Application of AI Large Models in Complex Data Processing: Analysis of Algorithm Optimization and Efficiency Improvement Paths

Jing Liu
Information Center, Hunan Post and Telecommunication College, Changsha 410000, Hunan, China
Corresponding author: Jing Liu (e-mail: jing_hnyd@163.com)

ABSTRACT With the rapid development of the digital economy, complex data characterized by high dimensionality, heterogeneity, sparsity, and dynamics has become increasingly prevalent across industrial and scientific applications, making efficient processing a critical challenge for intelligent systems. In particular, the growing volume of multi-source sensing data generated in electromagnetic wave analysis, antenna systems, and wireless communication networks places higher demands on large-scale intelligent computing frameworks. This paper systematically investigates the major bottlenecks of AI large models in complex data processing and proposes a cross-scenario adaptive optimization framework integrating algorithm optimization and multi-dimensional efficiency enhancement strategies. At the algorithm level, adaptive attention architecture and multi-level regularization are introduced to improve feature representation and generalization capability. At the efficiency level, a collaborative "model–data–engineering" optimization scheme combining structured pruning, mixed-precision quantization, and knowledge distillation is developed. Experimental results demonstrate that the optimized 32B model reduces storage volume by 85% to 4.8 GB, decreases memory usage by 80%, improves inference speed by tenfold, and limits task-specific accuracy loss to less than 3%. Practical validations in medical, industrial, and astronomical scenarios further confirm simultaneous improvements in accuracy and computational efficiency. The proposed framework provides an effective solution for complex data processing while offering valuable technical references for electromagnetic signal intelligence, multisource sensor data fusion, and intelligent wireless information processing.

INDEX TERMS AI large models, complex data processing, algorithm optimization, electromagnetic signal processing, multisource data fusion

I. INTRODUCTION

A. RESEARCH BACKGROUND

WITH the rapid development of information technology, data volume within the textile manufacturing and apparel design sectors has experienced explosive growth, and the complexity of fabric performance metrics and spinning process parameters has also increased significantly. This surge in digital information necessitates advanced analytical tools to manage the intricate variables of textile material science and automated production cycles, clinical diagnosis and astronomical exploration effectively. Complex data refers to data collections with high complexity in dimensions, structures, and feature distributions, whose value is difficult to mine using traditional methods; typical datasets in medical, astronomical and textile engineering fields all fully embody the core characteristics of complex data (high dimensionality, heterogeneity, sparsity, dynamics) and are typical carriers for the study of complex data processing methods, which is the logical basis for selecting these datasets as research objects in this paper. Common types of complex data include high-dimensional data, multi-collinear data, and non-linear data, with core characteristics such as high dimensionality, heterogeneity, sparsity, and dynamics, facing processing challenges like the curse of dimensionality and cross-modal fusion difficulties [1].

Fields such as medical care, finance, and scientific research are confronted with the challenge of processing large amounts of text data, and datasets generally have problems such as unbalanced data categories, all of which require complex data processing to support their core businesses [2-4]. However, in terms of data processing, there are pain points including insufficient processing accuracy, low efficiency, and poor adaptability, making it difficult for a single solution to adapt to multiple types of complex data.

AI large models are suitable for complex data processing due to their strong feature extraction capabilities, but they

have core bottlenecks such as insufficient generalization ability, high training costs, excessive memory overhead, high power consumption, and low inference throughput, and weak real-time processing capabilities for dynamic data [5, 6].

B. REVIEW OF DOMESTIC AND FOREIGN RESEARCH STATUS

Domestic and foreign large models have been applied in complex data processing in multiple fields [7-9], but existing solutions mostly focus on a single scenario and lack cross-scenario general solutions. Existing algorithm optimization mainly concentrates on the architecture and training levels [10-12], mostly targeting specific data types, making it difficult to meet the multi-feature fusion needs of complex data. Quantization, pruning, and other technologies are mainstream efficiency improvement methods [13], but the effect of a single technology is limited, and there is a lack of multi-dimensional collaborative optimization mechanisms and unified adaptation standards.

C. RESEARCH CONTENT AND METHODS

The core research content of this paper includes analyzing the core problems of large models in processing complex data, proposing cross-scenario adaptive algorithm optimization and efficiency improvement schemes, and verifying the effectiveness of the schemes through empirical evidence, and analyzing the failure scenarios and performance limitations of the optimized model in processing specific complex data. The research methods adopt a combination of literature research, simulation experiment, and case analysis methods to systematically sort out relevant achievements, select typical datasets to design comparative experiments, and evaluate the feasibility of the schemes through practical cases.

D. RESEARCH SIGNIFICANCE

At the theoretical level, this paper deeply explores the algorithm optimization logic of large models adapting to complex textile engineering data and fabric performance parameters, constructing a multi-dimensional efficiency improvement system that enhances the accuracy of spinning and weaving process modeling. By breaking through the limitations of a single technology, the research provides a robust theoretical framework for the intelligent analysis of fiber characteristics and apparel manufacturing variables, significantly enriching the digital transformation theory of the textile industry.

