# 能動学習 勉強会

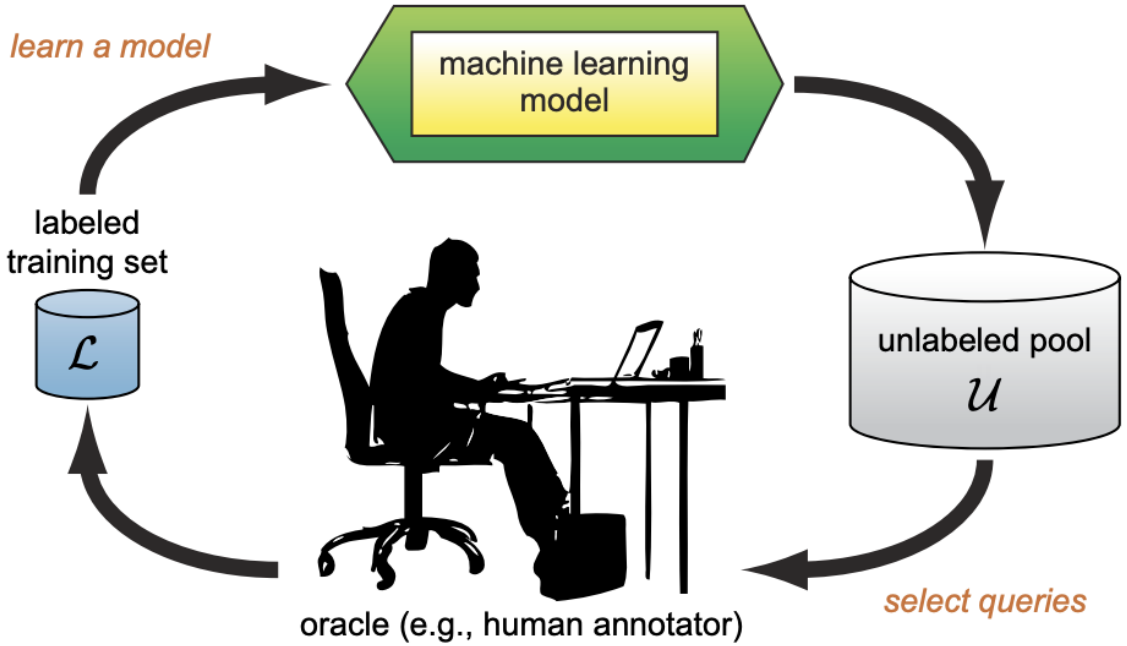
B4 上野詩翔

2022/07/07

# 振り返り

# 振り返り｜DAL(Deep Active Learning)とは

- ◆DeepLearning(DL)モデルに有効なデータに優先的にアノテーションすることでモデルの精度を保ちながらアノテーションコストを下げる技術
- ◆現在の主流はラベルなしプールから最適なデータを選択するプールベースAL
  - ➢1サイクル：選択データにアノテーション，モデルを学習(DL)，次のデータを選択



# 振り返り｜クエリ戦略

◆アノテーションするデータを決める際の選び方

- 不確実性ベース(現在の主流)
  - モデルの不確実性が高いデータを優先的に選択
  - 決定境界付近を優先的に選択しやすい(データが偏る)

- 代表性／多様性ベース
  - 選択したデータがデータセット全体を表すようにデータを選択
  - 不確実性が考慮されないことが多い他，計算量も多く，データセットに依存しやすい

- ハイブリット
  - 不確実性，多様性を考慮してデータを選択
  - 計算量が多く，データセットに依存しやすい

# 振り返り｜DALの現状

- ◆クエリ戦略では精度が頭打ち
  - ➢ラベル無しデータを活用するフレームワークが人気
    - ➢ 半教師あり学習，自己教師あり学習(SSL)
- ◆均等なサンプリングも注目の的
  - ➢選択するデータ枚数は偏らないほうが良い??
    - ➢ Balanced Sampling

