Amazon SageMaker Ground Truth

第 6 回 Amazon SageMaker 事例祭り

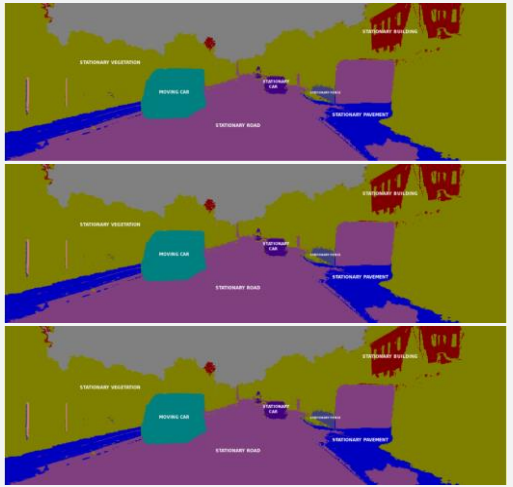
#sagemaker_fes

アマゾン ウェブ サービス ジャパン株式会社

独自のデータを利用した機械学習の流れ

ラベル付きデータ準備 → モデル開発 → モデル学習 → 推論

大量の高品質なラベル付けが重要



独自のデータを利用した機械学習の流れ

アノテーション（データへのラベル付け）にはコスト・時間がかかる

- 進捗管理・作業割り振り
- 効率の良いラベリングツールの作成
- 作業を割り当てるワーカーの募集
- これらを用意した上で数万個のデータへのラベル付け…

これらの課題を解決するのが Amazon SageMaker Ground Truth

Amazon SageMaker Ground Truth

New!

データにラベル（Ground Truth）を付与するアノテーション作業の支援サービス

- アノテーションの一般的なワークフローをサポート
- 4種類の組み込みラベリングツールを提供
- アノテーション作業を行うワーカーとの連携・管理機能を提供
- 大規模データセットに対しては自動ラベリング機能で最大70%のコスト削減

迅速なラベル付け

ワーカーとの連携が容易

高精度

aws

Ground Truth 利用のワークフロー

1. アノテーション対象の
データをアップロード

S3

Amazon
SageMaker
Ground Truth

2. ラベリング
ジョブの作成

3. タスクはワーカーに
自動で割り振られる

4. ラベリングツールで
ワーカーがアノテーションを行う

5. アノテーション結果が
S3に格納される

複数人の
結果をマージ

S3

6. 学習・推論
に利用

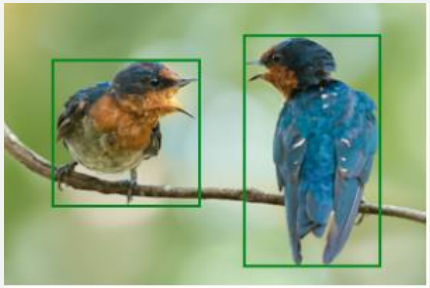
Amazon
SageMaker

ビルトインアルゴリズムを
そのまま適応可能

aws

組み込みラベリングツールの利用も 独自実装も可能

画像分類



物体検出

セマンティック
セグメンテーション

文章分類

カスタム

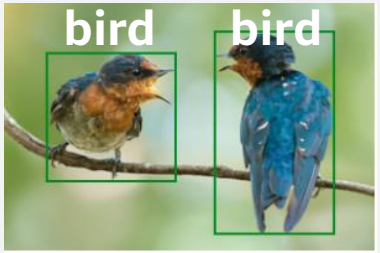
組み込みのラベリングツールの使用例

画像分類（画像から1つのラベルを推定）

- 撮影した画像から商品を推定（ラベルは100種類の商品）
- 顔画像から感情を推定（ラベルは喜怒哀楽+無表情の5種類）
- 独自の基準での不適切画像判定（ラベルは適切/不適切の2種類）

物体検出（画像から物体のラベルと矩形領域を推定）

- 製品や道路の画像から損傷箇所を検出
- レントゲン写真からある病気の箇所を検出する



組み込みのラベリングツールの使用例

セマンティックセグメンテーション（ピクセル単位のラベリング）

- 走行画像の分析（車や人、走路情報などの領域を抽出）

文章分類（文章から1つのラベルを推定）

- レビュー文章のポジティブ/ネガティブ判定
- 文章のカテゴリ分類（政治・スポーツ・芸能などのラベル）

※ 分類対象の文章には日本語をセット可能
（日本語が理解できるワーカーが必要なことに注意）

組み込みのラベリングツール

ワーカーは以下の3種類から選択可能

パブリック

- クラウドソーシングサービスの Amazon Mechanical Turk を利用
- 非言語依存で機密性の低いタスク向き

プライベート

- 友人や社員をワーカーとして登録出来る
- 機密性の高いタスク向き
- ワーカーの管理にCognitoを利用（SAMLでの連携も可）

ベンダー

- SageMaker Ground Truthに登録済みのアノテーション専門ベンダーに依頼
- 現時点では日本のベンダーは登録されていない

自動ラベリング

データの一部をワーカーがラベル付けするだけで、
残りのラベル付けが自動化され、時間とコストを大幅に削減

※ 5000データ以上の大規模データセットに対して利用可能なオプション機能

ラベリングジョブの作り方

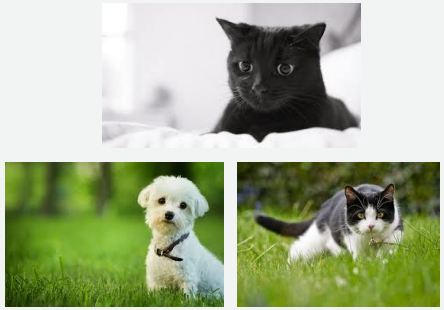
aws

ラベリングジョブ作成の4ステップ

1. データセットの準備
2. タスクの定義
3. ワーカーの選択
4. ラベリングツールの設定

1. データセットの準備

- データと結果を保存するためのS3バケットを用意
- アノテーション対象の画像群をS3バケットに保存
  （テキスト分類の場合は Text/CSV ファイル）
- 画像群のパスを記述したマニフェストファイルを作成し、S3バケットに保存
  （マニフェストファイルは自動生成も可能）



1. データセットの準備 (マニフェストファイル)

input.manifest

```
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00001.JPG"}
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00002.JPG"}
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00003.JPG"}
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00004.JPG"}
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00005.JPG"}
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00006.JPG"}
{"source-ref": "s3://sagemaker-groundtruth-demo/SSDB00007.JPG"}
```

