Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(3), 844-852
Published Online May 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.153078

基于BiLSTM-CAM的新闻文本分类研究

黄盼望，郑　浩

西京学院计算机学院，陕西　西安

收稿日期：2026年4月12日；录用日期：2026年5月17日；发布日期：2026年5月27日

摘　要

随着信息技术的迅速发展，新闻文本的数量急剧增加，怎么有效地对这些文本进行分类已成为一个重要的研究课题，新闻文本的多样性和复杂性使得传统的文本分类方法面临诸多挑战。本文提出一种BiLSTM-CAM模型旨在解决深度学习算法存在的问题，其在处理序列数据方面的优势，能够有效捕捉文本中的上下文信息，更好地识别新闻文本中的重要特征。在实验中，我们首先对数据集进行了预处理，包括文本清洗、分词和向量化等步骤，对BiLSTM-CAM模型进行了训练和调优。进行了对比实验与消融实验，研究结果表明，所提出的模型在处理数据时，能够有效提升分类性能，改善了传统方法的不足。

关键词

文本分类，新闻文本，BiLSTM-CAM

A Study on News Text Classification Based on BiLSTM-CAM

Panwang Huang, Hao Zheng

School of Computer Science, Xijing University, Xi’an Shaanxi

Received: April 12, 2026; accepted: May 17, 2026; published: May 27, 2026

Abstract

With the rapid development of information technology, the volume of news texts has surged dramatically. How to effectively classify these texts has become a significant research topic; however, the diversity and complexity of news texts pose numerous challenges to traditional text classification methods. This paper proposes a BiLSTM-CAM model designed to address the limitations of deep learning algorithms. Leveraging the advantages of deep learning in processing sequential data, this model can effectively capture contextual information within the text and better identify key features in news texts. In the experiments, we first preprocessed the dataset, including steps such as text cleaning, word segmentation, and vectorization, and then trained and fine-tuned the BiLSTM-

文章引用: 黄盼望, 郑浩. 基于 BiLSTM-CAM 的新闻文本分类研究[J]. 人工智能与机器人研究, 2026, 15(3): 844-852.
DOI: 10.12677/airr.2026.153078

CAM model. We conducted comparative and ablation experiments. The results indicate that the proposed model effectively improves classification performance when processing data, addressing the shortcomings of traditional methods.

Keywords

Text Classification, News Text, BiLSTM-CAM

1. 引言

随着互联网的飞速发展，网络上会产生大量信息，文本分类是解决数据过载的技术之一。新闻文本分类的目标是将新闻文章自动分配到预定义的类别中。随着社交媒体和在线新闻平台的普及，新闻文本的种类和数量不断增加，传统的文本分类方法面临着许多挑战[1]。在当今这个互联网爆炸式发展的时代，网络导致了新闻传播的途径愈加广泛，也使得新闻文本的数量不断增加，导致了用户缺少精力从海量的新闻中查找特定的新闻类别[2]。面对自然语言处理领域中的文本分类问题，在大数据的筛选和预测方面具有明显优势的神经网络方法能起到重要的作用。因此，采用深度学习技术来改善这一问题显得尤为重要。

2. 相关工作与知识

2.1. 新闻文本分类研究现状

目前存在对于新闻文本分类的研究，如王驰宇，针对新闻文本标注数据不足的问题，提出将传统学习模型与变分贝叶斯推断结合，构建轻量化的概率神经网络。该方法通过引入权重的后验分布来建模参数不确定性，在小样本条件下有效提升了模型泛化能力。实验在 THUCNews 数据集上进行，结果显示其 F1 值相比确定性基线模型提升至少 9.09% [3]。徐朋与沈子宁将新闻分类任务重新定义为标题与类别标签之间的语义匹配问题。他们设计了一个孪生神经网络结构，分别编码新闻标题和类别描述，通过对比学习优化。该方法在 THUCNews 数据集上验证了其有效性，分类性能优于传统端到端分类模型[4]。乔京等人提出一种结合 BERT、TextCNN 和 BiLSTM 的混合模型。首先利用 BERT 获取上下文敏感的词向量以增强语义表示，再通过 TextCNN 提取局部关键词特征，同时用 BiLSTM 捕获长距离依赖关系，最后融合三者特征进行分类。在搜狐新闻数据集上的实验表明，该方法准确率达到 92.4%，显著优于单一模型[5]。郝婷与冯赛赛聚焦于新闻标题这一短文本场景，构建了 BERT + LSTM 模型。该模型先使用 BERT 对标题进行深度语义编码，生成每个词的上下文向量，再输入 LSTM 层捕捉序列动态特征，最终通过全连接层完成分类[6]。