# 目次

◆振り返り

◆最近のクエリ戦略

◆自己教師あり学習を利用した手法

◆クラスバランスを考慮したサンプリング手法

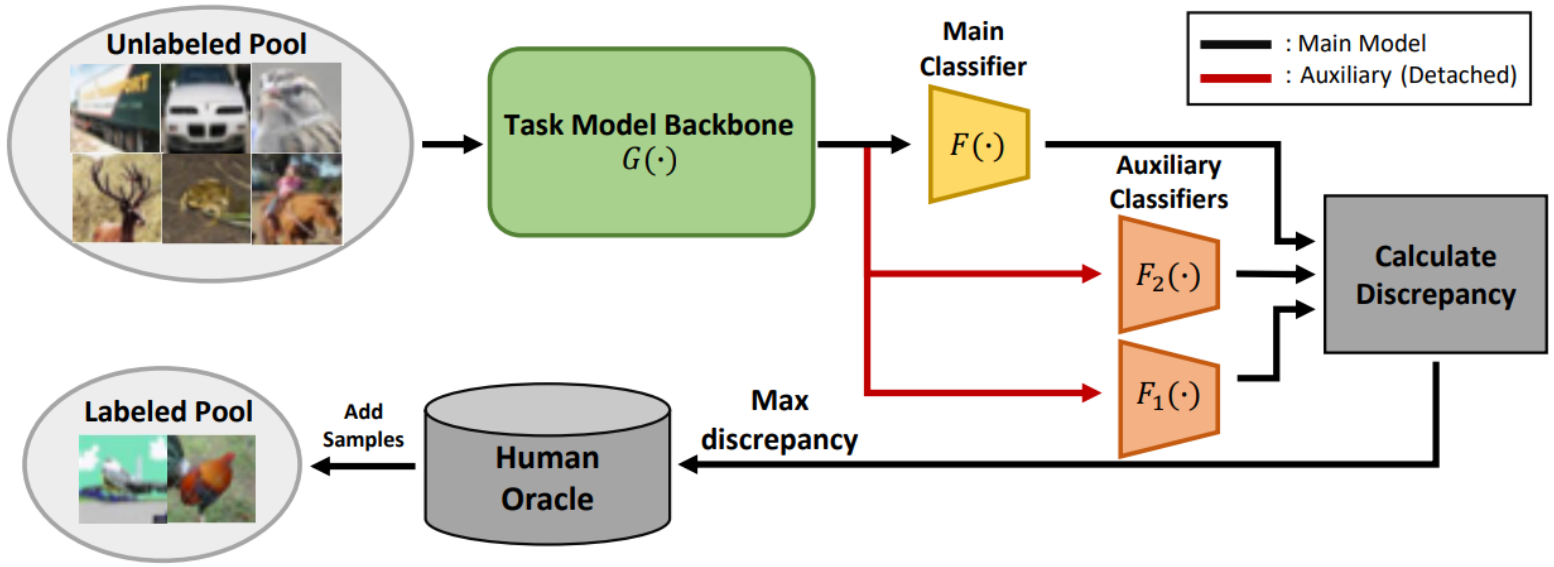
◆AL全体の問題

# 最近のクエリ戦略

# MCDAL:Maximum Classifier Discrepancy for Active Learning (2021 IEEE)

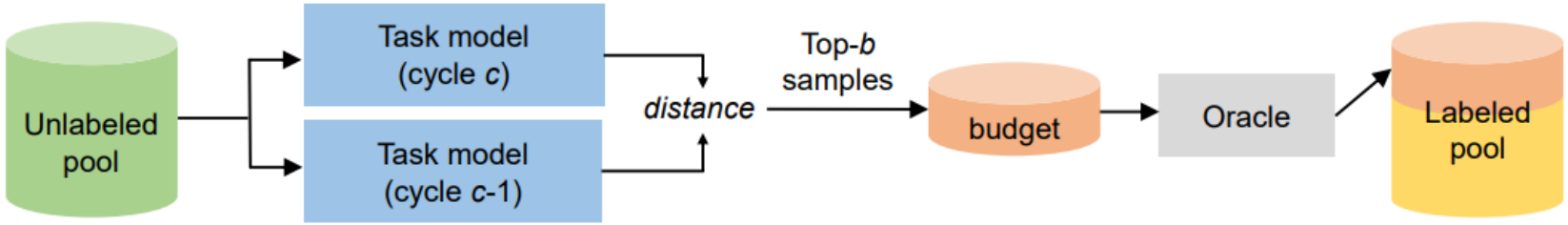
◆Encoderの出力層を1つから3つに増やし，予測間の不一致から不確実性を計算

