AdaCompress:

Adaptive Compression for Online Computer Vision Services

Hongshan Li, Yu Guo, Zhi Wang*, Shutao Xia, Wenwu Zhu*

Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

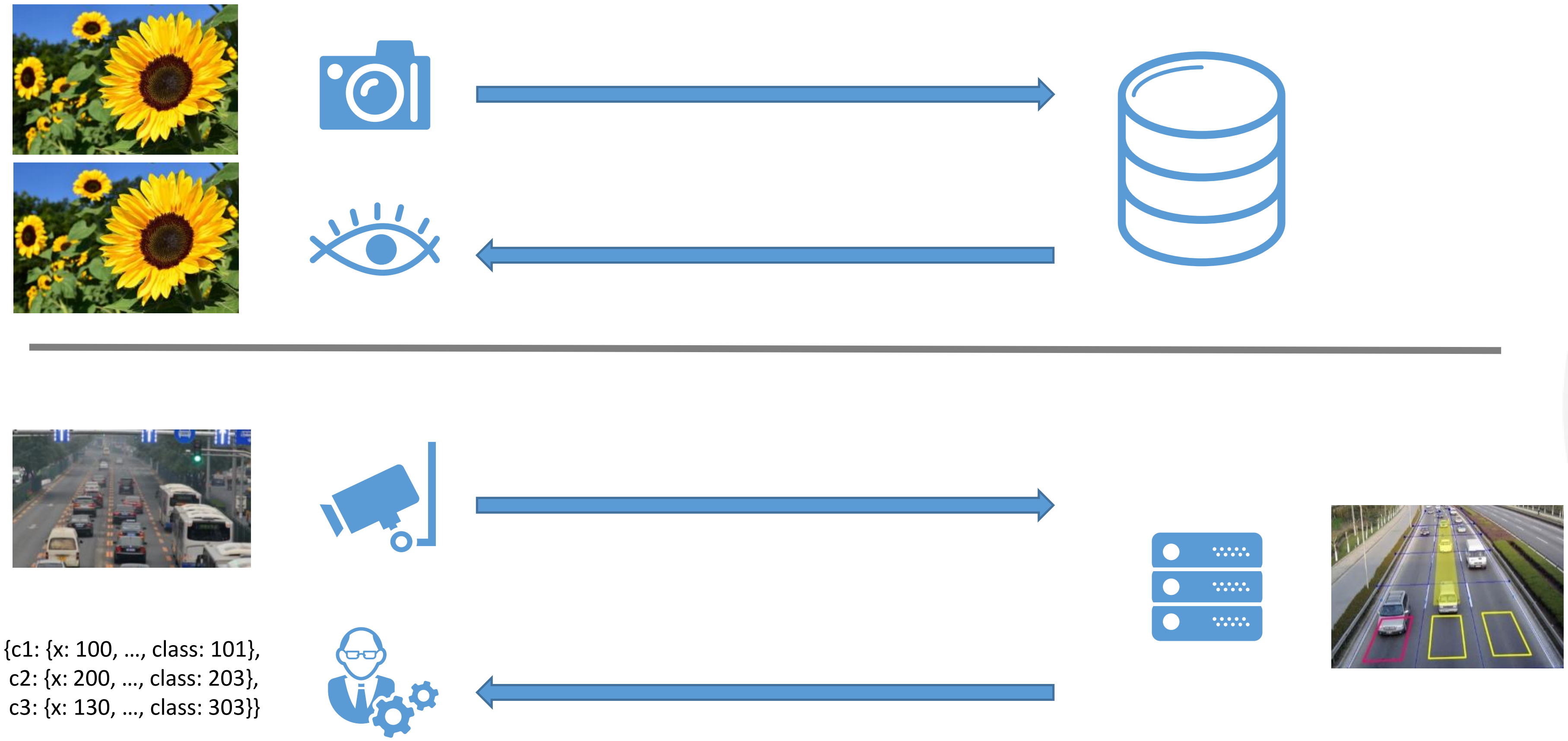
Graduate School in Shenzhen, Tsinghua University

Department of Computer Science and Technology, Tsinghua University

P5C-05

Background

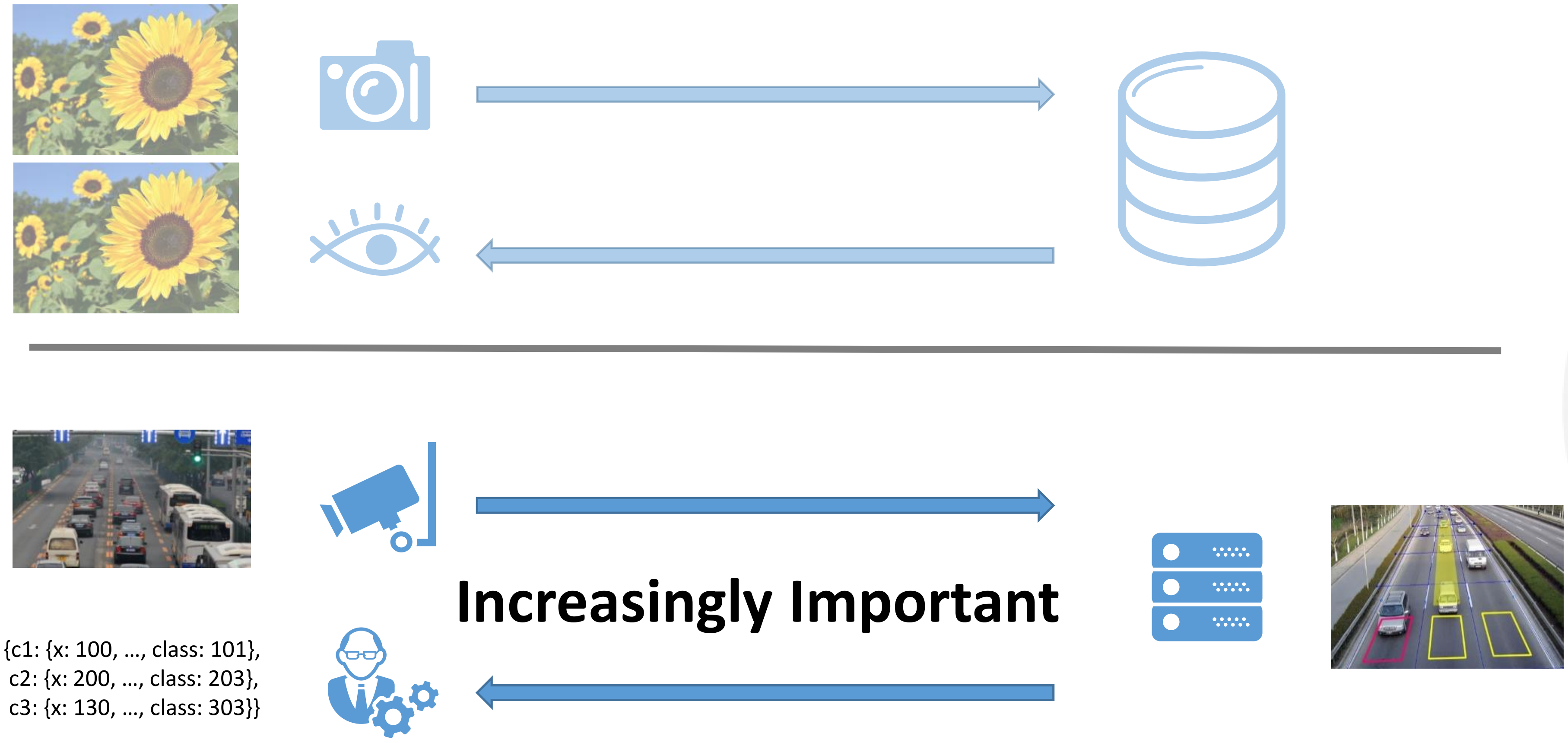
- **More images are uploaded to DL services rather than human**



{c1: {x: 100, …, class: 101},
c2: {x: 200, …, class: 203},
c3: {x: 130, …, class: 303}}

Background

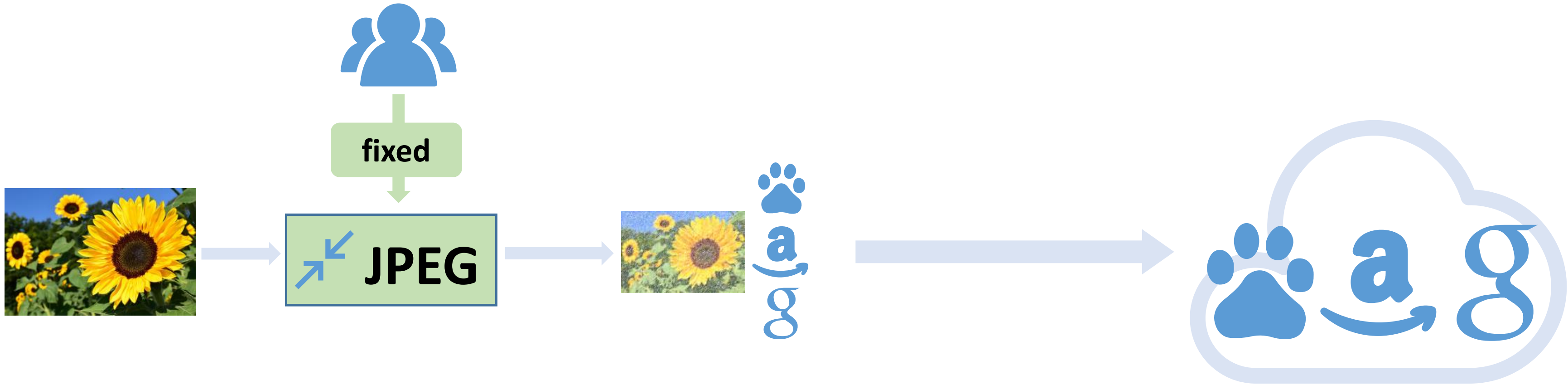
- More images are uploaded to DL services rather than human



Increasingly Important

{c1: {x: 100, …, class: 101},
c2: {x: 200, …, class: 203},
c3: {x: 130, …, class: 303}}

Background



Conventional computer vision application framework

- JPEG etc..
- Fixed compression degree for all images
- Same compression strategy for different service providers

Background

- **Is the conventional solution efficient enough?**

Background

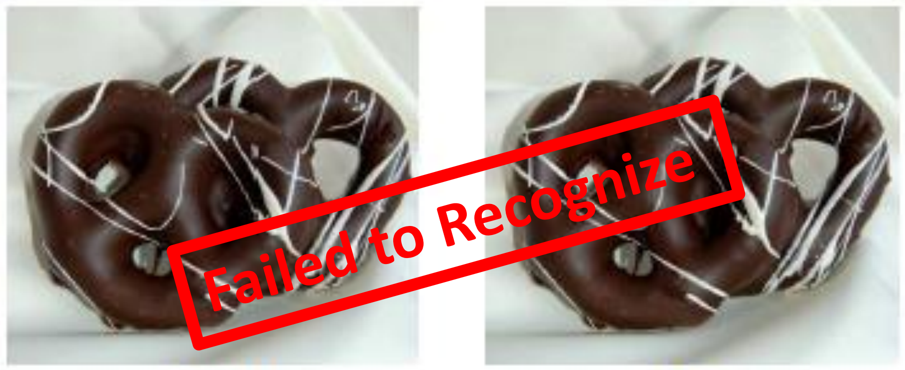
- **Is the conventional solution efficient enough?**

(1a) Q=75
Face++ prediction = ["donut"]

(1b) Q=55
Face++ prediction = []

Background

- Is the conventional solution efficient enough?



