Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 6 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i06.54703

PAPER

# Enhancing Translation Teaching Efficiency and Learning Experience through Mobile Technology Applications

**Hongjing Chang, Jing Zhao(✉)**

School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

lelezhao0412@gmail.com

## ABSTRACT

With the advancement of globalization, the significance of English translation in cross-cultural communication has been increasingly emphasized. Traditional methods of translation teaching are becoming insufficient to meet the demands of modern learners, particularly in the context of the rapid development of mobile internet and intelligent devices. Consequently, there is an urgent need to introduce new technological approaches to improve teaching efficiency and the learner experience in translation education. The widespread use of mobile technology presents new opportunities for translation teaching, especially in the areas of real-time feedback, personalized learning, and translation aids. However, existing translation tools and models are often too complex and unsuitable for lightweight devices, and they exhibit certain limitations in enhancing learning efficiency and translation quality. Thus, the design of lightweight translation models suited for mobile devices has become a critical issue in translation education. This study aims to explore the application of optimized translation model designs on mobile devices in translation teaching. The study focuses on two core components: first, the design of a lightweight translation model based on attention mechanisms to improve computational efficiency and translation quality; second, the investigation of encoder and decoder designs suitable for mobile devices to enhance model performance. Through these optimizations, the study aims to provide a more efficient and personalized tool for translation teaching, thereby improving teaching outcomes and the learning experience, while offering theoretical support and practical guidance for the innovation of translation education models.

## KEYWORDS

mobile technology, translation teaching, lightweight translation model, attention mechanism, encoder, decoder, learning experience

Chang, H.J., Zhao, J. (2025). Enhancing Translation Teaching Efficiency and Learning Experience through Mobile Technology Applications. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(6), pp. 75–89. https://doi.org/10.3991/ijim.v19i06.54703

Article submitted 2024-11-14. Revision uploaded 2025-01-23. Final acceptance 2025-02-04.

## 1 INTRODUCTION

With the acceleration of globalization, English, as one of the primary languages of international communication, has assumed a crucial role across various sectors [1, 2]. Particularly in the field of translation, English translation is not only an essential component of language learning but also a bridge for cross-cultural communication. The rapid development of information technology, especially the widespread use of mobile internet and intelligent devices, has led to continuous innovations in translation tools and methods [3–6]. The extensive application of mobile technology presents unprecedented opportunities for translation teaching, particularly in terms of enabling learners to access translation aids and real-time feedback anytime and anywhere. Efficiently leveraging mobile technology to enhance the effectiveness of translation teaching and improve the learning experience has become a focal issue in current educational research.

The significance of this study lies in exploring how mobile technology can be utilized to improve both efficiency and the learning experience in translation teaching, particularly in the context of English translation education [7, 8]. As the demands for translation learning become more diverse and teaching methods continue to evolve, traditional translation teaching methods and tools are gradually becoming inadequate for meeting the needs of modern learners. Mobile technology can not only increase the interactivity and flexibility of learning but also assist learners in mastering translation skills more efficiently through intelligent and personalized translation support [9–13]. Therefore, researching the application of mobile technology in translation teaching is expected to provide new theoretical support for translation education and offer guidance for the innovation of teaching design and methods in practice.

Although previous studies have explored the integration of translation teaching and mobile technology, many studies still exhibit certain limitations. First, existing mobile translation tools and models are often too large or complex for practical use, making them unsuitable for lightweight devices and convenient mobile applications, thereby restricting the user experience of learners [14–18]. Second, most current model designs prioritize generalizability and accuracy, neglecting the personalized needs of learners in specific teaching contexts, which results in limited improvements in translation teaching efficiency. Furthermore, existing attention mechanisms and encoder-decoder designs have yet to be optimized for the hardware constraints of mobile devices, leading to suboptimal system performance and response speeds, which negatively impact learning outcomes. As such, designing efficient translation models suitable for mobile devices has become a major challenge in current translation education research.

This study aims to propose a lightweight translation model based on mobile devices, with a focus on exploring the attention mechanism and optimizing the design of encoders and decoders. Specifically, the research first investigates the design of an attention mechanism for the lightweight translation model suitable for mobile devices. By reducing computational load and enhancing computational efficiency, the goal is to achieve superior translation quality and user experience. Additionally, the study examines how to design efficient encoder and decoder structures for mobile devices, taking into account the computational capabilities and network environments of mobile platforms to ensure the efficient completion of translation tasks. Through these design optimizations, this study seeks to provide a more efficient, personalized, and adaptive teaching tool for translation education, thereby promoting innovation and enhancement in translation education models.

## 2 ATTENTION MECHANISM DESIGN FOR THE LIGHTWEIGHT TRANSLATION MODEL ON MOBILE DEVICES

With the continuous improvement of mobile device hardware performance and their widespread use in daily life, the field of education, particularly translation teaching, has gradually shifted towards mobile and personalized approaches. Compared to traditional translation tools based on personal computers, mobile devices are characterized by convenience and availability, enabling learners to engage in translation practice and receive instant feedback more flexibly. However, existing translation models and tools often require complex computational resources and large-scale data processing capabilities, which are difficult to fully exploit within the limited hardware constraints of mobile devices. To meet the application demands of mobile platforms, the design of a lightweight translation model becomes critical. A lightweight translation model reduces computational resource usage and enhances processing efficiency, which not only improves the response speed of the translation system on mobile devices but also effectively lowers energy consumption and extends device lifespan. This results in a smoother and more efficient learning experience for learners.

