Dynamic Model – Graph Dispatcher

AWS AI

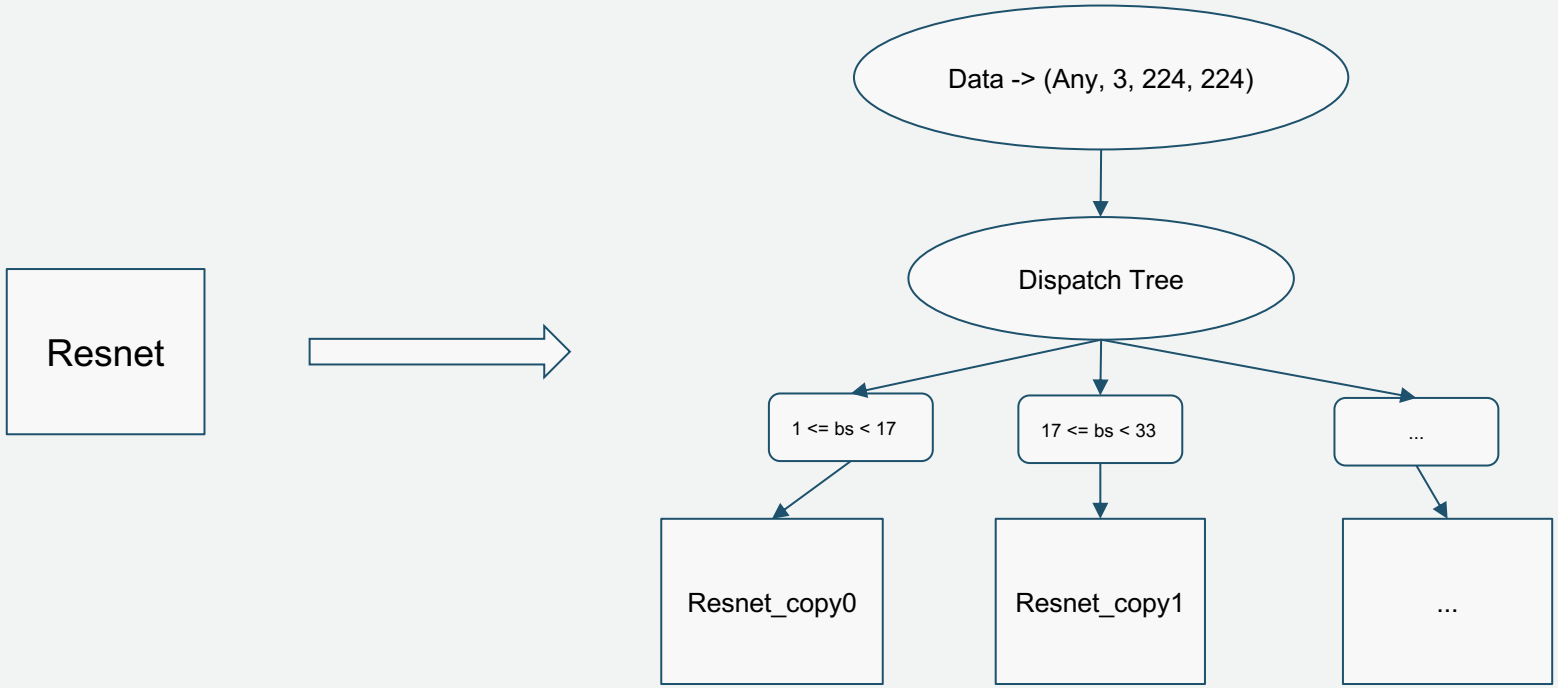
Presenter: Yao Wang
Amazon SageMaker Neo