➢メインの識別器の出力とラベル間のLoss，予測間の不一致によるLossを用いて学習

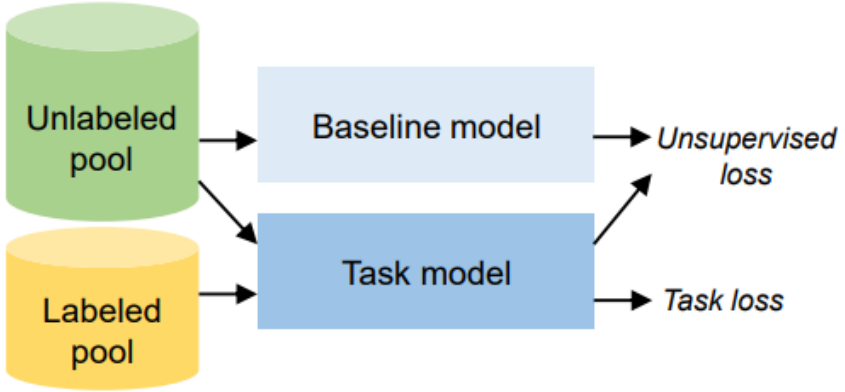


# Semi-Supervised Active Learning with Temporal Output Discrepancy (2021 ICCV)

◆1サイクル前と現在の出力間の不一致を基に不確実性を計算