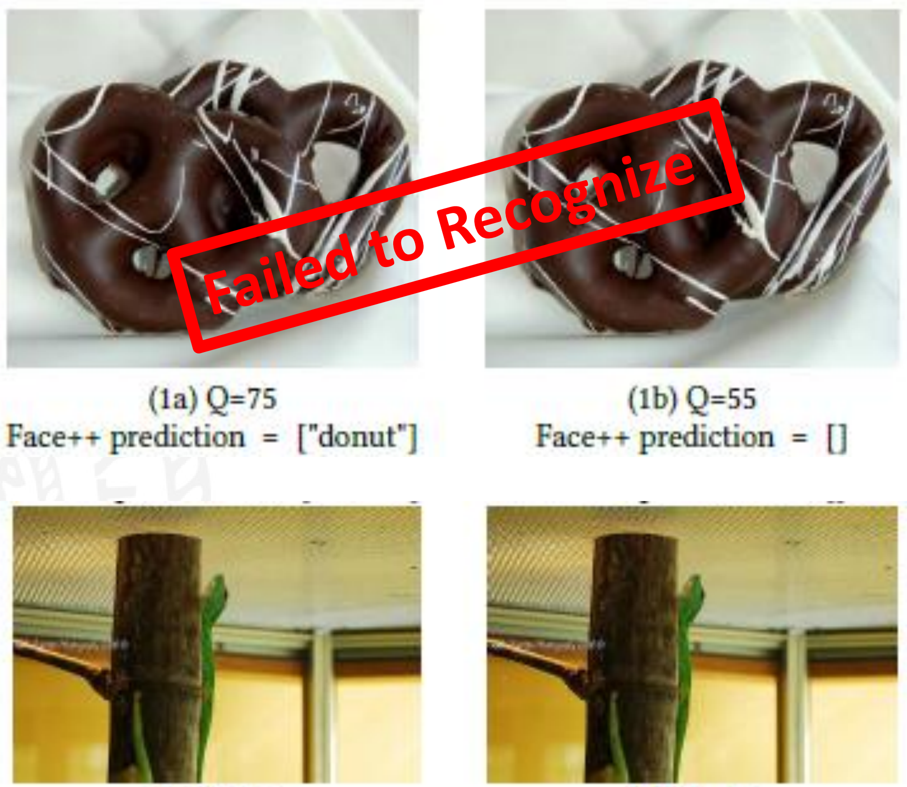
Failed to Recognize

(1a) Q=75
Face++ prediction = ["donut"]

(1b) Q=55
Face++ prediction = []

Background

- **Is the conventional solution efficient enough?**



Failed to Recognize

(1a) Q=75
Face++ prediction = ["donut"]

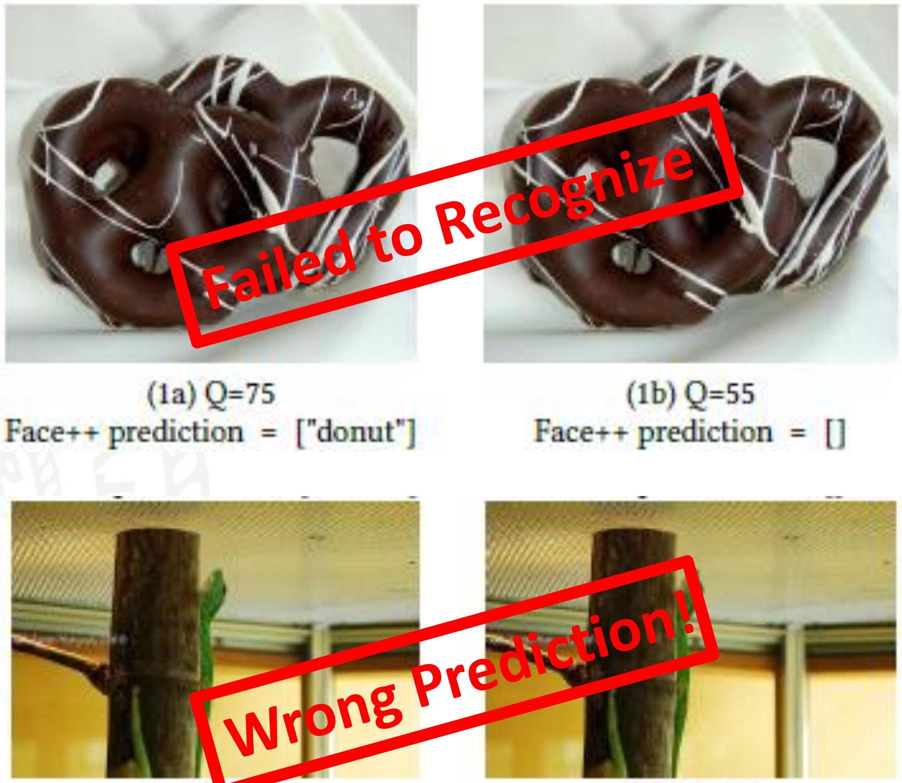
(1b) Q=55
Face++ prediction = []

(2a) Q=75
Baidu prediction = ["chameleon"]

(2b) Q=55
Baidu prediction = ["electric fan"]

Background

- Is the conventional solution efficient enough?



Failed to Recognize

Wrong Prediction!

(1a) Q=75
Face++ prediction = ["donut"]

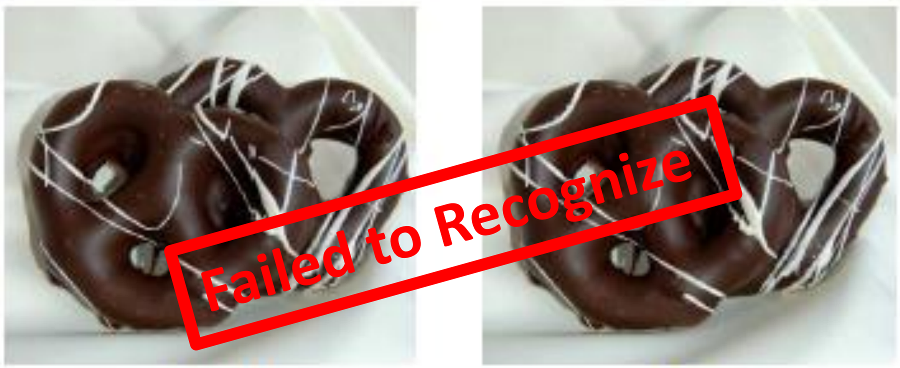
(1b) Q=55
Face++ prediction = []

(2a) Q=75
Baidu prediction = ["chameleon"]

(2b) Q=55
Baidu prediction = ["electric fan"]

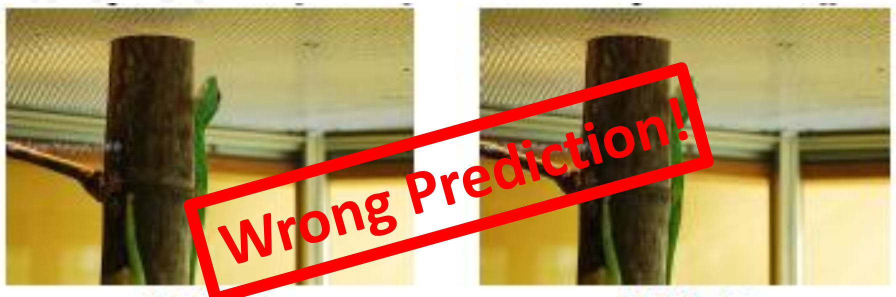
Background

- Is the conventional solution efficient enough?



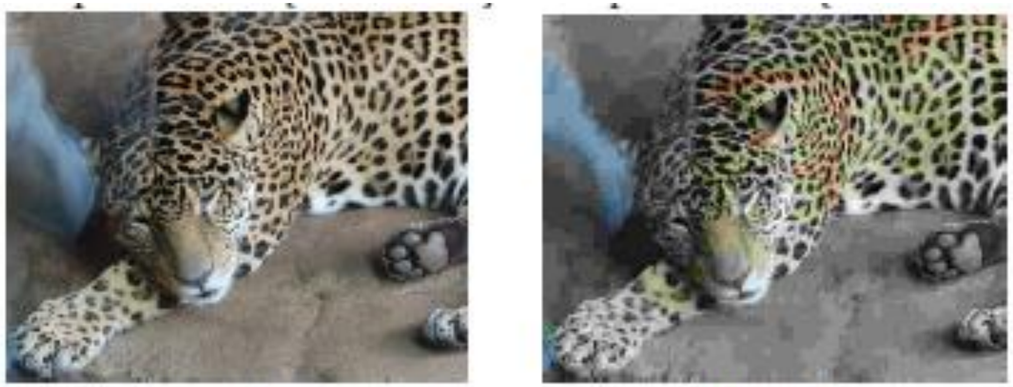
(1a) Q=75
Face++ prediction = ["donut"]

(1b) Q=55
Face++ prediction = []



(2a) Q=75
Baidu prediction = ["chameleon"]

(2b) Q=55
Baidu prediction = ["electric fan"]

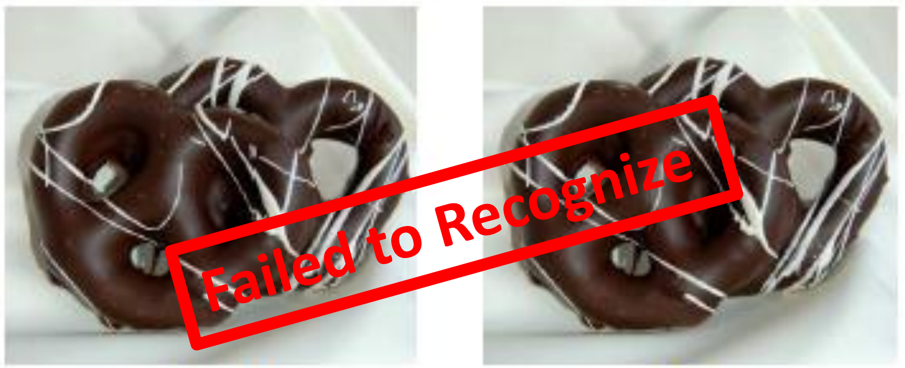


(3a) Q=75
Baidu prediction = ["leopard"]

(3b) Q=5
Baidu prediction = ["leopard"]

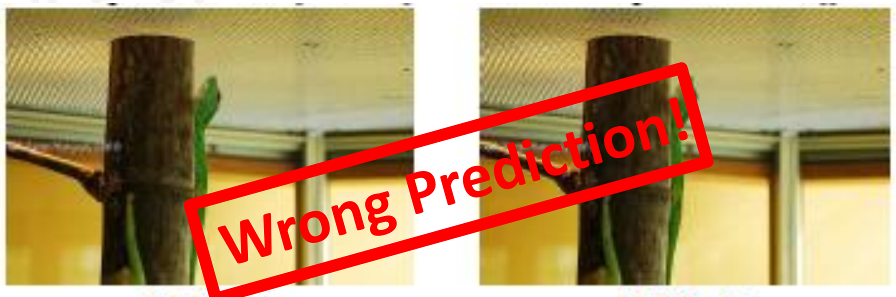
Background

- Is the conventional solution efficient enough?



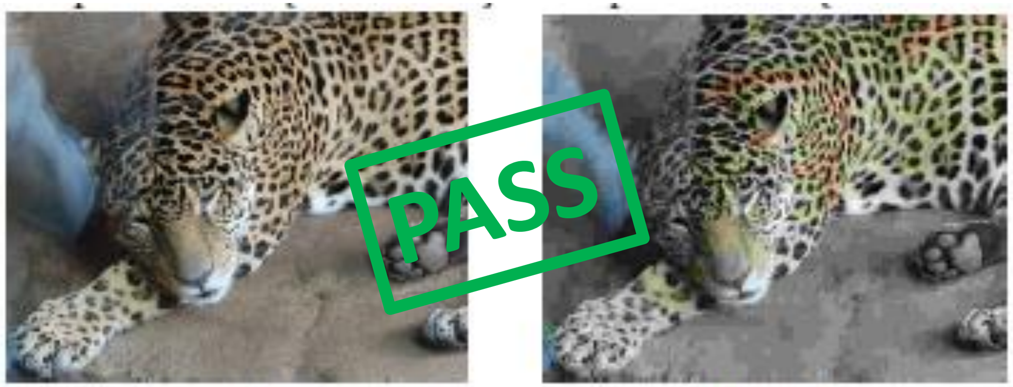
(1a) Q=75
Face++ prediction = ["donut"]

(1b) Q=55
Face++ prediction = []



(2a) Q=75
Baidu prediction = ["chameleon"]

(2b) Q=55
Baidu prediction = ["electric fan"]

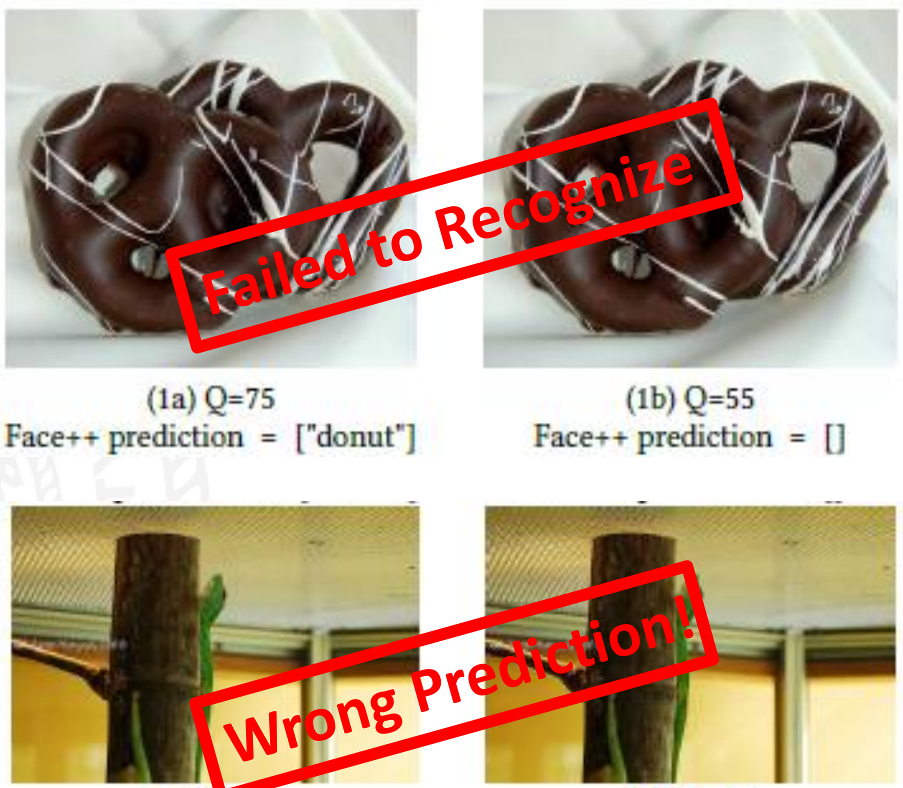


(3a) Q=75
Baidu prediction = ["leopard"]

(3b) Q=5
Baidu prediction = ["leopard"]