At the practical level, the proposed optimization scheme can solve the contradiction between accuracy and efficiency, reduce deployment costs, and has important industrial application value in fields such as scientific research, industry, and medical care and textile engineering.

II. CURRENT APPLICATION STATUS AND PROBLEM ANALYSIS OF AI LARGE MODELS IN COMPLEX DATA PROCESSING

A. ANALYSIS OF TYPICAL APPLICATION SCENARIOS

The typical application scenarios of AI large models in complex data processing cover multiple fields, all of which are typical embodiments of the general complex data definition in specific industries:

Textile engineering field: High-dimensional heterogeneous data processing of fabric performance parameters and spinning process metrics, the core is multi-type production data fusion, with pain points of low crossmodal alignment and fusion efficiency for material parameters and production data;

High-dimensional heterogeneous data processing in medical and financial fields: the core is multi-type data fusion, with pain points of low cross-modal alignment and fusion efficiency [14, 15];

Dynamic time-series data processing in industrial and meteorological fields: data is real-time and its distribution changes dynamically, with core problems of high update costs and difficulty in real-time processing [16];

Sparse and unstructured data processing in social media and remote sensing fields: the proportion of effective information is low, and the key difficulty is the extraction of effective features [17, 18].

B. CORE PROBLEMS IN EXISTING APPLICATIONS

Existing applications have core problems in multiple aspects:

At the algorithm level: insufficient generalization ability, high overfitting risk, and inaccurate feature capture, making it difficult to adapt to multiple types and dynamic scenarios;

At the efficiency level: showing the the characteristics of high training costs, excessive memory overhead, high power consumption, and low inference throughput, which are difficult to meet real-time requirements;

At the data adaptation level: facing problems such as poor compatibility, weak dynamic adaptation ability, and difficulty in small-sample adaptation, which increase computational costs and affect accuracy.

C. ANALYSIS OF PROBLEM CAUSES

The causes of the above problems mainly include three aspects:

(1) Most model architectures are designed for general purposes, with insufficient matching degree with complex data features, making it difficult to adapt to high dimensionality, heterogeneity and other needs;

(2) Unreasonable resource allocation in training and inference: training focuses on parameter update while neglecting preprocessing, and inference does not dynamically allocate resources, resulting in poor efficiency and effectiveness;

(3) Lack of targeted optimization schemes: existing schemes adopt a "one-size-fits-all" approach and fail to

design differentiated strategies combined with data features, making it difficult to balance accuracy and efficiency.

III. ALGORITHM OPTIMIZATION STRATEGIES FOR AI LARGE MODELS IN COMPLEX DATA PROCESSING

A. MODEL ARCHITECTURE OPTIMIZATION FOR COMPLEX DATA FEATURES

1) Adaptive Architecture Optimization Based on Attention Mechanism

Aiming at the high dimensionality and sparsity of complex data, a dynamic sparse attention architecture is proposed:

(1) By calculating feature importance scores, dynamically retain attention weights of key features and eliminate attention calculations of redundant features, reducing the computational complexity of attention from $O(n^2)$ to $O(n \log n)$;

(2) Introduce cross-feature attention mechanism to realize correlation modeling of features in different dimensions and improve the feature capture ability of high-dimensional data.

Experimental results show that in high-dimensional gene sequencing data processing, this architecture can reduce computational complexity by 40% while ensuring that the accuracy drop does not exceed 2%.

2) Multi-modal Fusion Architecture Optimization

For heterogeneous complex data, a two-stage multi-modal fusion architecture is designed:

The first stage is the feature alignment layer, which converts different types of data (text, images, audio, textile material parameters) into feature vectors of uniform dimension through modal adaptive encoding to solve the problem of cross-modal feature differences;

The second stage is the deep fusion layer, which introduces cross-attention mechanism and gating unit to dynamically adjust the weights of different modal features and strengthen the contribution of effective features.

In medical multi-modal case processing and textile engineering multi-source production data processing, the fusion accuracy of this architecture is improved by more than 15% compared with traditional architectures.

3) Lightweight Architecture Design

Aiming at the demand for complex data processing on edge devices, an architecture mode of "lightweight sub-model (1.2B parameters) derived from the 32B main model + task-specific head" is proposed: The backbone network adopts a design combining depthwise separable convolution and sparse Transformer to reduce basic computational load;

The task-specific head is customized according to specific complex data scenarios (such as time-series, images, textile production defect detection) to improve task adaptability.

At the same time, a dynamic channel pruning module is introduced to real-time adjust the number of network channels to adapt to data of different complexities. In industrial equipment monitoring data processing and textile weaving process data processing, the model volume is reduced by 60% compared with the standard model.

