Optimizing Gemma2B for Language Translation on Edge

Rishik Mishra
(rm6397)

Chen Xu
(cx2214)

NYU

Summary

Problem:
Develop a multilingual translation system for edge devices

Solution:
Leverage Gemma2B large language model (LLM) with LoRA adaptation for efficient Hindi-Chinese translation. Utilize pruning for model compression and performance optimization. Integrate text-to-speech (TTS) and speech-to-text (STT) for comprehensive functionality.

Benefits:
Provide natural-sounding translations, even with limited internet availability.

Deployment:
Built a fully interactive Hi-Ch translation voice assistant on Raspberry Pi

Challenges

Limited Device Resources:
Our initial approach involved deploying the Gemma2B LLM directly on a Raspberry Pi device. However, the LLM's large size significantly exceeded the Pi's limited RAM, hindering its ability to run other processes simultaneously.

Natural Language Nuance Capture:
Training the model on comprehensive datasets that are rich in natural language elements remains crucial to capture the subtleties of both Hindi and Chinese for natural-sounding translations.

Approach

Fine-tuning with LoRa:
We leveraged LoRa (Low Rank Adaptation) to fine-tune a pre-trained model on a parallel Chinese-Hindi dataset. This technique improves the model's ability to capture the semantic relationships between the two languages.

Layer Selection:
We experimented with fine-tuning the model using different combinations of self-attention and linear layers. This helped us determine which yielded the best translation accuracy for our specific use case.

Approach

Pruning:
We employed iterative pruning to remove redundant connections within the model. This technique assumes that many weights, especially those crucial for complex tasks like coding or mathematical calculations, are not essential for our translation model. By iteratively removing these, we aimed to achieve a smaller model size with minimal impact on accuracy.

Quantization:
We further attempted to reduced model size through quantization, converting the model's weights from high-precision floating-point numbers to lower precision formats.

Results

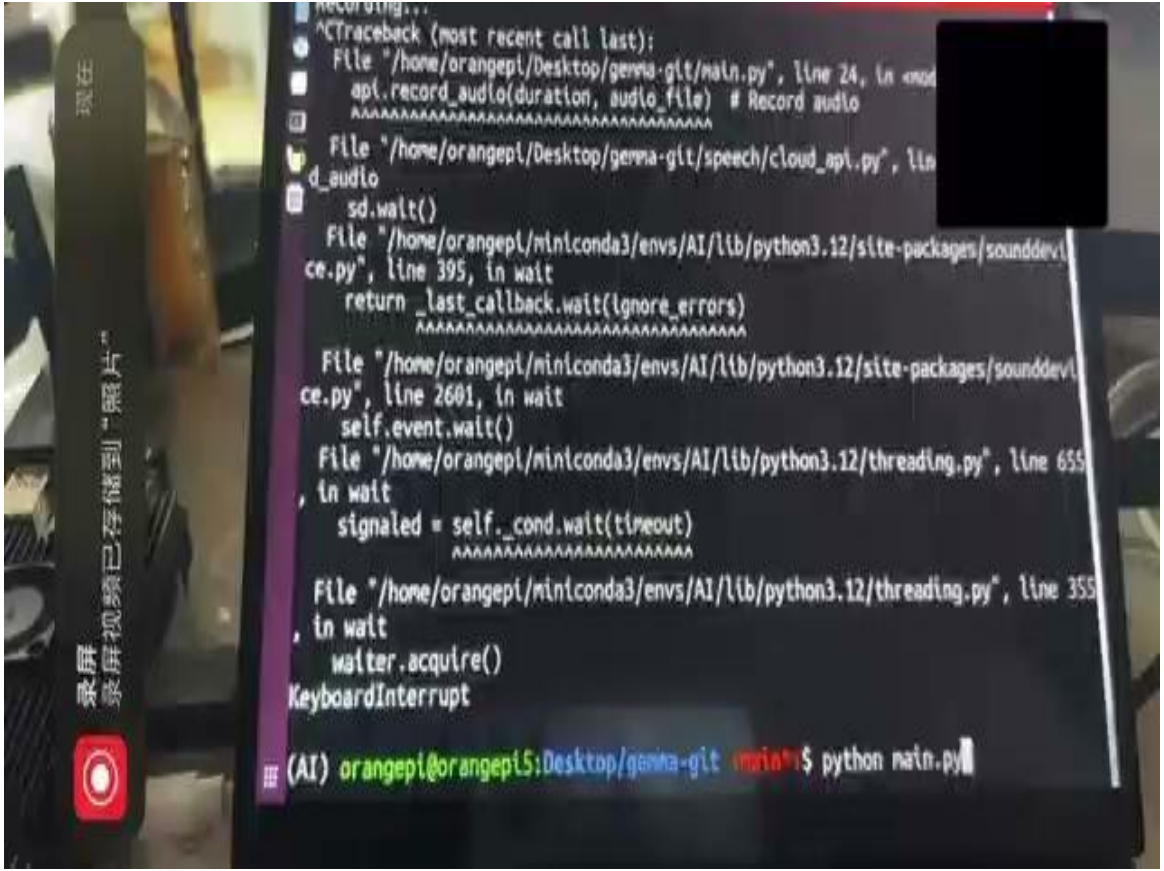
LoRA fine-tuning on a parallel corpus proved particularly effective, achieving accurate translations with fewer unnecessary words compared to standard fine-tuning. Additionally, focusing on Unicode character handling during training improved the model's ability to process both Hindi and Chinese languages effectively.

