AdaCompress: Adaptive Compression for Online Computer Vision Services

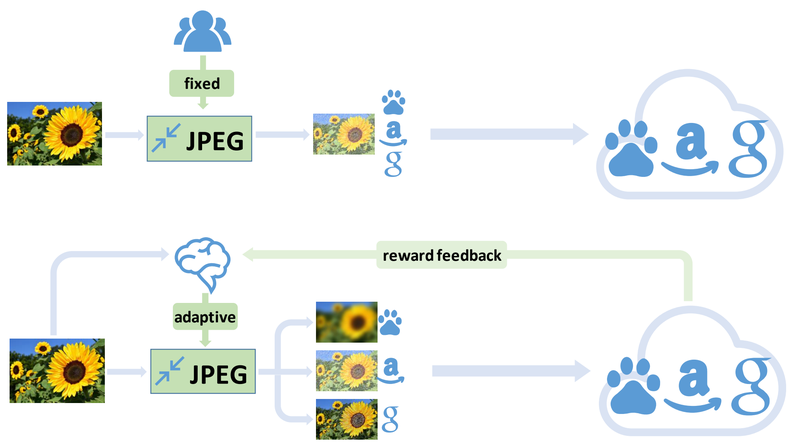
Hongshan Li[1], Yu Guo[2], Zhi Wang[2], Shutao Xia[2], Wenwu Zhu[1]

[1] Tsinghua-Berkeley Shenzhen Institue, Tsinghua University, Shenzhen,China

[2] Guaduate School in Shenzhen, Tsinghua University, Shenzhen, China

Program ID： P5C-05

1 Background: More images are viewed by DL services rather than human



Conventional image upload framework

- fixed compression degree for all images
- same compression strategy for all services

Conventional compression solutions are for human vision systems

- visual similarity is not necessarily related to model prediction accuracy
- mismatch in current compression framework and computer vision tasks

- **Problem Definition:** An adaptive image compression solution aims at minimizing data size while ensuring that the prediction accuracy is close to that of inference from the original images. An agent iteratively interact with the cloud services to learn an optimal compression strategy.

2 Challenges & Related Works

Challenges

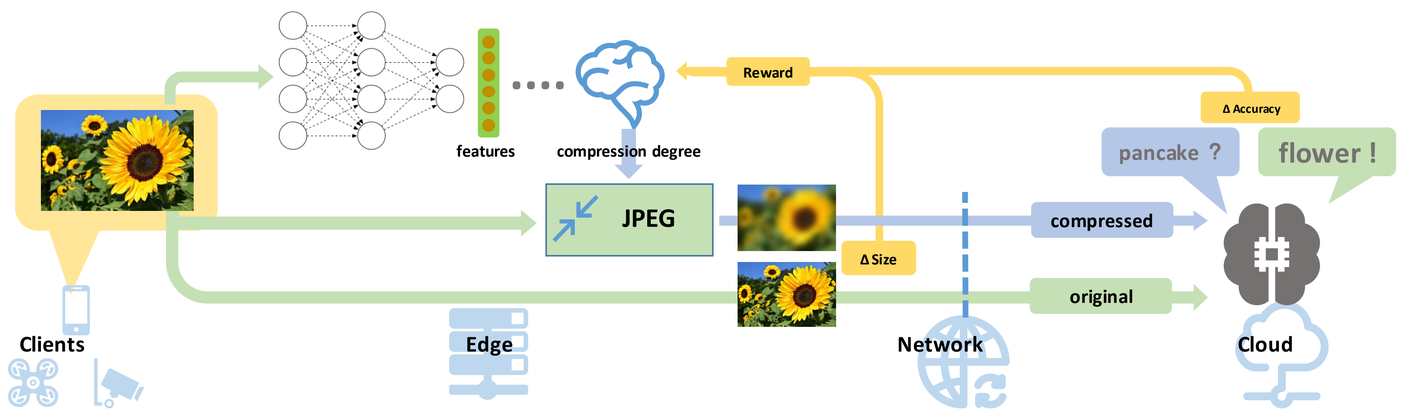
- Uncertain DNN prediction behaviors for visually similar images
- Mismatch between compression and prediction tasks, human visually optimized images to DNN prediction tasks
- Online computer vision services are black boxes for clients

Related researches

- Researches about relationship between compression and prediction accuracy
- Inference from compressed representations of auto-encoders
- Current researches need the backend models to generate compression strategies

3 Reinforcement learning based adaptive compression system for online computer vision services

Reinforcement learning framework design



Feature extractor

Extract image features using a pre-trained neural network

Reward feedback

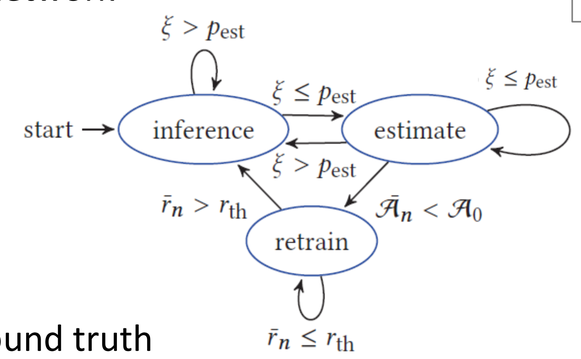
Size difference and accuracy difference feed back

