均. 此外，为了满足大分子的计算需求，如计算 30 个电子的双环丁烷以及 26 个电子的乙醇时，文献[13]中的费米网络采用了固定数量的隐藏单元. 但经过测试，在计算较小的分子时，适当削减网络的隐藏单元数量对结果精度的影响可以忽略. 受限于计算资源，本研究主要用 10 个电子左右的小分子进行计算. 为了方便对比，费米网络的参数量被控制在与 TFN 相近的范围.

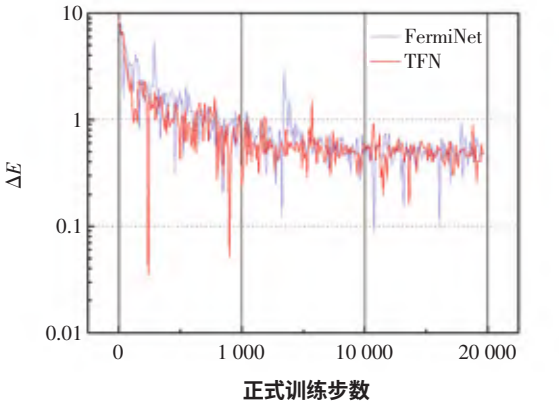

图 9 TFN 与费米网络的优化过程

Fig. 9 Optimization progress of the TFN and FermiNet

图 9 中，TFN 和费米网络训练状况的对比被分为 3 个阶段. 在 1~1 000 步的第 1 阶段，TFN 和费米网络均在尝试快速将能量下降，但费米网络由于参数较多，下降得更加流畅连贯. 经历第 2 阶段 (1 001~10 000步) 时，二者均开始逐渐稳定到平衡值附近. 在第 3 阶段 (10 000 后)，二者的平均能量基本不再下降，且结果的波动性逐渐变小. 随着后期参数的学习率减小，结果的波动性会进一步降低，在此阶段，TFN 和费米网络的表现相似. 即在正式训练时，经过前期快速的参数调整阶段后，TFN 的表现能力与费米网络相似.

为了清晰对比费米网络与 TFN 的计算结果，从图 9 第 3 阶段后期的收敛结果中截取数据并取均值，结果如图 10 所示. 由图 10 可知，在随机的抽样对比中，TFN 数据偏离均值程度略大，但其均值与费米网络结果相当.

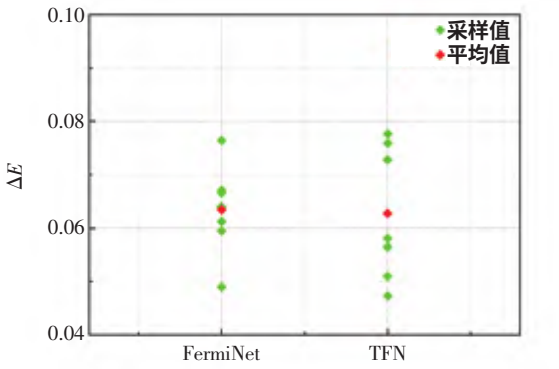

图 10 基态能的采样均值对比

Fig. 10 Comparisons of sampling mean values of the ground-state energy

此外，费米网络在计算小分子时所得结果的精度

已高于量子化学中的化学精度，但受限于计算资源，实际计算结果并未达到文献[13]中的精度. 本研究所用资源是一张 GTX 1650（3G）显卡和一张 Titan xp（12G）显卡. 此时，使用 Transformer 结构后的 TFN 达到等同费米网络的精度，其中间层使用的参数量却不足费米网络的 2/3. 在单电子流首层中 TFN 使用了HT，但因为 HT 只有一层，且只针对电子–原子核间距离的绝对值进行处理，其参数数量相对较小，对神经网络整体参数的量级影响不大，具体参数数量如表 2 所示. 此处，费米网络选用单电子流 32 个隐藏单元，双电子流 8 个隐藏单元. TFN 中 HT、ST 和 DT 的自注意力机制参数和线性连接单元数量分别选用（16，4）、（32，16）和（16，8）.

表 2　TFN 与费米网络参数的对比

Tab. 2　Comparisons of parameters of the TFN and FermiNet

神经网络	单电子流每层隐藏层参数数量	双电子流每层隐藏层参数数量	HT 参数数量	斯莱特行列式相关参数
TFN	2 144	112	560	同原费米网络
FermiNet	3 584	64	无	同原费米网络

4　结论与展望

本研究采用包含自注意力机制的 Transformer 结构替换网络中的线性连接结构，对费米网络进行了研究和改进，得到以下结论：

（1）研究了费米网络的结构并计算了引入不同自注意力机制的改进. 以小分子系统的基态能计算为例，对使用 Transformer 结构后系统内部参数对神经网络优化体系基态能的影响进行测试，并给出参数选择的建议，结果表明自注意力机制参数带来的优化效果优于线性连接参数的效果.

（2）对比了 TFN 和费米网络的计算结果. 在小分子的计算中，TFN 使用了比费米网络更少的参数，但得到了与费米网络相同精度的结果，TFN 的单电子流参数数量约缩减为费米网络的 60%，双电子流参数数量为费米网络的 175%，总体参数数量缩减约为费米网络的 75%. 上述结果说明 TFN 对费米网络结构的改进是有效的.

此外，Transformer 结构的自注意力机制中存在一些中间变量需要存储，在经过权值共享沿电子数的维度拓展后，这些中间变量占用的显存会迅速膨胀. 对此，在计算资源允许的条件下，可以采用 DeepSpeed 提供的 ZeRO-OffLoad 框架，将中间变量存储在内存中，降低显存压力，这使得 Transformer 结构有可能被推广到较大系统中计算.