Results

Results with fine tuned Gemma 2B

CPU time	Mem Usage	Execution time(stt + generation + tts)
2.98 sec	6.38 GB	18.12 sec

NYU


录屏
录屏视频已存储到"照片"
^CTraceback (most recent call last):
File "/home/orangepi/Desktop/gemma-git/main.py", line 24, in <mod
api.record_audio(duration, audio_file) # Record audio
File "/home/orangepi/Desktop/gemma-git/speech/cloud_api.py", lin
d_audio
sd.wait()
File "/home/orangepi/miniconda3/envs/AI/lib/python3.12/site-packages/sounddevi
ce.py", line 395, in wait
return _last_callback.wait(ignore_errors)
File "/home/orangepi/miniconda3/envs/AI/lib/python3.12/site-packages/sounddevi
ce.py", line 2601, in wait
self.event.wait()
File "/home/orangepi/miniconda3/envs/AI/lib/python3.12/threading.py", line 655
, in wait
signaled = self._cond.wait(timeout)
File "/home/orangepi/miniconda3/envs/AI/lib/python3.12/threading.py", line 355
, in wait
waiter.acquire()
KeyboardInterrupt
(AI) orangepi@orangepi5:Desktop/gemma-git (main)$ python main.py

Experimental Evaluation

LoRA parameter selection

Rank	Alpha	Loss
4	4	1.3627
4	32	1.9830
8	4	0.8363
8	32	0.4753
32	4	0.7653
32	32	0.5193

NYU

Experimental Evaluation

Layer selection for fine tuning

layers	loss
q_proj	0.9847
q_proj+v_proj	2.8254
all linear	1.2714
q_proj+v+proj+k_proj	0.4753
o_proj+v_proj	2.6398
gate_proj	1.9679
gate_proj + up_proj	1.1325

NYU

Experimental Evaluation

Model size comparison

Pruning	Parameters	Loss
0.0	2506172416	0.49566
0.4	1515268096	1.02843
0.5	1342155576	2.05524
0.2	2088498733	1.00341
0.1	2281118133	0.91843
0.6	1021340445	3.0552

NYU

Experimental Evaluation

Translation results

model	Input	Output	Input meaning	Output meaning
Pre-trained Gemma	Translate this Hindi to Chinese: आज टीम जीत गयी	昨我们希望通过这项计划，在2010年11月1日之前	The team has won	We hope this plan can before 11.1.2010
	Translate this Chinese to Hindi: 我是纽约大学的学生	मेरा नाम ओरियन प्रेस है	I am a student at New York University	My name is Orion Press
Fine-tuned	Translate this Hindi to Chinese: आज टीम जीत गयी	球队获得了胜利	The team has won	The team has gain victory
	Translate this Chinese to Hindi: 我是纽约大学的学生	मैं एक न्यूयॉर्क विश्वविद्यालय छात्र हूं।	I am a student at New York University	I am a New York University student.

Conclusion

Our approach successfully addressed the challenge of deploying a real-time multilingual translation system on edge devices with limited resources. By leveraging LoRA fine-tuning and pruning techniques, we were able to significantly reduce the size of the Gemma2B LLM while maintaining good translation accuracy. This allowed for Hindi-Chinese translation on a Raspberry Pi device.

There is still room for improvement. Future work could involve exploring more sophisticated pruning techniques for further model size reduction while maintaining the accuracy. Additionally, incorporating a larger and more diverse parallel corpus for training could potentially enhance translation quality and naturalness.

Github Repository

https://github.com/cyrus-xc/Gemma2b-voice-translator-on-edge

NYU