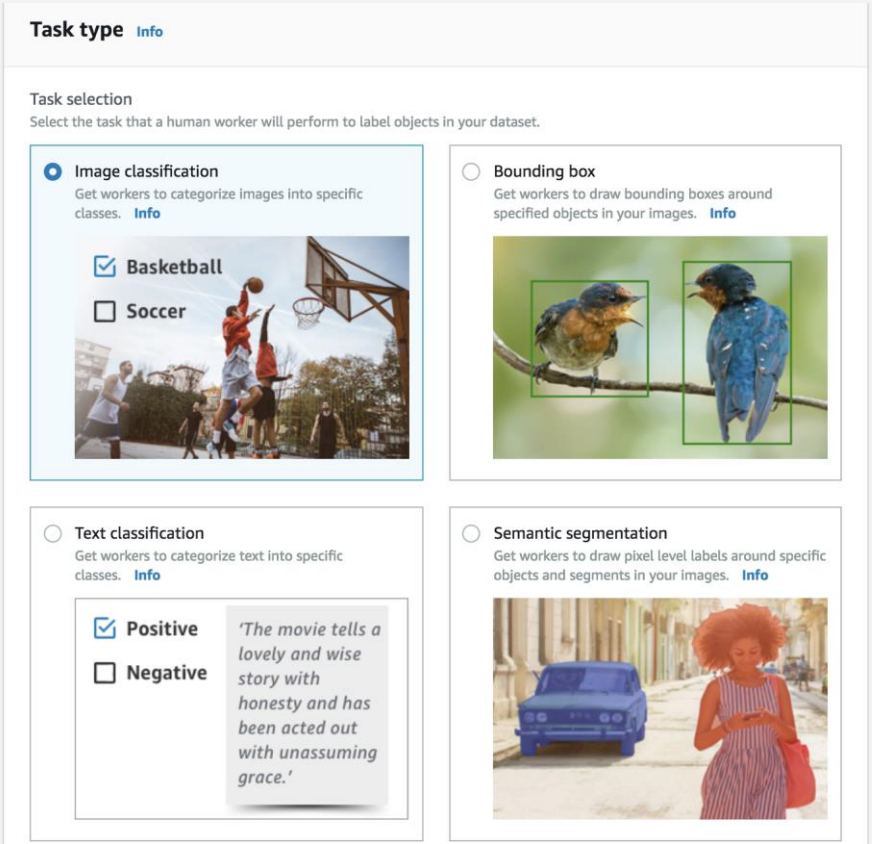
“source-ref” で S3 のパスを指定。
あるいは “source” に直接テキストを書くこともできる