致谢　感谢中国人民大学卢仲毅教授建议的费米网络（FeimiNet）研究方向以及利用自注意力机制（self-attention）优化费米网络的研究思路. 感谢中国人民大学贺荣强副教授的讨论和指导. 感谢中国人民大学物理系提供的 GPU 计算资源.

参考文献：

[1] KOHN W，SHAM L J. Self-consistent equations including exchange and correlation cffects[J]. Physical Review，1965，140（4A）：A1133-A1138.

[2] CALLAWAY J，MARCH N H. Density functional methods：Theory and applications[J]. Solid State Physics，1984，38：135-221.

[3] KRYACHKO E S，LUDEÑA E V. Density functional theory：Foundations reviewed[J]. Physics Reports，2014，544（2）：123-239.

[4] BARTLETT R J，MUSIAŁ M. Coupled-cluster theory in quantum chemistry[J]. Reviews of Modern Physics，2007，79（1）：291-352.

[5] FOULKES W M C，MITAS L，NEEDS R J，et al. Quantum Monte Carlo simulations of solids[J]. Reviews of Modern Physics，2001，73（1）：33-83.

[6] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60（6）：84-90.

[7] SILVER D，HUANG A，MADDISON C J，et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature，2016，529（7587）：484-489.

[8] SCHÜTT K T，GASTEGGER M，TKATCHENKO A，et al. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions[J]. Nature Communications，2019，10：5024.

[9] MILLS K，SPANNER M，TAMBLYN I. Deep learning and the Schrödinger equation[J]. Physical Review A，2017，96（4）：042113.

[10] SINITSKIY A V，PANDE V S. Physical machine learning outperforms "human learning" in Quantum Chemistry[EB/OL]. 2019：arXiv：1908.00971v2[physics.chem-ph]. https：//arxiv.org/abs/1908.00971v2.

[11] CARLEO G，TROYER M. Solving the quantum many-body problem with artificial neural networks[J]. Science，2017，355（6325）：602-606.

[12] CHOO K，MEZZACAPO A，CARLEO G. Fermionic neural-network states for ab-initio electronic structure[J]. Nature Communications，2020，11：2368.

[13] PFAU D，SPENCER J S，MATTHEWS A G D G，et al. Ab initio solution of the many-electron Schrödinger equation with deep neural networks[J]. Physical Review Research，2020，2（3）：033429.

[14] HERMANN J，SCHÄTZLE Z，NOÉ F. Deep-neural-network solution of the electronic Schrödinger equation[J]. Nature Chemistry，2020，12（10）：891-897.

[15] SPENCER J S，PFAU D，BOTEV A，et al. Better，faster Fermionic neural networks[EB/OL]. 2020：arXiv：2011.07125[cs.LG]. https：//arxiv.org/abs/2011.07125.

[16] FIRAT O，CHO K，BENGIO Y. Multi-way，multilingual neural machine translation with a shared attention mechanism[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. San Diego，California. Stroudsburg，PA，USA：Association for Computational Linguistics，2016.

[17] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Minneapolis，Minnesota，USA：Association for Computational Linguistics，2019.

[18] LEE J S，HSIANG J. Patent claim generation by fine-tuning OpenAI GPT-2[J]. World Patent Information，2020，62：101983.

[19] BOOTH G H，ALAVI A. Approaching chemical accuracy using full configuration-interaction quantum Monte Carlo：A study of ionization potentials[J]. The Journal of Chemical Physics，2010，132（17）：174104.

[20] JASTROW R. Many-body problem with strong forces[J]. Physical Review，1955，98（5）：1479-1484.

[21] HORNIK K，STINCHCOMBE M，WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks，1989，2（5）：359-366.

[22] KINGMA D P，BA J. Adam：A method for stochastic optimization[EB/OL]. 2017：arXiv：1412.6980[cs.LG]. https：//arxiv.org/abs/1412.6980v9.

[23] MARTENS J，GROSSE R. Optimizing neural networks with Kronecker-factored approximate curvature[C]//Proceedings of the 32nd International Conference on Machine Learning：Proceedings of Machine Learning Research. Lille，France. ICML，2015.

（责任编辑 亢原彬）

《天津师范大学学报：自然科学版》投稿须知

1. 来稿须内容新颖、论点明确、论证科学、数据可靠、文字精炼.

2. 请附中英文摘要、中图分类号和关键词（3~8 个）. 摘要内容包括目的、方法、结果、结论.中英文摘要须一致.

3. 稿件须符合编辑出版标准化要求，量和单位应符合国家标准的有关规定.

4. 文中只附必要的图表. 插图须清晰，线条均匀，字符易辨认并符合量和单位的国家标准. 表格采用三线表. 同时要标出中英文图序、图题和表序、表题.

5. 文中需列出近几年国内外有关本课题的研究文献，只列出文中引用的、公开发表的文献作为参考文献.

6. 文中注明基金资助项目名称和编号，作者姓名、性别、出生年月、学历、职称、研究方向、联系电话、E-mail 等. 投稿时请提交 Word 文档.

7. 本刊已加入清华光盘版(中国知网)、万方数据——数字化期刊群、重庆维普中文科技期刊数据库等多家权威数据库，作者发表文章的同时会被以上网站转载，著作权使用费含在本刊所付的稿酬中. 若有异议，请投稿时说明.

8. 编辑部对拟刊发的稿件有权作必要的技术性和文字性修改.

本刊编辑部