A lightweight convolution module was first designed for the mobile device translation model in this study. Unlike traditional convolutional neural networks, deep convolution decomposes convolution operations to process input features, enabling effective local feature extraction with limited computational resources. Specifically, deep convolution employs only one convolution kernel per layer to perform convolution on input data, thereby reducing computational load and improving processing speed. Under the hardware limitations of mobile devices, deep convolution significantly reduces model complexity, avoids excessive computational redundancy, and maintains a high feature extraction capacity. By independently applying convolution kernels across different channels, deep convolution is able to capture finer features of each channel, thereby enhancing the model's ability to represent textual input. Let the input dimension be denoted by $f$, and the size of the convolution kernel by $j$. The calculation formula for deep convolution is as follows:

$$DZ(A,Q_Z,:,u,z)=\sum_{k=1}^{j}Q_{z,k}*A_{\left(u+k-\left\lfloor\frac{j+1}{2}\right\rfloor\right),z} \tag{1}$$

Lightweight convolution, on the other hand, determines the importance of input elements by fixing the context window and using a set of shared weights that do not vary with time steps, thereby reducing the model complexity. Compared to deep convolution, the lightweight convolution used in this study applies the same convolution kernel across multiple channels when processing each input feature, avoiding redundant computations that would arise from performing independent convolutions on each channel. This approach significantly reduces the number of parameters in the model. This design principle not only alleviates the computational burden on mobile devices but also enhances the computation speed and execution efficiency of the model under limited hardware resources. Let $f/G$ represent the number of channels in each head with shared weights, and the calculation formula for the lightweight convolution output channel $z$ and the $u$-th element is as follows:

$$LC\left(A,Q_{\left\lceil\frac{zG}{f}\right\rceil,:},u,z\right)=DC\left(A,\mathrm{softmax}\left(Q_{\left\lceil\frac{zG}{f}\right\rceil,:},u,z\right)\right) \tag{2}$$

In this study, the proposed lightweight attention mechanism module was optimized in two ways, as illustrated in Figure 1. First, a dual-branch structure was introduced to optimize the traditional attention mechanism to accommodate the computational limitations of mobile devices and enhance the performance of the translation model. Unlike the traditional attention mechanism, the lightweight attention mechanism employs a more advantageous combination method in feature extraction. That is, by integrating dynamic convolution with the attention mechanism, it enhances the model's feature capture ability at both the global and local levels. At the global level, the lightweight attention mechanism continues to use the traditional attention mechanism, effectively focusing on the global dependencies of the input sequence. At the local level, to address the limitations of the attention mechanism in capturing local features, the lightweight attention mechanism incorporates a dynamic convolution network, which enables more precise extraction of local features within the input sequence.

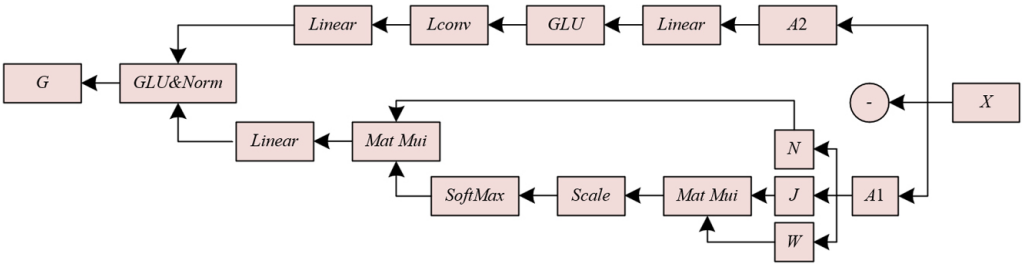


**Fig. 1.** Structure diagram of the lightweight attention mechanism module

In the design of the lightweight attention mechanism module, the input sentence was first processed through an encoder to obtain the sentence representation, which was then dimensionally split into $A_1$ and $A_2$, where $f1 = f2$ and $f = f1 + f2$. This splitting method allows the model to simultaneously utilize information from both dynamic convolution and the attention mechanism. Given that the Transformer architecture employs multi-head attention, the lightweight attention mechanism sets the number of heads $G$ for both the attention mechanism and dynamic convolution, ensuring that each head's dimension is $f_{mod} = f/2G$, thus facilitating efficient feature extraction and fusion. This design ensures that the model, while maintaining a low number of parameters, retains powerful feature extraction capabilities and efficient global feature modeling capabilities, thus optimizing the translation task on mobile devices. Assuming that the multiplication of $A_1$ with $Q_w$, $Q_j$, and $Q_n$ results in $W$, $J$, and $N$, and the learnable parameters are denoted by $Q_w$, $Q_j$, and $Q_n$, the following holds:

$$ATT(W, J, N) = \text{softmax}\left(\frac{WJ^S}{\sqrt{f_{MOD}}}\right)N \tag{3}$$