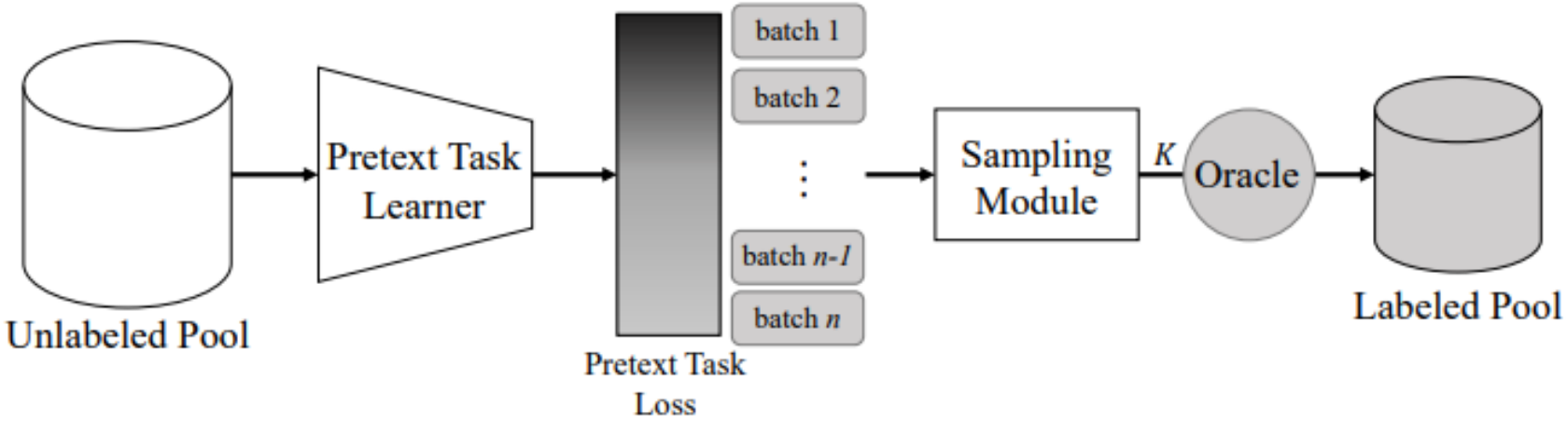

◆ラベルなしサンプルは予測器の出力との距離が最小になるようにモデルを学習



# 自己教師あり学習を利用した手法

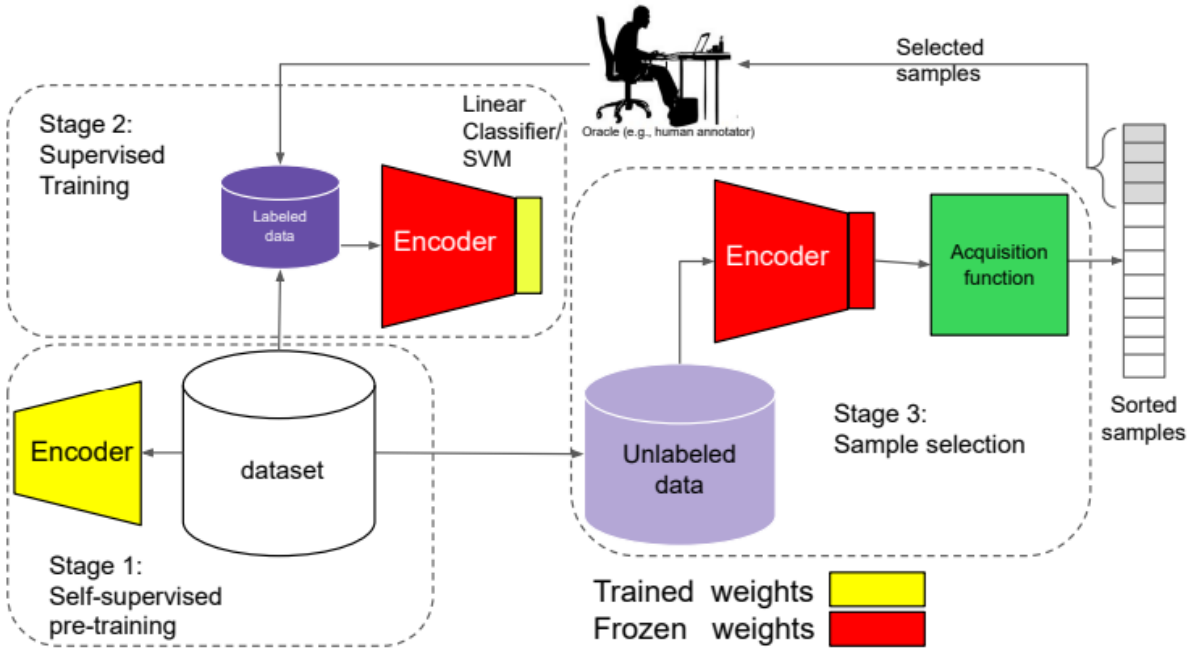
# Using Self-Supervised Pretext Tasks for Active Learning (2022)

