研究与开发

绿色AI效率评价模型的构建与应用

马晓亮[1,2,3]，刘英[2,3]，赵汝强[3,4]，杨邦兴[3,5]，高洁[2,3]，邓从健[3,4]

（1. 西安电子科技大学，陕西 西安 710071；
2. 中国电信股份有限公司广州分公司，广东 广州 510620；
3. 马晓亮劳模与工匠人才创新工作室， 广东 广州 510620；
4. 广州云趣信息科技有限公司，广东 广州 510665；
5. 中数通信息有限公司，广东 广州 510630）

摘　要：随着人工智能（artificial intelligence，AI）的兴起，大模型（large language model，LLM）日益成为知识推介和多轮对话的核心技术。伴随而来，AI大模型在数据处理、模型训练和部署过程中的高能耗问题亟须有效评估，以便在模型优化后进行前后量化对比。提出一种AI大模型能耗的评估方法，旨在量化评估AI模型的服务效率（efficiency，E）。该模型使用训练收敛时间（time，T）、模型参数规模（parameter，P）和浮点运算量（floating-point operations，F）等多维度因素，通过构建能源消耗函数C（T，P，F）实现量化分析；同时，运用非线性最小二乘法，得出模型参数。该分析方法不仅适用于电信运营商客服AI模型的运行效率分析，也可泛化于其他行业的AI模型能耗评估。

关键词：绿色AI；智能客服；能源效率评估；大语言模型

中图分类号：TP393

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2024184

Construction and application of a quantitative efficiency assessment model for green AI in intelligent customer service systems

MA Xiaoliang[1,2,3], LIU Ying[2,3], ZHAO Ruqiang[3,4], YANG Bangxing[3,5],
GAO Jie[2,3], DENG Congjian[3,4]

1. Xidian University, Xi'an 710071, China
2. China Telecom Corporation Guangzhou Branch, Guangzhou 510620, China
3. Ma Xiaoliang Innovation Studio for Model Workers and Creative Talents, Guangzhou 510620, China
4. Guangzhou Yunqu Information Technology Co., Ltd., Guangzhou 510665, China
5. China DataCom Co., Ltd., Guangzhou 510630, China

Abstract: The rise of AI and the increasing prominence of LLM are positioned as core technologies for knowledge dissemination and multi-turn dialogues. Alongside their growth, the high energy consumption associated with data

收稿日期：2024−04−29；修回日期：2024−06−12

通信作者：刘英，Liuying2.gd@chinatelecom.cn

processing, model training, and deployment in AI large models is necessitating effective evaluation to facilitate quantitative comparisons before and after model optimization. An assessment method for the energy consumption of AI large models was introduced, aimed at quantitatively evaluating the service efficiency (E) of AI models. This model was incorporated with multiple dimensions such as training convergence time (T), model parameter size (P), and floating-point operations (F), and quantitative analysis was achieved through the construction of an energy consumption function C(T, P, F). Furthermore, by employing the nonlinear least squares method, model parameters were derived. This analysis method was not only applicable to the operational efficiency analysis of AI models used by telecommunications operators but can also be generalized for energy consumption assessment of AI models across various industries.

Key words: green AI, intelligent customer service, energy efficiency assessment, LLM

0　引言

在当前信息化时代背景下，消费者越来越倾向于追求快速、个性化且高效的服务体验[1]。人工智能（artificial intelligence，AI）技术，如自然语言处理（natural language processing，NLP）和机器学习（machine learning，ML），已被广泛应用于智能客服系统中，旨在实现咨询的自动化处理、即时响应以及个性化服务方案的推荐[2]。随着AI技术尤其是大语言模型的广泛应用，AI在数据处理和模型训练等环节所产生的能源消耗问题开始显现，这违背了节能减排的发展趋势，也导致企业运营成本增加。绿色AI作为一种新兴的研究领域，其目标是在不牺牲性能的前提下减少AI系统的能源消耗[3]。

目前，绿色AI领域尚缺乏一种可量化评估其效率的数学模型[4]。而其他行业中的效率评估模型往往只关注单一维度的能耗数据，例如清华大学环境学院王春艳课题组所发布的白色家电能效等级，其评估维度主要是电量消耗[5]。这种单维度的评估方法可能导致对产品真实能效的误解[6]，且未考虑AI技术所特有的训练收敛时间、模型参数规模、浮点运算次数等因素。