aws

2. タスクの定義

- 画像分類
- 物体検出
- テキスト分類
- セマンティック
  セグメンテーション

あるいは、
- ユーザ定義のカスタムタスク



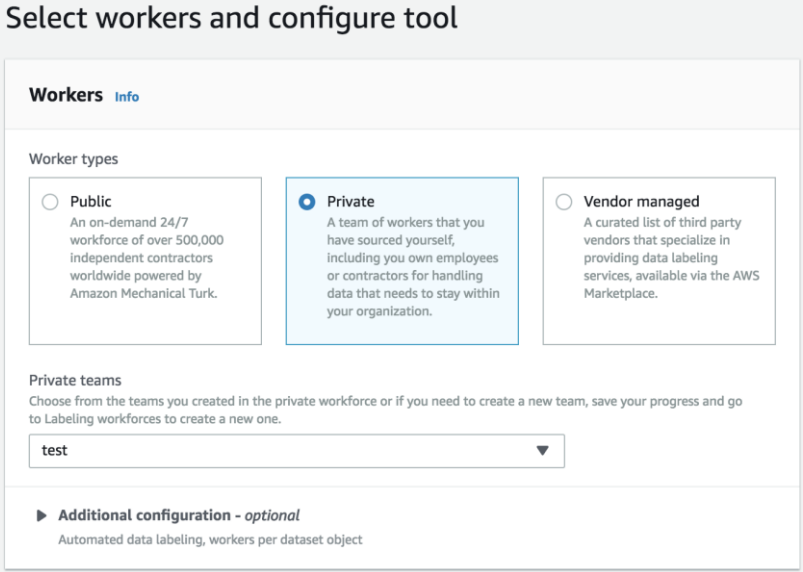
3. ワーカーの選択

- Public
    - Amazon Mechanical Turkを利用
- Private
    - 社員などワーカーを自ら調達

•ベンダー

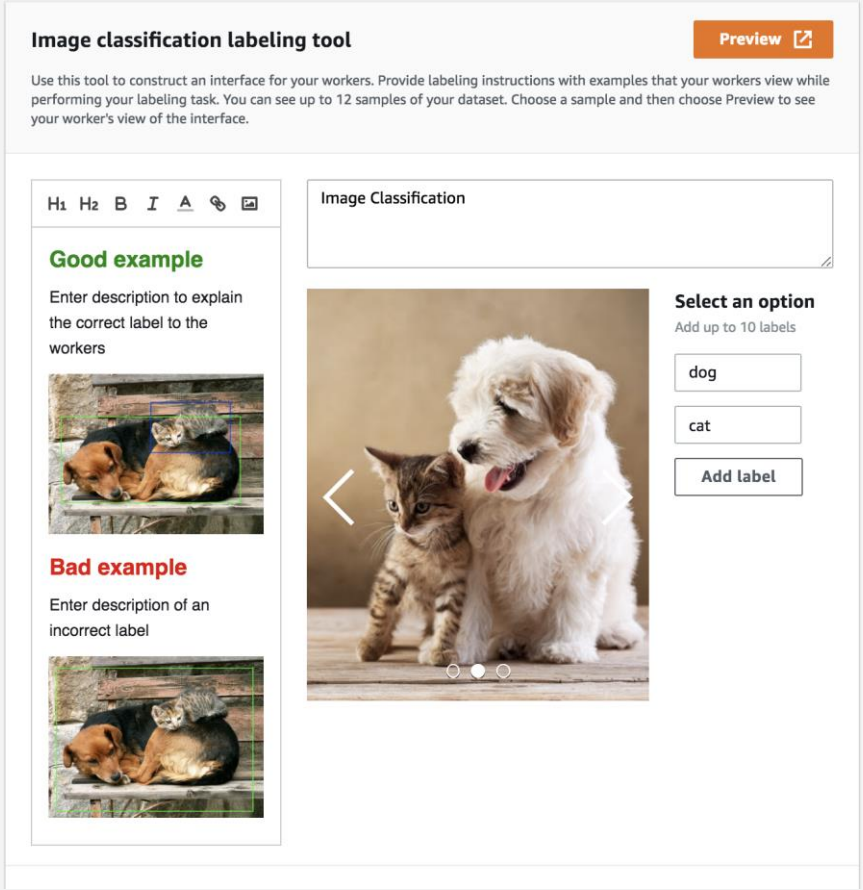
- 3rd パーティーベンダーに依頼

- 追加オプションも選べる
    - 自動ラベリング
    - 複数ワーカーによる
      ラベル付け

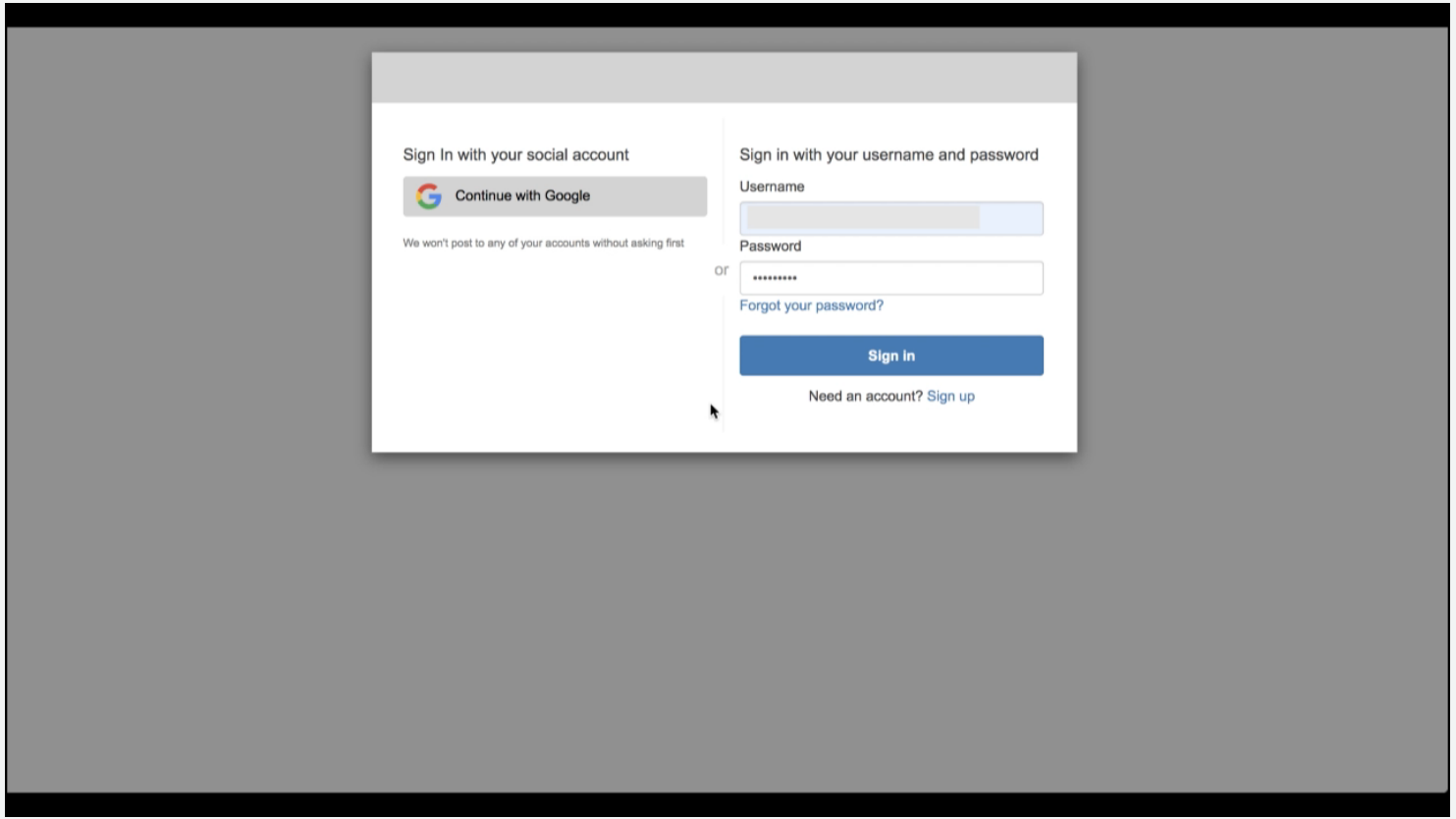


4. ラベリングツールの設定

- アノテーションの指示書を書く
- 良い例・悪い例を記載
- ラベルを設定



ワーカー側の利用の流れ



出力 (拡張マニフェストファイル)

output.manifest

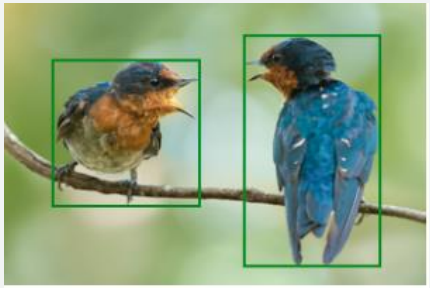

```
{ "source-ref": "s3://sagemaker-groundtruth-demo/SSDB00001.JPG",
  "GroundTruthDemo": {
    "annotations": [
      {"class_id": 0, "width": 54, "top": 482, "height": 39, "left": 337},
      {"class_id": 0, "width": 69, "top": 495, "height": 53, "left": 461},
      {"class_id": 0, "width": 52, "top": 482, "height": 41, "left": 523} ],
    "image_size": [{"width": 1280, "depth": 3, "height": 960} ] },
  "GroundTruthDemo-metadata": {
    "job-name": "labeling-job/groundtruthdemo",
    "class-map": {"0": "Car"},
    "human-annotated": "yes",
    "objects": [
      {"confidence": 0.94},
      {"confidence": 0.94},
      {"confidence": 0.94},],
    "creation-date": "2018-11-26T04:01:09.038134",
    "type": "groundtruth/object-detection" } }
```