当前新闻文本分类研究虽已取得显著进展，但仍存在诸多不足，类别间语义重叠使边界模糊，影响判断准确性，模型在不同媒体或领域间迁移能力弱，泛化性有限，对上下文和外部知识的依赖未被充分建模等，基于此本文提出 BiLSTM-CAM 模型。

2.2. 相关知识

2.2.1. 文本分类

文本分类是自然语言处理中的一项核心任务，旨在依据特定标准或属性将文本自动归入预设类别。

其基本原理是通过提取文本中的有效特征并训练分类模型，实现对新文本的自动判别。该技术广泛应用于新闻主题划分、情感倾向分析、垃圾邮件过滤等场景[7]。如图 1 所示，典型的文本分类流程通常包括以下几个关键步骤：文本预处理、文本向量化、特征筛选、分类模型训练以及分类效果评估。

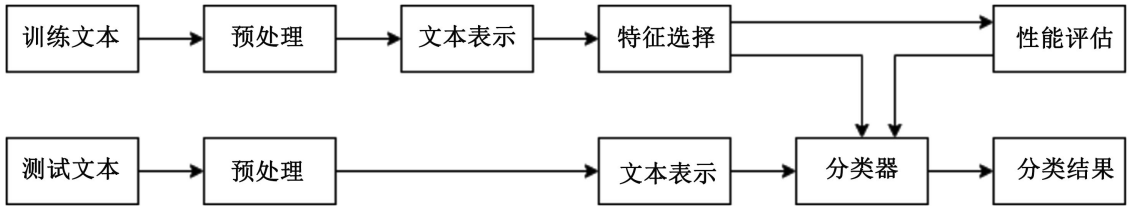

Figure 1. Text classification process
图 1. 文本分类流程

预处理模块主要负责对原始文本数据进行清洗，包括去除噪声、分词、过滤停用词等操作，为后续的特征提取与模型训练奠定基础。特征提取模块则通过为关键词赋予相应权重(如 TF-IDF、词频等)，从文本中筛选出具有判别性的信息。文本表示模块的作用是将经过处理的文本转化为可供算法处理的数值形式，通常表现为特征向量或矩阵，从而支持机器学习方法的应用。随后，文本分类器基于这些结构化的特征数据，利用传统机器学习算法(如朴素贝叶斯、支持向量机)或深度学习模型(如卷积神经网络、循环神经网络)进行训练，最终实现对新文本的自动分类[8]。

在文本表示方法中，词嵌入技术扮演着关键角色。它通过将词汇映射到连续的低维实数向量空间，有效捕捉词语间的语义和语法关系。其核心理念在于：语义相近的词语在向量空间中的距离也较近。相较于传统的稀疏表示(如 One-Hot 编码)，词嵌入能够更高效地表达语言的内在结构。目前主流的词嵌入方法包括 Word2Vec、GloVe、FastText，以及由搜狗等机构发布的中文词向量模型等[9]。这些技术显著提升了文本分类任务中语义理解的准确性和模型的泛化能力。

2.2.2. 循环神经网络

循环神经网络是一种能够处理序列数据的神经网络。与传统的前馈神经网络不同，循环神经网络具有循环连接，这使得它能够在时间维度上保持信息。循环神经网络的核心思想是通过隐藏状态(Hidden State)来记忆之前的信息，从而在处理当前输入时考虑到过去的上下文[10]，如图 2。

