Open Research Challenges in DCAI

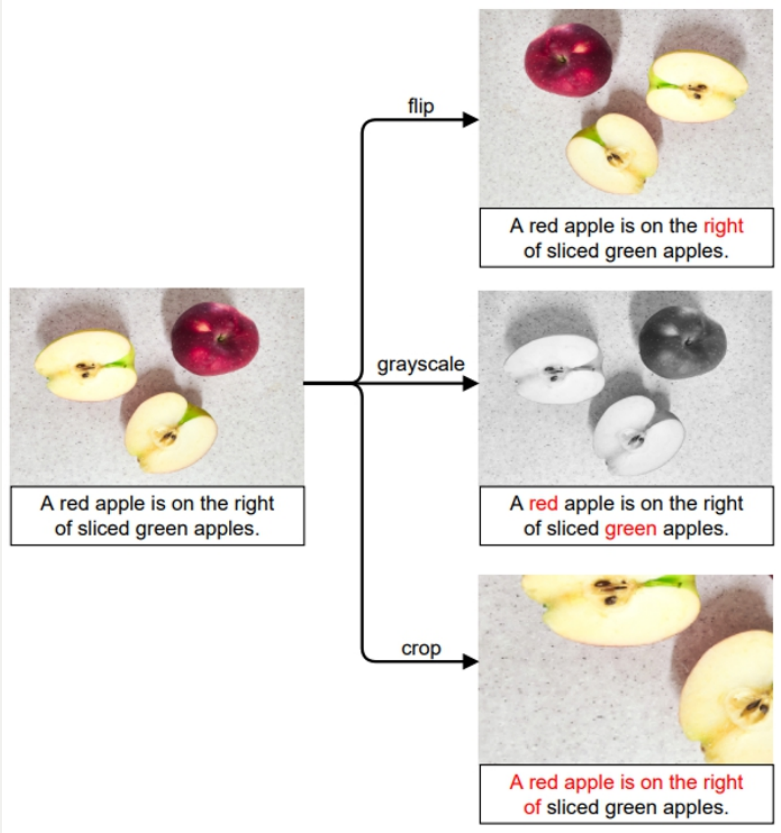
Inference Data & Data Maintenance

Challenging because it's open-ended: rather than optimizing performance metrics (=training data management), needs a comprehensive understanding of the performance and continuous support of data.

Cross-task Techniques

Broader DCAI view is needed: different DCAI tasks could have an interaction effect.

💡 Example 1) Augmentation method & the collected data



💡 Example 2) Training data & Evaluation set construction strategy

💡 Example 3) The training data could be adjusted based on the evaluation results

💡 Example 4) Data maintenance strategies must be designed based on the training/evaluation data characteristics

→ AutoML

Data-model Co-design

DCAI doesn't imply that the model has to stay unchanged.

→ Predict that the future advancements will come from co-designing data pipelines and models

→ Co-evolvement of data and model!

Data Bias

Discrimination issues in data because of the biased distributions for specific sensitive variables in data.

1) How to mitigate the bias in the training data

2) How to construct evaluation data to expose the unfairness issue

3) How to continuously maintain data unbiasedness in a dynamic environment

Benchmarks

Benchmarks are lacking for DCAI.

Existing benchmarks only focus on a specific DCAI task (e.g., feature selection)

DCAI benchmarks are needed to understand the overall data quality and comprehensively evaluate various DCAI techniques, thus accelerate the research progress.