Cloud-based or On-device:

An Empirical Study of Mobile Deep Inference

Tian Guo

WPI

mobile deep inference

Real-time translation

Image recognition

Personal assistant

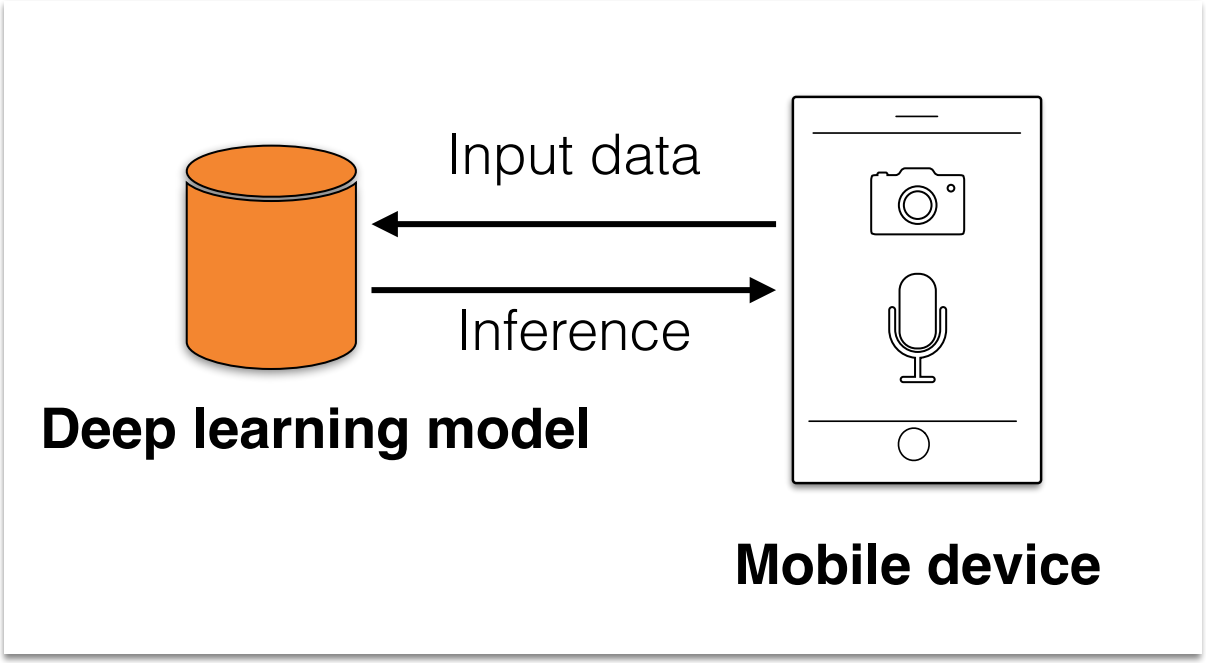
tian@cs.wpi.edu

mobile deep inference

Real-time translation

Image recognition

