Enabling Cross-Linguistic Compatibility in Image Generation: Text Embedding Alignment Techniques for CLIP Models

Stanford CS224N Custom Project

Bofei Zhu
zhu.bofei@gmail.com

Abstract

In this work, we tackle the challenge of achieving cross-linguistic compatibility in image generation, with a focus on incorporating Chinese language inputs into models predominantly designed for English. Our strategy emphasizes text embedding alignment techniques between text encoding models to facilitate the accurate generation of images from Chinese text prompts. By aligning the Chinese text encoder embeddings output with those from the original English text encoder models, we enable a seamless transition between languages in the image generation process. Our methodology explores both the use of Low-Rank Adaptation (LoRA) for model fine-tuning and an approach that selectively targets deeper model layers for fine-tuning. Remarkably, our training protocol requires approximately 100 times less computational resources compared to existing methods, yet achieves comparable results. This significant reduction in resource demand, combined with our method's effectiveness, marks a substantial advancement towards more accessible and linguistically flexible image generation technologies. Our contributions not only demonstrate the potential of text embedding alignment for enhancing cross-linguistic capabilities but also offer a more efficient pathway for expanding the generative models' language support, thereby making multilingual image generation more achievable.

1 Introduction

The proliferation of image generation models has marked a significant advancement in artificial intelligence and creative technologies. Among these, models capable of generating images from textual descriptions have seen remarkable development, driven by the integration of advanced natural language processing techniques. However, a substantial challenge within this domain is achieving cross-linguistic compatibility, particularly in extending the capabilities of these models to understand and process inputs in languages other than English. This challenge is not only technical, involving the generation and processing of text embeddings across languages, but also practical, as the computational resources required for training such models from scratch are immense.

A notable attempt in this direction is the training of a Chinese-specific image generation model, as exemplified by IDEA-CCNL's Taiyi-Stable-Diffusion-1B-Chinese (Zhang et al., 2022). Their method involves training the model from scratch to understand and generate images based on Chinese textual prompts, a process that, while effective, demands significant computational power and time. The necessity for such extensive resources poses a barrier to entry for many researchers and developers, making the exploration of more efficient alternatives a crucial area of inquiry.

Our approach seeks to address these challenges by proposing a method that not only requires substantially less computational effort but also maintains, if not enhances, the performance of cross-linguistic image generation. By focusing on the alignment of text embeddings between Chinese and English, we leverage existing models' capabilities, extending their utility without the need for exhaustive retraining. This strategy not only presents a more resource-efficient model but also contributes to the broader effort of making AI more accessible and versatile across linguistic boundaries.

2 Related Work

The development of models that generate images from textual descriptions has greatly benefited from advancements at the intersection of natural language processing and image generation. The pioneering work on CLIP (Radford et al., 2021) introduced a novel approach for aligning text and image embeddings, facilitating the intuitive generation of images from textual prompts. This foundational research has paved the way for subsequent innovations in text-to-image generation models such as DALL-E(Ramesh et al., 2021) and Stable Diffusion(Rombach et al., 2022).

Recent advancements in the efficiency of model adaptation techniques have addressed the challenges associated with training Vision-Language models for languages beyond English. Specifically, the creation of non-English languages large-scale datasets and CLIP models signify a crucial step forward. In a notable study(Yang et al., 2023), a comprehensive dataset of image-text pairs in Chinese was constructed, and a series of Chinese CLIP models were pretrained on this dataset. This work introduced a two-stage pretraining method that initially freezes the image encoder and subsequently optimizes all parameters, leading to enhanced model performance.

This backdrop of research emphasizes the challenge of expanding image generation models to accommodate a diversity of languages, highlighting both the potential and the computational demands of such endeavors. The Taiyi-Stable-Diffusion-1B-Chinese model(Zhang et al., 2022), which trains a model specifically for Chinese text inputs, illustrates one approach to this challenge. Despite its effectiveness, the project brings to light the need for resource-efficient methodologies that can adapt existing models to new linguistic contexts without excessive computational costs.

Our work engages with these developments by leveraging the insights gained from both the foundational CLIP model(Radford et al., 2021) and the more recent advancements in Chinese CLIP models(Yang et al., 2023). We focus on text embedding alignment strategies that aim to mitigate computational challenges while enhancing the model's ability to generate images from Chinese prompts. This approach not only seeks to lower the barriers for deploying advanced image generation models across different languages but also contributes to the growing body of research advocating for more scalable and efficient solutions to achieve linguistic diversity in AI applications.

3 Approach

Our approach to enabling cross-linguistic compatibility in image generation models, specifically between English and Chinese, integrates and innovates upon existing methods to align text embeddings and enhance model performance with minimal computational resources.