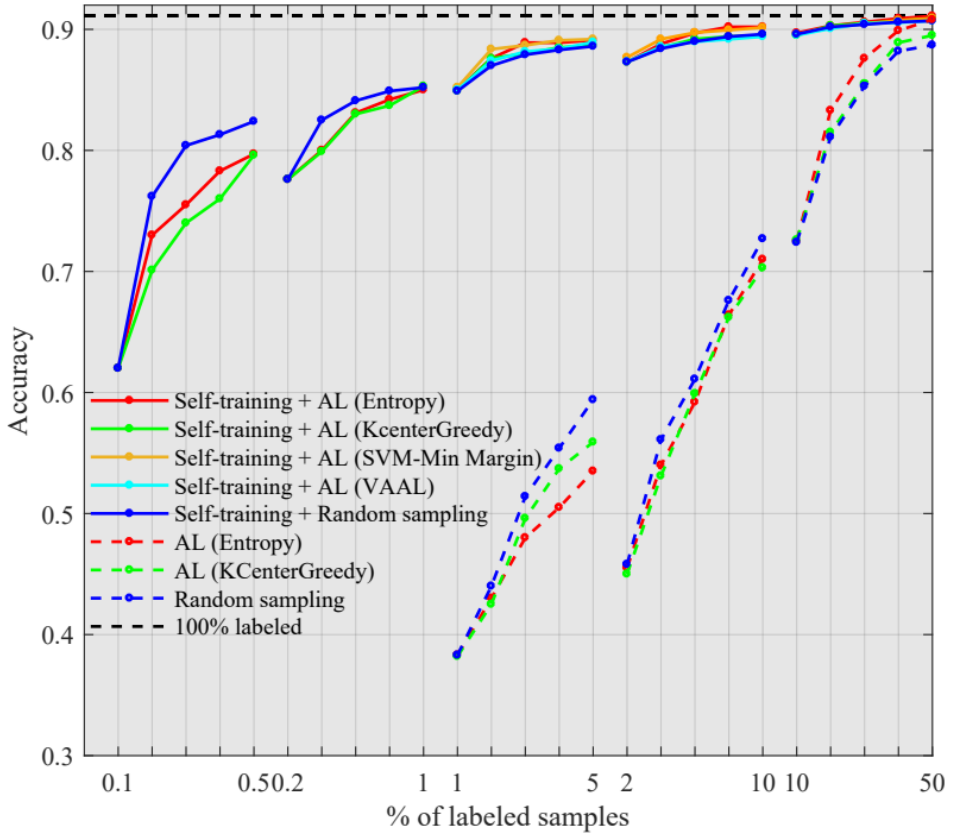
- ◆訓練データ全体でSSL(回転予測のPretext Task)を実行
- ◆SSLのLossでデータ全体をソートし，以下を繰り返し実行
  - ➢ ALのサイクル数のバッチにデータを分割
  - ➢ 各サイクルで各バッチから不確実性サンプリング
  - ➢ サンプリングデータにアノテーションし，モデルを学習