Challenge in Dynamic Code Generation

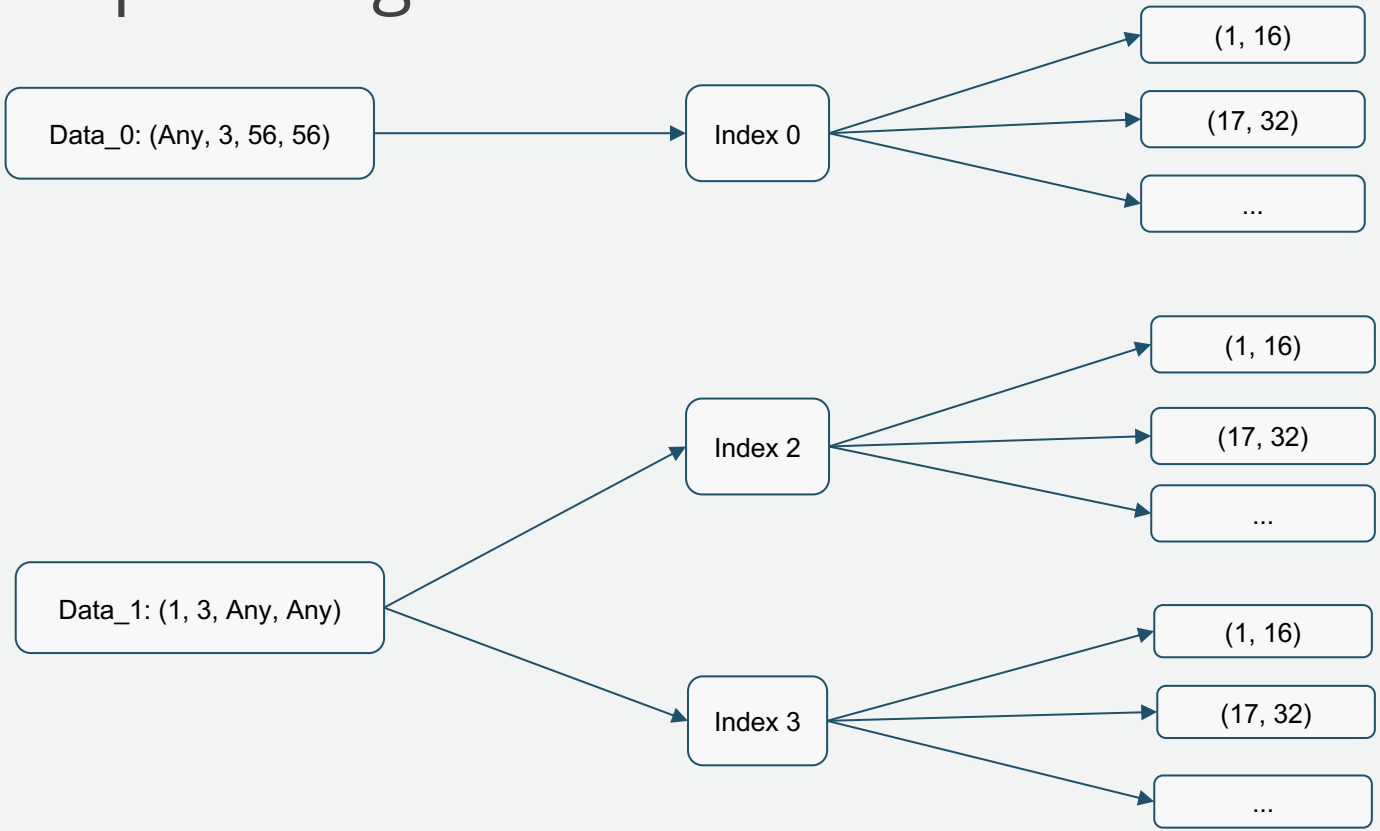
1. Extra overhead for kernel dispatching: dispatching operation is required for **every** kernel.
2. Requires runtime layout tracking system for operator involving layout transformation, such as conv2d_NCHWc.
3. Difficult to apply graph tuning, which can cause performance loss.

To avoid these disadvantages, we introduce graph dispatching strategy.

Dispatch a Whole Graph



Dispatching Function



Acknowledgement

Thank you!