Demo

カスタムテンプレート詳細

組み込みラベリングツールの利用も 独自実装も可能

画像分類



物体検出

セマンティック
セグメンテーション

文章分類

カスタム

組み込みラベリングツールの利用も 独自実装も可能

画像分類

物体検出

セマンティック
セグメンテーション

文章分類

カスタム

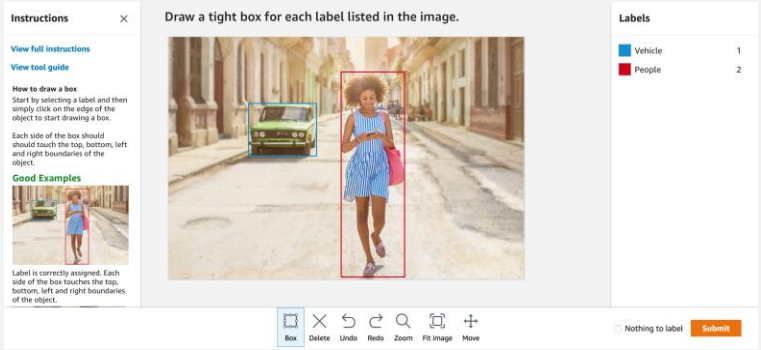
カスタムテンプレート

前・後処理のLambda関数とラベリングツールのHTMLテンプレートを変更することで、様々なタスクに対応できる機能

前処理用
Lambda

ラベリング
HTMLテンプレート

後処理用
Lambda



aws

カスタムテンプレート

前処理用Lambdaの作成

- 入力データの記述されたマニフェストファイルの各項目を読み込み、それをテンプレートエンジンに返す処理を記述

HTMLテンプレート作成

- テンプレートエンジンのLiquidを採用
- 簡単なサンプルも多数用意されている

後処理用Lambda作成

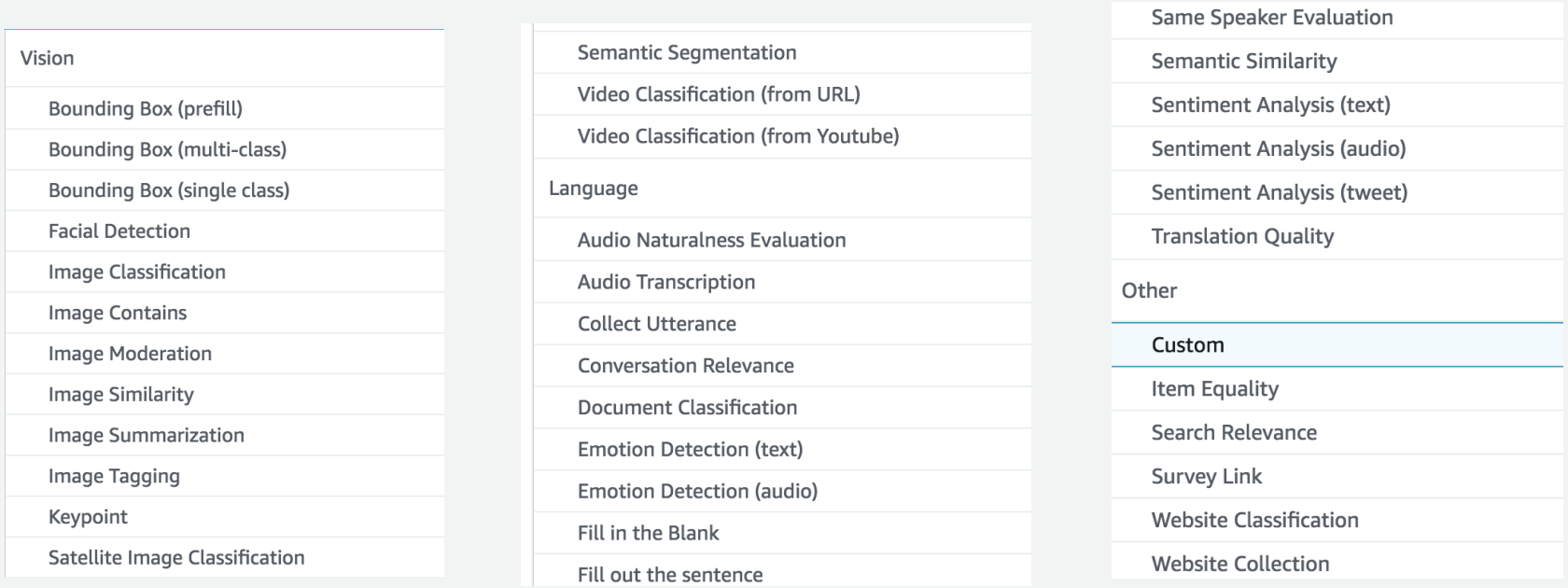
- ワーカーが処理を終了した際の後処理を記述

※ Lambdaはドキュメントのサンプルコードをベースに作成する必要あり
HTMLテンプレートは多数のサンプルの中から選択できる

aws

カスタムテンプレート

HTMLテンプレートのサンプル例



カスタムテンプレート

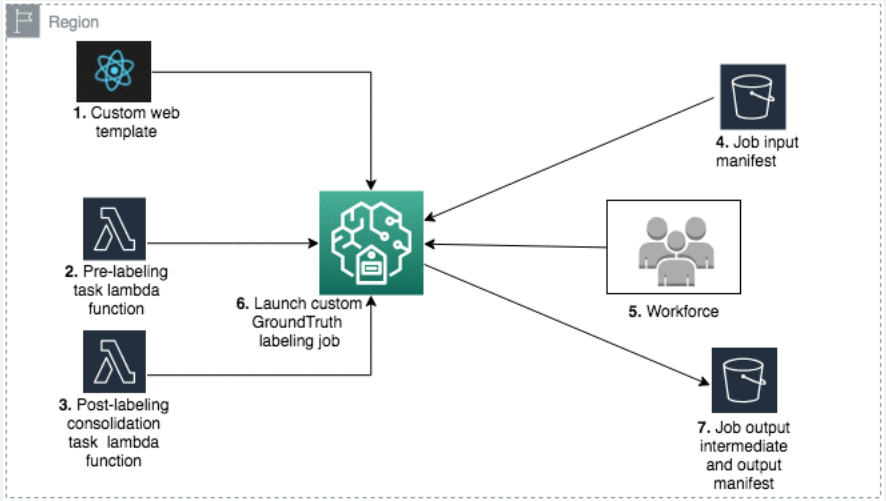
アノテーション対象のデータ形式

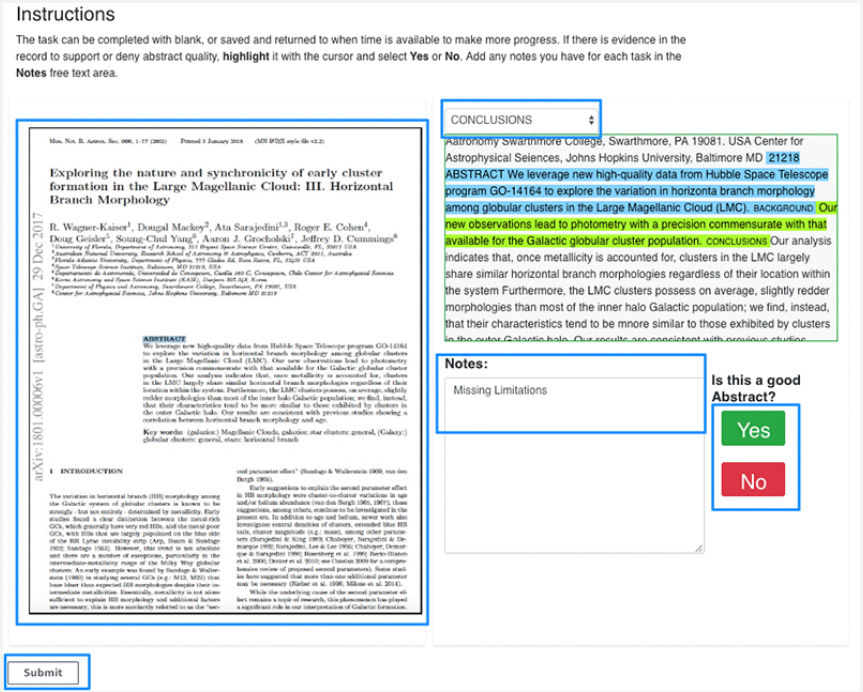
- 画像、映像、音声、文章など様々なファイルをUIに表示できる
- 複数のデータを扱うことも可能 （2画像の類似度推定等）

ラベルの形式

- 既存テンプレートにあるラベル
    - 多クラス分類
    - 画像のセマンティックセグメンテーション
    - 画像の物体検出（カスタムテンプレートであれば複数クラスも可能）
- テキスト（画像に説明文を付与するなど）
- 詳細はドキュメントに

Amazon SageMaker Ground Truth を使ったカスタムデータラベリングワークフローの構築





https://aws.amazon.com/jp/blogs/news/build-a-custom-data-labeling-workflow-with-amazon-sagemaker-ground-truth/