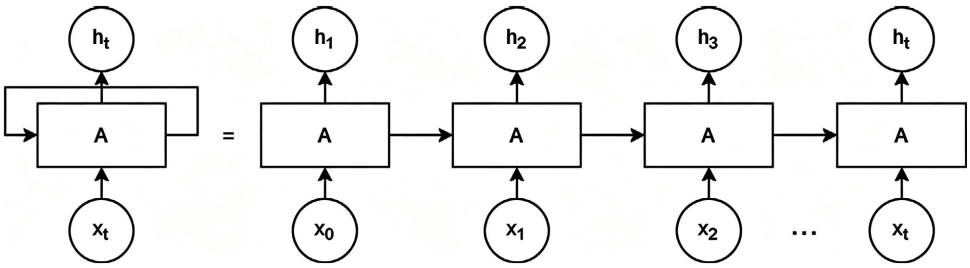

Figure 2. Recurrent neural network
图 2. 循环神经网络

2.2.3. LSTM

尽管 LSTM 在建模长距离依赖方面表现优异，但在处理超长序列时仍可能面临信息衰减的问题。为此，研究者陆续提出了多种改进模型和替代方案。例如，GRU (Gated Recurrent Unit)通过简化 LSTM 的门控结构，在保持较好性能的同时提升了训练效率；而 Transformer 架构则完全摒弃了循环机制，转而采

用自注意力(Self-Attention)机制，显著提高了并行计算能力，并在诸多自然语言处理任务中取得了超越LSTM 的效果[11]。

但在资源受限或序列长度适中的场景下，LSTM 因其结构稳定、易于理解和调优，依然具有较强的实用价值。特别是在文本分类任务中，LSTM 能够逐词处理输入序列，动态捕捉上下文语义，并将最终的隐藏状态或时间步的聚合表示作为文档的语义特征，供后续分类层使用。此外，LSTM 还可与词嵌入技术(如 Word2Vec 或 BERT)结合，构建端到端的深度分类模型，进一步提升分类准确率，故 LSTM 凭借其独特的门控机制和对时序信息的有效建模能力，成为处理序列数据的重要工具之一，尽管新型架构不断涌现，LSTM 仍在许多实际应用中发挥着不可替代的作用[12]，LSTM 结构图，如图 3。

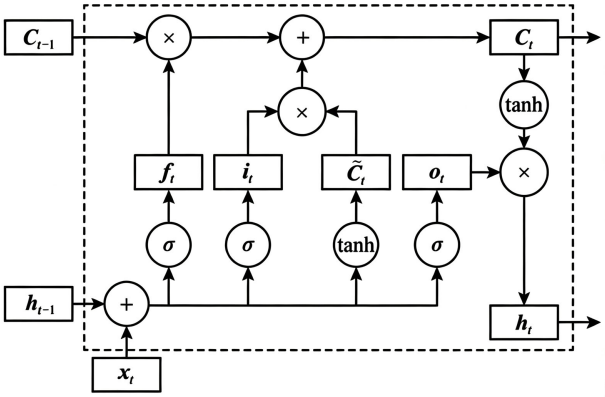

Figure 3. Diagram of the LSTM architecture
图 3. LSTM 结构图

2.2.4. 基于 BiLSTM-CAM 的新闻文本分类模型

为了提高分类精度提出 BiLSTM-CAM 深度学习模型，模型结构如图 4 所示。

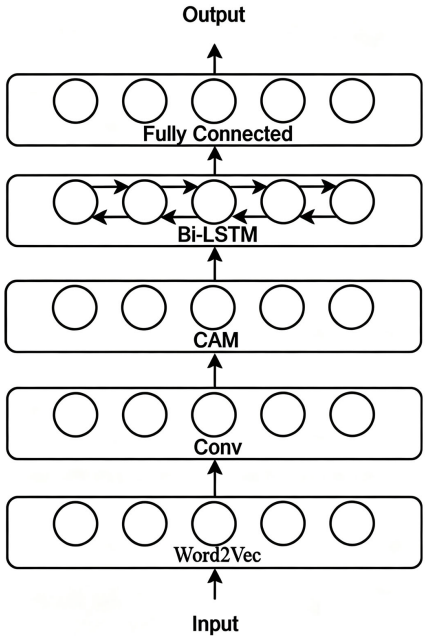

Figure 4. BiLSTM-CAM model structure
图 4. BiLSTM-CAM 模型结构图

1) Convolution Layer

卷积层(Conv)作为编码器结构的关键组成部分，其核心作用在于通过局部感受野机制提取输入序列的时空特征。具体而言，卷积层利用可学习的滤波器对输入数据进行滑动窗口操作，捕捉相邻时间步或空间位置之间的依赖关系，从而实现对高维特征的有效压缩与抽象表示。该过程不仅保留了原始信号中的关键模式信息，还增强了模型对局部结构变化的敏感性。此外，卷积层在编码阶段与后续的双向长短期记忆网络(BiLSTM)和通道注意力机制(CAM)协同工作，进一步提升了特征表达的层次性和语义丰富性，为后续的重构与分类任务提供了更具判别力的潜在表示。