B. TRAINING ALGORITHM OPTIMIZATION FOR COMPLEX DATA PROCESSING

1) Adaptive Learning Rate Optimization

Combined with the distribution characteristics of complex data, an adaptive learning rate strategy based on data complexity is proposed:

Quantify data complexity by calculating the variance and sparsity of batch data; adopt a small learning rate when the complexity is high to ensure training stability, and a large learning rate when the complexity is low to improve training efficiency.

At the same time, introduce learning rate warm-up and decay mechanisms to avoid non-convergence problems caused by excessively large learning rates in the initial training stage. In dynamic time-series data training, the convergence speed is improved by 30% compared with fixed learning rates.

2) Regularization and Anti-noise Algorithm Optimization

Aiming at the noise problem in complex data, a multi-level regularization strategy is proposed:

(1) Adopt an adaptive noise filtering module at the input layer to eliminate obviously noisy data;

(2) Introduce DropBlock regularization at the feature layer to randomly mask some redundant feature channels;

(3) Adopt label smoothing technology at the output layer to reduce the impact of noisy data on label prediction.

In the processing of noisy industrial monitoring data, this strategy improves model robustness by more than 20% and reduces overfitting rate by 15%.

3) Transfer Learning and Incremental Learning Algorithm Optimization

Aiming at cross-domain and dynamic data adaptation problems, a two-stage transfer-incremental learning framework is constructed:

The first stage transfers knowledge from source domain large models to target domains through domain-adaptive transfer learning to solve the training problem of small-sample complex data;

The second stage adopts incremental learning algorithm, and only updates the parameters related to new data in the model through the combination of parameter freezing and fine-tuning, realizing efficient adaptation of dynamic data and reducing model update costs by 70%.

C. INFERENCE ALGORITHM OPTIMIZATION BASED ON COMPLEX DATA CHARACTERISTICS

1) Dynamic Inference Path Optimization

Design a dynamic inference path selection mechanism based on data complexity:

Real-time evaluate the complexity of input data (dimensionality, heterogeneity, noise level) through a preprocessing module; adopt a complete inference path for high-complexity data to ensure accuracy, and a simplified

inference path (skipping some redundant layers) for low-complexity data to improve efficiency. In social media text data processing, this mechanism can dynamically adjust the inference process according to text length and complexity, improving inference speed by 50%.

2) Approximate Inference Algorithm Optimization

For complex data scenarios with moderate accuracy requirements, an approximate inference algorithm based on probability is introduced: Construct a feature probability distribution model; in the inference process, only perform precise calculations on high-probability effective features and approximate calculations on low-probability redundant features, greatly improving inference efficiency within the acceptable accuracy loss (not exceeding 3%).

In meteorological data prediction, this algorithm can shorten the inference time from 2 seconds to 0.3 seconds.

3) Multi-task Inference Fusion Optimization

Aiming at the demand for multi-task collaborative processing in complex scenarios, a multi-task inference architecture separating shared and specific features is proposed:

(1) Different tasks share the basic feature extraction module to reduce redundant calculations;

(2) Design specific inference heads for each task to ensure task accuracy.

Coordinate the resource occupation of different tasks through the dynamic attention weight allocation mechanism. In medical multi-task diagnosis (disease identification + prognosis prediction), the inference efficiency is improved by 45%, and the accuracy of each task is maintained above 90%.

D. EFFECTIVENESS VERIFICATION LOGIC OF ALGORITHM OPTIMIZATION

Construct a "scenario-objective-indicator" three-dimensional verification system:

(1) Divide verification scenarios according to the types of complex data (high-dimensional, heterogeneous, dynamic, etc.);

(2) Clarify optimization objectives under each scenario (accuracy improvement, generalization ability enhancement, etc.);

(3) Set targeted evaluation indicators, such as accuracy-dimensionality compression ratio for high-dimensional data, fusion accuracy-fusion efficiency for heterogeneous data, and adaptation speed-accuracy loss for dynamic data.

Comprehensively verify the effectiveness of optimization strategies through comparative experiments (models before and after optimization) and ablation experiments (individual contributions of each optimization module).

IV. EFFICIENCY IMPROVEMENT PATHS OF AI LARGE MODELS IN COMPLEX DATA PROCESSING

A. MODEL-LEVEL EFFICIENCY IMPROVEMENT PATHS

At the model level, the core is to "reduce volume and complexity", adopting a synergistic multi-technology fusion strategy of structured pruning, mixed-precision quantization (INT8/INT4) and knowledge distillation instead of a single technology, which avoids the significant degradation of model performance caused by a single extreme compression method. The adaptation scenarios and effect comparison are shown in Table 1.

This paper proposes a collaborative scheme of "structured pruning-mixed-precision quantization-knowledge distillation-dynamic memory sparsification". Tests on Tongyi Qianwen QwQ-32B (official open-source basic version) show that the model volume is reduced by 85% to 4.8GB, memory usage is decreased by 80%, inference speed is increased by 10 times, and the task-specific accuracy loss is only 3% (the general reasoning capability decreases by about 5%).