自動ラベリング詳細

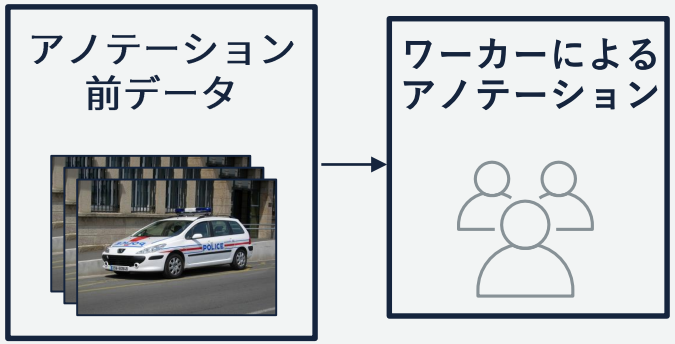
自動ラベリング

データの一部をワーカーがラベル付けするだけで、
残りのラベル付けが自動化され、時間とコストを大幅に削減

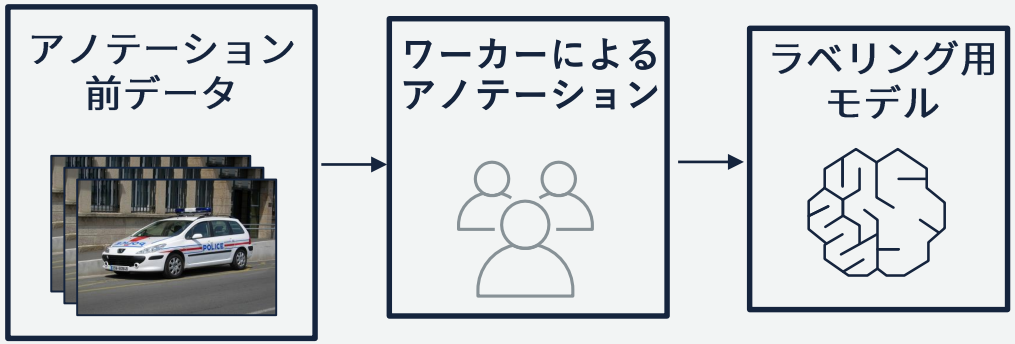
※ 5000データ以上の大規模データセットに対して利用可能なオプション機能

自動ラベリングの仕組み

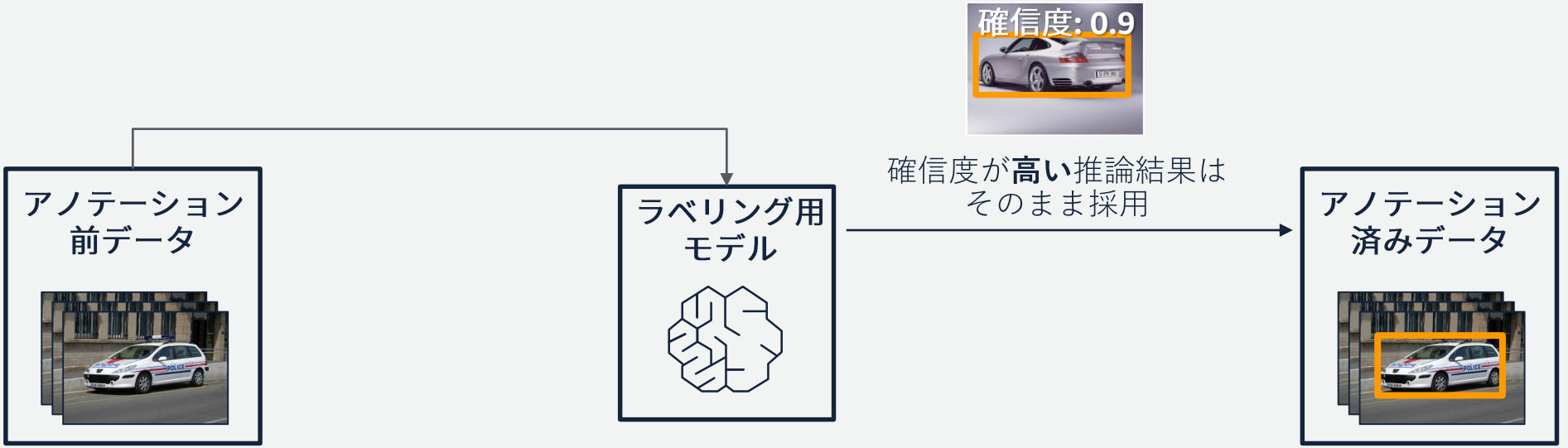
一部のデータ



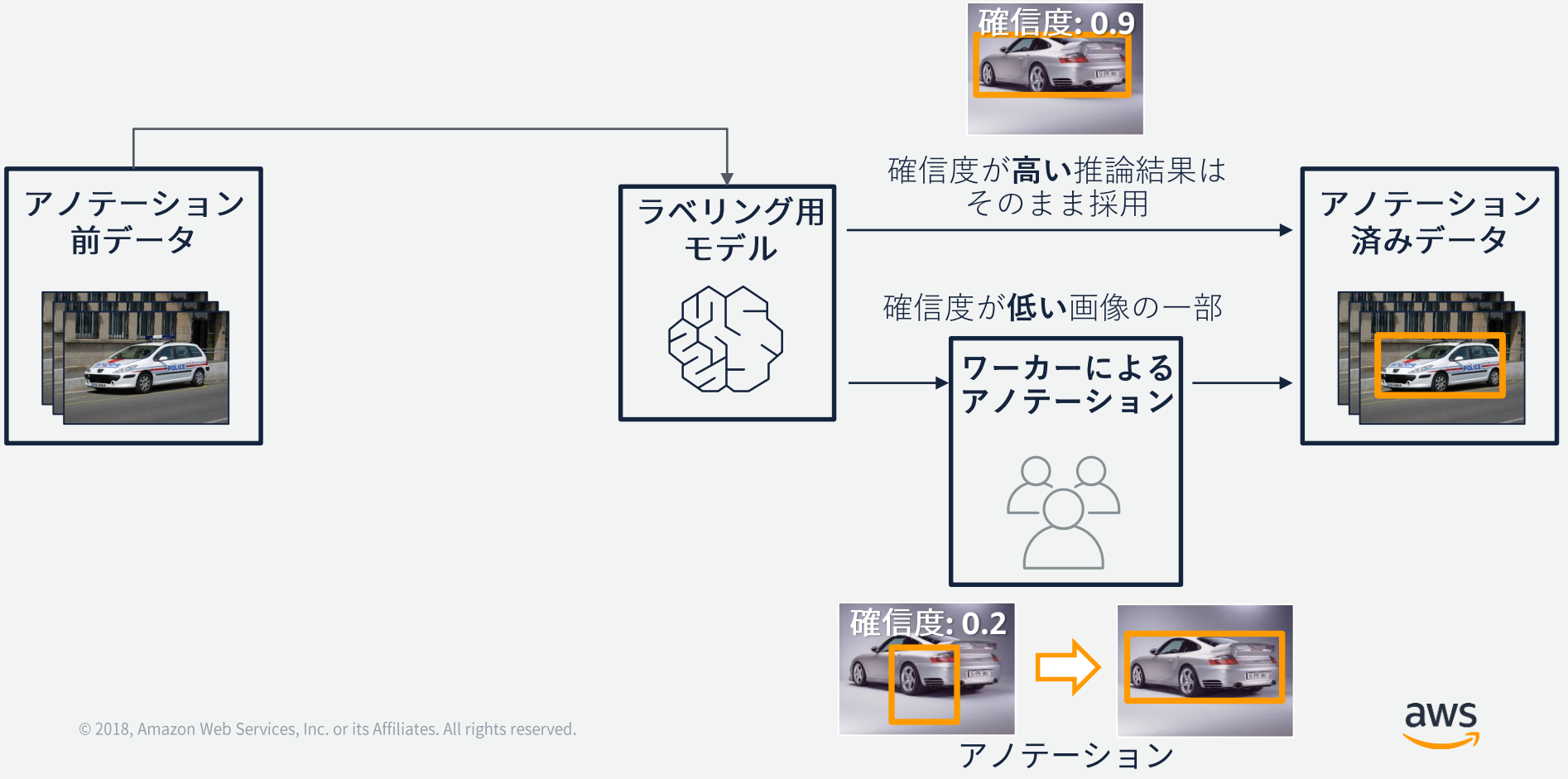
自動ラベリングの仕組み



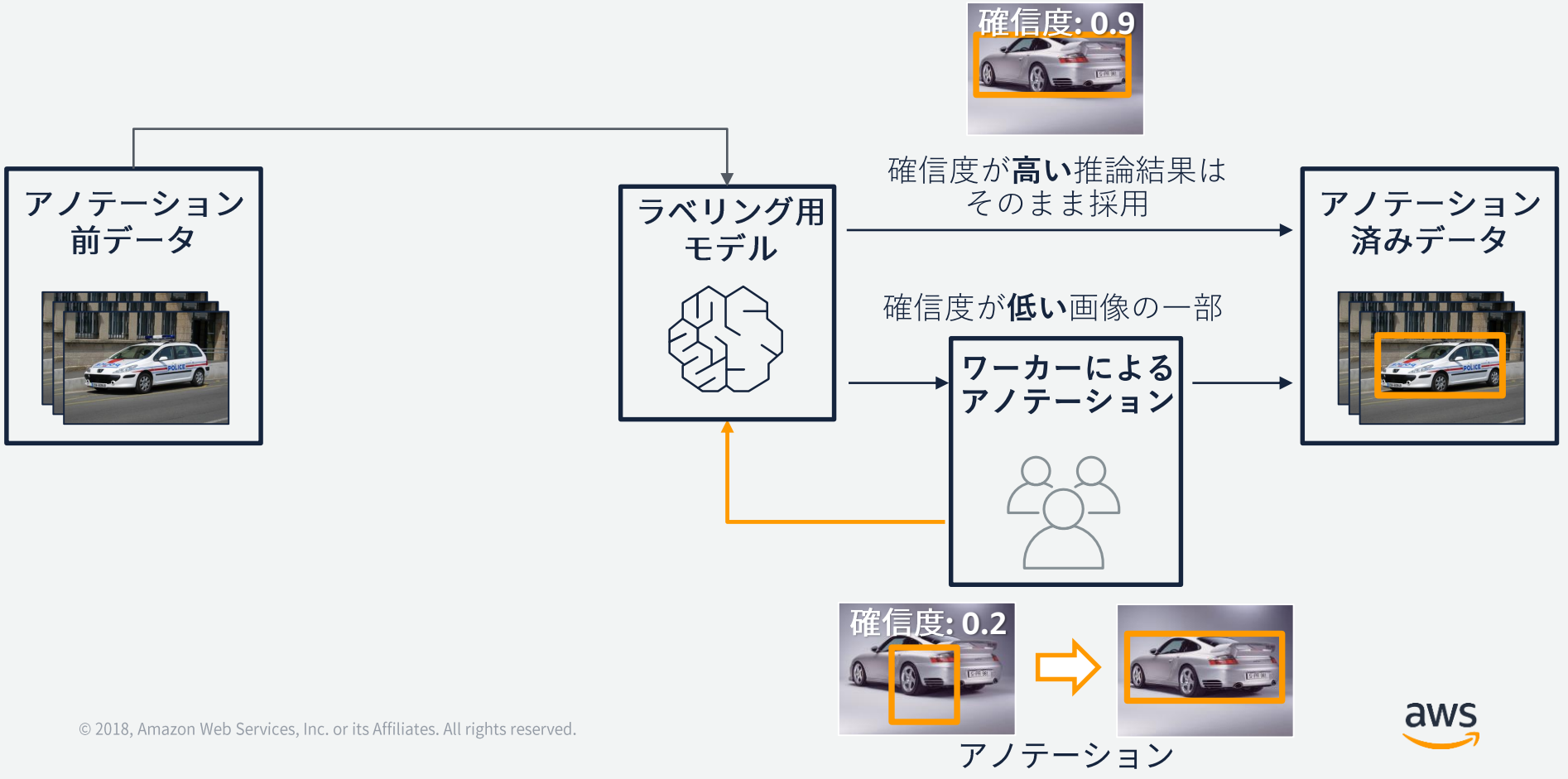
自動ラベリングの仕組み



自動ラベリングの仕組み



自動ラベリングの仕組み



価格と提供リージョン

価格

- ラベル付けした対象の数に応じた利用料　$0.08 / 個 (5万個以上はより安価)
- Amazon Mechanical Turk および 外部ベンダを利用する際の利用料
- 自動ラベリング利用時は、裏で動くSageMakerの学習/推論の利用料

提供リージョン

- バージニア北部 / オレゴン / オハイオ / アイルランド / **東京**

aws

Appendix

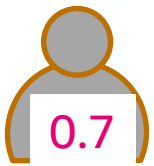
ラベルの決定方法

ラベルの決定 (Label Consolidation): 多数決の場合

bulldog　sharpei　bulldog　bulldog → **bulldog**

ラベルの決定 (Label Consolidation): 正答率による評価

(犬に詳しい人)



bulldog

sharpei