2) Channel Attention Mechanism Layer

注意力机制源于人类认知过程中的选择性信息处理方式。在面对海量复杂信息时，人类依靠有限的认知资源选择性关注重要内容。受此启发，注意力机制使神经网络能够自主学习并聚焦输入数据中的关键特征，显著提升模型性能。

本文采用改进的通道注意力机制(Channel Attention Mechanism, CAM)，输入特征图同时经过全局平均池化和全局最大池化，生成两个 1 × 1 × C 维度的通道描述符，分别提取通道的均值特征和峰值特征，激励阶段(Excitation)，压缩后的特征通过共享的多层感知机(Shared MLP)进行非线性变换，最终经 Sigmoid 激活函数生成通道权重向量。CAM 相当于让模型自动判断哪些语义主题(由不同卷积/BiLSTM 通道编码)对当前新闻类别更相关，从而实现语义级别的特征选择与聚焦。CAM 的计算公式为：

$$M(F)=\sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(F)\right)+\mathrm{MLP}\left(\mathrm{MaxPool}(F)\right)\right) \tag{1}$$

3) BiLSTM Layer

为了进一步提升对序列数据的上下文理解能力，本文运用双向长短期记忆网络(Bidirectional LSTM, BiLSTM)。与单向 LSTM 仅从前向后处理序列不同，BiLSTM 由两个独立的 LSTM 网络组成：一个按时间正向处理输入序列(从过去到未来)，另一个按时间反向处理序列(从未来到过去)。这两个方向的 LSTM 在每个时间步共享同一输入序列，并分别输出对应的隐藏状态。最终的输出是两个方向隐藏状态的拼接或组合，从而使得当前时刻的输出同时融合了历史信息和未来信息，拼接后，融合了序列的双向上下文信息，通过充分利用序列的双向信息，显著提升了模型对复杂时序模式的建模能力，能够全局增强上下文感知与特征表达，提升预测精度。

4) Fully Connected Layer

全连接层(Fully Connected Layer)主要承担特征整合与非线性映射的功能。经过卷积层和 BiLSTM 等模块提取的高维局部及时序特征被展平后输入至全连接层，该层通过全局连接方式对所有特征进行加权组合，实现从局部到整体的信息融合。同时，全连接层引入非线性激活函数，增强模型的表达能力，使其能够学习复杂的非线性关系。在文本分类任务中，其输出作为最终决策依据。

本模型充分利用了局部模式识别(Conv)、长程依赖建模(BiLSTM)和自适应特征选择(CAM)的优势，在保留文本细粒度语义的同时，动态聚焦于任务相关的高层语义通道，从而显著提升新闻文本分类的准确率与鲁棒性。

3. 实验

3.1. 数据集

本研究使用了一个包含 10 种新闻类别的数据集。在实验中，本文将数据集分为训练集和测试集。本文数据集由爬虫技术爬取到的搜狗新闻文本数据组成。数据集包含 180,000 条新闻标题文本，10 个类别，每个类包含约 18,000 条。其中 120,000 条文本作为训练集，60,000 条作为测试集。

3.2. 数据预处理

字符转化为向量，循环神经网络模型和卷积神经网络在进行文本分类时，不能接受变长序列，因为给定的权重参数是固定的不能改变输入规格，所以第一件是就是在数据预处理的时候把每个文本数转化成相同字或词的个数，若定为 32 个词，那么把超过 32 个词的文本标题后面的词去掉，做一个截断操作，直接去前 32 个词，若不够 32 个词，则要进行添加操作，把不够 32 个词的文本标题，添加索引指向空值的 0，长度必须一致，然后把每个词化为向量，每个词对应一个相同维数的向量。

3.3. 实验设置

本文对 BiLSTM-CAM 模型进行了多次训练，并记录每次训练的分类精度。对深度学习网络的参数设置有 Batch-Size (每次处理文本的个数)，Max Sequence Length (最大序列长度)，Number of Epochs (训练次数)等，模型详细超参数设置如表 1。