本文提出一个量化评价绿色AI效率的数学模型，阐述了该评价模型的构建过程，包括模型的理论基础、参数选择和求解方法，并通过实际案例分析，验证了该评价模型的实用性，为绿色AI的研究与应用提供了一种新的量化评估工具。

1　绿色AI效率评价模型

绿色AI效率评价模型应考虑多个因素，包括训练收敛时间、模型参数规模、浮点运算量、硬件使用效率（例如CPU/GPU的使用效率和内存需求）以及电力消耗指标（例如实时功耗和累计能耗）。在众多因素中，以下几个方面较重要[7]。

（1）训练收敛时间（time，T）：训练收敛时间指的是模型从接收输入到输出稳定且得到准确结果所需的时间跨度。通过应用迁移学习和模型蒸馏等技术，能够有效利用预先训练的知识和简化模型结构，加速模型收敛过程，并减少计算资源消耗，这对于提升系统能效具有重要意义。

（2）模型参数规模（parameter，P）：参数规模直接反映了AI模型的复杂度，与能源消耗呈正相关。模型参数规模变大往往能够提升其处理复杂查询的准确性，但同时也会带来更高的能源和计算成本。采用模型简化技术，如权重剪枝、知识蒸馏和网络量化等，能够有效减少模型参数，降低能源消耗，而不损害模型性能。

（3）浮点运算量（floating-point operations，F）：浮点运算量是衡量AI模型计算复杂度的重要指标。通过采纳更加高效的算法和优化网络结构，可以在保证模型准确率的前提下，减少所需

浮点运算量，进而降低能源消耗。

1.1 模型构建

本文将绿色AI的服务效率E定义为产出与投入之比：服务质量（quality，Q）作为产出的代T，即AI系统所提供的服务价值；能源消耗（consumption，C）则代表了投入，即提供相应服务所需的能源代价。据此，服务效率E被量化为每单位能源投入所能创造的服务价值，即在消耗最低限度的能源下实现最大程度的工作或服务产出，其数学模型表述如下：

$$E=\frac{Q}{C(T,P,F)}\lambda \tag{1}$$

（1）服务质量Q：在评估服务质量时，知识推荐采纳率、用户满意度等指标是关键评价指数。更高的服务质量是AI系统的目标，但不应以牺牲可持续性为代价。

（2）训练收敛时间T：更短的收敛时间通常意味用户体验更好，但这可能会增加能源消耗。

（3）模型参数规模P：描述模型的规模和复杂度。参数越多，模型可能越强大，但也需要更多的计算资源，增加能源消耗。

（4）浮点运算量F：直接与模型运行时的计算量有关。它跟模型复杂度、输入维度和样本数量有关。

（5）常数因子λ：在评价模型中引入了线性缩放技术，即通过常数因子λ调整服务效率值E的数值范围，本文设定常数因子λ为10^5。鉴于原始服务效率值E在比较能耗密集型AI模型时可能呈现极小数值，采用线性缩放技术对其进行调整，以避免数值下溢和计算误差的发生。此种处理优化了数据的可视化展示，确保了在数据表达和成果报告中，效率值差异能被更加直观地捕捉。同时，线性缩放在调整数值范围的过程中保持了数据间的相对比例，为不同AI模型效率的评估与对比提供了有效基准。

为了构建能源消耗函数$C(T,P,F)$，本文借鉴索洛增长模型（Solow growth model）表达式[8]：

$$\gamma=K^{\alpha}\times(AL)^{1-\alpha} \tag{2}$$

其中，γ代表经济总产出，K代表资本存量，L指劳动力投入，A为量化单位劳动力的技术水平，α是资本产出弹性系数，体现了资本投入在产出形成中的相对贡献度。

在能源消耗函数中，训练收敛时间T、模型参数规模P以及浮点运算量F构成了AI系统所需的资源投入，它们代表了提供服务所必须承担的能源成本。这些投入在某种程度上可以类比于传统经济模型中的资本存量K和劳动力L投入。基于此类比，本文提出索洛式的能源消耗函数：

$$C(T,P,F)=A_T T^{\beta_1}\times A_P P^{\beta_2}\times A_F F^{\beta_3} \tag{3}$$

（6）效率系数：A_T、A_P、A_F代表训练收敛时间T、模型参数规模P以及浮点运算量F的效率系数，例如，A_T代表技术效率与训练收敛时间的关系，即随着算法优化、计算架构改善或者更高效的训练策略的实施，在单位时间内通过训练获得的模型性能提升的技术效率。