bulldog

bulldog

正しい ラベルを選ぶ確率

bulldog 0.1
sharpei 0.9

- 単純な多数決ではなく、ワーカーの正答率をもとにラベルを決定することでアノテーションの質を担保
- 何人のワーカーの結果をまとめるかはジョブ作成時に指定

aws

パブリックワーカー詳細

パブリックなワーカーを利用する

プライベートとの違いは以下の3つのみ

- ジョブ作成時にワーカータイプ > パブリック を選択
- タスクあたりの料金を設定
- 確認項目にチェック

ワーカーの選択とツールの設定

パブリックなワーカーを利用する

あとはワーカーがアサインされ、タスクが終了するのを待つだけ

パブリックなワーカーを利用する

- タスクに対して設定する料金が安すぎると、
  人が集まらない場合やアノテーションの質が低下する場合あり

- アノテーションの質を向上させるためには、タスクの説明文章の改善や、
  １データにアノテーションするワーカーの数を増やすことで対応

その他

Amazon SageMaker Ground Truth と自動化されたデータのラベル付けによる低コストでのデータのアノテーション

Amazon Web Services ブログ

Amazon SageMaker Ground Truth と自動化されたデータのラベル付けによる低コストでのデータのアノテーション

by Krzysztof Chalupka | on 07 FEB 2019 | in SageMaker | Permalink | Share

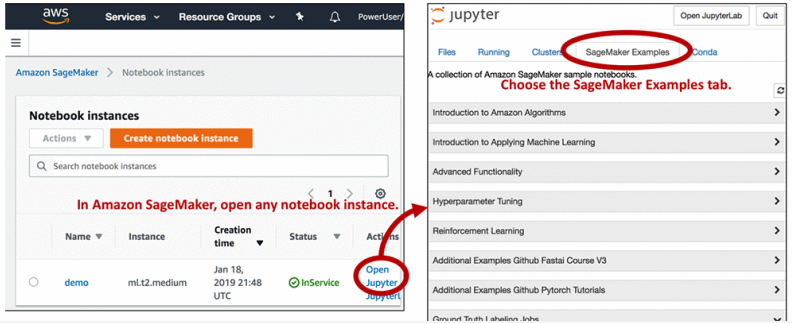