Table 1. Model hyperparameter settings
表 1. 模型超参数设置

超参数	取值
Batch size	64
Max sequence length	32
Number of epochs	50
Learning rate	0.001
BiLSTM units	256 (每方向)
Conv filters	128
Conv kernel size	3
Dropout	0.5
Optimizer	Adam
Loss function	Cross-Entropy Loss

3.4. 评价标准

衡量算法的评价标准为准确率、精确率、召回率和 F1 值。在评估算法性能时，准确率、精确率、召回率 F1 值是四个至关重要的评价标准。

准确率评估分类模型性能的一个基本指标，通过将模型正确分类的样本数量除以整个数据集的样本总数来计算。其计算方式如下：

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

精确率是指在算法提取出的所有关键词中，正确的关键词所占的比例。具体来说，精确率衡量的是算法在提取关键词时的准确性。高精确率意味着算法能够有效地减少错误提取的情况，从而提高结果的可信度和准确性。

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

召回率则是指在所有实际存在的关键词中，算法成功提取出的正确关键词所占的比例，召回率关注

的是算法的全面性，反映了算法在识别所有相关关键词方面的能力。

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

F1 值是精确率和召回率的加权调和平均值，旨在综合考虑这两个指标的表现。

$$\mathrm{F1}=\frac{2*\mathrm{Pre}*\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{5}$$

3.5. 实验结果

3.5.1. 模型训练实验

模型的训练分别对经济、房地产、股票、教育、科学、社会、政治、体育、游戏与娱乐新闻进行多分类，模型通过训练后，模型新闻分类结果实验结果如表 2。

Table 2. Model news categories
表 2. 模型新闻分类结果

模型类别	Pre (%)	Rec (%)	F1 (%)
Finance	92.05	89.50	90.76
Realty	91.92	93.70	92.80
Stocks	89.80	81.30	85.34
Education	93.35	95.40	94.36
Science	92.46	87.90	90.12
Society	90.13	90.90	90.51
Political	89.16	87.10	88.12
Sports	94.39	97.50	95.92
Game	93.57	92.30	92.93
Entertainment	92.87	92.50	92.68

实验结果表明，融合 BiLSTM 与通道注意力机制的模型在分类精度上能够达到良好表现，BiLSTM 能够有效捕捉文本序列中的长距离依赖关系，而通道注意力机制则进一步增强了模型对关键特征通道的聚焦能力，从而提升了整体分类性能。

3.5.2. 对比与消融实验

Table 3. Comparison of model results
表 3. 模型结果对比

模型类别	ACC (%)	Pre (%)	Rec (%)	F1 (%)
LSTM	76.70	76.68	76.70	76.69
TextCNN	83.58	85.82	84.49	85.14
BiGRU	85.06	87.50	85.10	85.10
BiLSTM + CAM	91.44	91.97	90.81	91.38

从表 3 可以看出，LSTM 作为经典的循环神经网络结构，具备一定的序列建模能力，但在处理长文

本时容易出现梯度消失或信息稀释问题，TextCNN 模型在提取局部特征有着优势，然而，其仍受限于固定感受野和缺乏对关键特征通道的自适应加权，导致部分重要信息被忽略。BiGRU 采用双向结构，能够同时利用前向和后向上下文信息，在一定程度上弥补了单向 LSTM 的不足。但 GRU 结构相比 LSTM 在长期依赖建模上略显薄弱。

BiLSTM 有效增强了对文本序列中长距离依赖关系的建模能力，双向结构使得模型能更全面地理解上下文语义。通道注意力机制(CAM)则通过学习不同特征通道的重要性权重，自动筛选出与当前分类任务最相关的语义特征，抑制冗余或噪声信息。

Table 4. Ablation experiment
表 4. 消融实验

模型类别	ACC (%)	Pre (%)	Rec (%)	F1 (%)
LSTM	76.73	76.68	76.82	76.69
BiLSTM	80.48	80.43	80.46	80.44
LSTM + CAM	86.65	86.58	86.70	86.59
BiLSTM + CAM	91.44	91.97	90.81	91.38