To further enhance the local feature extraction ability, dynamic convolution was introduced in the right branch of the lightweight attention mechanism. Similar to lightweight convolution, dynamic convolution utilizes time-step-dependent convolution kernels. By incorporating dynamic weights associated with the time steps, the convolution operation adjusts its parameters according to the input sequence's time steps, thus better adapting to varying input features. This dynamic convolution design allows local features to be effectively fused with global features through a gating mechanism while processing the sentence, and the output feature map can be

subsequently passed to the lightweight convolution module for further processing. In this dual-branch architecture, dynamic convolution not only enhances the accuracy of local feature extraction but also reduces the model's computational complexity, thereby lowering the computational resources required by mobile devices during runtime. Assuming that the learnable linear parameters are represented by $Q^{W}_{g,k,z}$, the dynamic convolution formula is as follows:

$$DC(A, u, z) = LC(A, d(A_u)_{g:}, u, z) \tag{4}$$

$$d(A_u) = \sum_{z=1}^{f} Q^{W}_{g,k,z} A_{u,z} \tag{5}$$

In response to the issues of computational complexity and excessive parameter volume in the feedforward neural networks of traditional Transformer models, a design for a lightweight attention mechanism module based on a gating mechanism was proposed in this study. In the standard Transformer, after the input sentence is processed by the attention mechanism, it is typically passed through a feedforward neural network for nonlinear fusion to enhance the model's expressive power and to prevent the potential collapse phenomenon that may occur within the attention mechanism. However, the structure of the feedforward neural network usually expands the intermediate dimension by a factor of four before scaling it back to the original dimension, a process that requires substantial computation and parameters. This is particularly pronounced when processing shorter sentences, where the computational complexity and parameter volume of the feedforward neural network become especially prominent. As a result, the high complexity and large parameter volume of the feedforward neural network are not suitable for tasks such as English-Chinese machine translation, which deal with low-resource languages. This is especially true in scenarios with limited resources on mobile devices, where it may adversely affect the model's operational efficiency.

To address this issue, a strategy replacing the traditional feedforward neural network with a gating mechanism was proposed in this study. In the lightweight attention mechanism module, a dual-branch structure was employed, and a gated recurrent unit (GRU) structure was used to replace the traditional feedforward neural network, achieving the same nonlinear fusion functionality. Specifically, the dimensions of the left and right branches in the dual-branch structure were designed as $f1 = f2 = f/2$, and these branches were effectively fused through the GRU structure. Compared to the traditional feedforward neural network, the GRU structure not only has fewer parameters but also performs more efficiently in the processing of sequential data. This approach allows the model to maintain comparable functionality to the feedforward neural network without increasing the computational and storage burdens, thus effectively reducing model complexity and minimizing the consumption of computational resources on mobile devices. Figure 2 shows the vector dimension diagram of the lightweight attention mechanism module. The learnable linear transformation is denoted by $Q_h$, the concatenation operation is represented by [·;·], the activation function is denoted by $\delta(\cdot)$, and $\otimes$ represents element-wise multiplication. The GRU structure can be formulated as follows:

$$c = \delta(Q_h[ATT(\cdot); DC(\cdot)]) \tag{6}$$

$$g = (1 - c) \otimes ATT(\cdot) + c \otimes DC(\cdot) \tag{7}$$

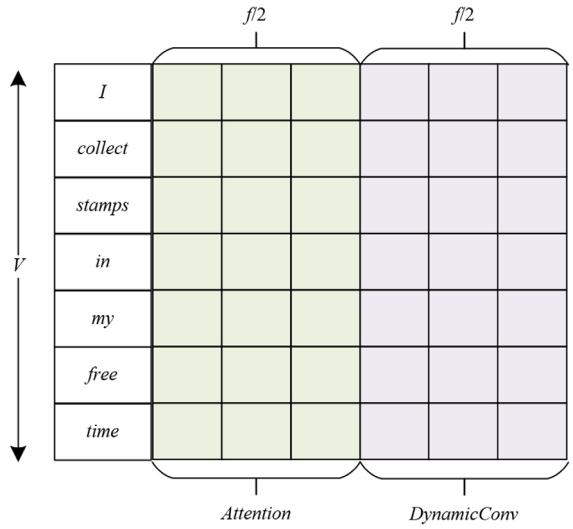


**Fig. 2.** Vector dimension diagram of the lightweight attention mechanism module

To accelerate the convergence of the GRU, layer normalization was applied as follows:

$$G = (\text{LayerNorm}(g)) \tag{8}$$

## 3 ENCODER AND DECODER DESIGN FOR THE LIGHTWEIGHT TRANSLATION MODEL ON MOBILE DEVICES

The encoder and decoder of the lightweight translation model for mobile devices were further designed in this study, as shown in Figures 3 and 4.