DQN based reinforcement learning agent

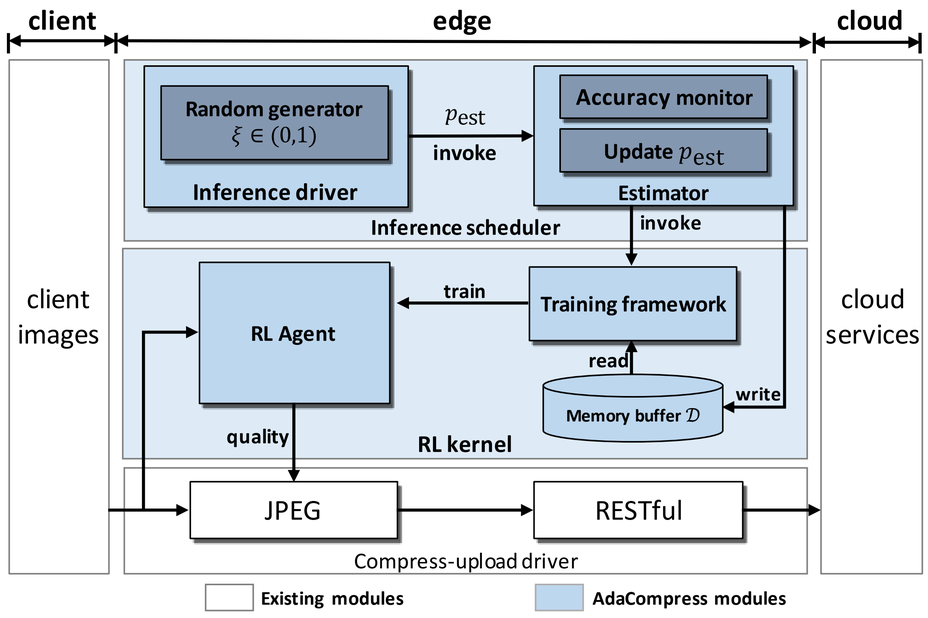
Continuous state space, discrete action space

Gold ground-truth

Using the prediction from the original image as ground truth



Scenery change capturing system design



Inference state

Normal inference, upload compressed images only

Estimate state

Upload origin images to estimate current compression strategy

Retrain state

Once the accuracy is too low, retrain the agent with current images

4 Insights

Grad-Cam importance heat map

Showing which part of the image causes the agent chooses a specific compression level

Acknowledgements

This work is supported in part by NSFC under Grant 61872215, 61531006, 61771273 and U1611461, National Key R&D Program of China under Grant 2018YFB1800204 and 2015CB352300, SZSTI under Grant JCYJ20180306174057899 and JCYJ20180508152204044, and Shenzhen Nanshan District Ling-Hang Team under Grant LHTD20170005.

5 Results

Experiment setup

Benchmark:

JPEG compression, Q=75

Dataset:

Imagenet & night-time images from DNIM

Edge:

Desktop PC with NVIDIA 1080ti

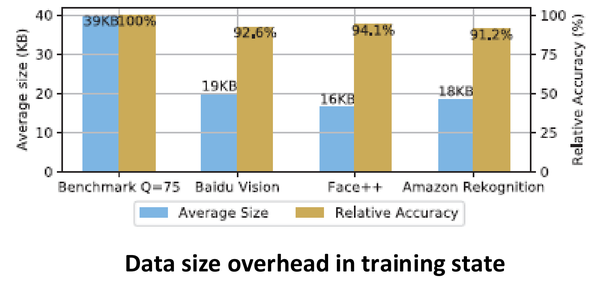
Cloud services:

Baidu Vision, Face++, Amazon Rekognition

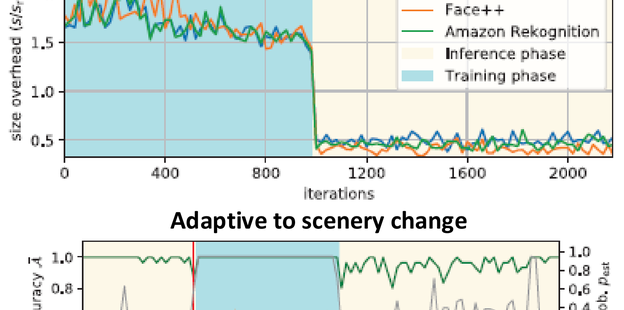
	Benchmark	AdaCompress
Average upload size	42.68 KB	18.46 KB
Inference latency	0 s	2.09 ms
Transmission latency	12.35 ms	5.34 ms
Overall latency	12.35 ms	7.43 ms

40% overall latency reduction (simulation)

50% traffic load with similar relative accuracy



Data size overhead in training state



Adaptive to scenery change

Contact: Hongshan Li, Mail: [redacted] Wechat: h[redacted]782