B. DATA-LEVEL EFFICIENCY IMPROVEMENT PATHS

1) Efficient Preprocessing Scheme for Complex Data

Design a parallel preprocessing framework to process different types of complex data in parallel using multithreading:

(1) Separate processing of structured and unstructured data; perform standardization and missing value filling for structured data, and format conversion and feature extraction for unstructured data;

(2) Introduce an incremental preprocessing mechanism to only process changed parts of new data, avoiding repeated calculations.

In medical multi-modal data preprocessing, the efficiency of the parallel framework is 6 times higher than that of serial processing.

2) Data Sampling and Filtering Optimization

Propose a hierarchical sampling strategy based on feature importance:

(1) First calculate the importance score of data features through a lightweight model, and divide the data into core layer (high importance), general layer (medium importance), and redundant layer (low importance);

(2) Fully retain core layer data, sample general layer data proportionally, and eliminate redundant layer data. On the premise of ensuring feature integrity, the data volume is reduced by more than 50%.

In sparse data processing of recommendation systems, this strategy improves training efficiency by 40% with an accuracy loss of less than 2%.

3) Data Caching and Reuse Strategy

Construct a three-level data caching system:

L1 cache stores recently processed high-frequency data (such as real-time industrial equipment monitoring data), L2 cache stores common feature templates, and L3 cache stores historical processing results; Introduce a cache elimination

TABLE 1. Comparison of Adaptation Scenarios and Effects

Improvement Technology	Core Principle	Optimization Effect	Adaptable Complex Data Scenarios	Limitations
Quantization	Convert FP32 parameters to INT8/INT4 low precision	Model volume reduced by 75%, inference speed increased by 3-5 times	High-dimensional heterogeneous data, edge-side time-series data processing, textile engineering production data processing	Large accuracy loss in high-precision scenarios (such as medical diagnosis) when used alone
Pruning	Eliminate redundant parameters/channels with weights less than the threshold	Model volume reduced by 40-60%, computational load decreased by 30-50%	Sparse unstructured data, industrial monitoring data, textile fabric parameter data processing	Excessive pruning leads to decreased model robustness when used alone
Knowledge Distillation	Transfer knowledge from teacher models to small student models	Model volume reduced by 90%, accuracy loss less than 5%	General complex data processing, edge-side deployment scenarios, edge-side textile defect detection	High computational cost in the distillation process
Dynamic Memory Sparsification	Dynamically eliminate redundant tokens in KV cache and transfer useful information	Memory usage reduced by 75%, inference accuracy improved by 5-12%	Long-sequence time-series data, multi-turn inference scenarios, textile production time-series data processing	Complex token selection logic design

mechanism to dynamically remove cached data that has not been used for a long time, improving cache utilization.

At the same time, support cached data reuse; directly call cached results when processing the same type of complex data, improving efficiency by 70% in repeated scenarios (such as daily meteorological data processing).

C. ENGINEERING AND HARDWARE-LEVEL EFFICIENCY IMPROVEMENT PATHS

1) Distributed Training and Inference Architecture Design

Adopt a "multi-node + multi-GPU" distributed architecture for the training and cloud-side inference of the 32B main model, and match the lightweight sub-model (1.2B) for edge-side deployment to realize the hierarchical application of the model:

(1) In the training phase, split complex datasets into multiple nodes through data parallelism, each node is responsible for training part of the data, and synchronize parameters through a parameter server;

(2) In the inference phase, adopt a combination of model parallelism and data parallelism, split the 32B large model into multiple GPUs, and process multiple batches of complex data simultaneously.

(3) In the edge-side inference phase, deploy the 1.2B lightweight sub-model on low-power edge hardware to meet real-time processing needs.

In massive astronomical observation data processing and large-scale textile manufacturing production data processing, the efficiency of the distributed architecture is 20 times higher than that of single-machine processing.

2) Hardware Adaptation Optimization

Customize and optimize model calculation logic according to the computational characteristics of different hardware (GPU, TPU, FPGA) , and propose differentiated hardware adaptation strategies for cloud-side and edge-side deployment:

(1) Cloud-side GPU/TPU adaptation adopts mixed-precision computing, based on FP16 computing, and uses FP32 computing for key layers to ensure accuracy;

(2) Edge-side FPGA/Jetson AGX Orin adaptation adopts dedicated lightweight computing unit design and low-power optimization, strengthening the computational efficiency of core operations such as convolution and attention while controlling power consumption within 15W;

(3) FPGA adaptation adopts dedicated computing unit design to strengthen the computational efficiency of core operations such as convolution and attention.

When the optimized 32B model is deployed on cloud-side FPGA, power consumption is reduced by 60% and inference speed is increased by 8 times; when the 1.2B lightweight sub-model is deployed on edge-side FPGA/Jetson AGX Orin, power consumption is controlled within 15W and inference speed meets real-time requirements.