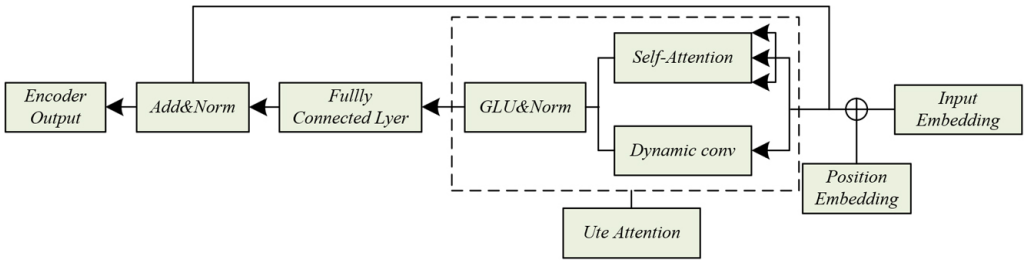


**Fig. 3.** Encoder structure of the lightweight translation model for mobile devices

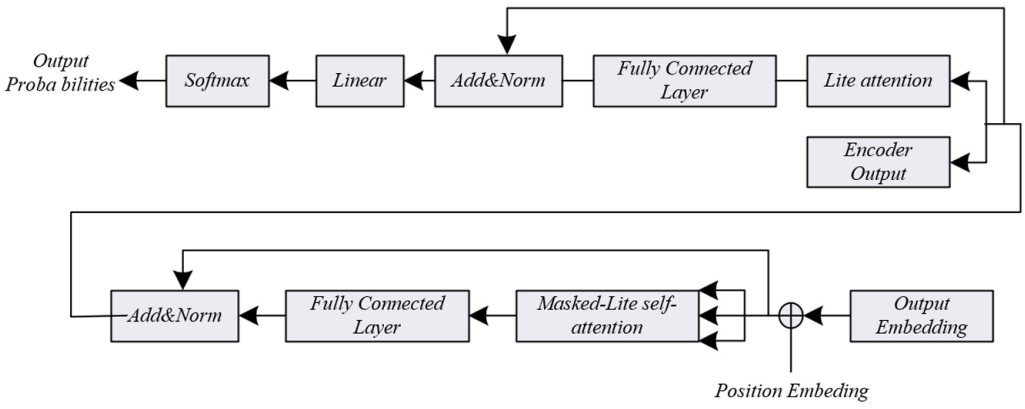


**Fig. 4.** Decoder structure diagram of the lightweight translation model for mobile devices

The encoder design of the proposed lightweight Transformer model aims to optimize machine translation tasks on mobile devices, particularly in environments with limited computational resources, to enhance both translation efficiency and quality. Similar to the traditional Transformer encoder structure, the proposed model uses the addition of word embeddings and positional information as the input structure, ensuring that both semantic and positional information within the sentence can be effectively encoded. The inclusion of positional information allows the model to better understand the word order relationships in the sequential data, which is particularly important for processing long sentences or complex structures in translation tasks. Under this input structure, the sentence vector $X$ is first passed through positional encoding, after which it enters the lightweight attention mechanism module for further processing. Specifically, the embedding information of the word $\tilde{a}_j$ is represented by $R_j^Q$, and for the $u$-th English word $\tilde{a}_u$, the input is given by:

$$\tilde{a}_j = OR\left[j,u\right] + R_j^Q \tag{9}$$

Assuming the position of the word in the sentence is represented by $j$, and the embedding dimension of the word is denoted by $f$, where $2u$ is an even dimension, and $2u + 1$ is an odd dimension. $OR[j, u]$ is expressed as:

$$OR\left[j,u\right] = SIN\left(\frac{j}{10000^{\frac{2u}{f}}}\right) \tag{10}$$

$$OR\left[j,u\right] = COS\left(\frac{j}{10000^{\frac{2u+1}{f}}}\right) \tag{11}$$

The formula for inputting $A$ into the lightweight attention mechanism module is as follows:

$$G = L\_ATT\,(A) \tag{12}$$

Unlike the attention mechanism in the traditional Transformer encoder, the lightweight Transformer presented in this study adopts a lightweight attention mechanism structure as its core. The lightweight attention mechanism module integrates two feature extraction methods: the attention mechanism and dynamic convolution, and processes the input sentence through a dual-branch structure. In the lightweight attention mechanism module, the sentence vector is processed by two branches: the left branch uses the traditional multi-head attention mechanism to capture global dependencies, while the right branch employs dynamic convolution to enhance the extraction of local features. This dual-branch design enables the model to process both global and local features efficiently with fewer parameters, achieving more effective feature extraction. The output $G$ of the lightweight attention mechanism module combines the features from both branches. However, since the output dimension $f_{OUT} = f/2$, it is necessary to expand the output dimension back to the input dimension $f$ through a fully connected layer to ensure consistency between the input and output. Assuming that the parameter matrices are represented by $Q_1$ and $y_1$, the calculation formula is given as:

$$G = FCL(G) = \mathrm{ReLU}(aQ_1 + y_1) \tag{13}$$

To further enhance the model's stability and training performance, an Add & Norm structure was introduced after the output of the lightweight attention mechanism module. This structure incorporates residual connections and layer normalization operations, aimed at addressing the vanishing gradient problem in deep network training. Residual connections help facilitate the transmission of information by adding the input and output, preventing the potential vanishing or exploding gradient issues during backpropagation in deep networks. Layer normalization is applied to the output of each layer, which aids in accelerating training and improving the model's generalization ability. The calculation formula is as follows:

$$OUT = (\text{LayerNorm}(A + G)) \tag{14}$$

After processing through the lightweight attention mechanism module, the fully connected layer for dimension expansion, and the Add & Norm structure, the encoder ultimately outputs the hidden state representation of the English sentence. This representation serves as one of the inputs for the decoder and continues to participate in the subsequent translation task training.

In this study, the design of the lightweight Transformer model's decoder also follows the structure of the standard Transformer, but was optimized to meet the requirements of mobile device lightweight translation. The design goal of the decoder is to handle Chinese sentence translation tasks while maintaining high translation quality under resource-constrained environments. While the general structure of the decoder is similar to that of the encoder, several key differences exist, including the different input language, the use of masked attention mechanisms, and adjustments in computation.

Firstly, the input to the decoder is the Chinese sentence, rather than the English sentence. This directly impacts the training and translation process of the model, as the decoder's task is to generate the Chinese sentence translation based on the hidden state of the English sentence output from the encoder. During the decoding process, in order to prevent the model from accessing future word information when predicting the current word, a masked operation was introduced in the first lightweight attention mechanism layer. The masked operation ensures that each word in the target language only depends on the previous words, without access to words that follow the current word. This allows the decoder to progressively generate the translation output, preventing information leakage and ensuring compliance with the generation process.