# Reducing Label Effort: Self-Supervised meets Active Learning (2021 ICCV)

- ◆自己教師あり学習(SimSiam)の事前学習をALに組み込む
- ◆事前学習＋既存のクエリ戦略 が高精度と発見

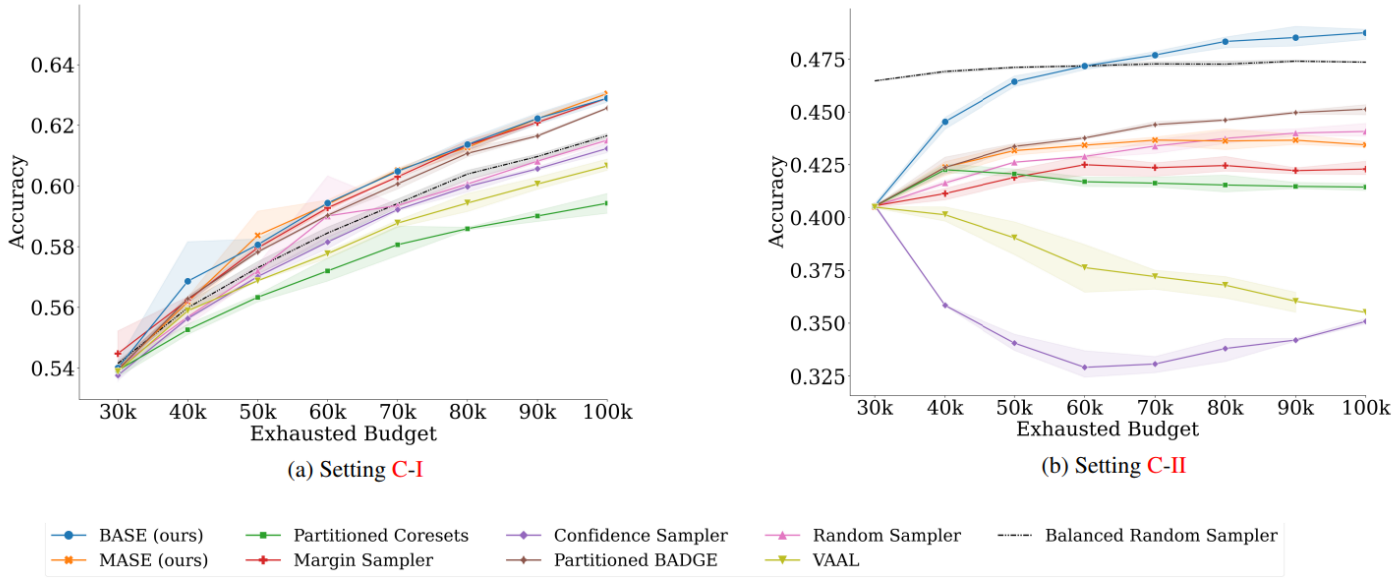


# クラスバランスを考慮したサンプリング手法

# Active Learning at the ImageNet Scale (2022)

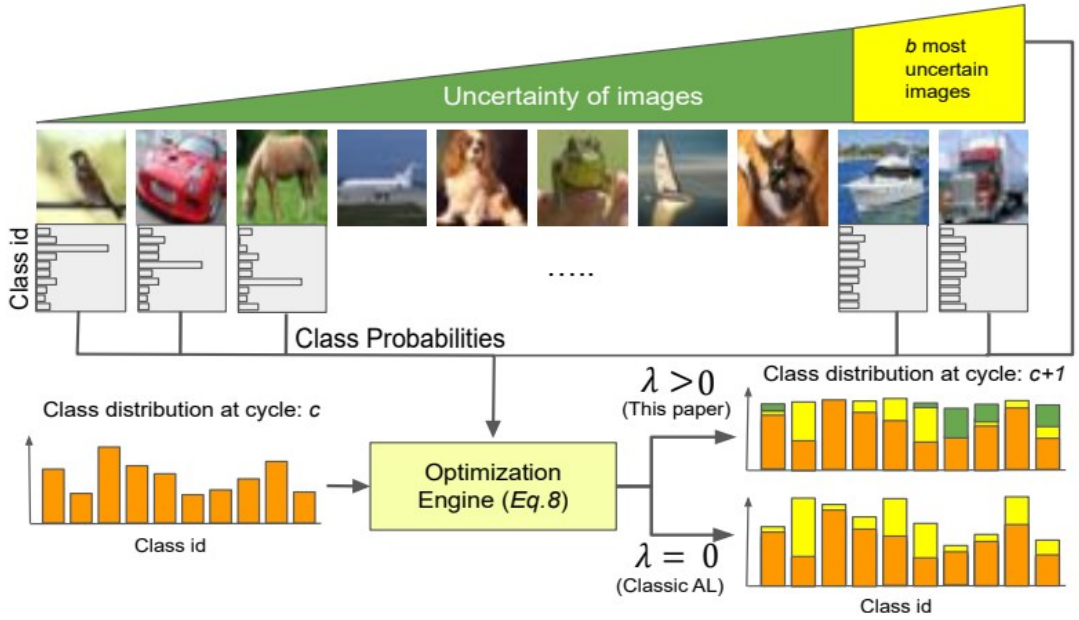
- ImageNetほどの規模ではサンプリングが不均衡になる不確実性クエリ戦略が機能しないと指摘
- 均衡サンプリング手法BASEを提案し，既存手法と精度を比較
  - 左：SSL➡ファインチューン　右：SSL➡線形評価



