Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 2, 2025

Analysis on the Method of Improving the Performance of Natural Language Processing Model Driven by Artificial Intelligence

Zenghua Wang *

Yancheng Advanced Vocational School of Economics and Trade, Yancheng, Jiangsu, China
*** Corresponding author Email:** 494119434@qq.com

Abstract: Natural language processing (NLP), as the core field of artificial intelligence (AI), has made remarkable progress with the help of Transformer to pre-train large models, but it is still restricted by bottlenecks such as strong data dependence, high computational cost, and insufficient generalization and robustness. Aiming at the problem that the existing research is mostly limited to one-dimensional optimization, this study proposes a hierarchical optimization framework (HOF) covering data, model, training and deployment, and improves the performance of NLP model through full link collaborative design. In the data layer, an antagonistic knowledge injection (AKI) method is proposed, which uses external knowledge maps to guide text generation and verification, thus alleviating the problem of data shortage in low-resource scenes. In the model layer, Dynamic Sparse Gating Transformer (DSGT) is designed to balance accuracy and reasoning efficiency through dynamic sparse gating mechanism. In the training layer, the Meta-Adaptive Multitask Learning (MAMTL) method driven by meta-learning is adopted to enhance the cross-language and cross-domain generalization ability. In the deployment layer, an Optimal Transport Alignment (OTA) method based on optimal transport is proposed to achieve efficient multimodal semantic fusion. The experimental results show that the BLEU of HOF is 47.5 on FLORES-200 data set, the macro F1 is 84.7 on X-Cross data set, the reasoning delay of edge devices is reduced to 62ms, and the gender and race prediction bias is significantly reduced. This research fills the gap of NLP full link collaborative design, and provides a new paradigm for promoting efficient and robust NLP technology and building a trusted AI system.

Keywords: Natural Language Processing; Artificial Intelligence; Performance Improvement; Hierarchical Optimization Framework.

1. Introduction

As the core field of artificial intelligence (AI), natural language processing (NLP) has made remarkable progress in tasks such as machine translation and text generation, relying on pre-trained large models based on Transformer, such as BERT and GPT, which has promoted the landing of many kinds of intelligent applications [1-2]. However, its development is still limited by key bottlenecks such as strong data dependence, high computational cost, insufficient generalization and robustness, poor interpretability and weak multimodal fusion ability, especially in low-resource scenes and cross-modal interaction.

This study focuses on the systematic challenge of improving the performance of NLP model. Aiming at the problem that the existing research is mostly limited to single-dimensional optimization, it is devoted to the collaborative design of the whole link from data, model, training to deployment. The core of this study is to explore how to build a low-resource-friendly data enhancement and knowledge injection method to reduce the dependence on labeled data, design an efficient and lightweight model architecture to balance accuracy and reasoning efficiency, develop adaptive training strategies to enhance the generalization ability of the model in cross-language and cross-domain scenarios, and realize the effective fusion of multimodal semantic information to break through purity. On the theoretical level, a hierarchical optimization framework (HOF) covering data, model, training and deployment is proposed, which fills the research gap of full link collaborative design in NLP field and provides a new paradigm for performance improvement. On the practical level, through the experimental verification in low-resource language translation, real-time reasoning of edge devices and multimodal tasks, the application of efficient, lightweight and robust NLP technology in real scenes is promoted; At the social level, it is committed to improving the interpretability and fairness of the model, reducing the risk of AI bias and helping to build a credible and responsible AI system.

2. Core Methodology

2.1. Performance Improvement Technical Framework

HOF improves the performance of NLP model through systematic collaborative design. As shown in Figure 1, the framework covers data layer, model layer, training layer and deployment layer: in the data layer, low-resource data enhancement and knowledge injection technology are adopted to reduce the dependence on large-scale annotation data [3]; Design a dynamic sparse compression architecture in the model layer, taking into account the model accuracy and computational efficiency; Adaptive multi-task joint learning strategy is introduced in the training layer to enhance the generalization ability of the model in cross-domain and cross-task scenarios [4-5]; Multi-modal semantic alignment reasoning is realized in the deployment layer, which supports efficient and accurate interaction in complex real scenes.