3) Memory and Storage Optimization

(1) Introduce a memory fragmentation arrangement mechanism to real-time optimize memory allocation and reduce memory waste;

(2) Adopt a storage hierarchy strategy: store frequently accessed complex data in high-speed SSDs and infrequently accessed data in mechanical hard disks;

(3) Compress stored data through data compression technologies (such as ZIP, LZ4) to reduce storage costs. Combined with dynamic memory sparsification technology, it can effectively solve the problem of KV cache expansion. In long-sequence time-series data processing, memory usage is reduced by 75%.

V. INTEGRATED SCHEME OF MULTI-PATH COLLABORATIVE IMPROVEMENT

Construct a "model-data-engineering" three-dimensional collaborative optimization system, clarifying the collaboration mechanism and priority of each dimension:

(1) Model compression as the core to reduce basic computational complexity;

(2) Data preprocessing as the support to reduce computational overhead of invalid data;

(3) Engineering optimization as the guarantee to improve the adaptability efficiency of hardware and architecture.

Design differentiated collaborative strategies for different complex data scenarios:

(1) For high-dimensional heterogeneous data, prioritize "architecture optimization + quantization + parallel preprocessing";

(2) For dynamic time-series data, prioritize "dynamic inference + cache reuse + distributed inference";

(3) For edge-side sparse data, prioritize "knowledge distillation + pruning + FPGA adaptation".

Achieve the maximum efficiency improvement effect through collaborative optimization.

VI. EMPIRICAL ANALYSIS AND CASE VERIFICATION

A. EXPERIMENTAL DESIGN

1) Experimental Data Selection

Three types of typical complex datasets are selected to cover different scenario requirements:

(1) Dataset A (medical multi-modal data): contains 100,000 cases, including text, CT images, and physiological signals, with a dimension of 5000;

(2) Dataset B (industrial time-series data): equipment monitoring data from a factory, with a total of 1 million time-series records and a sampling frequency of 10Hz;

(3) Dataset C (astronomical sparse data): solar observation data, including 100,000 satellite images with an effective information ratio of 5%.

The basic information of the datasets is shown in Table 2.

2) Experimental Models and Comparative Schemes

The experimental model adopts an improved Transformer-based model (integrating the algorithm optimization strategy proposed in this paper), denoted as Opt-Model (32B parameters, with a 1.2B lightweight sub-model for edge-side deployment).

Comparative schemes include: baseline model (standard Transformer, 32B), single optimization model (only adopting quantization or pruning, 32B), and existing mainstream models (Tongyi Qianwen QwQ-32B (official open-source basic version), GPT-3 (text-davinci-003 version)). All models are deployed on the same hardware environment (Intel Xeon Gold 6330, NVIDIA A100 (80GB), 256GB memory, 2TB SSD) with the same software configuration (Python 3.9, PyTorch 2.0, CUDA 11.7), ensuring the fairness of comparison and the standardization of benchmarking for efficiency indicators such as inference latency.

3) Evaluation Index Setting

Evaluation indexes are set from three dimensions: accuracy, efficiency, and robustness:

(1) Accuracy indexes include Accuracy (Acc), F1-score (F1), and Recall (Rec);

(2) Efficiency indexes include Training Time (Tt), Inference Latency (Ti), Memory Usage (M), and Model Volume (S);

(3) The robustness index adopts the accuracy loss rate (Lr) after adding 10% noise.

The specific definitions of indexes are shown in Table 3.

4) Experimental Environment and Parameter Settings

Hardware environment: CPU is Intel Xeon Gold 6330, GPU is NVIDIA A100 (80GB), memory is 256GB, storage is 2TB SSD;

Software environment: Python 3.9, PyTorch 2.0, CUDA 11.7.

Model parameters: Both Opt-Model and the baseline model have a parameter scale of 32B, the initial learning rate is 1e-4, the batch size is 32, and the number of training epochs is 100.

B. EMPIRICAL VERIFICATION OF ALGORITHM OPTIMIZATION STRATEGIES

The accuracy and robustness test results of different models on the four datasets are shown in Table 4. It can be seen from the results that Opt-Model performs the best on all datasets in task-specific accuracy: On Dataset A, the accuracy reaches 92.3%, which is 12.1 percentage points higher than the baseline model, and the F1-score is increased by 11.8 percentage points;

On Dataset B, the recall rate reaches 94.5%, which is 8.7 percentage points higher than GPT-3;

On Dataset C, the accuracy loss rate after adding noise is only 4.2%, which is 10.3 percentage points lower than the single optimization model.

On Dataset D, the fabric performance parameter prediction accuracy reaches 90.8%, which is 11.5 percentage points higher than the baseline model.

This indicates that the algorithm optimization strategy proposed in this paper can effectively improve the model's feature capture ability and robustness for complex data, and adapt to the processing needs of different types of complex data.