Personal assistant



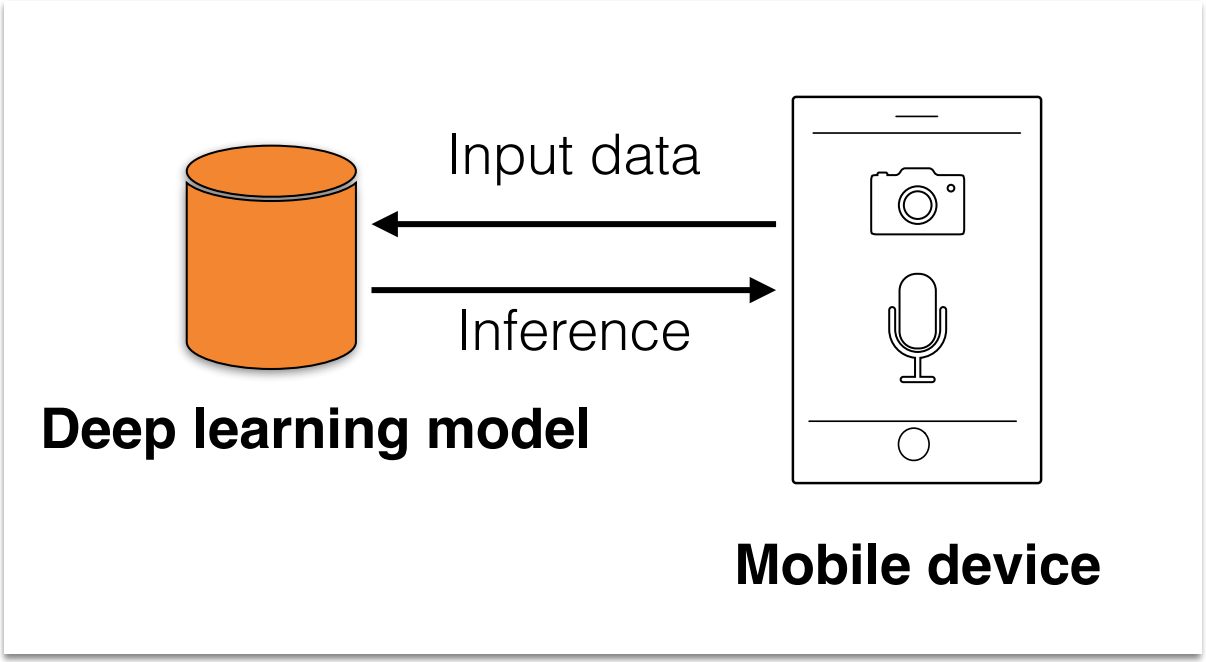
tian@cs.wpi.edu

mobile deep inference

Real-time translation

Image recognition

Personal assistant

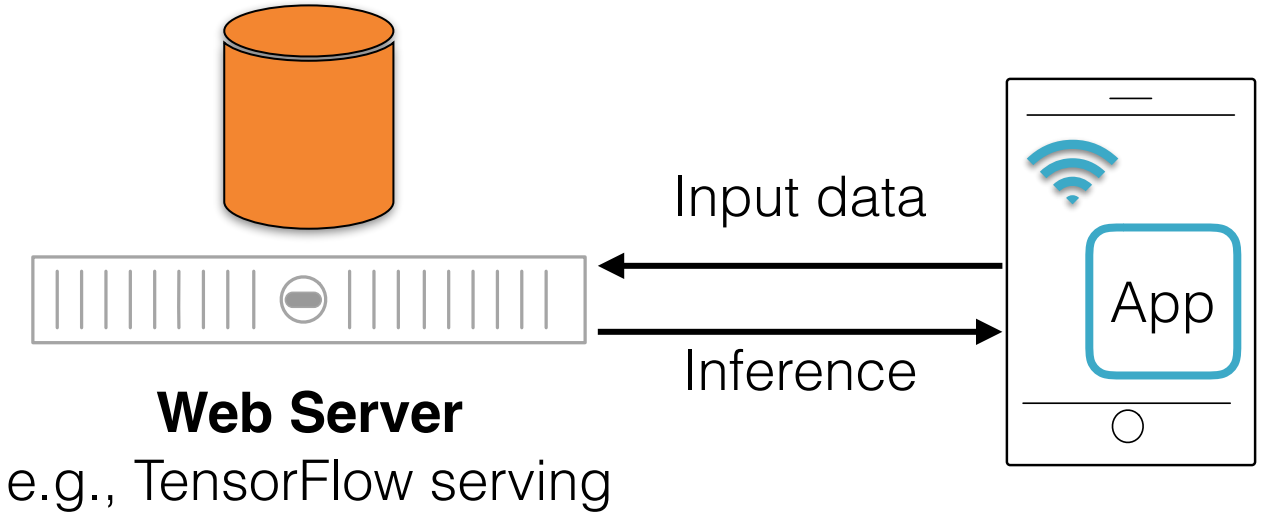


Executing inference tasks on deep learning models for mobile applications.