The second lightweight attention mechanism layer in the decoder is similar to the attention mechanism in the encoder, with one notable difference: the method of calculating the attention mechanism in the second lightweight attention mechanism layer of the decoder. Specifically, the keys and values in this layer are not directly computed from the output of the previous decoder layer, but instead are derived from the hidden states output by the encoder. This design enables the decoder to fully utilize the contextual information generated by the encoder, thereby dynamically adjusting the translation output during the decoding process and enhancing the contextual consistency of the translation results.

In the final part of the decoder, a SoftMax layer was employed to predict the probability of the next translation word. Due to the use of the masked operation in the decoding process, the SoftMax layer only considers all the information prior to the current word and is not influenced by subsequent words. The output dimension of the SoftMax layer matches the vocabulary size, generating a probability distribution of vocabulary size for each prediction, from which the word with the

highest probability is selected as the translation output. However, as the vocabulary size increases, the number of parameters that the SoftMax layer needs to update also increases, which may result in greater redundancy in the model and impact its computational efficiency and training time. Particularly on resource-constrained devices, a larger vocabulary can increase the complexity of the model, thus prolonging training and inference times and raising computational costs. Therefore, in practical applications, a balance must be struck between vocabulary size and model efficiency to meet the performance requirements of mobile devices.

## 4 EXPERIMENTAL RESULTS AND ANALYSIS

**Table 1.** Comparison of translations

| Sentence Type | Sentence Content |
|---|---|
| Sentence 1 | *Remember what should be remembered, and forget what should be forgotten. Alter what is changeable, and accept what is mutable.* |
| Reference translation | 记住该记住的，忘记该忘记的。改变能改变的，<br>接受不能改变的。 |
| $F_1$ dataset | 记住值得铭记的东西，忘记应该遗忘的事情；<br>改变可以改变的事情，接受无法改变的事实. |
| $F_1'$ dataset | 记住应该记住的，忘记应该忘记的。改变可变的，<br>接受不可变的。 |
| Sentence 2 | *You never really know a man until you stand in his shoes and walk around in them.* |
| Reference translation | 除非你穿上一个人的鞋子，像他一样走来走去，<br>否则你永远无法真正了解一个人。 |
| $F_2$ dataset | 除非你站在一个人的立场上走来走去，<br>否则你永远不会真正了解他 |
| $F_2'$ dataset | 如果你不感同身受，你永远不会真正地了解一个人。 |

Table 1 presents examples of the comparison of translations. In this table, the $F_1$ and $F_2$ datasets correspond to the training and validation sets, respectively, while the $F_1'$ and $F_2'$ datasets correspond to the pseudo-parallel corpora for the training and validation sets. When analyzing the experimental data in Table 1, the discussion can focus on three key aspects: translation quality, learning efficiency, and improvement in user experience, particularly in light of the characteristics of translation models on mobile devices. First, according to the experimental results, the translation quality has generally improved for both the $F_1$ and $F_2$ datasets. The translation performance of the $F_1$ dataset is relatively precise, especially for sentences such as "remember what should be remembered, and forget what should be forgotten," where the model effectively preserves both the semantic and syntactic structure of the sentence, with an accurate understanding of the context. In the $F_1'$ dataset, the translation quality does not significantly decline compared to the $F_1$ dataset, indicating that the pseudo-parallel corpus has had a positive impact on the training of the translation model. Notably, for the translation of "alter what is changeable, and accept what is mutable," the model demonstrates flexibility in adjusting word order, thereby improving the fluency and naturalness of the sentence. It is evident that the lightweight translation model on mobile devices has achieved notable improvements in the learning experience. In particular, the performance of the $F_2$ and $F_2'$ datasets reveals that through optimizing the attention mechanism and the

encoder/decoder structure, both the accuracy and processing speed of the translation can be significantly enhanced. This structural optimization allows the model to perform high-quality translations with reduced computational resources, thus improving the user interaction experience.