C. EMPIRICAL VERIFICATION OF EFFICIENCY IMPROVEMENT PATHS

The test results of different efficiency schemes are shown in Table 5. The inference latency comparison (1500ms vs. 150ms) mentioned in the paper is based on the same hardware environment (NVIDIA A100 (80GB)) for the baseline model (standard Transformer, 32B) and the Opt-Model lightweight sub-model (1.2B) in industrial equipment fault prediction task, which is a standardized comparison of the same task on the same hardware.

It can be seen from the results that the multi-path collaborative optimization proposed in this paper has a significant effect:

The model volume is reduced by 85%, memory usage is decreased by 80%, inference speed is increased by 10 times, training time is shortened by 5 times, and the task-specific accuracy loss is only 3% (the general reasoning capability decreases by about 5%), which is far superior

TABLE 2. Detailed Information of Experimental Datasets

Dataset	Data Type	Data Volume	Core Features	Processing Task
A (Medical)	Multi-modal heterogeneous data	100,000 cases	High-dimensional, heterogeneous	Disease diagnosis and prognosis prediction
B (Industrial)	Dynamic time-series data	1 million records	Dynamic, high real-time performance	Equipment fault prediction
C (Astronomical)	Sparse image data	100,000 images	Sparse, unstructured	Solar flare recognition

TABLE 3. Definitions and Optimization Objectives of Experimental Evaluation Indexes

Index Type	Index Name	Core Definition	Optimization Objective
Accuracy Index	Accuracy (Acc)	(Number of correctly processed samples / Total number of samples) × 100%	Maximization
Accuracy Index	F1-score (F1)	2 × (Precision × Recall) / (Precision + Recall)	Maximization
Efficiency Index	Inference Latency (Ti)	Average time for single sample processing (ms)	Minimization
Efficiency Index	Memory Usage (M)	Maximum memory usage during model operation (GB)	Minimization
Robustness Index	Accuracy Loss Rate (Lr)	(Accuracy before noise - Accuracy after noise) / Accuracy before noise × 100%	Minimization

TABLE 4. Accuracy and Robustness Test Results of Different Models

Model	Dataset A (Medical)		Dataset B (Industrial)		Dataset C (Astronomical)		Dataset D (Textile Engineering)		Average Accuracy Loss Rate (Lr)
	Acc (%)	F1 (%)	Acc (%)	Rec (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	
Baseline Model (Transformer)	80.2	79.5	83.1	83.5	78.6	77.3	79.3	78.6	14.5%
Single Optimization Model (Quantization)	85.7	84.3	86.9	87.1	82.4	81.5	84.2	83.5	12.8%
GPT-3	88.5	87.9	86.3	87.4	85.9	84.7	87.8	86.9	8.7%
Opt-Model (Proposed in This Paper)	92.3	91.3	90.7	94.5	91.2	90.5	90.8	89.7	4.2%

TABLE 5. Test Results of Different Efficiency Improvement Schemes

Efficiency Improvement Scheme	Model Volume (S)	Memory Usage (M)	Inference Latency (Ti)	Training Time (Tt)	Accuracy Loss Rate (Lr)
No Optimization (Baseline)	32GB	64GB	2000ms	100h	0%
Single Quantization (INT8)	8GB	32GB	667ms	80h	5%
Quantization + Pruning	6.4GB	25.6GB	400ms	60h	4.5%
Collaborative Optimization (Proposed in This Paper)	4.8GB	12.8GB	200ms	20h	3%

to the effect of a single technology optimization. The extreme compression effect is achieved by the synergistic effect of multiple technologies, which avoids the significant performance degradation caused by a single compression method.

D. ANALYSIS OF PRACTICAL APPLICATION CASES

1) Case 1: Industrial Equipment Fault Prediction

A certain auto parts factory needs to real-time process dynamic time-series data such as equipment vibration and temperature to realize fault prediction.

The original scheme adopted a standard Transformer model, which had problems of long inference latency (1500ms) and low accuracy (Acc=82%).

After adopting the Opt-Model and collaborative efficiency improvement scheme proposed in this paper:

(1) The model inference latency is shortened to 150ms, meeting real-time requirements;

(2) The fault prediction accuracy is increased to 94.5%, and the false positive rate is reduced by 20%;

(3) The equipment failure rate is reduced by 15%, saving 3 million yuan in annual maintenance costs.

2) Case 2: Solar Flare Recognition

The National Astronomical Observatory needs to process solar observation sparse image data to realize accurate recognition of solar flares.

The original scheme adopted a traditional CNN model, with a recognition accuracy of only 83.5% and low processing efficiency.

After adopting the optimization scheme proposed in this paper:

(1) The model recognition accuracy is increased to 91.2%;

(2) The processing speed is increased by 10 times;

(3) It can support real-time analysis of observation data from the "Kuafu-1" satellite, providing efficient support for solar activity prediction.

3) Case 3: Medical Multi-modal Diagnosis

A Grade A tertiary hospital needs to process multi-modal data such as case texts and CT images to assist disease

diagnosis.

