AUTO-LABELING

Chen Chen

Oct 14, 2023

Problem: Cost of Labeling

- Image classification, object detection, text classification...

- Labeling data manually is a time-consuming process

- Expensive

Solution: Auto-labeling

Trained algorithms	Process
● Trained algorithms label each and every sample of the data automatically ● Sometimes the trained model is tuned using reinforcement learning with human feedback or Reinforcement learning from human feedback(RLHF).	● *Collecting large amounts of data* ○ *Dividing the dataset into training and testing subsets.* ● *Train a machine learning model on the training set using manual labeling* *(GPT-4, Grounded SAM, OWL ViT, DETIC...)* ● Use the trained model to automatically label the test set. ● Validate the accuracy of the auto-labeled data using a subset for manual verification.

Why Auto-label?

Faster & Efficient

Accuracy and Consistency

Ability to Handle Large Datasets

Cost-Effective

Challenges:

- The algorithms trained to label the data are only good as the data fed into it

- Quality control and validation
  - labeling task is complex or the algorithm is not interpretable

- Generate false positives
  - Need human assigned in the loop

- Interpretability and transparency

Some Hands-on Tools

Refuelai-Autolabel

Roboflow-Autodistill

References & Resources

https://blog.roboflow.com/autodistill/#step-3-train-a-new-model-using-the-annotated-images

https://blog.roboflow.com/how-to-train-yolov8-on-a-custom-dataset/

https://docs.refuel.ai/