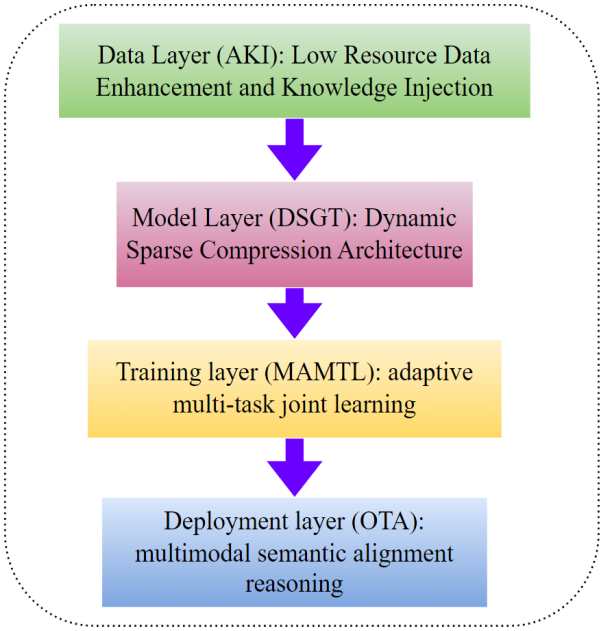

Figure 1. HOF frame structure

2.2. Detailed Description of Innovative Methods

Propose Adversarial Knowledge Injection (AKI) method at the data layer to alleviate the problem of annotated data scarcity in low resource scenarios [6]. This method introduces an external structured knowledge graph as a semantic prior to guide the generation and validation of high-quality text. The core mechanism is to build a dual channel adversarial generator: the generator uses a conditional variational autoencoder (CVAE) to generate semantically reasonable text based on knowledge constraints, while the discriminator jointly judges the authenticity and knowledge consistency of the generated text, ensuring that the output is both close to the real language distribution and conforms to the known knowledge structure.

In order to realize the effective fusion of knowledge, AKI designs a knowledge constraint loss function:

$$L_{KC} = E_{x\sim p_{data}}\left[\log D(x)\right] + \lambda KL\left(q(z|x)\|p(z|K)\right) \quad (1)$$

Where $q(z|x)$ represents the potential semantic distribution of text x, $p(z|K)$ is the prior distribution provided by knowledge graph K, and λ controls the knowledge constraint strength (set to 0.8 in the experiment).

In the model layer, Dynamic Sparse Gating Transformer (DSGT) is proposed to solve the contradiction between accuracy and reasoning efficiency of large models. The core of DSGT is to introduce the dynamic sparse gating mechanism, and calculate the gating vector through the learnable gating weight W_g, combining the previous state h_{l-1} of the current layer and the calculated cost prediction vector C_l:

$$g_l = \sigma\left(W_g\left[h_{i-1}; C_l\right]\right) \quad \in \{0,1\}^d \quad (2)$$

Where σ is a Sigmoid function, and then binarization processing (threshold 0.5) is performed to generate a binary gating signal to determine whether each layer in the network is activated.

The gating vector g_l realizes the selective activation mechanism: only when the corresponding position of g_l is 1, the complete calculation of this layer is performed, otherwise, the layer is skipped to save calculation resources [7]. This dynamic sparse strategy enables the model to adaptively allocate the computation according to the input complexity, quickly reason on simple samples, and reserve the deep processing ability on complex samples, thus greatly reducing the overall calculation cost and delay without significantly sacrificing the accuracy of the model, especially suitable for edge devices and real-time application scenarios with limited resources.

The training layer is based on the Meta-Adaptive Multitask Learning (MAMTL) method, which can improve the generalization ability of the model in cross-language and cross-domain scenarios [8]. Through the task routing mechanism, multiple tasks are divided into K clusters by domain, and the organizational management of task structure is realized. On this basis, an adaptive weight distribution mechanism is introduced to dynamically adjust the loss weight of each task in joint training according to the semantic similarity between the current target task and the target domain, so that the model pays more attention to the tasks with high correlation, thus enhancing the effect of knowledge transfer.