Method To enable the Stable Diffusion model (Rombach et al., 2022) to interpret Chinese inputs, we focus exclusively on modifying its text encoder component. Specifically, we replace the English-based CLIP (Radford et al., 2021) text encoder with the Chinese-CLIP (Yang et al., 2023) text encoder. The critical task lies in fine-tuning the Chinese-CLIP text encoder to ensure that its output text embeddings closely align with the original embeddings. This alignment is crucial for the model to accurately generate images from Chinese prompts, as it allows the model to effectively map the semantic meaning of the Chinese text to the corresponding visual representations in the latent space. By carefully fine-tuning the Chinese-CLIP text encoder, we aim to preserve the model's generative capabilities while extending its applicability to Chinese language inputs.

To achieve this alignment, we explore two different strategies:

1. **Selective Layer Unfreezing:** Our first approach to fine-tuning the Chinese-CLIP(Yang et al., 2023) text encoder involves selectively unfreezing and fine-tuning the deeper layers of the encoder. By focusing on these specific layers, we can make precise modifications to align the encoder's output text embeddings with the original embeddings of the Stable Diffusion model. This targeted fine-tuning allows us to leverage the foundational knowledge captured in the pre-trained Chinese-CLIP model while making strategic adjustments to adapt it to

the specific requirements of the Stable Diffusion architecture. By carefully selecting which layers to unfreeze and fine-tune, we can strike a balance between preserving the model's understanding of the Chinese language and ensuring that the generated text embeddings are compatible with the Stable Diffusion model's latent space. This approach enables us to make fine-grained improvements to the encoder's performance, ensuring accurate mapping between the Chinese text prompts and the corresponding visual representations within the Stable Diffusion framework, without overwriting the valuable language understanding capabilities of the Chinese-CLIP model.

2. **LoRA Fine-Tuning:** Alternatively, we employ Low-Rank Adaptation (LoRA)(Hu et al., 2021) to fine-tune the Chinese-CLIP text encoder. LoRA is a parameter-efficient fine-tuning technique that introduces a low-rank decomposition of the encoder's weight matrices. By adding a small number of trainable parameters to the existing layers, LoRA allows for fine-grained adjustments to the encoder's behavior without modifying the original pre-trained weights. This approach enables us to adapt the Chinese-CLIP text encoder to generate text embeddings that match the original embeddings of the Stable Diffusion model while minimizing the computational overhead and storage requirements associated with fine-tuning. By leveraging LoRA, we can efficiently fine-tune the encoder, focusing on the specific adjustments needed to align the text embeddings with the Stable Diffusion architecture. This method provides a balance between adaptation performance and resource efficiency, making it a compelling choice for integrating the Chinese-CLIP text encoder into the Stable Diffusion model.

Both strategies use the cosine similarity loss between the Chinese text embeddings and the original English text embeddings as the primary metric for fine-tuning. By minimizing this loss, we bring the Chinese embeddings closer to the original ones, ensuring their alignment within the Stable Diffusion model's latent space.

During fine-tuning, the Chinese-CLIP text encoder adjusts its parameters to generate embeddings that are more compatible with the original embeddings. A lower cosine similarity loss indicates successful adaptation, as the Chinese embeddings become more aligned with the visual representations in the latent space.

Monitoring the cosine similarity loss allows us to assess the effectiveness of the adaptation process. A significant reduction in the loss demonstrates the encoder's improved ability to process Chinese inputs and generate embeddings that seamlessly integrate with the Stable Diffusion model, maintaining the quality and consistency of the generated images.

Baseline Our baseline is the performance of the Stable Diffusion v1.5(Rombach et al., 2022) model, which utilizes the original CLIP text encoder for English text prompt processing. This provides a reference for evaluating the success of our adapted encoder with Chinese text inputs.

Original Contributions Our work introduces a novel approach to enable cross-linguistic compatibility in image generation models by directly matching text embeddings between encoders for different languages. This approach represents a focused and efficient method to extend the model's linguistic reach with minimal alterations to its structure.

4 Experiments

Data Our initial attempts involved leveraging the OPUS dataset(Tiedemann and Thottingal, 2020), a parallel corpus of English and Chinese sentences. The OPUS dataset has been instrumental in various NLP tasks involving translation and language understanding due to its diverse range of sentences and contexts. However, when applied to our specific task of aligning text embeddings for image generation, the performance fell short of our expectations. This discrepancy led us to hypothesize that the core issue lay in the dataset's composition; primarily, the OPUS corpus is rich in conversational language that lacks the visual descriptiveness required for effective image generation tasks.