tian@cs.wpi.edu

cloud-based vs. on-device

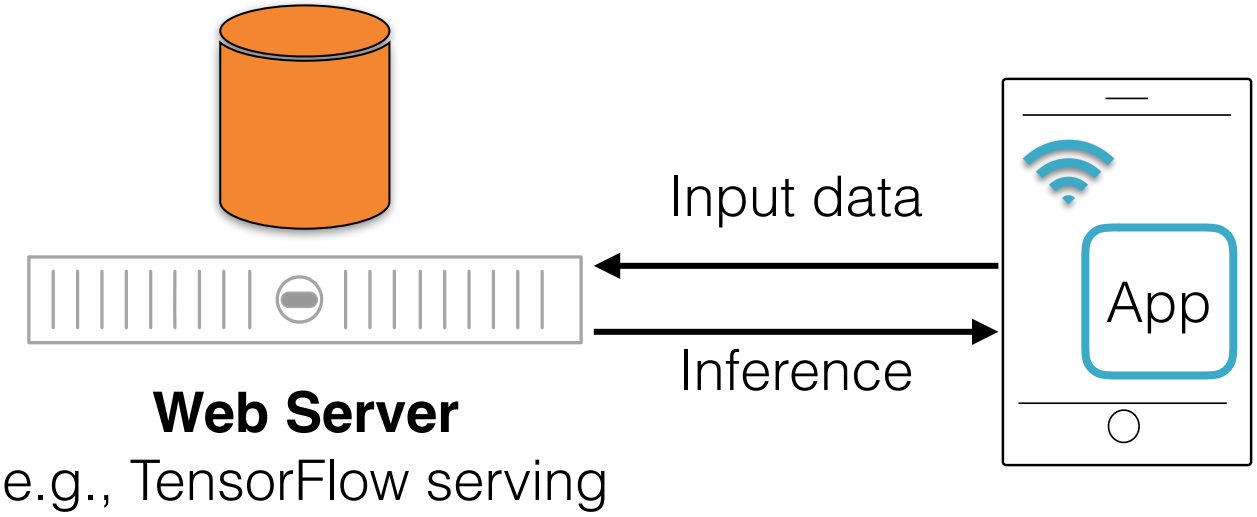
Cloud-based deep inference



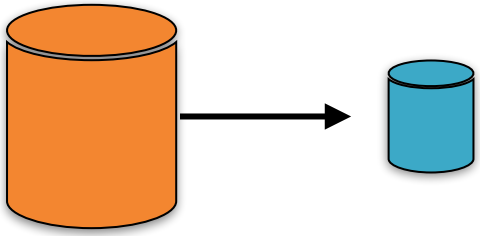
tian@cs.wpi.edu

cloud-based vs. on-device

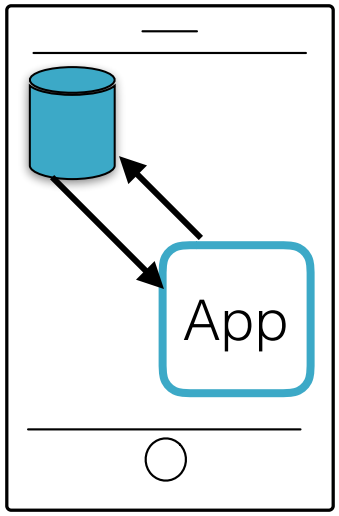
Cloud-based deep inference



On-device deep inference



Model optimizations
e.g., Deep Compression



tian@cs.wpi.edu

$ Which inference mode is **better** for mobile deep learning applications?

tian@cs.wpi.edu

measurement methodology

> Android benchmark app

- Caffe-based deep learning environments: CaffeLib, CNNDroid

> Deep learning models

- AlexNet
- NIN
- SqueezeNet

Similar top-5 error rates
Different model sizes

> Setup

- On-device: a late-2013 mobile device on university WiFi
- Cloud-based: GPU instances hosted in Amazon Virginia

tian@cs.wpi.edu

measurement methodology

> Android benchmark app

- Two deep learning frameworks: CaffeLib, CNNDroid

> Deep learning models

- AlexNet
- NIN
- SqueezeNet

> Setup

- On-device: a late-2013 mobile device on university WiFi
- Cloud-based: GPU instances hosted in Amazon Virginia

~20ms RTT, well provisioned
good mobile network condition

tian@cs.wpi.edu

inference time comparison

inference mode
Cloud+ CPU
Cloud+ GPU
Device+ CaffeLib
Device+ CNNDroid

tian@cs.wpi.edu

inference time comparison

inference mode	load model
Cloud+ CPU	0
Cloud+ GPU	0
Device+ CaffeLib	2422ms
Device+ CNNDroid	61256ms

Models are preloaded

tian@cs.wpi.edu

inference time comparison

inference mode	load model
Cloud+ CPU	0
Cloud+ GPU	0
Device+ CaffeLib	2422ms
Device+ CNNDroid	61256ms

Models are preloaded

Average across three models and all inference tasks

Impact factors:
Model size
Mobile capacity

tian@cs.wpi.edu

inference time comparison

scale bitmap: ~76ms

inference mode	load model	upload bitmap
Cloud+ CPU	0	37ms
Cloud+ GPU	0	37ms
Device+ CaffeLib	2422ms	0
Device+ CNNDroid	61256ms	0

Impact factors:
network condition
input data size

tian@cs.wpi.edu

inference time comparison

inference mode	load model	upload bitmap	compute probability
Cloud+ CPU	0	37ms	239ms
Cloud+ GPU	0	37ms	19ms
Device+ CaffeLib	2422ms	0	8911ms
Device+ CNNDroid	61256ms	0	2132ms

Impact factors:
Server resource
inference load

Impact factors:
Mobile capacity
Model complexity

tian@cs.wpi.edu

inference time comparison

inference mode	load model	upload bitmap	compute probability	inference
Cloud+ CPU	0	37ms	239ms	352ms
Cloud+ GPU	0	37ms	19ms	132ms
Device+ CaffeLib	2422ms	0	8911ms	11413ms
Device+ CNNDroid	61256ms	0	2132ms	63458ms

tian@cs.wpi.edu

inference time comparison

Cloud-based deep inference is up to **67 times** faster than performing on-device.

inference mode	load model	upload bitmap	compute probability	inference
Cloud+ CPU	0	37ms	239ms	352ms
Cloud+ GPU	0	37ms	19ms	132ms
Device+ CaffeLib	2422ms	0	8911ms	11413ms
Device+ CNNDroid	61256ms	0	2132ms	63458ms

tian@cs.wpi.edu

takeaways

> Deep learning powered mobile applications are gaining huge popularity

> Cloud-based vs. on-device inferences are complementary with different impact factors

- Server location, server resource, inference load

- Mobile network, mobile resource, model complexity

> Beneficial to **dynamically switching** between two inference modes

tian@cs.wpi.edu

$ Questions?

tian@cs.wpi.edu