ABSTRACT

Harun Arif. 52419757
PENGGUNAAN TRITON UNTUK GENERATIVE AI BERBASIS LARGE LANGU-AGE MODEL.
Skripsi, Fakultas Teknologi Industri, Jurusan Informatika, Universitas Gunadarma, 2023.

This research addresses the use of Triton Inference Server to manage and perform inference on the ResNet50 model, a Convolutional Neural Network (CNN) in the context of the development of Generative AI and Large Language Models. The main objective of this research is to design the deployment process and perform inference on the ResNet50 model using Triton Inference Server. This research aims to generate classification scores from the ResNet50 model for images with 1000 predictions, in the format of confidence_score and classification_index. This research method adapts the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a framework for systematically managing data analysis projects. The research used the CRISP-DM method in all stages, from initial understanding to implementation of analysis results. The results showed that the Triton Inference Server was very helpful in improving the efficiency and effectiveness of the ResNet50 model inference process, especially when running it on GPUs. The inference process of 1000 predictions on 10 images took an average time of only 0.029 seconds. This research makes an important contribution to the management of large models in AI, especially in image processing.

Keywords : Triton Inference Server, ResNet50, Inferensi Model, Confidence Score, Classification Index, Cross-Industry Standard Process for Data Mining (CRISP-DM)

(x + 47 + attachment)

Reference (2006-2023)