After adopting the scheme proposed in this paper:

(1) The model diagnosis accuracy is increased from 80.2% to 92.3%, and the diagnosis time is shortened from 5 minutes to 30 seconds;

(2) At the same time, the model's ability to recognize rare diseases is significantly enhanced, providing important reference for clinical diagnosis.

VII. CONCLUSION

A. CORE PROBLEMS AND SOLUTIONS

This paper clarifies three core problems faced by AI large models in processing complex textile engineering data and fabric performance variables: insufficient algorithm generalization ability across diverse fiber types and medical feature types, high training costs, excessive memory overhead, high power consumption, and low inference throughput in real-time weaving defect detection and medical diagnosis, and poor data adaptability to the high heterogeneity of textile manufacturing parameters and astronomical observation data. The root causes lie in insufficient matching degree between model architecture and complex data features, unreasonable resource allocation, and lack of targeted optimization schemes.

The solutions are:

(1) Optimize model architecture and training-inference logic at the algorithm level to adapt to the multi-feature needs of complex data, and design a 1.2B lightweight sub-model for edge-side deployment on the basis of the 32B main model;

(2) Construct a multi-dimensional collaborative optimization system at the efficiency level based on structured pruning, mixed-precision quantization and knowledge distillation to balance task-specific accuracy and efficiency, and distinguish the application scenarios of distributed architecture (cloud-side training/inference) and lightweight deployment (edge-side inference);

(3) Design differentiated schemes combined with specific scenarios (textile engineering, medical, industrial, astronomical) at the application level to improve feasibility, and analyze the performance limitations of the model in extreme complex data scenarios.

B. CORE POINTS OF ALGORITHM OPTIMIZATION AND EFFICIENCY IMPROVEMENT

At the algorithm optimization level, customized architectures such as adaptive attention and two-stage multi-modal fusion are proposed to enhance the model's feature capture ability for complex textile defect patterns and fabric performance parameters. Combined with adaptive learning rates and dynamic inference, these strategies significantly improve robustness when processing heterogeneous textile manufacturing data and medical multi-modal data, ensuring high precision in the automated identification of fiber irregularities, weaving inconsistencies, disease features and astronomical target features. At the same time, the paper clearly points out that the 3% accuracy loss is only for task-specific performance, and the general reasoning capability of the compressed model decreases by about 5%, which objectively reflects the trade-off between model compression and performance.

At the efficiency improvement level, a "model compression-data preprocessing-engineering optimization" collaborative system is constructed, integrating technologies such as structured pruning, mixed-precision quantization, dynamic memory sparsification, and distributed architecture. The system realizes hierarchical deployment of the model: the "multi-node + multi-GPU" distributed architecture is used for cloud-side training and inference of the 32B main model, and the 1.2B lightweight sub-model is deployed on edge hardware (FPGA/Jetson AGX Orin, power consumption $<$15W) to realize real-time processing of complex data on edge devices. On the premise that the task-specific accuracy loss does not exceed 3%, the 32B model volume can be reduced by 85% and the inference speed can be increased by 10 times.

C. EMPIRICAL VERIFICATION CONCLUSIONS

Experimental and case verifications show that the optimization scheme proposed in this paper performs excellently in complex data processing in multiple fields such as medical care, industry, and scientific research, which can effectively improve processing accuracy, efficiency, and robustness, and has broad application value. It should be noted that the Opt-Model is mainly optimized for typical complex data (dimensionality 5000-8000, noise ratio $\leq$10%, sufficient sample size) in mainstream engineering scenarios, and has certain limitations in processing extreme complex data: it is difficult to maintain high performance in ultra-high-dimensional sparse data (dimensionality $>$ 10000), strong-noise heterogeneous data (noise ratio $>$ 30%) and cross-domain extreme small-sample complex data (sample size $<$ 100), which is mainly due to the fixed threshold setting of the model's feature screening and noise filtering modules. The follow-up research will focus on the failure mechanism analysis of the model in extreme complex data processing scenarios, and optimize the adaptive threshold adjustment mechanism and transfer learning framework to further improve the model's generalization ability. This research has laid a solid theoretical and experimental foundation for the subsequent in-depth research on extreme complex data processing.

AUTHOR CONTRIBUTIONS

Jing Liu designed, collected and analyzed the data, and drafted the manuscript. Jing Liu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jing Liu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that

questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by,

Hunan Province Education Science Planning Project: "Research and Practice on the Digital Transformation of Vocational Education Curriculum in Hunan under the '5G + Smart Education' Context", Project No.: ZJ234381.

Natural Science Foundation of Hunan Province Project: "Research on the Mechanism and Path of Quality Evaluation of Smart Education in Higher Vocational Colleges Based on 5G Technology", Project No.: 2025JJ80327.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Rahman, "Statistical analysis on complexly structured data," Bulletin of the Australian Mathematical Society, pp. 111(1), 2025, doi: 10.1017/S0004972724001060.