To address this challenge, we curated our dataset by translating 300,000 texts from the YFCC100M dataset(Thomee et al., 2016), specifically focusing on the subset known as YFCC15M for its rich collection of image and text pairs suitable for training image generation models. This dataset comprises descriptive texts associated with images, providing a more appropriate foundation for our task of image generation from textual descriptions. By translating these texts into Chinese, we created a parallel corpus that mirrors the visual descriptiveness of the original English dataset, ensuring that both the input (Chinese text prompts) and the output (the corresponding English text embeddings from the original CLIP model) are precisely aligned in terms of visual content representation. This tailored dataset significantly improved the relevance and effectiveness of the fine-tuning process, aligning more closely with our goal of generating contextually accurate images from Chinese text prompts.

Evaluation Method

- **Quantitative Evaluation:** We measure the cosine similarity loss between the embeddings from the English CLIP model and those from our fine-tuned Chinese-CLIP encoder. This metric quantifies the alignment of semantic meanings between the two sets of embeddings, with a lower loss indicating better alignment.
- **Qualitative Evaluation:** Additionally, we perform a qualitative comparison of images generated from equivalent English and Chinese prompts. This evaluation aims to visually assess how well the model maintains the semantic integrity of the prompts across languages, observing the relevance, creativity, and accuracy of the generated images in reflecting the intended concepts.

Experiment Details Our project aimed to adapt the CLIP text encoder for Chinese text inputs, focusing on two main strategies: selective layer unfreezing and LoRA fine-tuning. This adaptation utilized a dataset derived from translating 300,000 texts from the YFCC15M dataset, with the objective of aligning Chinese and English text embeddings more closely.

For our experiments, we employed the Hugging Face `transformers` and the `PEFT` library to implement LoRA fine-tuning on the pretrained Chinese-CLIP model. A critical step in our process was the preprocessing of English CLIP text embeddings, which was necessary for calculating the loss between the model outputs and these preprocessed embeddings.

Our experimental configuration was set as follows:

- **Selective Layer Unfreezing:** We specifically unfreezed the last three transformer layers of the Chinese-CLIP text encoder for fine-tuning. This approach allowed us to make targeted adjustments to the layers most directly involved in generating text embeddings, optimizing the encoder's ability to process Chinese text with minimal disturbance to the pre-trained model's foundational knowledge.
- **LoRA Fine-Tuning:** In implementing the LoRA fine-tuning strategy, we configured the model with specific parameters to fine-tune the text encoder efficiently. We set the rank `r` to 16 and the `lora_alpha` parameter to 64. These settings were chosen to balance the fine-tuning process's efficiency and effectiveness, allowing for nuanced adjustments to the model's parameters with a focus on enhancing its cross-linguistic compatibility.
- **Training Configuration:** Utilizing the original architecture of the CLIP text encoder, we applied selective layer unfreezing and LoRA fine-tuning strategies independently. The fine-tuning parameters included a learning rate of 5×10^{-3}, with a batch size of 1024, extending over 20,000 training steps.
- **Loss Metric:** The primary metric guiding our fine-tuning process was the cosine similarity loss, aiming to minimize this loss by aligning the Chinese text embeddings with the preprocessed English CLIP text embeddings. This method ensured a direct comparison and loss calculation, essential for evaluating the effectiveness of our adaptation strategies.
- **Computational Resources:** The experiments were conducted on a single NVIDIA A100 GPU for each strategy, each requiring around 12 hours of training time, highlighting the efficiency of our approach.

Results

- **Quantitative Results:** Our experiments were conducted to assess the effectiveness of two distinct strategies, selective layer unfreezing and LoRA fine-tuning, on aligning the text embeddings from the fine-tuned Chinese-CLIP encoder with those from the original English

CLIP model. The primary metric for this evaluation was the cosine similarity loss, with a lower loss indicating a tighter semantic alignment between the two sets of embeddings. The results from our experimental setup, as shown in table 1, demonstrate a notable reduction in cosine similarity loss for both the selective layer unfreezing and LoRA fine-tuning strategies when compared to the baseline. Interestingly, the selective layer unfreezing approach outperformed the LoRA fine-tuning strategy, achieving the lowest cosine similarity loss. This outcome suggests that the precision offered by directly adjusting the deeper layers of the text encoder leads to a more effective alignment of the Chinese text embeddings with the English counterparts.

Model	Training Loss	Evaluation Loss
Original Chinese-CLIP	-	1.02
Selective Layer Unfreezing	0.22	0.32
LoRA Fine-Tuning	0.33	0.39

Table 1: Comparison of training and evaluation loss across different models.

While the LoRA fine-tuning strategy also showed a significant improvement over the baseline, its slightly higher loss relative to the selective layer unfreezing method indicates that, in this context, the direct and targeted adjustments of the encoder layers are more advantageous for embedding alignment. This finding aligns with our hypothesis that specific modifications to the encoder can greatly enhance its cross-linguistic capabilities, albeit it slightly deviates from our initial expectations regarding the comparative effectiveness of LoRA fine-tuning.