Background

- **Is the conventional solution efficient enough?**

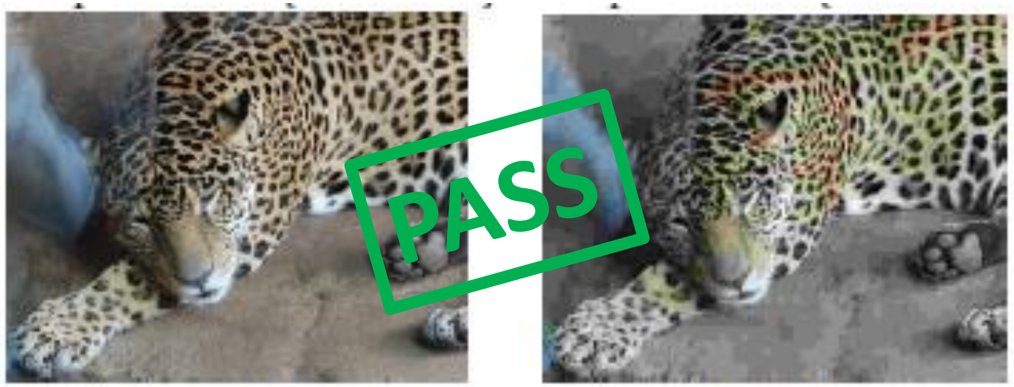


(1a) Q=75
Face++ prediction = ["donut"]

(1b) Q=55
Face++ prediction = []

(2a) Q=75
Baidu prediction = ["chameleon"]

(2b) Q=55
Baidu prediction = ["electric fan"]



(3a) Q=75
Baidu prediction = ["leopard"]

(3b) Q=5
Baidu prediction = ["leopard"]

- **In some cases, prediction performance does not related to image quality**

Related Works

- **Relationship between compression and accuracy**
  - **Severe compression does not always deteriorate the model inference accuracy (Delac et al.)**
  - **Four types of quality distortions can affect the performance in deep learning inference (Dodge et al.)**

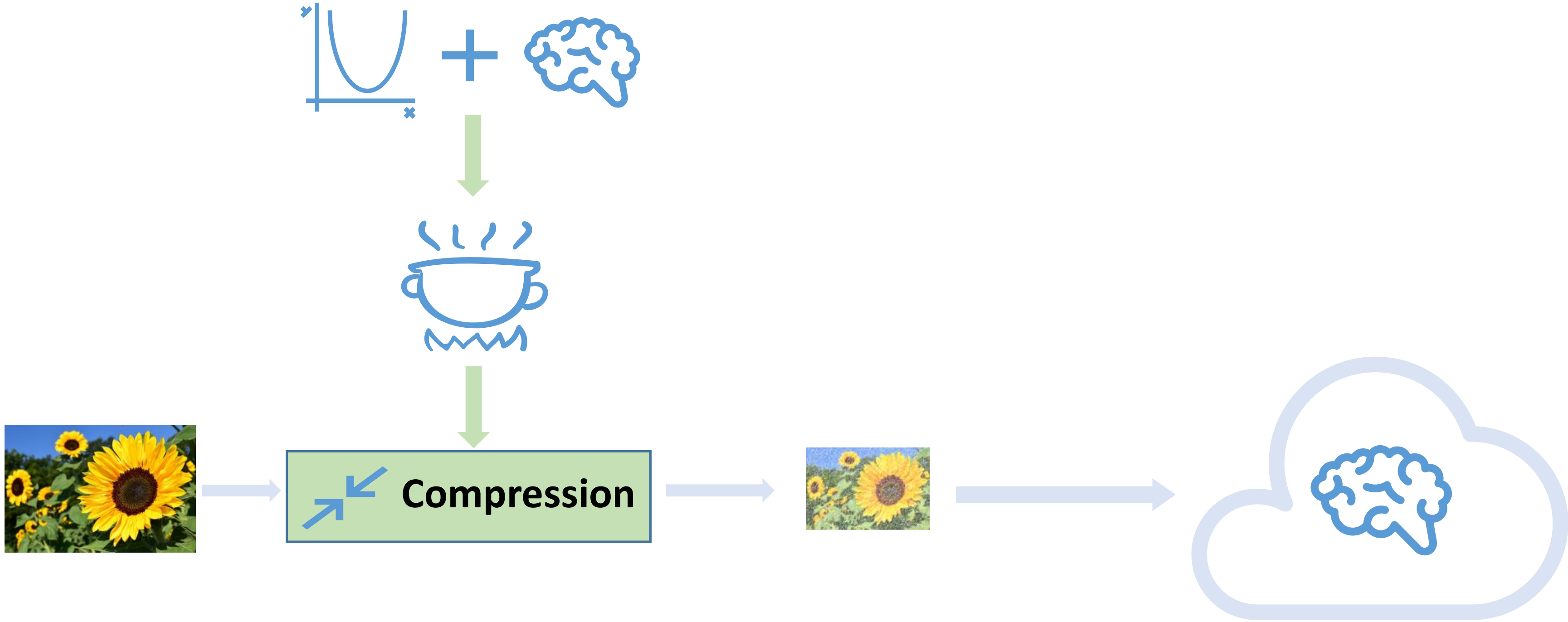
Related Works

- Relationship between compression and accuracy
  - Severe compression does not always deteriorate the model inference accuracy (Delac et al.)
  - Four types of quality distortions can affect the performance in deep learning inference (Dodge et al.)

- Dedicated compression for DNNs
  - Train DNNs from the compressed representations of auto-encoder (Robert et al.)
  - Linear JPEG quantization table learned from the dataset (Liu et al.)
  - DNNs inference from block-wise DCT coefficients in JPEG (Baluja et al.)