（7）能耗弹性：能耗弹性系数β_1、β_2和β_3量化了训练收敛时间T、模型参数规模P以及浮点运算量F对总能耗的相对影响。其数值大小反映了在其他条件不变的前提下，相应要素在能源消耗中的敏感性。若某一弹性系数大于1，则意味该要素的增加会导致能耗以超线性速度增长；相反，若弹性系数小于1，则表明能耗增长速度将低于要素增长速度。

考虑AI系统的性能和能耗模式可能呈现高度非线性和复杂的依赖关系，传统的线性参数估计方法可能无法准确捕捉这些特性，本文运用非线性最小二乘法拟合技术求解模型参数，非线性最小二乘法的优势在于不依赖传统参数模型对数据分布形态的限制性假设，能确保模型的准确性与鲁棒性。

1.2　数据来源

本文选取了某省会城市的电信运营商客服中心的通话记录数据集，该数据集涵盖了2023年5月1日至7月31日3个月的时间跨度，累计收录了500 000条客户服务通话记录，其中包括投诉和咨询等各类交互信息。在AI模型选择方面，采用了通用大语言模型ChatGLM2-6B作为基准模型，旨在评估其在电信运营商客服问答系统中的应用性能。在数据处理过程中，采取了多种数据预处理技术，包括数据清洗、去重、格式化以及特征提取等，以确保数据集的质量和一致性。此外，为了保护用户隐私，所有个人身份信息均经过匿名处理。

1.3　模型求解

绿色AI效率评估数学模型中的服务质量Q使用知识推荐的采纳率；能源消耗C函数中的参数T为模型相对于通用模型ChatGLM2-6B的收敛时间；参数P为大模型参数量级（为方便计算，以60亿计为一个量级；F为大模型浮点运算量（例如，INT8则取值为8，INT4则取值为4，而FP16则取值为16）。

（1）选取在浮点运算精度分别为FP16、INT8、INT4条件下的训练集，得出了相应的服务质量指标Q值和大模型服务效率E值。将绿色AI效率评价模型公式进行代码封装，能源消耗函数C关键代码如下：

```
#能源消耗计算函数
#T(训练收敛时间),P(模型参数规模), F(浮点运算量)
#A_T(训练收敛时间的效率系数),A_P(模型参数规模的效率系数), A_F(浮点运算量的效率系数)
#β_1(训练收敛时间的弹性系数), β_2(模型参数规模的弹性系数), β_3(浮点运算量的弹性系数)
def func_C( T, P, F, A_T, A_P, A_F, β_1, β_2, β_3):
```

而服务效率计算函数E关键代码如下：

```
#服务效率计算函数
#x(T(训练收敛时间), P(模型参数规模), F(浮点运算量)及Q(服务质量))
#A_t(训练收敛时间的效率系数),A_p(模型参数规模的效率系数),A_f(浮点运算量的效率系数)
#β_1(训练收敛时间的弹性系数),β_2(模型参数规模的弹性系数),β_3(浮点运算量的弹性系数)
def func_C( T, P, F, A_T, A_P, A_F, β_1, β_2, β_3):
    T, P, F, Q = x
    C = func_C( T, P, F, A_T, A_P, A_F, β_1, β_2, β_3)
    return Q / C
```

（2）为了降低非线性最小二乘法的计算开销，基于运营经验，预设定一组效率系数（A_T，A_P，A_F）和能耗弹性系数（β_1，β_2，β_3）的初始值，分别为5、2、3以及4、2、1。

（3）采用非线性最小二乘法计算，得出了各效率系数和能耗弹性系数的最优取值，非线性最小二乘法工程关键代码如下。

```
#参数值范围限定
params_bounds = ( [0, 0, 0, 1, 1, 1], [10, 10, 10, 10, 10, 10] )
#初始参数值
A_T = 5 #训练收敛时间的效率系数
Ap = 2 #模型参数规模的效率系数
A_F = 3 #浮点运算量的效率系数
β_1 = 4 #训练收敛时间的弹性系数
β_2 = 2 #模型参数规模的弹性系数
β_3 = 1 #浮点运算量的弹性系数
#能源消耗计算函数
#T(训练收敛时间), P(模型参数规模), F(浮点运算量)
#A_T(训练收敛时间的效率系数), Ap(模型参数规模的效率系数), A_F(浮点运算量的效率系数)
#β_1(训练收敛时间的弹性系数), β_2(模型参数规模的弹性系数), β_3(浮点运算量的弹性系数)
def func_ C( T, P, F, A_T, A_P, A_F, β_1, β_2, β_3):
    return A_T * T **β_1 * Ap * P **β_2 * A_F * F **β_3
```