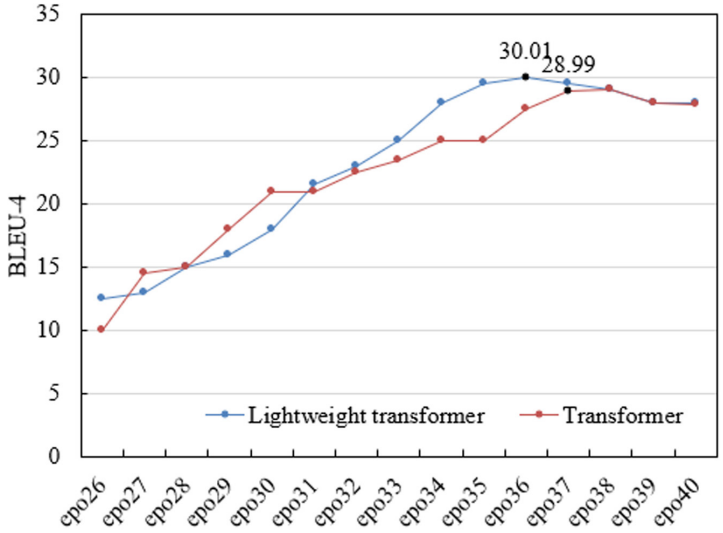


**Fig. 5.** Comparison of experimental results before and after lightweighting of the model under the $F_1'$ dataset

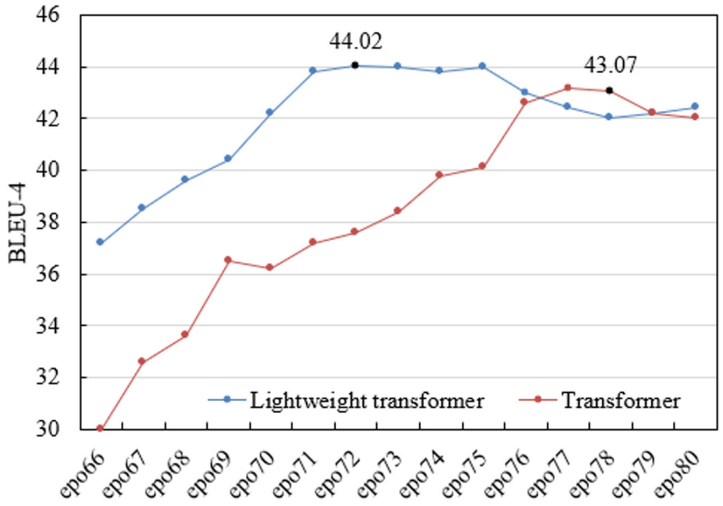


**Fig. 6.** Comparison of experimental results before and after lightweighting of the model under the $F_2'$ dataset

Based on the experimental results shown in Figures 5 and 6, which present a comparison of the performance before and after the lightweighting of the model under the $F_1'$ and $F_2'$ datasets, a significant improvement in performance can be observed after model lightweighting. Taking the $F_1'$ dataset as an example, the lightweight Transformer model increased from a score of 12.5 at epoch 26 to 28 at epoch 40, showing a marked performance improvement. Within the same number of training epochs, the standard Transformer model's performance increased from 10 to 27.92. In contrast, the lightweight version showed a more pronounced improvement in translation quality, especially during the later stages of training, where the lightweight model outperformed the standard model. Additionally, similar trends were observed

in the $F_2'$ dataset. The lightweight Transformer model increased from a score of 37.2 at epoch 66 to 42.4 at epoch 80, while the standard Transformer model increased from 30 to 42. Although both models reached similar scores at the end, the lightweight model exhibited a more noticeable performance gain during the training process. This suggests that the lightweight model improves training efficiency while maintaining translation quality. From the aforementioned experimental results, it can be concluded that through architectural optimization, the lightweight model achieves higher performance gains under the same computational resources. This not only reduces the computational load but also improves training speed and processing efficiency, making it particularly suitable for the computational constraints of mobile devices.

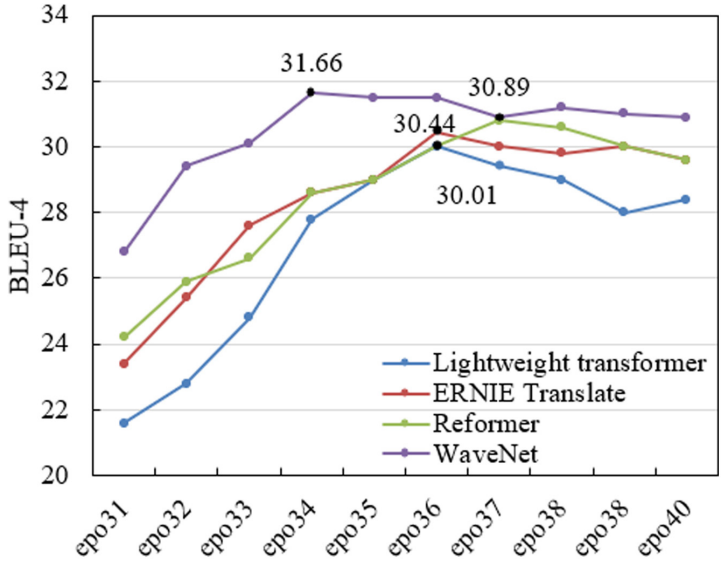


**Fig. 7.** Comparison of the experimental results of different machine translation models under the $F_1'$ dataset

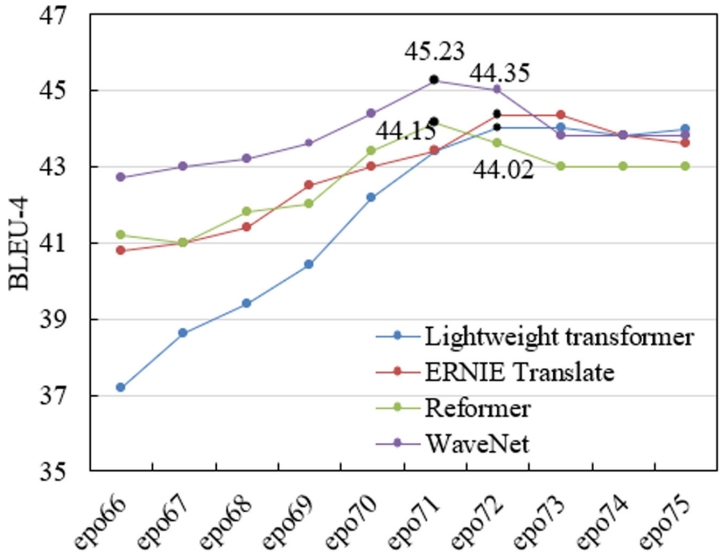


**Fig. 8.** Comparison of the experimental results of different machine translation models under the $F_2'$ dataset

Based on the comparison data for different machine translation models under the $F_1'$ and $F_2'$ datasets, as shown in Figures 7 and 8, it can be observed that the