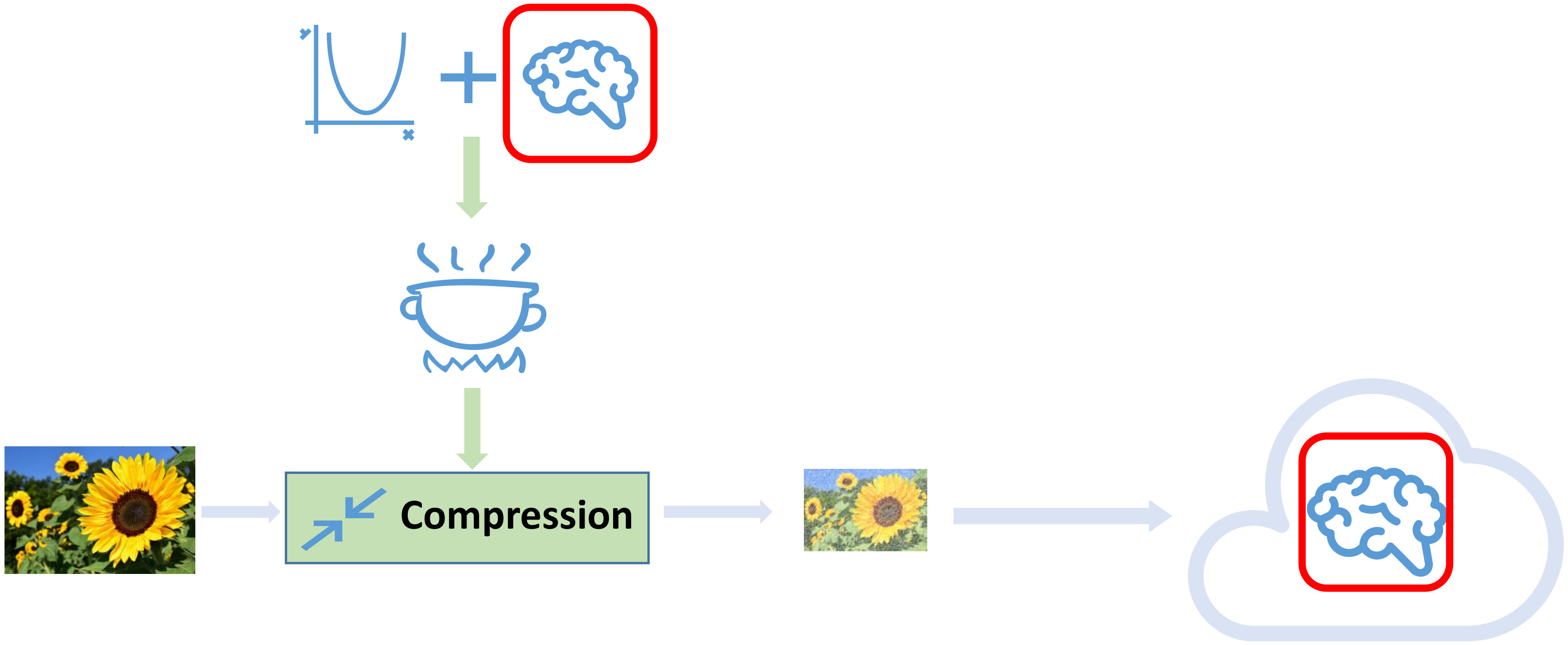
Limitation of related works

- **Pre-knowledge of original model**



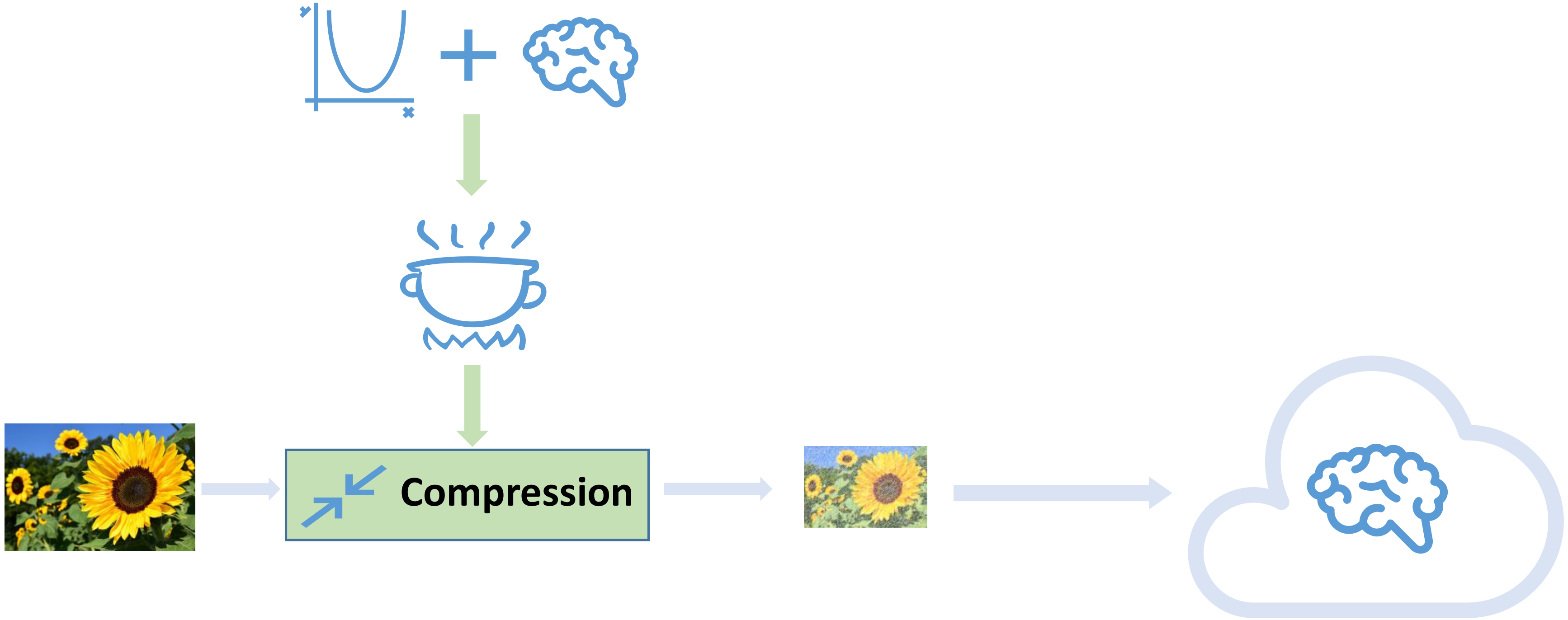
Limitation of related works

- **Pre-knowledge of original model**



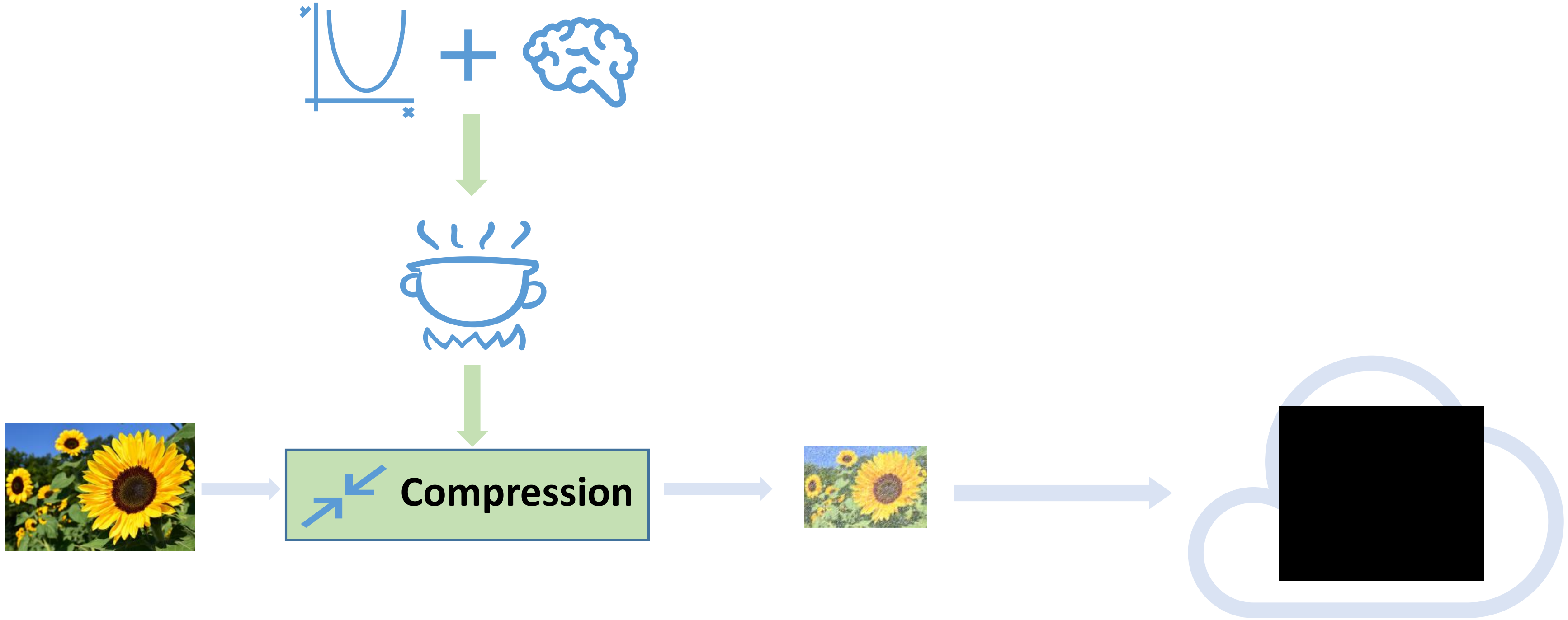
Limitation of related works

- **Pre-knowledge of original model**



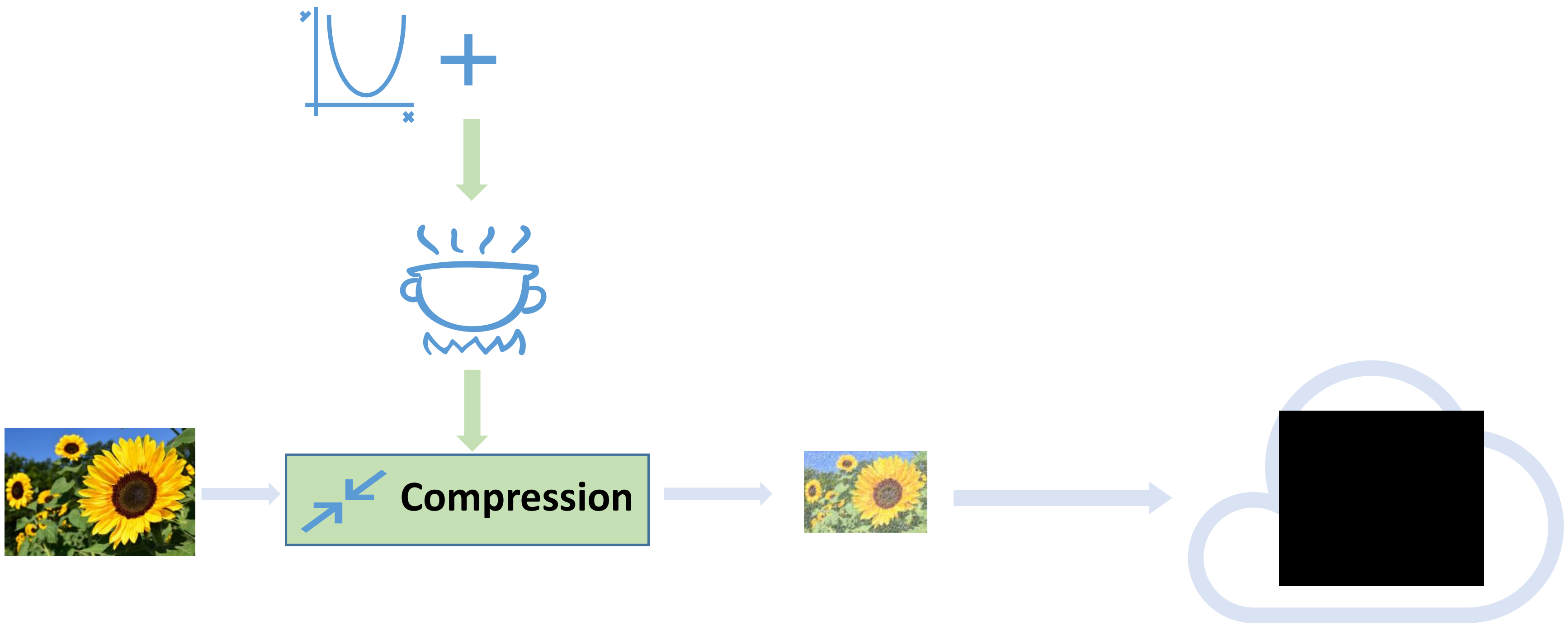
Limitation of related works

- **Pre-knowledge of original model**

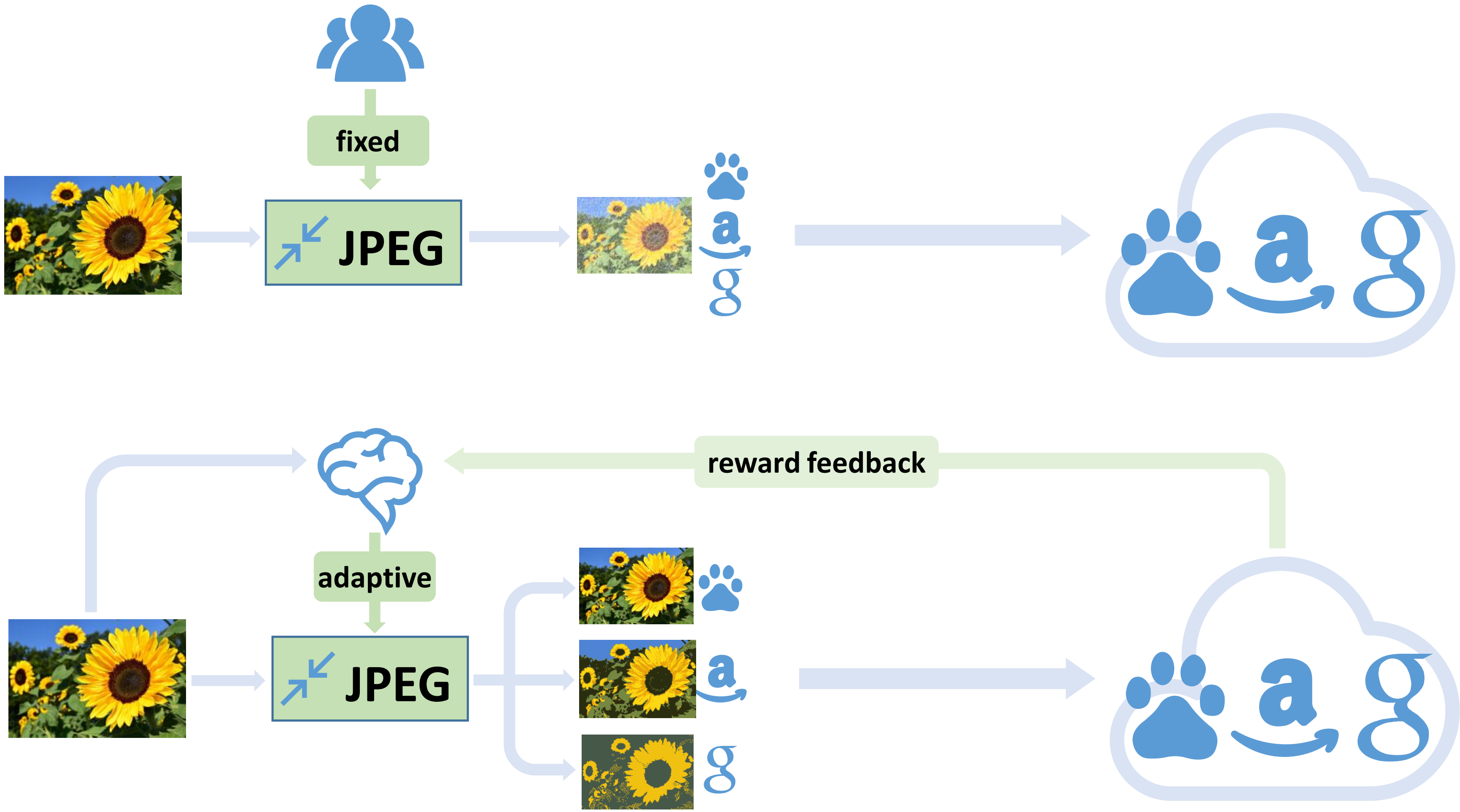


Limitation of related works

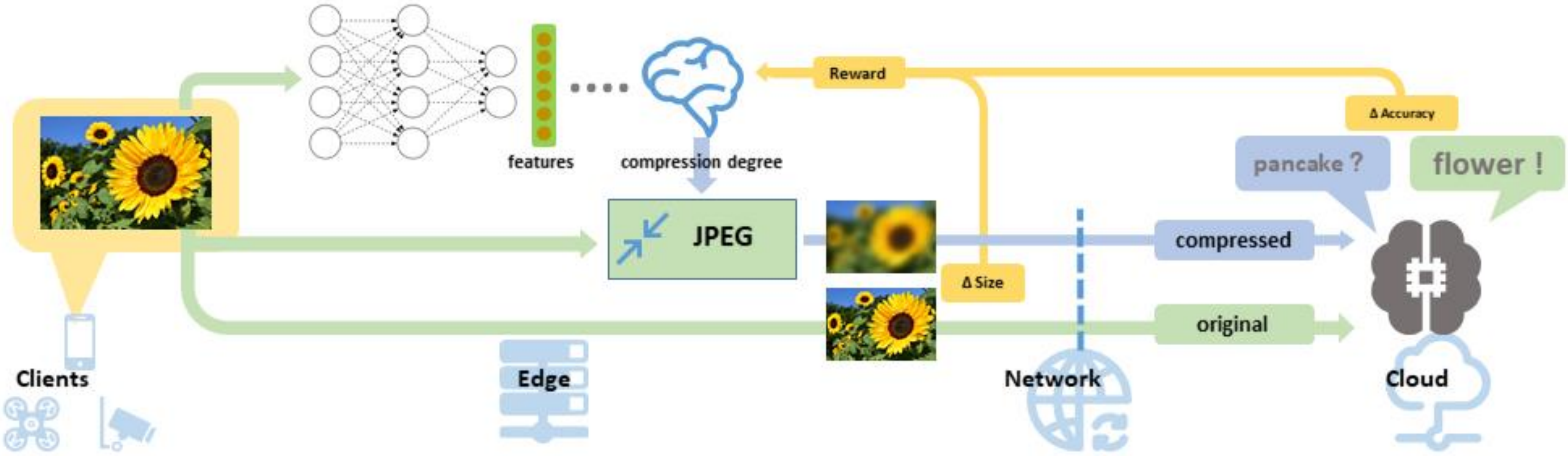
- **Pre-knowledge of original model**



Solution

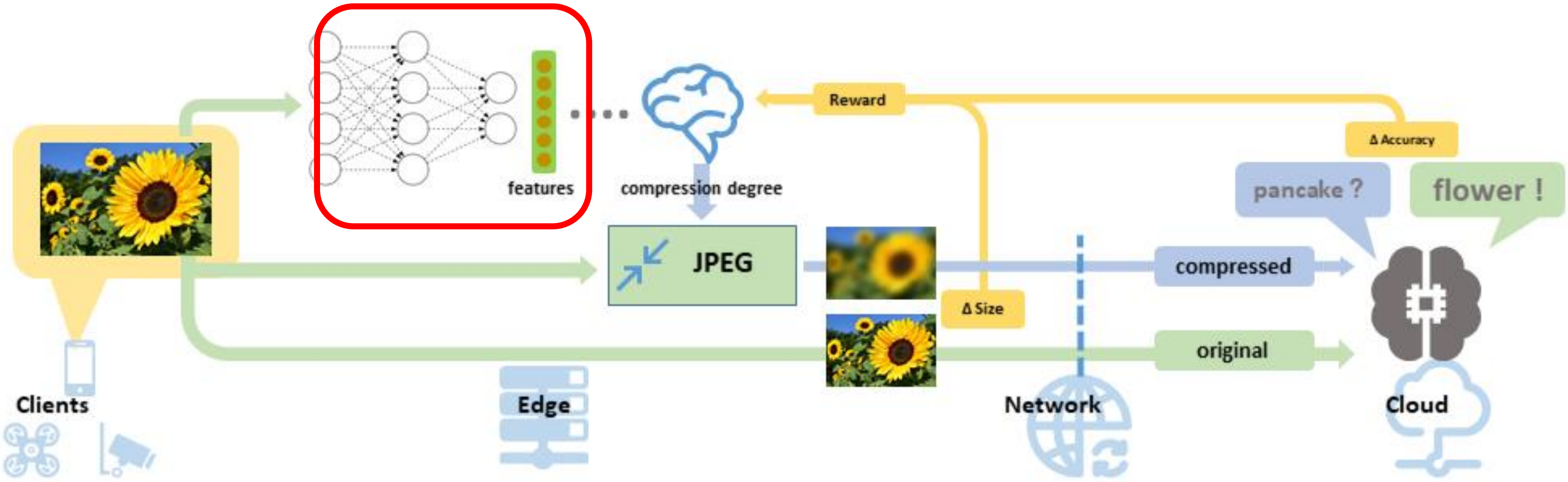


Solution



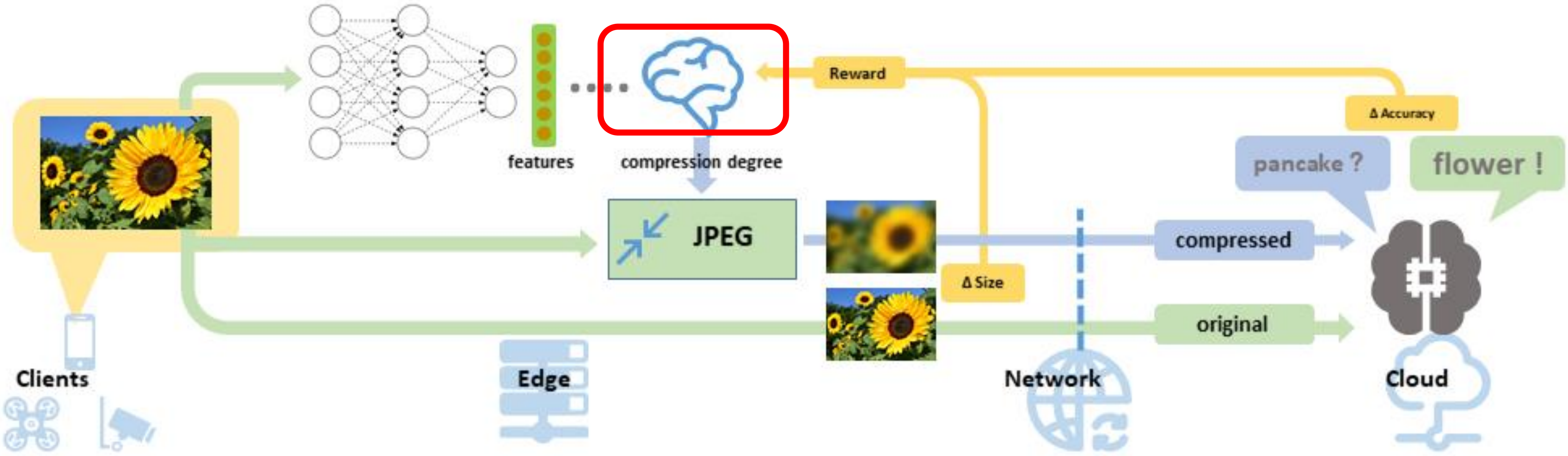
- **Reward**
  - **Δsize – Δaccuracy**
- **States**
  - **Features of the input images**
- **Actions**
  - **10 discrete compression levels**

Solution