Specifically, the total loss function is defined as:

$$L_{total} = \sum_{i=1}^{K} \alpha_i L_{task_i}, \alpha_i = \frac{e^{\beta sim\left(D_i, D_{target}\right)}}{\sum e^{\beta sim}} \quad (3)$$

Among them, the weight α_i is calculated by the Softmax mechanism, which depends on the semantic similarity $sim(\cdot)$ between the source domain D_i and the target domain D_{target} (calculated by SBERT) and the learnable adjustment factor β. Furthermore, the meta-learning strategy is adopted, and through the performance feedback on the verification set, the meta-gradient update mechanism is used to optimize β, so as to realize the dynamic and global regulation of task weight, which makes the model have stronger adaptive ability and cross-domain generalization performance.

In this study, an Optimal Transport Alignment (OTA) method based on optimal transmission is proposed at the deployment layer, aiming at the efficient integration of multimodal semantics such as text, image and voice [9]. OTA models the modal alignment problem as the optimal transmission problem between probability distributions, and realizes cross-modal semantic matching by minimizing the following loss functions:

$$L_{OTA} = \inf_{\gamma\in\prod(P_T, P_V)} E_{(t,v)\sim\gamma}\left[c\left(f_t(t), f_v(v)\right)\right] + \varepsilon OT_\varepsilon(\gamma) \quad (4)$$

Among them, f_t, f_v is the encoder of text and visual mode respectively, and $c(\cdot)$ uses cosine similarity as the cost function to measure the difference of cross-modal representation; $\prod(P_T, P_V)$ represents the joint distribution set between text and visual feature distribution.

The second term in the loss function is Sinkhorn regularization term $OT_\varepsilon(\gamma)$ (regularization coefficient $\varepsilon = 0.1$), which is used to improve the computational stability and efficiency of the optimal transmission problem. In the actual deployment, in order to further reduce the reasoning delay, a cross-modal attention caching mechanism is introduced: the cross-modal attention weight of previous time steps is cached in the sequence generation process, so that repeated calculation is reduced, and the reasoning speed of multi-modal tasks is significantly improved, thus achieving accurate and efficient multi-modal semantic alignment and real-time response.

3. Experiment and Analysis

FLORES-200 data set is used to evaluate the performance of low-resource translation, X-Cross data set covering eight fields, such as medical care, finance and law, is used to verify the cross-domain generalization ability, and COCO-Captions (image-text) and LibriSpeech (voice) data set are combined to test the multi-modal task performance. The model was trained on NVIDIA V100 cloud server, and the reasoning efficiency was tested on Jetson Xavier NX edge device.

The experimental results show that the proposed HOF is significantly superior to the baseline model in all indicators. As shown in Table 1, on FLORES-200, the BLEU reaches 47.5, on X-Cross, the macro F1 reaches 84.7, the reasoning delay of edge devices is reduced to 62ms, and the energy consumption is reduced to 2.9J, which verifies the effectiveness and superiority of collaborative optimization among data layer (AKI), model layer (DSGT) and training layer (MAMTL).

Table 1. Comparison of performance improvement of full link method

Method module	**BLEU (FLORES)**	**Macro F1(X-Cross)**	**Delay (ms)**	**Energy consumption (J)**
BERT-Large	42.1	78.3	210	9.8
+data layer (AKI)	**46.3** (+4.2)	-	-	-
+model layer (DSGT)	45.8	79.1	**68** (-67.6%)	**3.2** (-67%)
+training layer (MAMTL)	-	**82.9** (+4.6)	-	-
Full link (HOF)	**47.5**	**84.7**	**62**	**2.9**

The social impact of the proposed method is verified by the fairness test on BiasBench. Compared with BERT-Large, the HOF in this study significantly reduces the prediction bias of the model in terms of gender and race (gender bias is reduced from 0.32 to 0.19, and racial bias is reduced from 0.41 to 0.23), which shows that it is helpful to build a fairer and more credible AI system while improving its performance (as shown in Table 2).