# Class-Balanced Active Learning for Image Classification (2022 WACV)

◆ソフトマックスの予測を基にラベルなしデータのクラスを予測し，各クラスのデータが均等になるようにサンプリング

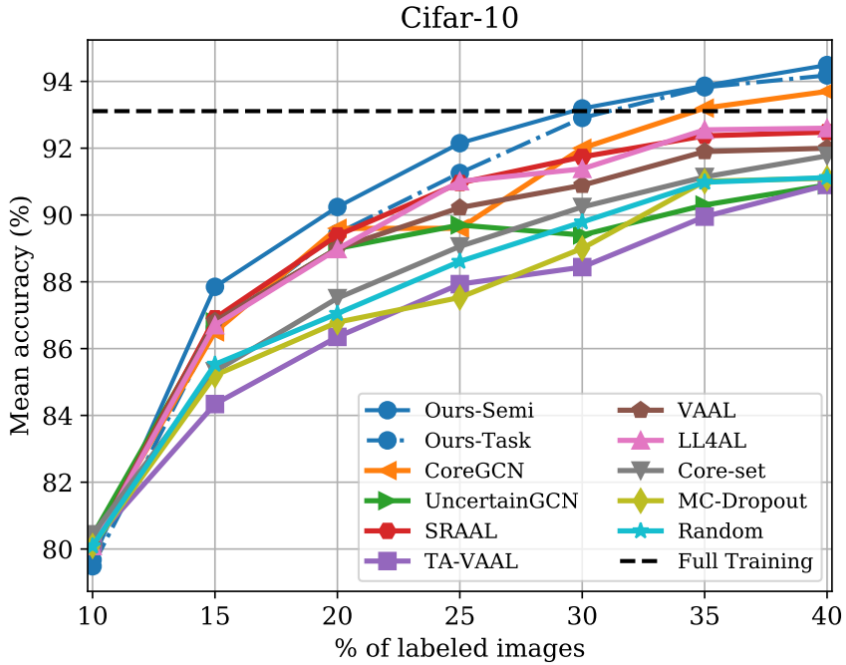
- ➢不確実性の高いデータをある程度集め，不確実性の低いデータで枚数の調整を行う
  - ➢不確実性の低いデータで多様性を考慮
  - ➢不確実性の指標は自由
- ➢実際に均等とは限らないので注意

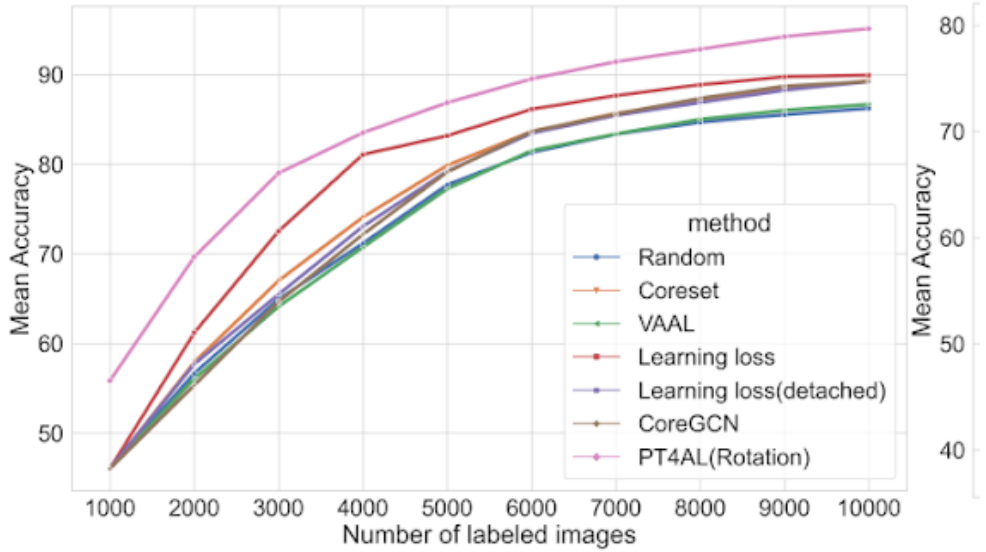


# AL全体の問題

# 実験設定がバラバラ

- ALにはサイクル数，クエリ数といったハイパラが存在
  - データ数にかかわるため，本来は統一されている必要がある
  - 実際は論文によってバラバラ
    - 精度の評価が困難



