Workgroup: CATS
Published: 11 August 2023
Intended Status: Standards Track
Expires: 12 February 2024
Authors: Q. An
Alibaba Group

Use Case of Computing-Aware AI large model

Abstract

AI models, especially AI large models have been fastly developed and widely deployed to serve the needs of users and multiple industries. Due to that AI large models involve mega-scale data and parameters, high consumption on computing and network resources, distributed computing becomes a natural choice to deploy AI large models.

This document desribes the key concepts and deployment scenarios of AI large model, to demonstrate the necessity of considering computing and network resources to meet the requirements of AI tasks.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 12 February 2024.

Table of Contents

1. Introduction

AI large model refers to a type of artificial intelligence model that is trained on massive amounts of data using deep learning techniques. These models are characterized by their large size, high complexity, and high computational requirements. AI large models have become increasingly important in various fields, such as natural language processing, computer vision, and speech recognition.

There are usually two types of AI large models, AI foundation model and customized model. AI foundation large model is a model that can handle multiple tasks and domains, and has wider applicability and flexibility, but may not perform as well as customized model in specific domain tasks. Customized model is trained for specific industries or domains, and more focused on solving specific problems, but may not be applicable to other domains. AI foundation model usually involve mega-scale parametere, while customized model involves large or middle-scale parameters.

Also, AI large model contains two key phases: training and inference. Training refers to the process of developing an AI model by feeding it with large amounts of data and optimizing it to learn and improve its performance. Training has high demand on computing and memory resource. On the other hand, inference is the process of using the trained AI model to make predictions or decisions based on new input data. Inference focuses more on the balance between computing resource, latency and power cost.

There are mainly four types of AI tasks:

*Text: text-to-text (conversation), text classification (e.g. sentiment analysis)

*Vision: image classification (label images), object detection.

*Audio: speech-to-text, text-to-speech

*Multimodal: text-to-image, image-to-text, text-to-video, image-to-image, image-to-video, etc.

Vison, audio, multimodal tasks often bring on high demand on network resource and computing resource.

There are two AI large model deployment cases that will benefit from the dynamic selection of service instances and the traffic steering.

Figure 1 shows the Cloud-edge co-inference AI model deployment. It can achieve low latency as the AI inference is deployed near to device. And it requires low demand on device resources. But when handling AI inference tasks, if traffic load between device and edge is high or edge computing resource is overloaded, traffic steering is needed to ensure the QoS.

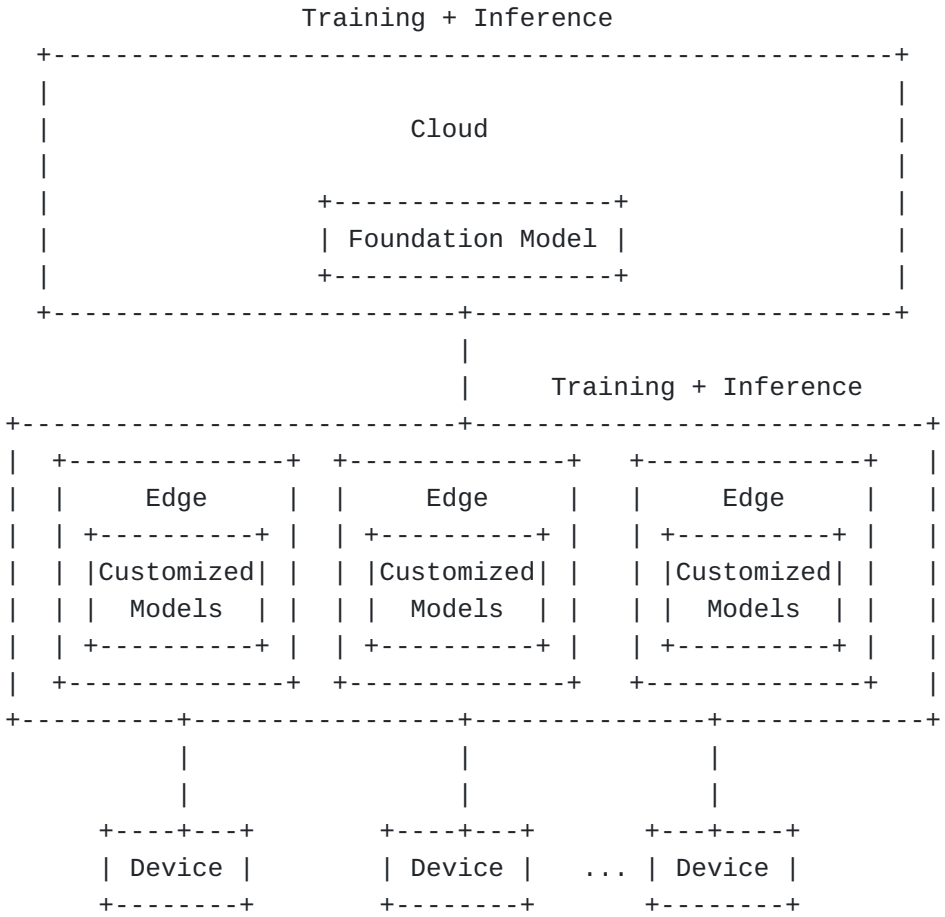

Figure 1: Cloud-edge co-inference

Figure 2 shows the Cloud-edge-device co-inference AI model deployment. It is a more flexible deployment (also more complex). It can achieve low latency as the AI inference is deployed locally or near to device. And device can work when edge isn't available.