Amazon SageMaker Ground Truth を使うと、正確にラベル付けされた機械学習データセットを簡単に低価格で構築することができます。ラベル付けのコストを削減するために、Ground Truth の機械学習を使用して、人によるアノテーションが必要な「困難な」画像と、機械学習で自動的にラベル付けできる「簡単な」画像を選択します。この記事では、自動化されたデータのラベル付けの仕組みと、その結果の評価方法について説明します。

自動化されたデータのラベル付けを伴う物体検出ジョブを実行する

以前のブログ記事では、Julien Simon が AWS マネジメントコンソールを使ってデータのラベル付けジョブを実行する方法を説明しました。このプロセスをより細かく制御するには、API を使用できます。 その方法をご紹介するため、今回は鳥の画像 1,000 個に対して バウンディングボックスアノテーション を生成する API を使用した Amazon SageMaker Jupyter ノートブックを使用します。

注意: デモノートブックの実行コストは約 200 USD です。

デモノートブックにアクセスするには、ml.m4.xlarge インスタンスタイプを使用して Amazon SageMaker ノートブックインスタンスを開始します。インスタンスは、このステップバイステップチュートリアルに従ってセットアップできます。ステップ 3 では、IAM ロールの作成時に「任意の S3 バケット」にチェックを入れるようにしてください。 以下にあるように、Jupyter ノートブックを開いて [SageMaker Examples] タブを選択し、object_detection_tutorial.ipynb を起動します。



https://aws.amazon.com/jp/blogs/news/annotate-data-for-less-with-amazon-sagemaker-ground-truth-and-automated-data-labeling/