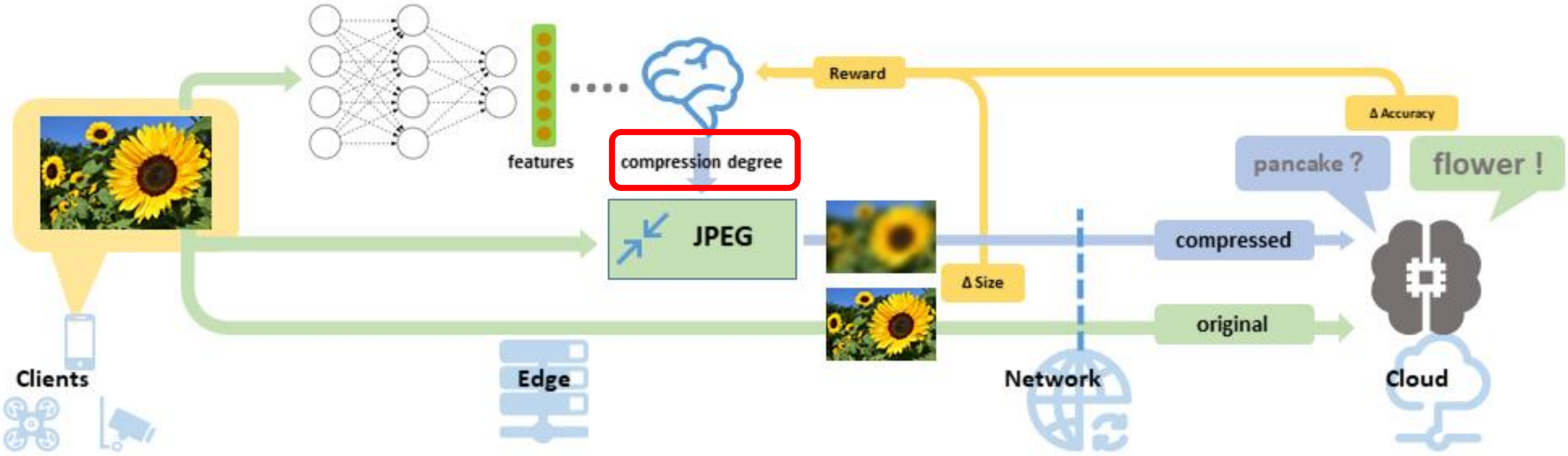
- Reward
  - Δsize – Δaccuracy
- States
  - Features of the input images
- Actions
  - 10 discrete compression levels

Solution



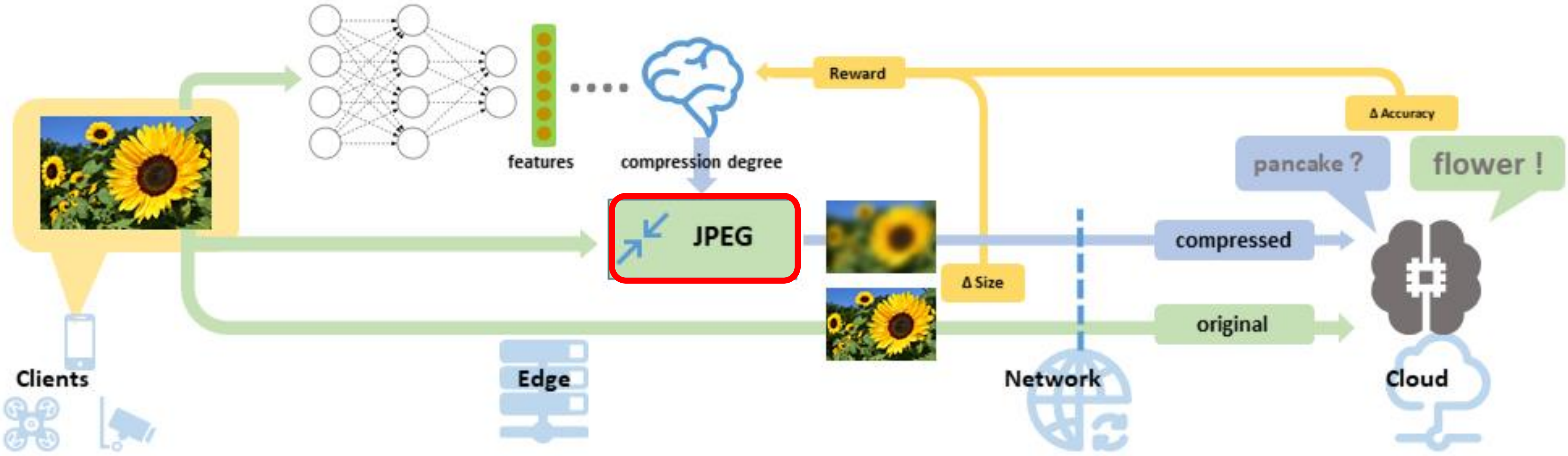
- Reward
  - Δsize – Δaccuracy
- States
  - Features of the input images
- Actions
  - 10 discrete compression levels

Solution



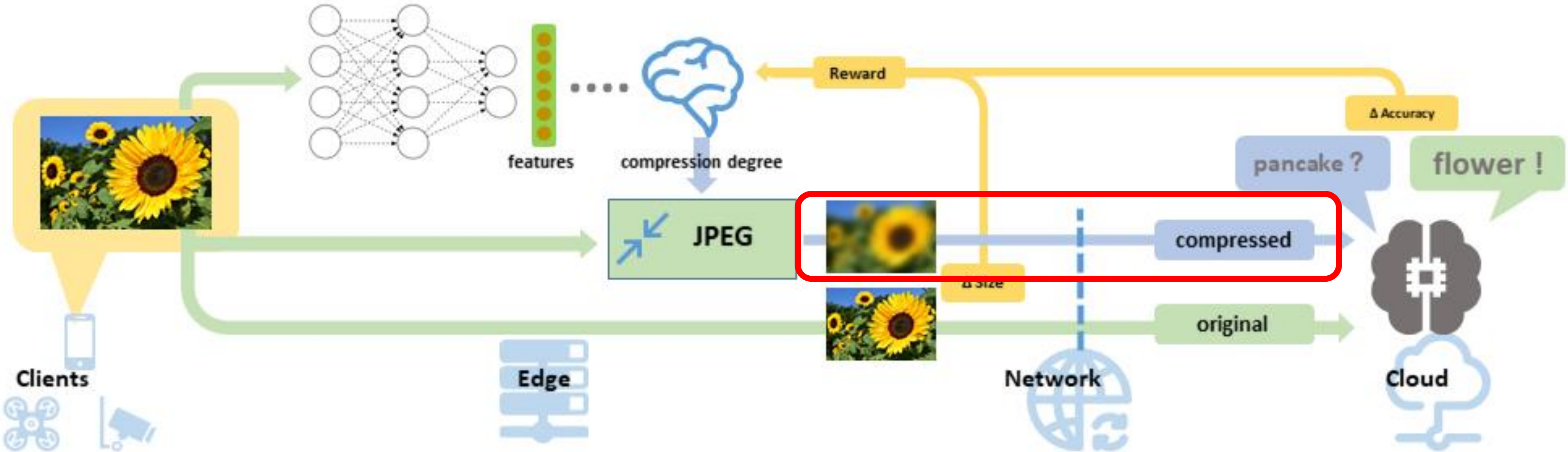
- Reward
  - Δsize – Δaccuracy
- States
  - Features of the input images
- Actions
  - 10 discrete compression levels

Solution



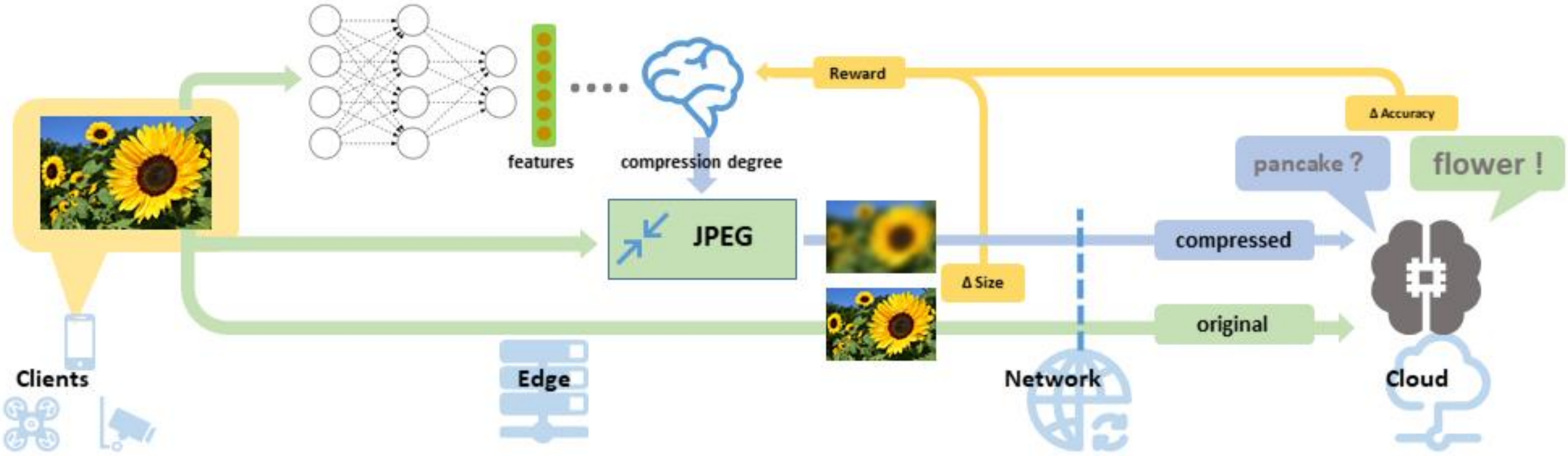
- **Reward**
  - **Δsize – Δaccuracy**
- **States**
  - **Features of the input images**
- **Actions**
  - **10 discrete compression levels**

Solution



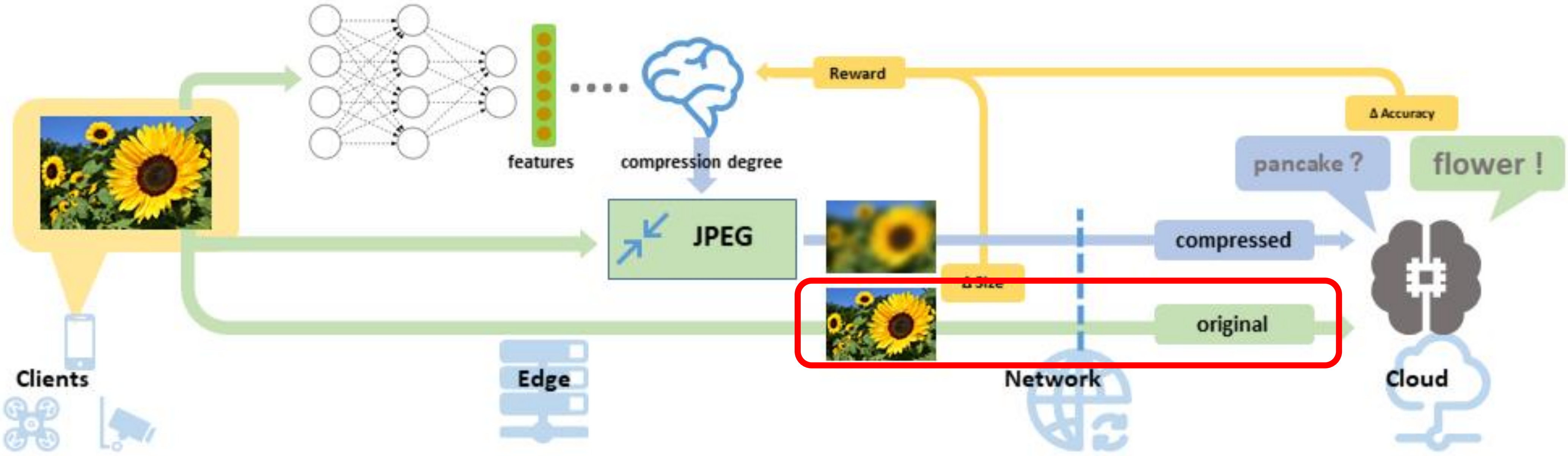
- **Reward**
  - **Δsize – Δaccuracy**
- **States**
  - **Features of the input images**
- **Actions**
  - **10 discrete compression levels**

