Claims

[Claim 1]

A processing system performed by using an edge device and a server device, wherein

the edge device includes:

a first inference unit configured to extract a feature amount of processing target data by using a first model and execute inference processing on the processing target data on a basis of the extracted feature amount; and

a determination unit configured to output an inference result by the first inference unit in a case where reliability of the inference result by the first inference unit exceeds a threshold, and output the feature amount of the processing target data to the server device in a case where the reliability is equal to or less than the threshold, and

the server device includes:

a second inference unit configured to execute inference processing on the processing target data on the basis of the feature amount of the processing target data output from the edge device by using a second model having higher inference accuracy than the first model.

[Claim 2]

The processing system according to claim 1, wherein

the first model includes

a first extraction unit configured to extract the

feature amount of the processing target data, and

a first execution unit configured to execute first inference processing on the basis of the feature amount of the processing target data, and

the second model includes

a second execution unit configured to execute second inference processing on the basis of the feature amount of the processing target data, and

the second execution unit executes the second inference processing on the basis of the feature amount of the processing target data extracted by the first extraction unit.

[Claim 3]

The processing system according to claim 1 or 2, wherein the determination unit outputs the quantized feature amount of the processing target data to the server device.

[Claim 4]

The processing system according to any one of claims 1 to 3, wherein the reliability is based on entropy of the inference result by the first inference unit.

[Claim 5]

A processing method executed by a processing system performed by using an edge device and a server device,

the processing method comprising:

a first inference step in which the edge device

extracts a feature amount of processing target data by using a first model and executes inference processing on the processing target data on a basis of the extracted feature amount;

a determination step in which the edge device outputs an inference result in the first inference step in a case where reliability of the inference result in the first inference step exceeds a threshold, and outputs the feature amount of the processing target data to the server device in a case where the reliability is equal to or less than the threshold; and

a second inference step in which the server device executes inference processing on the processing target data on the basis of the feature amount of the processing target data output from the edge device by using a second model having higher inference accuracy than the first model.

[Claim 6]

A computer-readable recording medium storing therein a processing program that causes a computer to execute a process comprising:

a computer as an edge device is caused to execute:

a first inference step of extracting a feature amount of processing target data by using a first model and executing inference processing on the processing target

data on a basis of the extracted feature amount; and

a determination step of outputting an inference result in the first inference step in a case where reliability of the inference result in the first inference step exceeds a threshold, and outputting the feature amount of the processing target data in a case where the reliability is equal to or less than the threshold, and

a computer as a server device is caused to execute:

a second inference step of executing inference processing on the processing target data on the basis of the feature amount of the processing target data output from the edge device by using a second model having higher inference accuracy than the first model.