#服务效率计算函数

#x(T(训练收敛时间), P(模型参数规模), F(浮点运算量)及Q(服务质量))

#A_T(训练收敛时间的效率系数), Ap(模型参数规模的效率系数), A_F(浮点运算量的效率系数)

#β_1(训练收敛时间的弹性系数), β_2(模型参数规模的弹性系数), β_3(浮点运算量的弹性系数)

def func_E(x, ,A_T, A_P, A_F, β_1, β_2, β_3):

$T, P, F, Q = x$

C = func_C($T, P, F, A_T, A_P, A_F, \beta_1, \beta_2, \beta_3$)

return Q / C

进行非线性最小二乘拟合

params, covariance=curve_fit(func_E, (T_arr, P_arr,F_arr,Q_arr),y_arr p0=[AT,Ap,AF,β1,β2,β3,], maxfev=10000,bounds=params_bounds)

在150次试验数据的支持下，非线性最小二乘法的结果显示R^2值达到0.81，说明该模型对数据有较好的拟合程度，其误差平方和解释了81%的因变量的变化，而拟合的效率系数和能耗弹性系数具体值为：A_T=6，A_P=5，A_F=5，β_1=1.0，β_2=2.0，β_3=1.0。模型参数求解值见表1。

表1 模型参数求解值

类型	参数名	求解值
效率系数	A_T	6
效率系数	A_P	5
效率系数	A_F	5
能耗弹性系数	β_1	1.0
能耗弹性系数	β_2	2.0
能耗弹性系数	β_3	1.0

（4）所得模型

$$E = \frac{Q}{C(T,P,F)}\lambda \quad (4)$$

$$C(T,P,F) = 6 \times T^1 \times 5 \times P^2 \times 5 \times F^1 \quad (5)$$

2 AI模型的节能策略

2.1 优化系统框架

研究者们已经探索并实施了一系列策略以提升框架效能，同时降低能源消耗。这些策略包括但不限于参数共享、知识蒸馏以及网络剪枝等先进技术。

（1）参数共享强化

参数共享允许模型不同部分共享相同参数，减少模型总参数数量。这种方法适用于循环神经网络（recurrent neural network，RNN）和卷积神经网络（convolutional neural network，CNN），这两种网络常用于处理自然语言处理（NLP）任务，如语音识别和文本理解[9]。参数共享能减少模型复杂性，降低过拟合风险，减少计算资源的需求。

（2）知识蒸馏法应用

知识蒸馏是将一个大型、复杂的模型（教师模型）知识传递给一个小型、更高效的模型（学生模型）。这种方法可以在保持模型性能的同时显著减少模型大小和计算需求[10]。通过知识蒸馏，智能客服系统可以更快速响应用户查询，同时减少能源消耗。

（3）网络剪枝的有效实践

网络剪枝通过去除模型中的冗余连接和权重来降低模型内存占用和计算需求，同时保持模型性能[11]。在智能客服系统中，网络剪枝可以提高数据传输效率，尤其在移动网络环境下，这有助于提供实时服务响应。

2.2 采用高效学习和训练方法

运用能量效率高的学习算法和训练方法，已成为提升整体系统能效的关键。通过这种方式，不仅优化了算法的计算效率，还在保证服务质量的前提下，降低了对计算资源的需求。这些方法包括高效权重初始化、优化的正则化策略以及自动化机器学习（automated machine learning，AutoML）的应用等。

（1）高效权重初始化

在深度学习中，合适的权重初始化对于训练的稳定性和速度至关重要。不当的初始化可能导致梯度消失或爆炸，阻碍模型学习过程。Xavier初始化和He初始化是两种常用的权重初始化方法，通过考虑输入和输出单元的数量来调整权重的初始规模[12]，有助于保持在网络各层的激活值和梯度的分布均衡，加速收敛并减少训练时间。

（2）优化的正则化策略

正则化技术有助于提高模型的泛化能力，并防止过拟合。Dropout是一种正则化技术，它在训练过程中随机关闭部分神经元，迫使网络学习具有更加鲁棒的特征。此外，L1和L2正则化通过对权重施加惩罚，促使模型学习更小、更分散的权重，降低模型复杂度和过拟合风险[13]。