lightweight translation model for mobile devices exhibits a significant advantage over other models. Regarding the performance on the $F_1'$ dataset, the lightweight Transformer showed a steady increase in performance during the training process from epoch 31 to epoch 40, improving from a score of 21.6 to 28.4. Although its final score was slightly lower than that of WaveNet (31.66) and Reformer (30.8), the lightweight Transformer demonstrated a smooth growth trend and provided a good cost-effectiveness ratio. When compared to the Enhanced Representation through Knowledge Integration (ERNIE) Translate model (with a maximum score of 30.44), although the lightweight Transformer showed some performance gaps at certain stages, it significantly outperformed these more complex models in terms of computational efficiency, highlighting the advantage of lightweight models in balancing translation quality and computational resources. The results from the $F_2'$ dataset further revealed improvements in both accuracy and efficiency for the lightweight model. The lightweight transformer improved from a score of 37.2 to 44. While WaveNet achieved the best performance in the later stages of training (a score of 45.23), the lightweight Transformer exhibited a faster rate of accuracy improvement, demonstrating higher training efficiency. This highlights the potential of the lightweight translation model, particularly in practical applications on mobile devices.

**Table 2.** Experimental results of different machine translation models across three metrics

| Dataset | Model | BLEU-4 | Parameters | Speed |
|---|---|---|---|---|
| $F_1'$ | Transformer | 27.59 | 17.22M | × 1.1 |
| | Lightweight Transformer | 31.26 | 6.79M | × 0.78 |
| | ERNIE Translate | 31.28 | 5.74M | × 0.61 |
| | Reformer | 31.65 | 6.34M | × 0.71 |
| | WaveNet | 32.56 | 7.24M | × 0.82 |
| $F_2'$ | Transformer | 42.59 | 21.52M | × 1.1 |
| | Lightweight Transformer | 43.15 | 6.47M | × 7.7 |
| | ERNIE Translate | 43.59 | 5.62M | × 6.2 |
| | Reformer | 43.22 | 6.34M | × 0.71 |
| | WaveNet | 44.59 | 7.24M | × 0.82 |

The experimental results presented in Table 2 demonstrate that the lightweight Transformer model exhibits significant advantages across several key metrics, especially in mobile device applications, where it can both provide high translation quality and significantly enhance translation efficiency. In the $F_1'$ dataset, the BLEU-4 score of the lightweight Transformer is 31.26, a notable improvement compared to the standard Transformer (27.59). Although WaveNet (32.56) and Reformer (31.65) slightly outperform in terms of translation quality, the lightweight Transformer's score is close to these models, while its parameter count is substantially lower at 6.79 M, compared to the standard Transformer's 17.22 M and WaveNet's 7.24 M, demonstrating its advantage in resource utilization. Moreover, the training speed of the lightweight Transformer is commendable; while slightly lower than that of WaveNet (× 0.82), it surpasses ERNIE Translate (× 0.61) and Reformer (× 0.71), showing a well-balanced performance in computational efficiency. For the $F_2'$ dataset, the BLEU-4 score of the lightweight Transformer is 43.15, slightly below that

of WaveNet (44.59), but higher than that of the standard Transformer (42.59) and Reformer (43.22). Furthermore, its translation speed has greatly improved, reaching × 7.7, far surpassing all other models, especially when compared to the standard Transformer's × 1.1 and ERNIE Translate's × 6.2, where the improvement in translation efficiency is particularly striking.

In conclusion, the lightweight translation model proposed in this study demonstrates significant advantages in both translation quality and efficiency, making it especially suitable for mobile devices with limited computational resources. By optimizing the attention mechanism and streamlining the encoder and decoder structures, the lightweight Transformer successfully reduces the parameter count while maintaining translation quality, significantly boosting computational efficiency. Compared to the standard Transformer, the lightweight model not only achieves higher translation accuracy but also shows superior adaptability in translation speed, better meeting the demands of translation tasks on mobile devices.

## 5 CONCLUSION

A lightweight translation model based on mobile devices was proposed in this study, aimed at enhancing translation teaching efficiency and the learning experience on mobile devices. The core of the research lies in optimizing the attention mechanism and streamlining the encoder and decoder structures within the translation model, thereby reducing computational complexity and resource consumption while improving translation quality and execution speed. Analysis of experimental results on the $F_1'$ and $F_2'$ datasets reveals that the lightweight Transformer model outperforms traditional machine translation models in terms of computational efficiency, translation quality, and mobile device compatibility. On both the $F_1'$ and $F_2'$ datasets, the BLEU-4 score of the lightweight Transformer is comparable to other models, with a notable advantage in computational efficiency, especially in mobile device environments, where it provides faster translation responses and a smoother user experience. Furthermore, despite a reduction in parameter count, the lightweight model maintains good translation performance, offering an efficient, accurate, and low-resource solution for translation applications on mobile devices, which holds significant practical value.

Although the lightweight translation model proposed in this study achieves promising results in improving translation efficiency and quality, some limitations remain. First, despite its excellent performance on mobile devices, the lightweight Transformer may experience a decline in accuracy when handling complex contexts or long sentences. Second, while improvements have been made in the training and inference speed of the model, it still lags behind certain more efficient complex models, such as WaveNet, particularly in larger-scale datasets, where its performance may not meet expectations. Future research could focus on further optimizing the model structure, particularly in improving its capability for long-text translation and fine-grained contextual understanding.