Table 2. Fairness test

Model	Gender prejudice	Racial bias
BERT-Large	**0.32**	**0.41**
HOF	**0.19**	**0.23**

4. Conclusion

In this paper, the method of improving the performance of NLP model driven by AI is systematically discussed, and a HOF covering the whole link collaborative design of data, model, training and deployment is proposed. Through data enhancement and knowledge injection technology in low-resource scenarios, dynamic sparse compression model architecture, adaptive multi-task joint learning strategy, and multimodal semantic alignment method based on optimal transmission, HOF successfully improved the performance of NLP model. The experimental results show that HOF achieves a BLEU score of 47.5 in the low-resource translation task and a macro F1 of 84.7 in the cross-domain generalization task, and the reasoning delay on the edge device is reduced to 62 milliseconds, and the energy consumption is reduced to 2.9 Joules. In addition, HOF shows the ability to significantly reduce gender and racial bias in fairness test, which is helpful to build a more credible and responsible AI system. HOF provides a new paradigm for improving the performance of NLP model, which is expected to promote the wide application of efficient, lightweight and robust NLP technology in real scenes.

References

[1] Sangmin Kim,Byeongcheon Lee,Muazzam Maqsood,Jihoon Moon & Seungmin Rho.(2025). Deep Learning-Based Natural Language Processing Model and Optical Character Recognition for Detection of Online Grooming on Social Networking Services.Computer Modeling in Engineering & Sciences, 143(2),2079-2108.

[2] Wang Dai, Kebiao Mao,Zhonghua Guo,Zhihao Qin,Jiancheng Shi,Sayed M. Bateni & Liurui Xiao. (2025). Joint optimization of AI large and small models for surface temperature and emissivity retrieval using knowledge distillation. Artificial Intelligence in Agriculture,15(3),407-425.

[3] Yang Juan,Bai Yu,Gong Jie & Han Menghui.(2025).The Financial Institution Text Data Mining and Value Analysis Model Based on Big Data and Natural Language Processing. Journal of Organizational and End User Computing (JOEUC), 37(1), 1-40.

[4] Stefanie Seo,Andy S Ding,Syed Ameen Ahmad,Kevin Z Xin,Max L Jiam, Vincent Xin... & Nicole T Jiam.(2025).A Novel Natural Language Processing Model for Triaging Head and Neck Patient Appointments..Otolaryngology--head and neck surgery: official journal of American Academy of Otolaryngology-Head and Neck Surgery,173(1),126-133.

[5] Yu Jing.(2024).Sentiment Analysis of Text Using Deep Learning-based Natural Language Processing Models. Education Reform and Development,6(12),180-185.

[6] Trgovac Ana Mulović,Mandić Antonija & Marković Biljana.(2024). Tools of Artificial Intelligence Technology as a Framework for Transformation Digital Marketing Communication. Tehnički glasnik,18(4),660-665.

[7] Laxmi Choudhary & Jitendra Singh Choudhary.(2024). Deep Learning Meets Machine Learning: A Synergistic Approach towards Artificial Intelligence.Journal of Scientific Research and Reports,30(11),865-875.

[8] Xianqiu Zheng,Zhidong Zhang,Liqin Wang,Jinhua Wu & Zuofeng Dong.(2024).Investigation on text generative model based on deep learning in natural language processing.Journal of Computational Methods in Sciences and Engineering, 24 (6), 4089-4100.

[9] Marcel C. Langenbach, Borek Foldyna,Ibrahim Hadzic,Isabel L. Langenbach, Vineet K. Raghu, Michael T. Lu... & Julius C. Heemelaar. (2024). Automated anonymization of radiology reports: comparison of publicly available natural language processing and large language models.European Radiology, 35(5),1-8.