Careful consideration to ensure that edge will only be used when the trade-offs are right. Similar to Cloud-edge co-inference AI model deployment, traffic steering is needed.

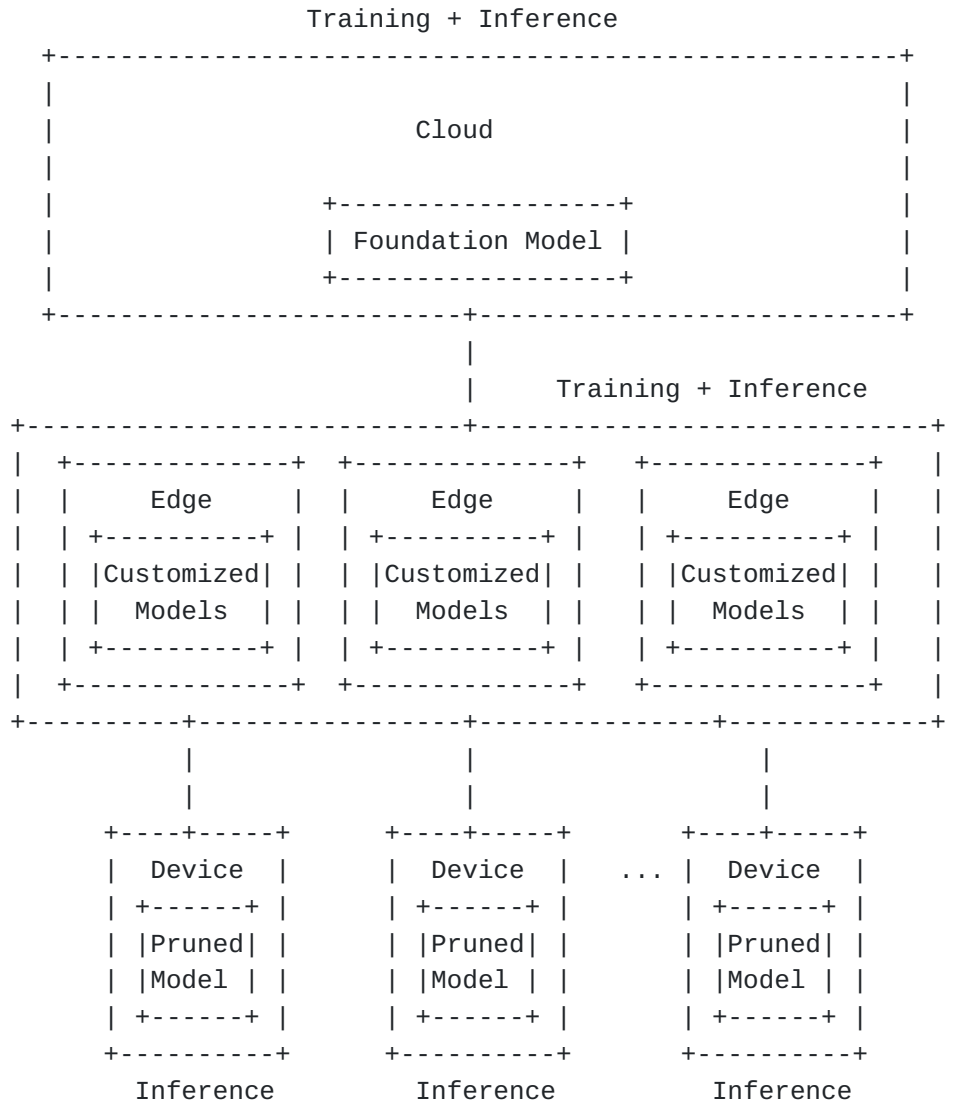

Figure 2: Cloud-edge-device co-inference

Many AI tasks brings on high demand on network resource and computing resource: vison, audio, multimodal. Also, it is common that same customized model is deployed in multiple edge sites to achieve load balance and high reliability.

The edge site’s computing resource and network info should be collectively considered to make suitable traffic steering decision. For example, if the available computing resource in nearest edge site is low, the traffic of AI tasks should be steered to another

edge with high resource. Also, if multiple AI tasks, delay-sensitive task (live streaming with AI-generated avatar) and delay-tolerant task (text-to-image) arrive in edge, delay-tolerate task should be steered to another edge if the nearest edge's resource is limited.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. IANA Considerations

This document makes no request of IANA.

4. Security Considerations

TBD

5. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Author's Address

Qing An
Alibaba Group
China

Email: anqing.aq@alibaba-inc.com