## 6 REFERENCES

[1] G. Chai and Q. Wen, "An interactive English–Chinese translation system based on GLA algorithm," *Journal of Information & Knowledge Management*, vol. 21, no. Supp02, p. 2240014, 2022. https://doi.org/10.1142/S0219649222400147

[2] G. Dong, Y. Yang, and Q. Zhang, "Application of feature extraction algorithm in the construction of interactive English Chinese translation mode," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 8881631, 2021. https://doi.org/10.1155/2021/8881631

[3] J. Yu and X. Ma, "English translation model based on intelligent recognition and deep learning," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 3079775, 2022. https://doi.org/10.1155/2022/3079775

[4] L. M. Rababah, N. Al-Khawaldeh, and M. A. Rababah, "Mobile-assisted listening instructions with Jordanian audio materials: A pathway to EFL proficiency," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 21, pp. 129–144, 2023. https://doi.org/10.3991/ijim.v17i21.42789

[5] M. Matsuhara, K. Araki, and K. Tochinai, "Machine translation method using inductive learning for mobile terminal," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 87, no. 9, pp. 33–47, 2004. https://doi.org/10.1002/ecjc.20100

[6] F. Wang and Y. Wang, "Optimizing offline mode and data synchronization techniques for literature translation applications on mobile devices," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 22, pp. 115–129, 2024. https://doi.org/10.3991/ijim.v18i22.52451

[7] F. Qiao and H. Wang, "Mobile interactive translation teaching model based on 'Internet+'," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 10, pp. 6705–6714, 2017. https://doi.org/10.12973/ejmste/78191

[8] H. Alotaibi and D. Salamah, "The impact of translation apps on translation students' performance," *Education and Information Technologies*, vol. 28, no. 8, pp. 10709–10729, 2023. https://doi.org/10.1007/s10639-023-11578-y

[9] Z. Tan, Z. Yang, M. Zhang, Q. Liu, M. Sun, and Y. Liu, "Dynamic multi-branch layers for on-device neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 958–967, 2022. https://doi.org/10.1109/TASLP.2022.3153257

[10] Y. C. Lee and C. W. Hsueh, "Hardware/Software co-design of memory page translation for mobile virtualization," *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 3070–3082, 2016. https://doi.org/10.1109/TC.2016.2519907

[11] S. Yun, Y. J. Lee, and S. H. Kim, "Multilingual speech-to-speech translation system for mobile consumer devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 508–516, 2014. https://doi.org/10.1109/TCE.2014.6937337

[12] C. K. Chang and C. K. Hsu, "A mobile-assisted synchronously collaborative translation–annotation system for English as a foreign language (EFL) reading comprehension," *Computer Assisted Language Learning*, vol. 24, no. 2, pp. 155–180, 2011. https://doi.org/10.1080/09588221.2010.536952

[13] R. Hu and K. Wu, "Edge computing and 5G based low-delay business English translation framework," *Internet Technology Letters*, vol. 6, no. 5, p. e321, 2023. https://doi.org/10.1002/itl2.321

[14] M. Li, J. Pang, F. Yue, F. Liu, J. Wang, and J. Tan, "Enhancing dynamic binary translation in mobile computing by leveraging polyhedral optimization," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 6611867, 2021. https://doi.org/10.1155/2021/6611867

[15] A. Panayiotou *et al.*, "The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study," *Journal of Clinical Nursing*, vol. 29, nos. 17–18, pp. 3516–3526, 2020. https://doi.org/10.1111/jocn.15390

[16] S. Moh, "Approaches to using a wireless mobile terminal to help severely hearing impaired people," in *Computers Helping People with Special Needs*, ICCHP 2004, in Lecture Notes in Computer Science, K. Miesenberger, J. Klaus, W. L. Zagler, and D. Burger, Eds., Springer: Berlin, Heidelberg, vol. 3118, 2004, pp. 1137–1143. https://doi.org/10.1007/978-3-540-27817-7_167

[17] Y. Yi, J. H. Cho, J. B. Kim, J. Y. Kim, S. Y. Park, and W. C. Lee, "Change in talar translation in the coronal plane after mobile-bearing total ankle replacement and its association with lower-limb and hindfoot alignment," *The Journal of Bone and Joint Surgery*, vol. 99, no. 4, p. e13, 2017. https://doi.org/10.2106/JBJS.15.01340

[18] E. Most *et al.*, "The kinematics of fixed- and mobile-bearing total knee arthroplasty," *Clinical Orthopaedics and Related Research*, vol. 416, pp. 197–207, 2003. https://doi.org/10.1097/01.blo.0000092999.90435.d1

## 7 AUTHORS

**Hongjing Chang** is a PhD candidate at the School of Languages, Literacies, and Translation in Universiti Sains Malaysia, Malaysia. Her research interests encompass translation theory and practice, machine translation, and media-translatology. To date, she has published 19 academic papers both domestically and internationally (E-mail: hong_qingting999@163.com).

**Jing Zhao** is a PhD candidate at the School of Languages, Literacies, and Translation in Universiti Sains Malaysia, Malaysia. Her research interests encompass cognitive translation, machine translation, and cross-cultural communication. She has published over 3 academic papers, authored 2 professional books, and completed 1 translation work (E-mail: lelezhao0412@gmail.com).