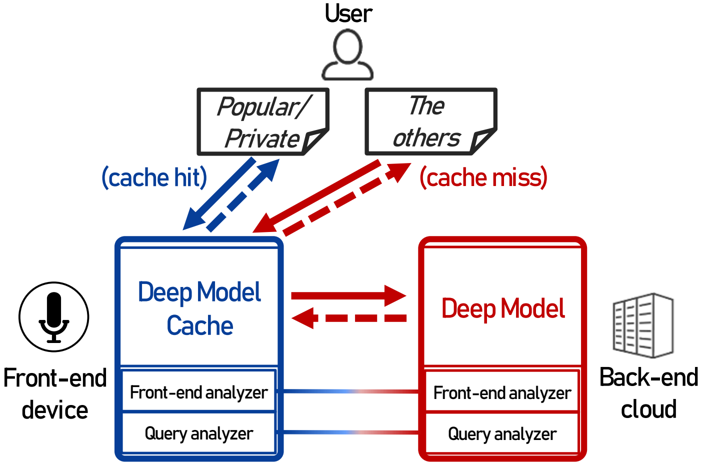
Online Adjustment of Two-stage Inference for Knowledge Caching

Geonha Park and Changho Hwang
Electrical Engineering, KAIST, South Korea

- Knowledge caching
  - Cache popular or private classes of a large deep model (DM) in local devices
- Motivations of online adjustment
  - Varying query popularity increases cache miss rate and response latency
  - Need to adapt DM cache depends on the popularity
- System components for online adjustment
  - Front-end analyzer decides DM cache size based on hardware constraints
  - Query analyzer decides whether to update DM cache



Overview of knowledge caching and online adjustment