[2] D. Wang, K. Yin, and H. Wang, "Intelligent classification of construction quality problems based on unbalanced short text data mining," Ain Shams Engineering Journal, vol. 15, no. 10, Art. no. 102983, 2024, doi: 10.1016/j.asej.2024.102983.

[3] C. Li, Y. Zhang, Q. Kuang, Liu, and Y, "Deep quantile forests for high-dimensional data," Statistical Theory and Related Fields, pp. 1-19, 2026, doi: 10.1080/24754269.2026.2616873.

[4] X. Gao, X. Chen, Y. Fang, and H. Cheng, "A Data Governance Model Based on Multi-Agents System," Proceedings of the 2025 IEEE 7th International Conference on Communications, Information System and Computer Engineering (CISCE); Piscataway, NJ, USA: IEEE; 2025, pp. 339-345, doi: 10.1109/CISCE65916.2025.11065402.

[5] J. Yu, J. Nie, W. Li, and W. Chen, "Dual-track Innovation Driven by AI Large Models: Mechanism and Practical Exploration," Technology Economics, vol. 43, no. 12, pp. 10-22, 2024, doi: 10.12404/j.issn.1002-980X.J24101816.

[6] S. Almanasra, "Applications of integrating artificial intelligence and big data: A comprehensive analysis," Journal of Intelligent Systems, pp. 33(1), 2024, doi: 10.1515/jisys-2024-0237.

[7] Q. Wang, C. Yao, D. Li, J. Zhu, and W. Cao, "Research on the Fusion and Analysis Method for the Multi-Source Heterogeneous Elevator Risk Data Based on Deep Learning," Proceedings of the 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS); Piscataway, NJ, USA: IEEE; 2024, pp. 630-634, doi: 10.1109/icedcs64328.2024.00118.

[8] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time Series Prediction using Deep Learning Methods in Healthcare," ACM Transactions on Management Information Systems, vol. 14, no. 1, pp. 1-29, 2023, doi: 10.1145/3531326.

[9] E. Anderson and A. Cheng, "Portfolio Choices with Many Big Models," Management Science, vol. 68, no. 1, pp. 690-715, 2022, doi: 10.1287/mnsc.2020.3876.

[10] H. Wu and Z. Wen, "Research on Adaptive Attention Dense Network Structure in Camera Source Recognition Method," Proceedings of the 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom); Piscataway, NJ, USA: IEEE; 2024, pp. 2238-2243, doi: 10.1109/TrustCom63139.2024.00308.

[11] Z. H. Tang, H. Lan, Z. Liu, Q. Wang, L. Zhao, and K. Li, "HiTrain: Heterogeneous Memory Offloading and I/O Optimization for Large Model Training," Journal of Computer Research and Development, vol. 63, pp. 1-13, 2025, doi: 10.7544/issn1000-1239.202550478.

[12] J. Chen, W. Yao, and J. Li, "Parallel optimization techniques for ultra-large scale refined numerical simulations on Chinese supercomputers," Cluster Computing, vol. 28, no. 8, pp. 1-24, 2025, doi: 10.1007/s10586-024-05057-3.

[13] L. Liu, G. Liu, B. Qi, X. Deng, D. Xue, and S. Qian, "Survey of Large Model Efficient Inference Technologies in Practical Application Scenarios," Computer Science, vol. 53, no. 1, pp. 12-28, 2026. Available: https://link.cnki.net/urlid/50.1075.TP.20250703.1601.002.

[14] C. Yao, G. Wang, P. Xiang, and G. Chen, "Opportunities, Challenges and Practices of Medical Large Models and Multi-modal Data Fusion in Smart Healthcare," Chinese Journal of Health Informatics and Management, vol. 22, no. 02, pp. 171-178, 2025, doi: 10.3969/j.issn.1672-5166.2025.02.02.

[15] Y. Pei, J. Cartlidge, A. Mandal, and D. Gold, "Cross-Modal Temporal Fusion for Financial Market Forecasting," arXiv preprint arXiv:2504.13522, 2025, doi: 10.3233/FAIA251474.

[16] S. Li, W. Yang, P. Zhang, X. Xiao, D. Cao, Y. Qin, et al., "Climatellm: Efficient Weather Forecasting via Frequency-aware Large Language Models," arXiv preprint arXiv:2502.11059, 2025, doi: 10.48550/arXiv.2502.11059.

[17] J. Li, L. Shi, M. Ding, Y. Lei, D. Zhao, and L. Chen, "Social Media Text Stance Detection Based on Large Language Models," Journal of Frontiers of Computer Science and Technology, vol. 19, no. 5, pp. 1302-1312, 2025, doi: 10.3778/j.issn.1673-9418.2408074.

[18] H. Vakharia and X. Du, "Efficient Multi-Resolution Fusion for Remote Sensing Data with Label Uncertainty," 2024, doi: 10.1109/IGARSS52108.2023.10282851.