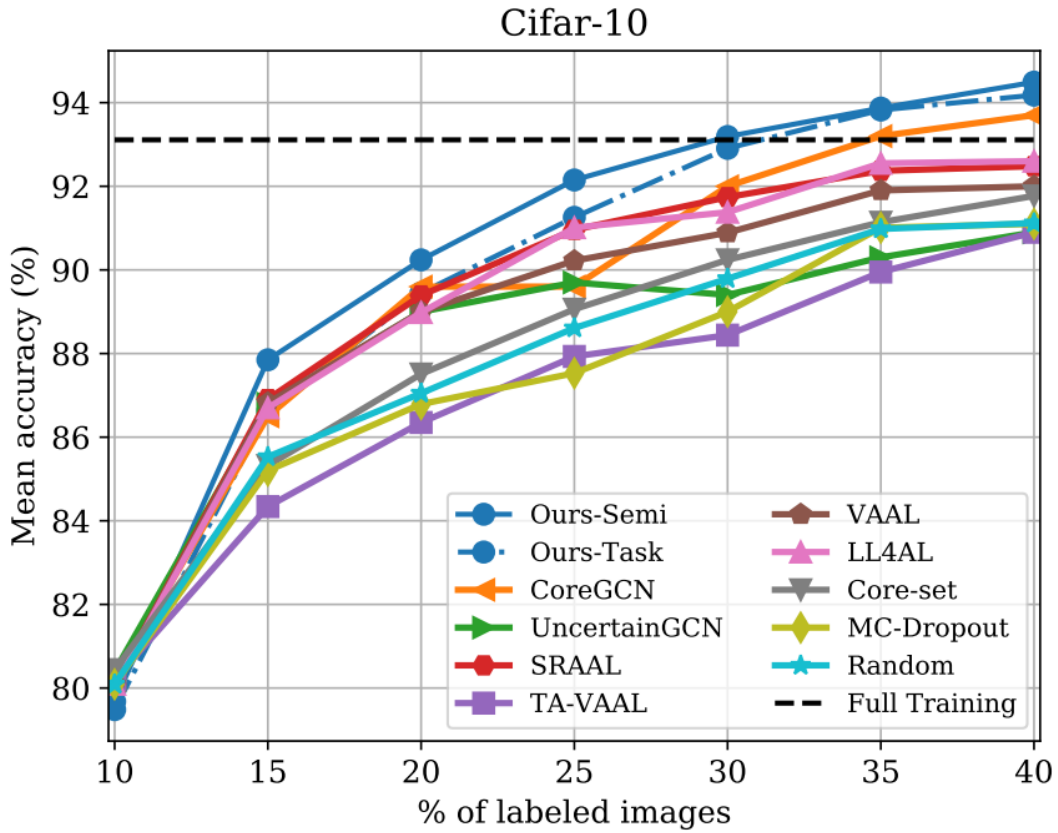




(a) CIFAR10

# 実験設定が曖昧

◆ALは事前学習がない場合，初めに数%のデータをランダムで選びアノテーションする手法が多い

- seedが固定されていれば同じ精度になるはず
- 既存手法がRandomを下回る状態で比較されるケースも
  - 論文の結果を信じて大丈夫？？



# まとめ

# まとめ

- ◆クエリ戦略
  - ➢不確実性ベースがメイン
    - ➢予測器の不一致を用いる手法も登場したが，既存のエントロピーやマージンも使われている
- ◆最近の手法
  - ➢自己教師あり学習
    - ➢既存のクエリ戦略をそのまま利用可能
  - ➢バランスを考慮したサンプリング
    - ➢不均衡データセットに対するパフォーマンスも注目されている
- ◆論文中の実験は設定に注意が必要
  - ➢ハイパラの設定や重みの使い方に注意
  - ➢自分で実験するのも良い