（3）自动化机器学习（AutoML）的应用

在实施高效学习和训练方法时，如何处理大量数据和复杂模型成为一大挑战。AutoML的目标是自动化机器学习工作流程中的关键决策环节，如模型选择、超参数调整和特征工程，以加速新任务的学习过程，并同时处理多个相关任务[14]，有助于发现更加高效的模型，降低能源消耗。

2.3　提升数据应用效率

数据处理效率的优化是实现绿色人工智能战略的核心。采纳创新性的数据处理技术，可以在不牺牲服务质量的前提下，降低数据处理阶段的能源消耗。这些技术包括主动学习方案的创新、小样本学习策略等，它们作用于数据的传输和处理流程，以减少所需的计算资源和能源投入。

（1）主动学习方案的创新

主动学习是一种选择性标注数据的方法，可以减少标注成本和时间。在智能客服系统中，主动学习算法能够识别对模型训练最有价值的数据样本，优先标注这些样本[15]。这种方法提高了数据利用率，减少了不必要的计算。

（2）小样本学习的挑战与应对

在应对高度特定化请求（如地区特有方言或少见问题）时，可能会遇到训练数据匮乏的问题。小样本学习通过元学习、模型微调和少量样本增强等技术，使得模型能够从有限数据中快速学习并适应新任务。这些技术提高了数据有效性，降低了对大量标注数据的依赖[16]。小样本学习技术解决了如何从有限数据中快速学习的问题，这对于应对偶发性事件具有价值。

3　应用案例

某省级电信运营商的客户服务系统部署大语言模型（LLM）时，面临了多项挑战，尤其是在图形处理单元（GPU）资源配置方面，高性能GPU能够提供必要的计算能力以支持LLM的运行，但持续运行能耗较高，这与推广节能减排的理念存在一定的冲突。

针对上述问题，该运营商采纳一套综合策略：采用轻量化的大语言模型（LightLLM）来优化网络结构，以此降低对GPU资源的依赖，引入FlashAttention算法，通过对注意力计算过程进行优化，提升模型推理的速度。LangChain LLM服务实施框架如图1所示。

该运营商采用了本文构建的绿色AI效率评价

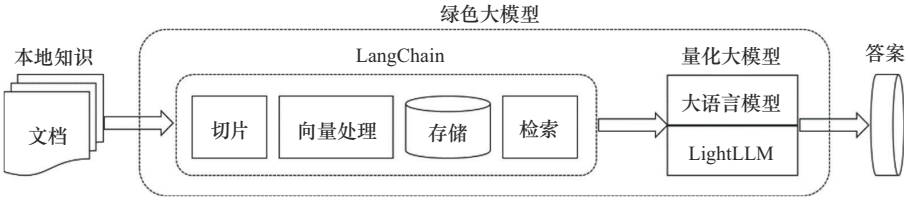

图1　LangChain LLM服务实施框架

模型，进行优化前后AI系统效率的量化对比。

优化前的AI系统效率：通用大模型ChatGLM2-6B的参数T相对收敛时间为1个单位，大模型参数量P为62亿（即约为1个量级），浮点运算量F为16，代入绿色AI效率评价模型公式后得出其能源消耗C函数值为2 400；而ChatGLM2-6B的采纳率Q为21.5%,求得优化前的服务效率E为8.96。

$$C=6\times1^{1}\times5\times1^{2}\times5\times16^{1}=2\,400 \tag{6}$$

$$E=\frac{0.215}{2\,400}\times10^{5}=8.96 \tag{7}$$

优化后的AI系统效率：绿色大模型的参数T相对收敛时间为0.8，大模型参数量P为ChatGLM2-6B的参数量加上客服本地知识的参数量约为62亿（即约为1个量级），浮点运算量F为8，得出其能源消耗C'函数值为960；而ChatGLM2-6B的知识推荐采纳率Q为74.8%,求得优化后的服务效率E'为77.9，即$E'=8.69E$。

$$C'=6\times0.8^{1}\times5\times1^{2}\times5\times8^{1}=960 \tag{8}$$

$$E'=\frac{0.748}{960}\times10^{5}=77.9 \tag{9}$$

对比结果显示：优化后的AI系统在服务效率E'上实现了769%的显著提升。优化前后AI系统的服务效率对比见表2。

表2 优化前后AI系统的服务效率对比

对比项	模型	服务效率E
优化前	ChatGLM2-6B	8.96
优化后	绿色大语言模型	77.9
提升		+769%

4 结束语

大型语言模型在智能信息处理中展现出卓越的性能，在理解和生成内容方面能为用户提供精准的服务。然而，这些模型的训练和运行往往消耗大量计算资源，并伴随高能耗的问题。在实际的智能服务场景中，高能耗不仅增加了运营成本，还对环境产生负面影响。因此，在部署大型模型提升智能客服系统性能时，须权衡其性能与能源效率之间的关系。