由表 4 得，消融实验通过逐步引入 BiLSTM 和通道注意力机制，分析不同部分对文本分类性能的影响。实验结果表明两个核心模块在提升模型表现中起到有效作用。基础模型 LSTM 的各项指标不高，其在捕捉长距离依赖和完整语义信息方面存在局限，当使用 LSTM 的改进版 BiLSTM 后，各项指标有所提高，这说明，BiLSTM 的双向结构能够同时利用前向和后向上下文信息，增强了对文本整体语义的理解能力，尤其在新闻类文本中，事件发展顺序和前后逻辑关系复杂，双向建模优势明显。在 LSTM + CAM 模型中，性能因为引入了 CAM 使性能得到了提升，因为通道注意力机制能够有效识别并加权重要特征通道，抑制噪声信息，增强了特征表达能力。

本文提出的 BiLSTM + CAM 模型在所有指标上均达到最高水平，这一结果表明，BiLSTM 与 CAM 不是简单叠加，形成了互补协同关系，BiLSTM 提供高质量的上下文感知表示，CAM 对这些表示进行通道级优化，突出关键语义特征，提升了模型的语义理解能力和分类精度。

4. 结语

在本研究中，探讨了深度学习技术在新闻文本分类中的应用，并且使用了 BiLSTM-CAM 模型。随着互联网和社交媒体的迅猛发展，新闻文本的数量和种类急剧增加，传统的文本分类方法面临着许多挑战，尤其是分类结果准确率有待提高。通过采用深度学习模型，能够更好地捕捉文本中的上下文信息和特征，从而提高分类的准确性和效率。实验结果表明，BiLSTM-CAM 模型在分类准确率上有着较好的提升，这不仅验证了深度学习在自然语言处理领域的潜力，也为未来的研究提供了新的方向。

未来的研究可以进一步探索其他深度学习架构，如 Transformer 模型，或结合多种模型的集成方法，以进一步提高文本分类的性能。

参考文献

[1] 李晓英, 杨名, 全睿, 等. 基于深度学习的不均衡文本分类方法[J]. 吉林大学学报(工学版), 2022, 52(8): 1889-1895.

[2] 刘晓琳, 宋营营, 李卓. 基于增强逐点图卷积网络的民航短文本组合分类方法[J/OL]. 北京航空航天大学学报: 1-18. https://doi.org/10.13700/j.bh.1001-5965.2024.0223, 2024-10-30.

[3] 王驰宇. 基于变分贝叶斯的小样本新闻文本分类方法[J]. 中国传媒科技, 2026(1): 149-153.

[4] 徐朋, 沈子宁. 基于孪生神经网络的新闻文本分类方法研究[J]. 计算机与数字工程, 2025, 53(10): 2831-2836.

[5] 乔京, 常承伟, 王哲, 等. 融合语义增强的新闻文本分类研究[J]. 计算机仿真, 2025, 42(6): 72-77.

[6] 郝婷, 冯赛赛. 基于深度学习的文本分类模型研究[J]. 信息记录材料, 2025, 26(6): 114-116+119.

[7] Shi, J., Wei, T. and Li, Y. (2024) Residual Diverse Ensemble for Long-Tailed Multi-Label Text Classification. *Science China Information Sciences*, **67**, 92-105. https://doi.org/10.1007/s11432-022-3915-6

[8] 季天瑶, 王挺韶. 基于词嵌入与卷积神经网络的建筑能耗预测[J]. 华南理工大学学报(自然科学版), 2021, 49(6): 40-48.

[9] Chen, G.Z., Liu, S. and Xu, J.T. (2023) Memory-Boosting RNN with Dynamic Graph for Event-Based Action Recognition. *Optoelectronics Letters*, **19**, 629-634. https://doi.org/10.1007/s11801-023-3028-7.

[10] Li, O., Lei, J., Qin, C., Zhang, Z., Tao, J. and Liu, C. (2024) A Novel Multi-Channel CNN-LSTM and Transformer-Based Network for Diesel Engine Misfire Diagnosis under Different Noise Conditions. *Science China Technological Sciences*, **67**, 2965-2967. https://doi.org/10.1007/s11431-023-2698-2

[11] Gong, J., Liu, X., Zhang, Y., Zhu, F. and Hu, G. (2024) Prediction of Single Cell Mechanical Properties in Microchannels Based on Deep Learning. *Applied Mathematics and Mechanics*, **45**, 1857-1874. https://doi.org/10.1007/s10483-024-3187-6

[12] 孙刘成, 黄润才. 融合 LSTM 和注意力机制的新闻文本分类模型[J]. 传感器与微系统, 2022, 41(9): 38-41.