Solution



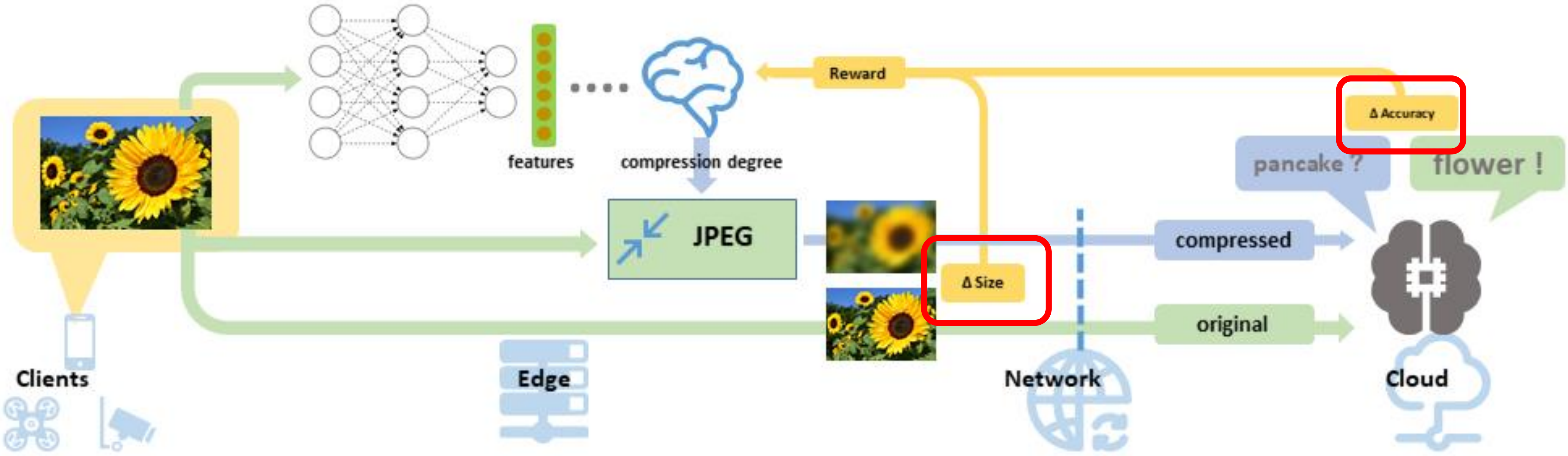
- Reward
  - Δsize – Δaccuracy
- States
  - Features of the input images
- Actions
  - 10 discrete compression levels

Solution



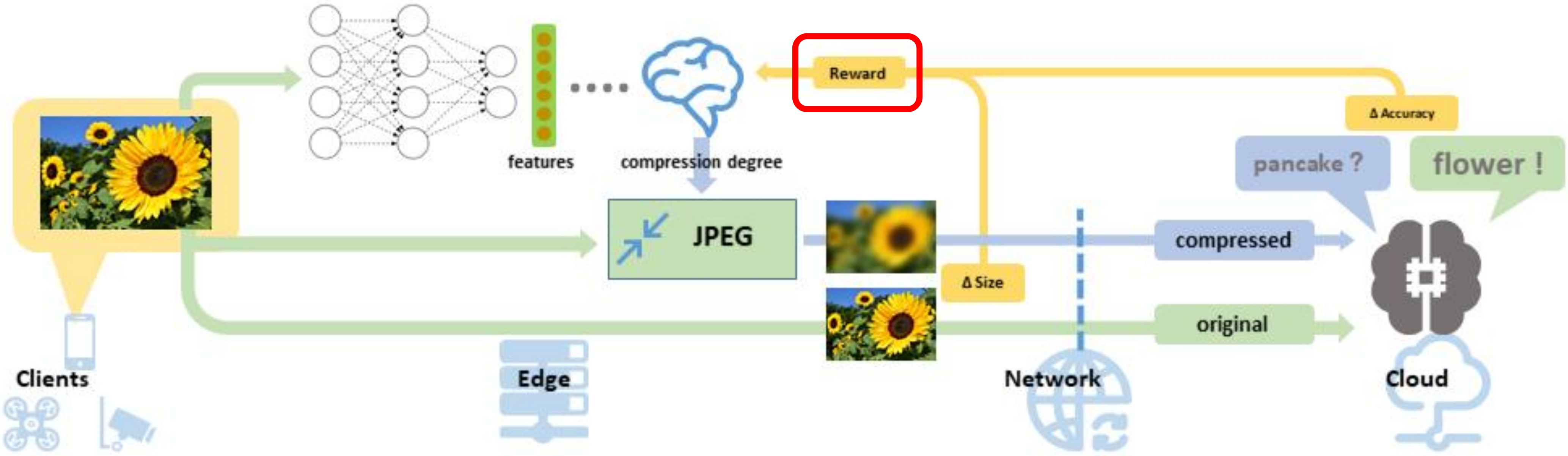
- Reward
  - Δsize – Δaccuracy
- States
  - Features of the input images
- Actions
  - 10 discrete compression levels

Solution



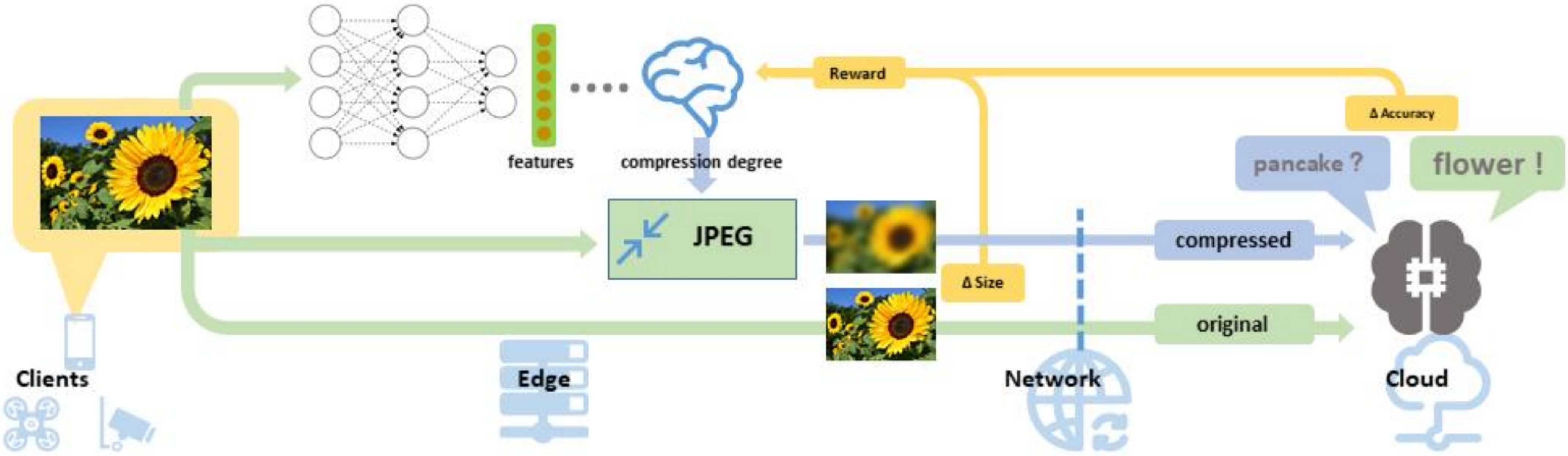
- **Reward**
  - **Δsize – Δaccuracy**
- **States**
  - **Features of the input images**
- **Actions**
  - **10 discrete compression levels**

Solution



- Reward
  - Δsize – Δaccuracy
- States
  - Features of the input images
- Actions
  - 10 discrete compression levels

Solution



- **Reward**
  - **Δsize – Δaccuracy**
- **States**
  - **Features of the input images**
- **Actions**
  - **10 discrete compression levels**

Performance

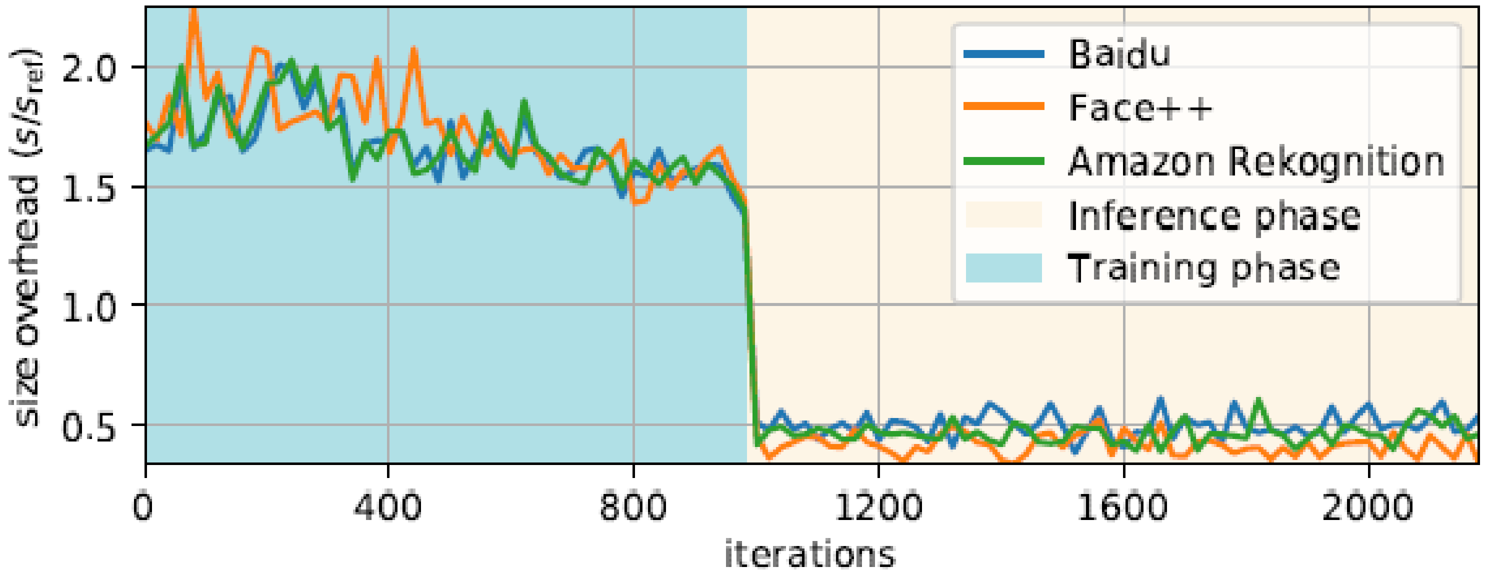


Figure 8: Upload size overhead in training and inference phase

- Size overhead in training phase
- Inference phase is longer than training phase

Performance

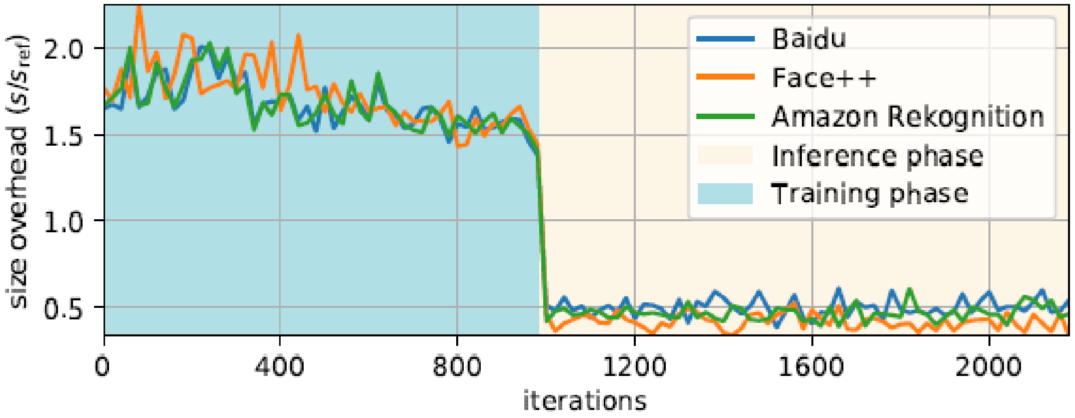


Figure 8: Upload size overhead in training and inference phase

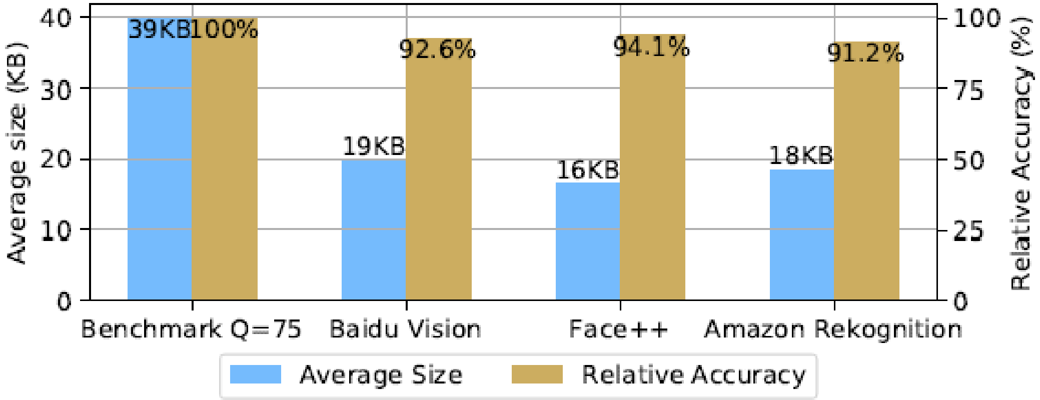


Figure 9: Average size and relative accuracy on different cloud services

- **Size overhead in training phase**
- **Inference phase is longer than training phase**