本文所提量化评价绿色AI效率的数学模型，考虑诸如训练收敛时间、模型参数规模和浮点运算量等多维度因素，为评估和比较不同AI模型的能效提供了新的量化工具，有助于在能源效率方面做出明智的决策，也为其他领域追求高效且能耗友好的AI应用提供参考和借鉴。

参考文献：

[1] BOCCARDI F, HEATH R W, LOZANO A, et al. Five disruptive technology directions for 5G[J]. IEEE Communications Magazine, 2014, 52(2): 74-80.

[2] WU Q. Optimization of AI-driven communication systems for green hospitals in sustainable cities[J]. Sustainable Cities and Society, 2021, 72: 103050.

[3] 马晓亮, 刘英, 杜德泉, 等. 运营商智能客服的关键技术和发展趋势[J]. 电信科学, 2023, 39(5): 76-89.
MA X L, LIU Y, DU D Q, et al. Key technologies and development trends of intelligent customer service for operators[J]. Telecommunications Science, 2023, 39(5): 76-89.

[4] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep learning in NLP[EB]. 2019: 1906.02243.

[5] 魏泽洋, 刘毅, 王春艳, 等. 环境计算：概念、发展与挑战[J]. 清华大学学报(自然科学版), 2022, 62(12): 1839-1850.
WEI Z Y, LIU Y, WANG C Y, et al. Environmental computing: concept, evolution, and challenges[J]. Journal of Tsinghua University(Science and Technology), 2022, 62(12): 1839-1850.

[6] ALLOUHI A, EL FOUIH Y, KOUSKSOU T, et al. Energy consumption and efficiency in buildings: current status and future trends[J]. Journal of Cleaner Production, 2015, 109: 118-130.

[7] SCHWARTZ R, DODGE J, SMITH N A, et al. Green AI[J]. Communications of the ACM, 2020, 63(12): 54-63.

[8] SOLOW R M. A contribution to the theory of economic growth The Quarterly Journal of Economics, 1956, 70(1): 65-94.

[9] DENIL M, SHAKIBI B, DINH L, et al. Predicting parameters in deep learning[J]. Advances in Neural Information Processing Systems, 2013, 26.

[10] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB]. 2015: 1503.02531.

[11] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network[J]. Advances in Neural Information Processing Systems, 2015, 28.

[12] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9:249-256.
[13] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15: 1929-1958.
[14] JIN H, SONG Q, HU X. Auto-keras: an efficient neural architecture search system[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2019: 1946-1956.
[15] REN P Z, XIAO Y, CHANG X J, et al. A survey of deep active learning[J]. ACM Computing Surveys, 2022, 54(9): 1-40.
[16] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[J]. Advances in Neural Information Processing Systems, 2016: 3637-3645.

[作者简介]

马晓亮（1973- ），男，博士，中国电信股份有限公司广州分公司副总经理、正高级工程师，华南理工大学工商管理学院讲席教授，马晓亮劳模和工匠人才创新工作室领衔人，主要研究方向为人工智能、自然语言处理和数据安全保护等。

刘英（1979- ），男，现就职于中国电信股份有限公司广州分公司，马晓亮劳模和工匠人才创新工作室成员，主要研究方向为智能客服平台组件、自然语言处理等。

赵汝强（1977- ），男，现就职于广州云趣信息科技有限公司，马晓亮劳模和工匠人才创新工作室成员，主要研究方向为智能服务体系搭建和应用、客服中心的系统建设和运营管理、客户投诉和体验管理等。

杨邦兴（1977- ），男，现就职于中数通信息有限公司，马晓亮劳模和工匠人才创新工作室成员，主要研究方向为智能客服平台组件、语义识别转型。

高洁（1985- ），女，现就职于中国电信股份有限公司广州分公司，马晓亮劳模和工匠人才创新工作室成员，主要研究方向为机器学习、算法运用等。

邓从健（1975- ），男，现就职于广州云趣信息科技有限公司，马晓亮劳模和工匠人才创新工作室成员，主要研究方向为自然语言处理、人机交互、算法开发、模式识别等领域。