LLM Pipelines: Seamless Integration on Embedded Devices

tinyAI Virtual Forum on Generative AI and Foundation Models on the Edge 2024

Enzo Ruedas

Equal contribution with:
- Tess Boivin
- Baptiste Pouthier
- Laurent Pilati

Voice & Audio Team, NXP
LLM Pipelines
What is the LLM Pipelines project?

LLM Pipelines

Automatic Speech Recognition (ASR)
Model size: 50M INT8

Large Language Model (LLM)

Text-To-Speech (TTS)
Model size: 2.3M INT8

Wake Word Engine (WWE)
VIT
What is the LLM Pipelines project?
What is the LLM Pipelines project?
What is the LLM Pipelines project?
What is the LLM Pipelines project?
What is the LLM Pipelines project?

**LLM Pipelines - LLM**

- **Fine-tuning**
- **Quantization**

Clients requirements → Fine-tuning → Quantization → Hardware requirements

**NXP MPU platforms:** i.MX 8M Plus, i.MX 93, i.MX 95
Neural Processing Unit (NPU) scales Machine Learning Solutions from MCX MCUs to i.MX 9 Applications Processors

NXP eIQ Neuron Neural Processing Unit (NPU) hardware scales from 32 to >2k operations/cycle, for use across the range of MCUs and MPUs

NXP MPU platforms:
- **i.MX 8M Plus**
  - 4x Arm® Cortex® - A53 (1.8Ghz)
  - NPU (2.3 TOPS)
- **i.MX 93**
  - 2x Arm® Cortex® - A55 (1.7Ghz)
  - NPU (0.5 TOPS)
- **i.MX 95**
  - 6x Arm® Cortex® - A55 (1.7Ghz)
  - eIQ Neuron NPU (2 TOPS)
Our metrics

**Similarity:**

**Fine-tuning**

- **RAG**
  - Query
  - Prompt + Chunks + Query
  - Context
  - Embedding Model
  - Embedding

**Quantization**

- **Reference model**
  - Answer
  - Embedding

- **Compared model**
  - Answer
  - Embedding

\[ \text{Similarity} = \cos(\text{context, model-answer}) \]

Measure the model ability to use the given context in formulating its answer.

\[ \text{Similarity} = \cos(\text{reference-answer, compared-answer}) \]

Measure the quality loss of the model after quantization.

**Time To First Token (TTFT):** The time in seconds before the model produces its first answer token.

**Token/s:** Average number of token generated per second during generation time.
Fine-tuning
Fine-tuning strategies

Ingestion
- Use case documentation
  - Chunking strategy
  - Chunks

LoRA
- Use case specific dataset
  - Fine-tuning
  - Foundation LLM

Use case application
- Vector database
- Top K chunks
- Query
- Question
- Answer
- User
- Fine-tuned LLM
  - Completion
  - Prompt
  - Synthesis
The Low Rank Adaptation (LoRA) for fine-tuning
The Low Rank Adaptation (LoRA) for fine-tuning

Results - automotive use case:

“What is the DIC?”

- Foundation LLM
  - The DIC (Drug Information Center) is a database of information on drugs and their interactions with each other and with the body’s systems.

- Fine-tuned LLM
  - The DIC stands for Driver Information Center and displays various vehicle information in real-time. It shows things like oil life, tire pressure, engine hours, and more.
The Low Rank Adaptation (LoRA) for fine-tuning
The Retrieval-Augmented Generation (RAG) for fine-tuning

**Fine-tuning**
- **Ingestion**
  - Use case documentation
  - Chunking strategy
  - Chunks

**LoRA**
- Use case specific dataset
- Fine-tuning
- Foundation LLM

**Use case application**
- **Retrieval**
  - Vector database
  - Top K chunks
  - Query

- **Synthesis**
  - Fine-tuned LLM
  - Completion
  - Prompt

- **User**
  - Question
  - Answer
The Retrieval-Augmented Generation (RAG) for fine-tuning

Results on i.MX 95 – TinyLlama 1B – automotive use case:

➢ Out of domain question:

```
User: What is the weather today?
Please ask something about the car?
```

| Similarity | 0.236 |
| TTFT       | Ø     |
| Time to detect | 0.08 s |

➢ In domain question:

- Without RAG:

```
User: What is the contact number for assistance?
The contact number for assistance is not mentioned in the given text.
```

| Similarity | 0.079 |
| TTFT       | 1.5 s |
| Token/s    | 5.1   |

- With RAG:

```
User: What is the contact number for assistance?
The contact number for assistance is 1-888-881-3392.
```

| Similarity | 0.818 |
| TTFT       | 3.9 s |
| Token/s    | 5.1   |

 Hallucination

 Real value
The Retrieval-Augmented Generation (RAG) for fine-tuning

Time To First Token (TTFT) vs. Prompt size

Prompt size (tokens) vs. Time To First Token (TTFT) for different devices:
- i.MX 8M Plus (4.7 token/s)
- i.MX 93 (2.6 token/s)
- i.MX 95 (5.1 token/s)

ℹ️ The lower the better
Quantization
Quantization strategies

Quantization

Dynamic Quantization:
(Q2 → Q8)

Static Quantization:
SmoothQuant

Fine-tuned LLM

or

Quantized LLM

No Quantization
About Quantization

Reduce the model memory footprint

Accelerate the inference computation on CPU & NPU

Necessary for NPU enablement with integer operations

Minimize the loss of response quality after quantization
Quantization strategies

Results on i.MX 95 – TinyLlama 1B – automotive use case:

"Can I play my music on the car?"

LLM

Without Quantization
(float16/32)

Yes, you can play your music on the car using a Bluetooth speaker or a USB drive.

20 tokens

Similarity 1
TTFT 12.6 s
Token/s 2.5

Quantized LLM

Dynamic Quantization
(int8)

You can play your favorite music on the car using a Bluetooth speaker or a USB drive.

19 tokens

Similarity 0.887
TTFT 3.8 s
Token/s 5.6

Quantized LLM

Static Quantization
SmoothQuant (int8)

To play your music through the car, you need to connect your smartphone or tablet to the car's Bluetooth system.

27 tokens

Similarity 0.804
TTFT 3.9 s
Token/s 5.1
Quantization strategies

Results on i.MX 95

Time To First Token (TTFT) vs. Prompt size

Prompt size (tokens)

- **Float (2.5 token/s)**
- **Dynamic (5.6 token/s)**
- **Static (5.1 token/s)**

The lower the better
Final architecture