Insight

- **Different compression strategies in different environments**

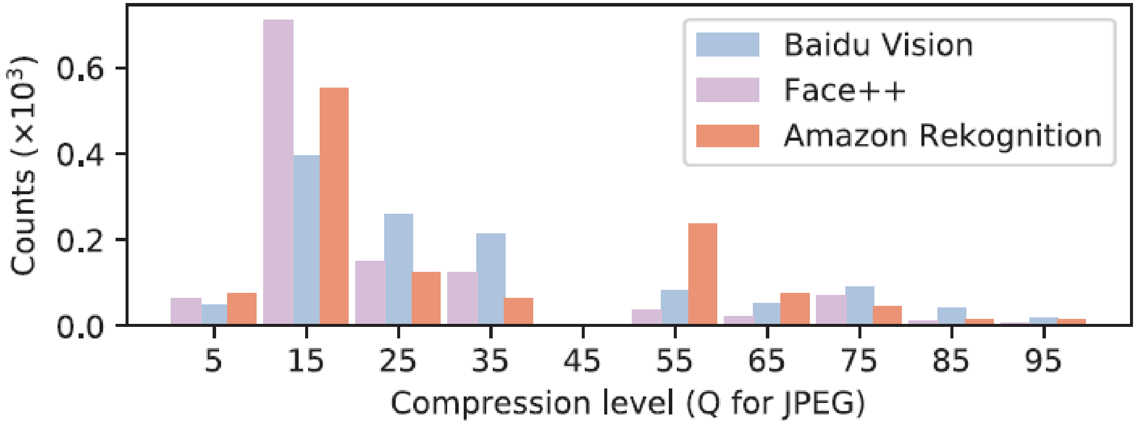


Figure 5: Histogram of RL agent's best compression level selection for different cloud services

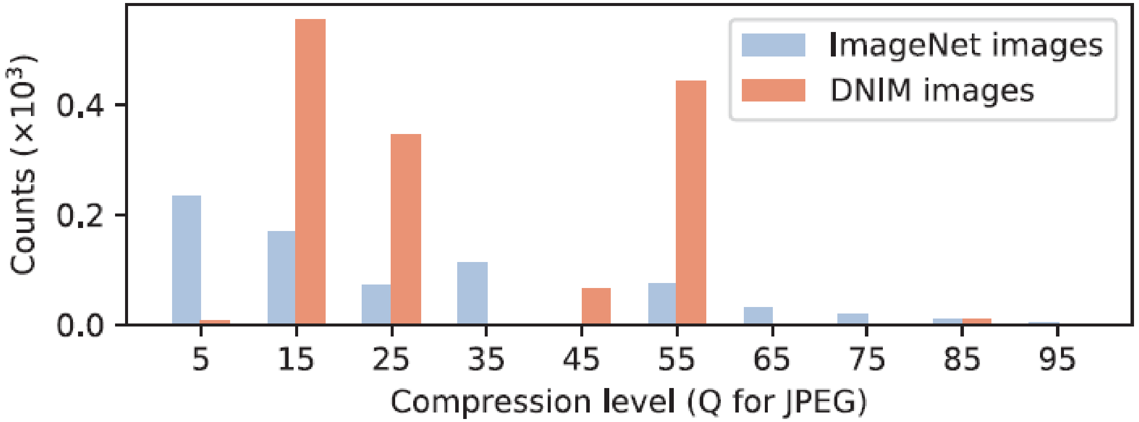
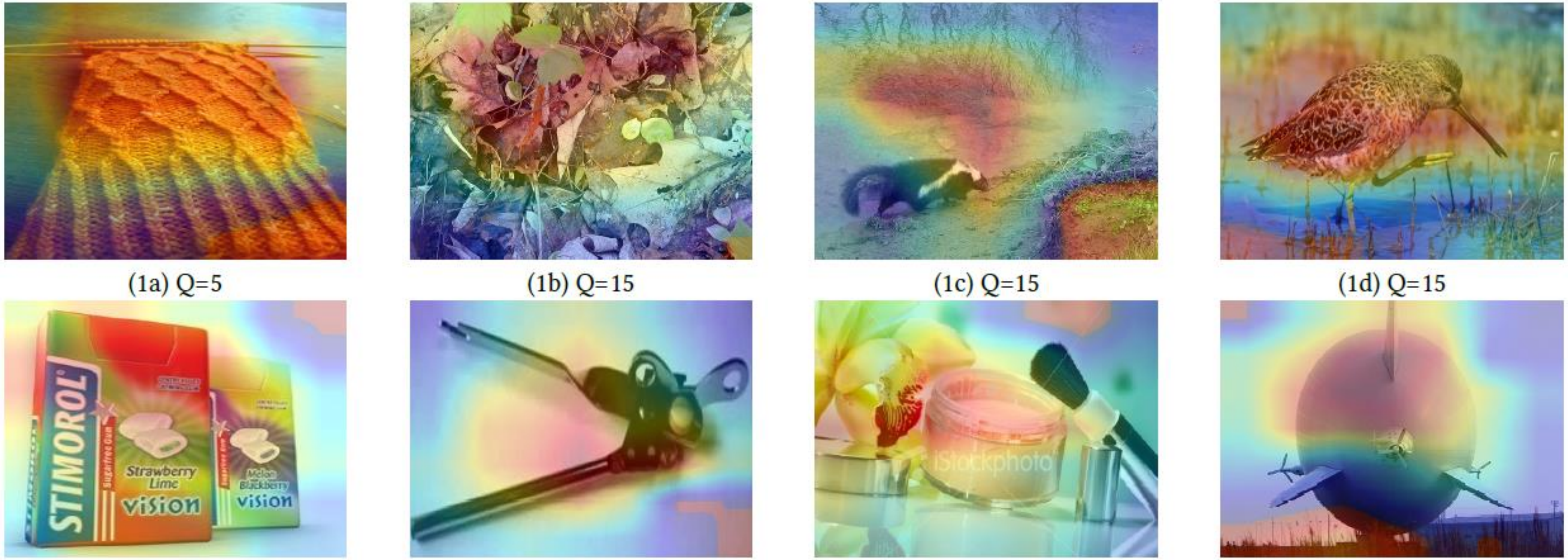


Figure 6: Histogram of RL agent's best compression level selection for different scenery image inputs

Insight

- **Grad-Cam**
- **High quality for smooth region**



(1a) Q=5 (1b) Q=15 (1c) Q=15 (1d) Q=15

(2a) Q=85 (2b) Q=85 (2c) Q=75 (2d) Q=75

Figure 7: Visualization of the importance map for the RL agent to choose a compression quality

Scenery change

- Scenery change (day to night, sunny to rainy etc.)
- State machine with 3 states
- Occasionally estimate system accuracy
- Retrain when necessary