- **Qualitative Results:** The qualitative evaluation of images generated from equivalent English and Chinese prompts further corroborates the quantitative findings, as shown in table 2. The selective layer unfreezing strategy led to the generation of images that closely matched the semantic integrity and creative intent of the prompts across both languages. This success underscores the approach's ability to maintain the conceptual fidelity of the generated images, reinforcing its superiority in aligning text embeddings for cross-linguistic image generation tasks.

  These results highlight the selective layer unfreezing strategy's potential in adapting existing language models for multilingual applications, particularly in the context of image generation from text prompts. This strategy's effectiveness in achieving a tighter semantic alignment between Chinese and English text embeddings presents a promising direction for future research and development in the field of multilingual image generation.

  In addition to the quantitative and qualitative evaluations presented, it's noteworthy to highlight the efficiency of our adaptation strategies in terms of computational resources and training duration. Both the selective layer unfreezing and LoRA fine-tuning methods required only a single NVIDIA A100 GPU and approximately 12 hours of training time. This contrasts sharply with the existing method of training a Chinese image generation model from scratch, such as the Taiyi-Stable-Diffusion-1B-Chinese-v0.1(Zhang et al., 2022), which necessitated a significantly larger computational commitment of 32 NVIDIA A100 GPUs for around 100 hours. Despite this vast difference in resource utilization, our adapted models demonstrate performance that is qualitatively comparable, if not better, to that of the extensively trained models. This efficiency does not only underscore the practicality and accessibility of our approach but also signifies a substantial advancement in developing multilingual capabilities for image generation models, making it a viable and cost-saving option for researchers and developers with limited access to large-scale computational resources.

5 Conclusion

Our exploration into adapting the CLIP text encoder for Chinese text inputs in image generation, via selective layer unfreezing and LoRA fine-tuning strategies, revealed that both methods effectively align Chinese text embeddings with those of the original English CLIP model. Notably, selective layer unfreezing not only slightly outperformed LoRA fine-tuning in reducing cosine similarity loss but also generated images from Chinese prompts with remarkable quality, comparable or even superior

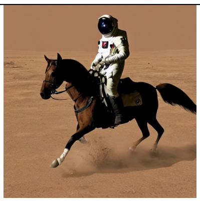
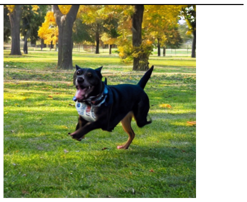
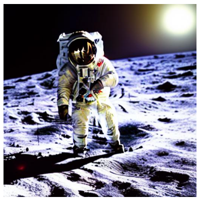
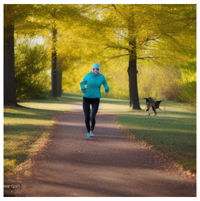
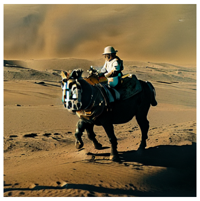
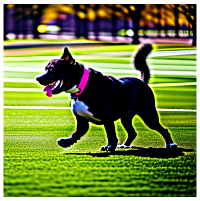
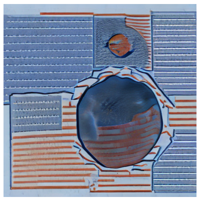
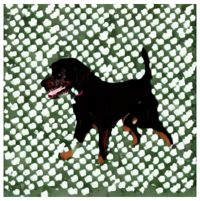
English Prompt	A photo of an astronaut riding a horse on mars	Dog running in the park
Chinese Prompt	宇航员在火星上骑马的照片	狗在公园里跑步
Stable Diffusion v1.5 (English + Seed 42)		
Taiyi-Stable-Diffusion-1B-Chinese-v0.1		
Selective Layer Unfreezing (Seed 42)		
LoRA Fine-Tuning (Seed 42)		

Table 2: Qualitative comparison of image outputs for different adaptation models.

to those produced by more resource-intensive models. This was achieved using significantly fewer computational resources, showcasing the potential for achieving advanced multilingual capabilities in a more accessible manner.

However, our approach, focusing on two languages and specific strategies, has its limitations, indicating the necessity for a broader quantitative analysis and exploration across various languages and fine-tuning techniques.

The particular success of selective layer unfreezing in producing quality outcomes emphasizes its utility, not just for image generation models but for enhancing multimodal models with text embedding layers. This finding encourages further research into developing linguistically inclusive and computationally accessible multimodal models, leveraging minimal resource investments to unlock significant advancements in generative AI.

References

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2023. Chinese clip: Contrastive vision-language pretraining in chinese.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.