Amazon SageMaker Ground Truth を使用して階層型ラベル分類法を作成する

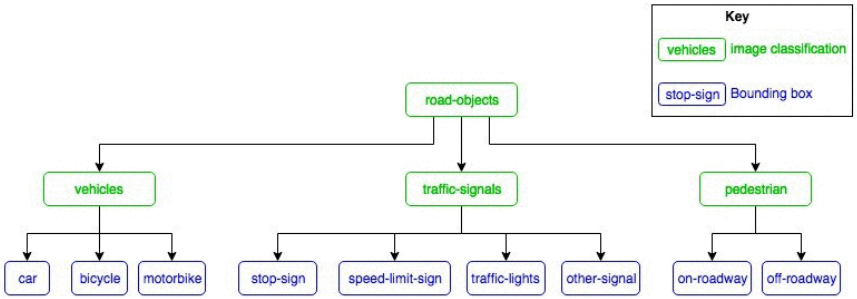
Amazon Web Services ブログ

Amazon SageMaker Ground Truth を使用して階層型ラベル分類法を作成する

by Doug Youd | on 19 FEB 2019 | in SageMaker | Permalink | Share

re:Invent 2018 で、私たちは Amazon SageMaker Ground Truth を発表しました。これは、機械学習を使用して非常に正確なデータセットを構築し、ラベル付けのコストを最大 70% 削減することができます。Amazon SageMaker Ground Truth を使用すると、パブリックおよびプライベートでラベル付けを行う人間の作業者に簡単なアクセスと、一般的なラベル付けタスクのための組み込みのワークフローとインターフェースが提供されます。さらに、Amazon SageMaker Ground Truth は自動データラベル付けを使用してラベル付けのコストを削減します。自動データラベル付けは、人間がラベルを付けたデータから Ground Truth をトレーニングし、サービスが独自にデータにラベルを付けることを学習することによって機能します。

路上でカメラで撮影した画像の大規模なコーパスがあるとしましょう。それぞれの画像には、無人自動車用のアルゴリズムを開発するために重要な多くの異なる対象物 (たとえば、車または交通標識) が含まれている可能性があります。最初に、画像から取得したい情報の階層表現を定義する必要があります (そのようなラベル分類法がどのように見えるかの例については、下記を参照してください)。次に、これらのラベルが付いていない未処理の画像を取得し、高レベルのクラス (「車」、「交通標識」、「歩行者」など) でラベル付けすることによってラベル付けプロセスを開始します。



このブログ記事では、ジョブをチェーンさせて拡張マニフェスト機能を利用することで、Amazon SageMaker Ground Truth を使用してこのような階層的なラベル付けを実現する方法を紹介します。

https://aws.amazon.com/jp/blogs/news/creating-hierarchical-label-taxonomies-using-amazon-sagemaker-ground-truth/

Amazon SageMaker Ground Truth でのラベリングジョブ用の優れた説明の作成

Amazon Web Services ブログ

Amazon SageMaker Ground Truth でのラベリングジョブ用の優れた説明の作成

by Tristan McKinney | on 06 APR 2019 | in SageMaker | Permalink | Share

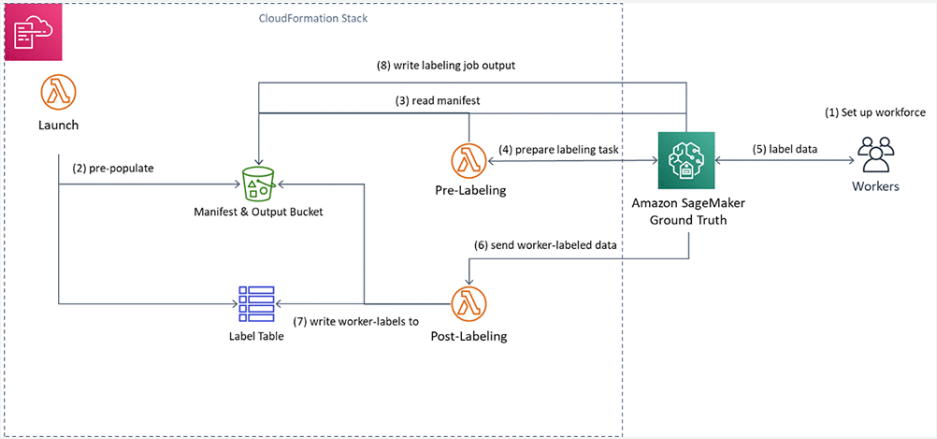
Amazon SageMaker Ground Truthは、機械学習 (ML) 用の高精度なトレーニングデータセットをすばやく構築するお手伝いをします。ご自身のワークフォース、データラベリングに特化したベンダー管理ワークフォースの選択、または Amazon Mechanical Turk が提供するパブリックワークフォースを使用して、人が生成するラベルを提供することができます。質の高いのラベルを取得するには、特にパブリックワークフォースを使用している場合、簡単かつ簡潔で明確な説明が必要です。良い説明が書ければ、アノテーションの品質を向上させることができます。正しく行えば、この作業に時間を費やす価値はあります。

このブログ記事では、パブリックワークフォースに効果的な説明を作成するためのベストプラクティスをご紹介します。ここで重要なポイントが 2 点あります。ワークフォースへの認知負荷をできるだけ減らすこと、そして説明を微調整して後で発生する問題を避けるためにもプロセスの早い段階で実験することです。たとえば実験で、データの一部に自分でラベルを付けたり、プロセス全体の中でも小規模なジョブをパブリックワークフォースに行ってもらうことができます。

以下のスクリーンショットは、ワーカーの観点から見て適切な説明のある Ground Truth のバウンディングボックスのラベリングタスク例を示しています。このタスク例では、Google Open Images Dataset から取得したイメージにある花の周りを囲む四角形の枠を描くようワーカーに伝えます。ワーカーがアノテーションを付けている作業中、イメージの左側にあるサイドバーには短い説明が表示されます。はっきりと要領を得た、かつタスクに特化した説明で、サンプルのイメージに焦点を当てています。

https://aws.amazon.com/jp/blogs/news/create-high-quality-instructions-for-amazon-sagemaker-ground-truth-labeling-jobs/

Amazon SageMaker Ground Truth を使って大量ラベル付けの品質保証を簡単に行う



https://aws.amazon.com/jp/blogs/news/easily-perform-bulk-label-quality-assurance-using-amazon-sagemaker-ground-truth-2/

Amazon SageMaker Ground Truth – A Deep Dive with an Interactive Workshop to Build High-Quality Training Datasets

Take an example: How can we improve

- Task "Draw a bounding box" makes people think.
  "Draw bounding boxes around objects of the specified class in this image."
- It is better if the instructions have an image similar to the annotated ones.
- Panel instruction is too long and should be shorted

Object detection

- Common solution: Use label from a single, trusted worker
- Our approach: Find corresponding boxes based on overlap, and average the coordinates, favoring smaller boxes. Reject unmatched boxes

aws