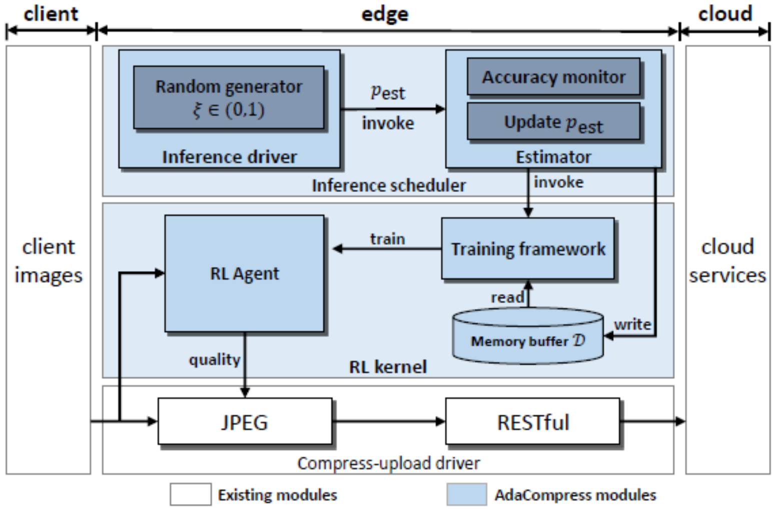


Figure 3: Diagram of AdaCompress architecture

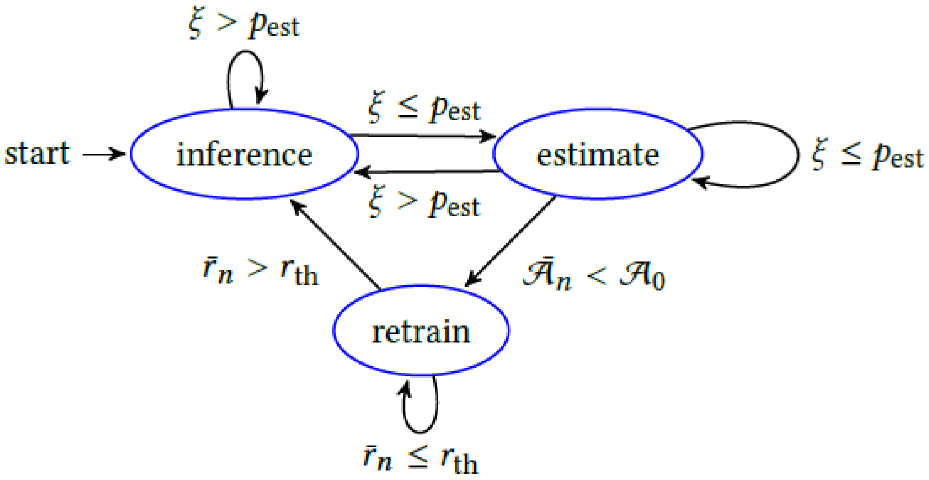


Figure 4: State switching policy

Performance upon scenery change

- **Imitate scenery change by changing dataset**

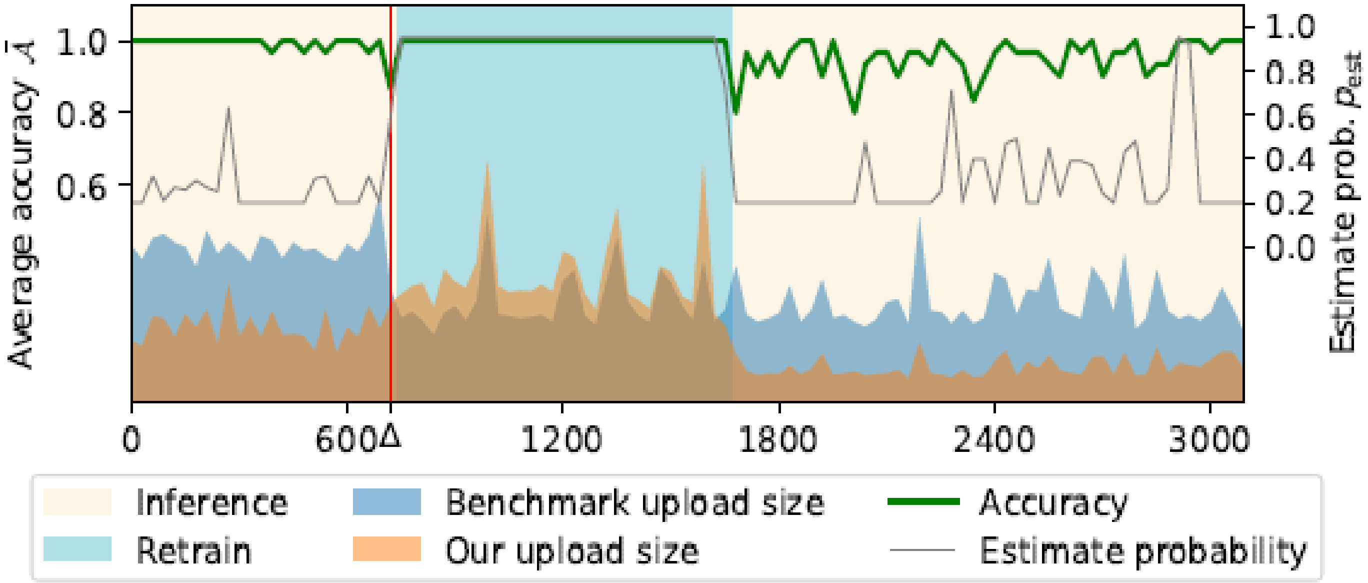


Figure 10: AdaCompress's reaction upon scenery change

Performance upon scenery change

- Imitate scenery change by changing dataset

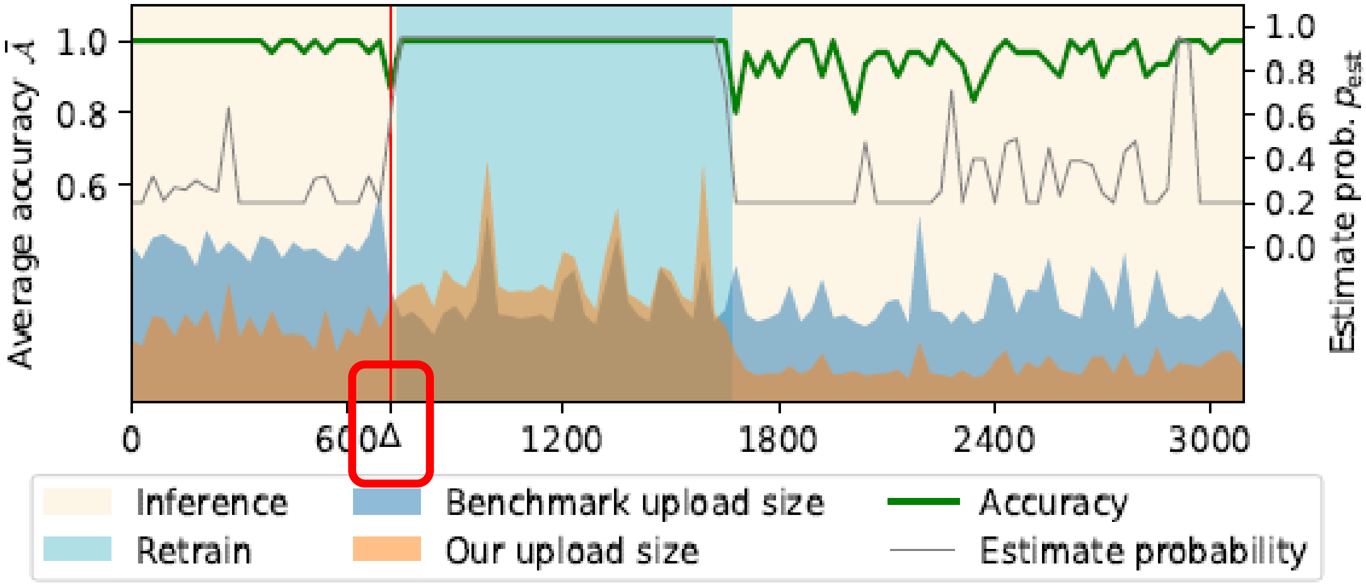


Figure 10: AdaCompress's reaction upon scenery change

Performance upon scenery change

- **Imitate scenery change by changing dataset**

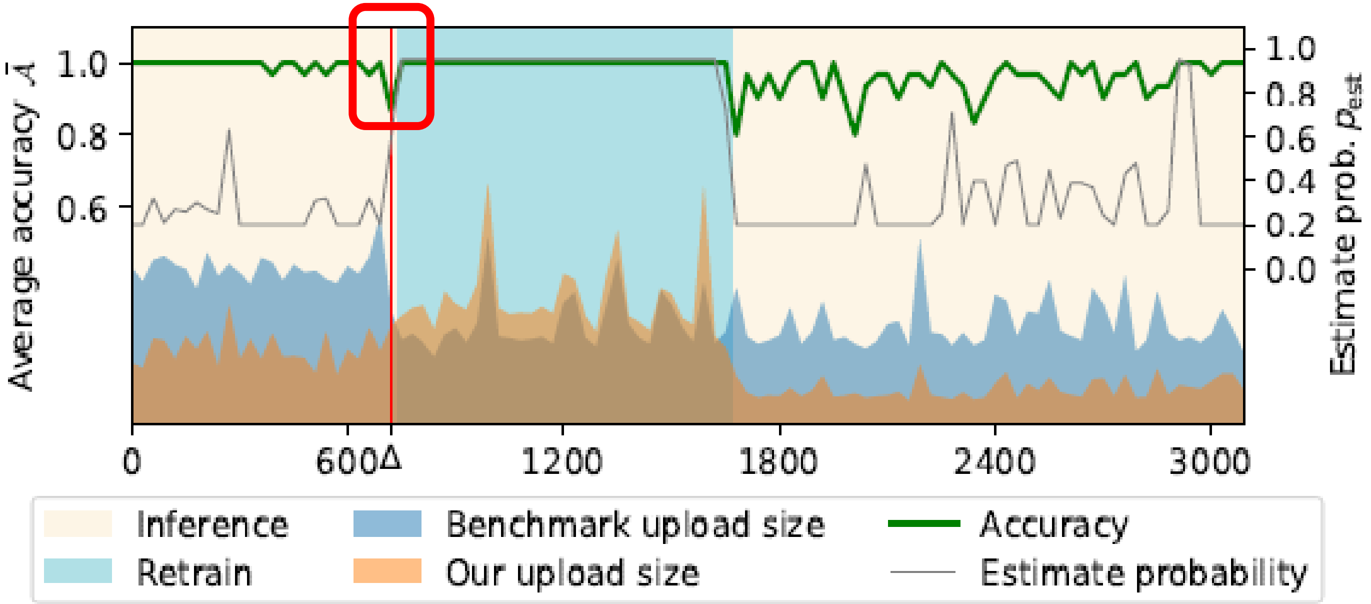


Figure 10: AdaCompress's reaction upon scenery change

Performance upon scenery change

- Imitate scenery change by changing dataset

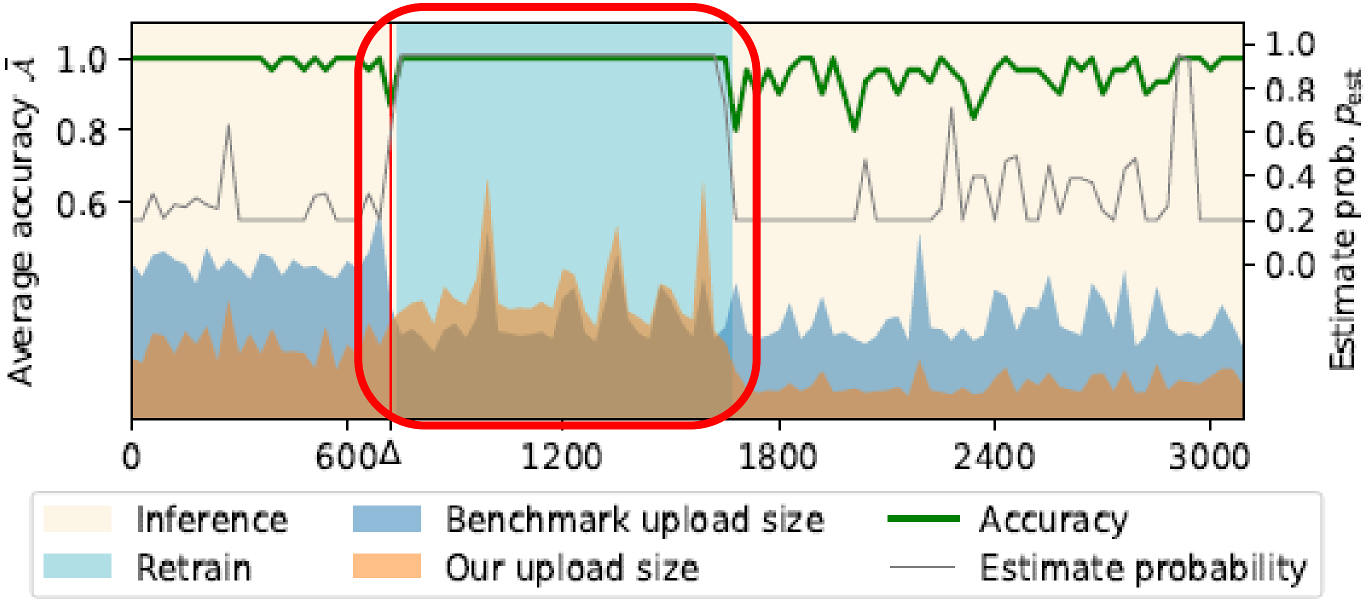


Figure 10: AdaCompress's reaction upon scenery change

Performance upon scenery change

- Imitate scenery change by changing dataset

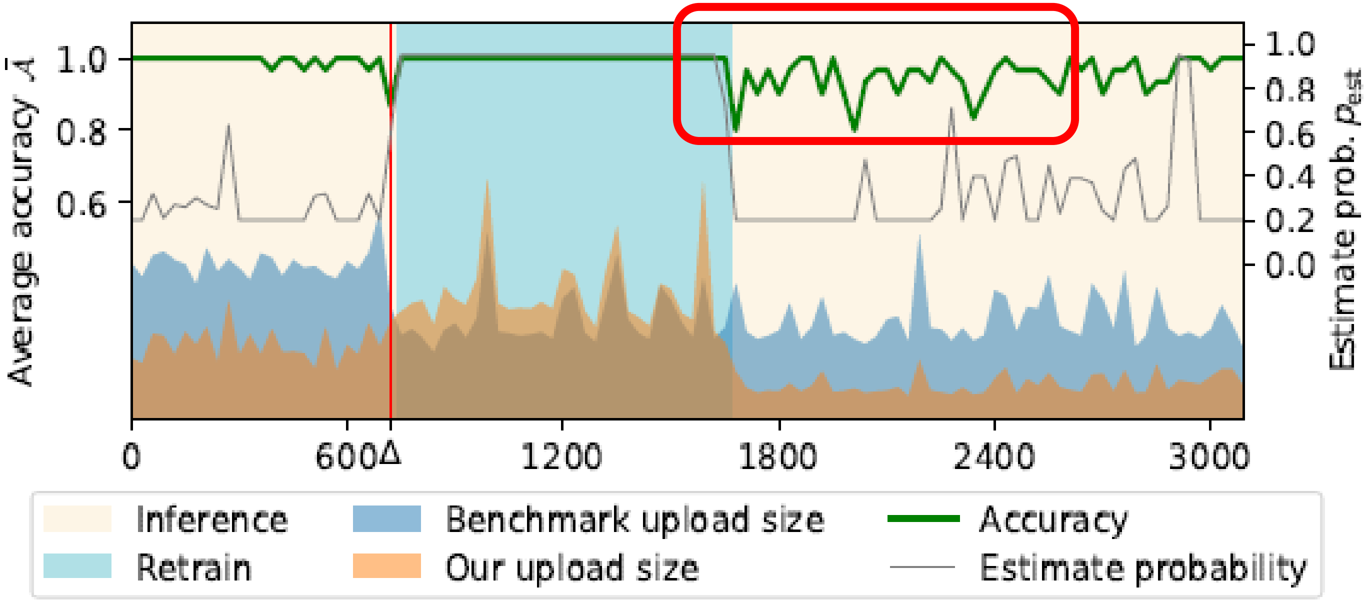


Figure 10: AdaCompress's reaction upon scenery change

Performance upon scenery change

- Imitate scenery change by changing dataset

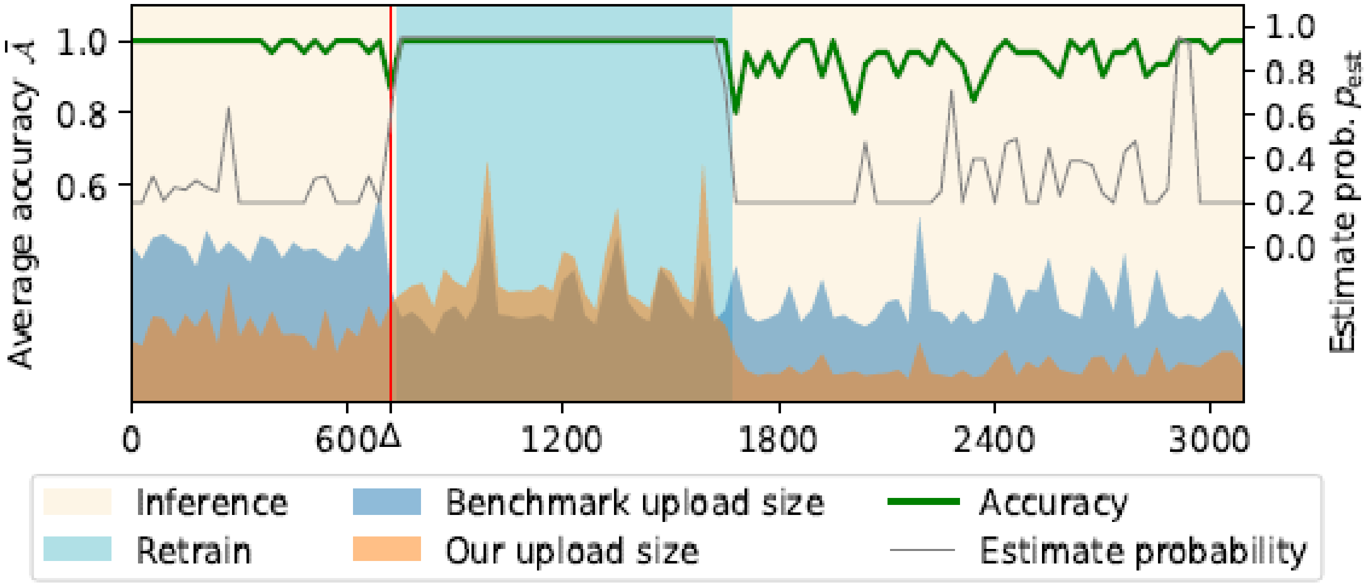


Figure 10: AdaCompress's reaction upon scenery change

Thank You

Source Code:

https://github.com/hosea1008/AdaCompress