

NVIDIA Nemo Curator と VLM を活用したデータキュレーションの効率化を検証 重複排除の検証効果を踏まえて AD/ADAS 開発向けの「FastLabel Data Curation」を提供開始

Data-centric AI^[※1]開発に必要なデータ収集・生成からアノテーション、モデル開発、DataOps 構築までの全工程を支援する FastLabel 株式会社（本社：東京都新宿区、代表取締役 CEO：鈴木健史、以下「当社」）は、AD/ADAS 開発向けのソリューション「FastLabel Data Curation」の提供開始をお知らせします。このソリューションは、[NVIDIA Nemo Curator](#) ^[※2] と VLM ^[※3] を活用したデータキュレーションの効率化の検証によって、重複排除とコスト削減の効果が得られています。

[※1] AI のデータセットを改善することでモデルの精度改善を実現するアプローチ

[※2] 高速でスケーラブルなデータセットの準備とキュレーションを行うために設計された Python 製のライブラリ

[※3] Vision Language Model（視覚言語モデル）の略称で画像とテキストを統合的に処理する AI モデルのこと



【データキュレーションにおける重複データ排除の意義】

近年の大規模言語モデル（LLM）などの発展は、大規模かつ高品質なデータセットに支えられています。その準備プロセスであるデータキュレーションは、AI 開発における最も困難で時間のかかる工程の一つです。データキュレーションの中でも特に重複データの排除は重要で、重複データはモデルの学習や評価プロセスに様々な悪影響を及ぼし、以下のような問題を引き起こします。

学習効率の低下と計算資源の浪費：重複データはモデルに同じ内容を繰り返し学習させ、GPU リソースといった貴重な計算資源を浪費させます。これにより、学習コストが不必要に増加します。

本プレスリリースに関するお問い合わせ

FastLabel 株式会社 広報担当 pr@fastlabel.ai

お問い合わせフォーム URL <https://fastlabel.ai/contact-pr>

バイアスの増幅と汎化性能の低下：特定のデータが過剰に重複すると、モデルはその特徴やバイアスを過学習し、出力が偏ってしまいます。結果として、未知のデータに対する汎化性能が著しく低下し、多様な状況に対応できないモデルになります。

評価の信頼性低下（データリーク）：学習データと評価用データが重複していると、モデルの性能スコアが不当に高く算出されるデータリークが発生します。これはモデルの真の実力を見誤らせる深刻な問題であり、実環境での性能不振に直結します。

このように、重複データは学習コスト、モデル品質、評価の信頼性という AI 開発の根幹を揺るがす問題を引き起こすため、大規模データセットから効率的に重複を除去する技術が不可欠です。

【検証のアプローチ】

今回の検証においては、通常想定されるアプローチとは異なる方法で重複排除を実施しました。そもそも Nemo Curator を活用して画像の重複排除を行う際、意味的に似ている画像を重複と判定して除外可能な Semantic Deduplication (SemDeDup) を利用できます。SemDeDup では、データの Embedding（埋め込みベクトル）を使用して、データ間の類似度を計算し、類似度が高いものを重複とみなして除去することができます。今回、生成される Embedding をある程度コントロールするために、通常想定されるアプローチである画像の Embedding ではなく、VLM を用いて抽出した画像キャプションの Embedding を利用して、画像内において指定した特徴に注目して重複を排除するアプローチを行いました（以下、イメージ）。

<イメージ>

	対象	重複判定の方法
通常のアプローチ	実際の画像データ	画像全体の視覚的な類似性に基づく
今回検証したアプローチ	VLM で付与した画像のキャプション	指定の視覚的特徴を反映したキャプション（言語）の類似性に基づく

【検証の内容】

今回、運転者視点で撮影された合計 1,076 枚のドライブレコーダー画像（停車状態 5 パターン、交差点、山間部など多様なシーンを含む）のうち、見た目があまり変わらない 242 枚の画像を対象に検証を行いました。また、重複判定において注目すべき特徴として複数の項目を設定し、これらを抽出できるプロンプトで VLM にキャプションを生成させています。同時に別途プロンプトを作成し、各条件を metadata として json 形式で出力させつつ、重複排除前後で各特徴の割合を確認できるようにしました。これらを前提に以下を行いました。

検証①：キャプションを用いて SemDeDup を実行することで、指定した画像の特徴に注目して重複排除ができるか確認する

検証②：画像の埋め込みを使用した場合には重複排除しやすい「見た目があまり変わらない画像」が、キャプションを使用した場合に重複排除できるか確認する

本プレスリリースに関するお問い合わせ

FastLabel 株式会社 広報担当 pr@fastlabel.ai

お問い合わせフォーム URL <https://fastlabel.ai/contact-pr>

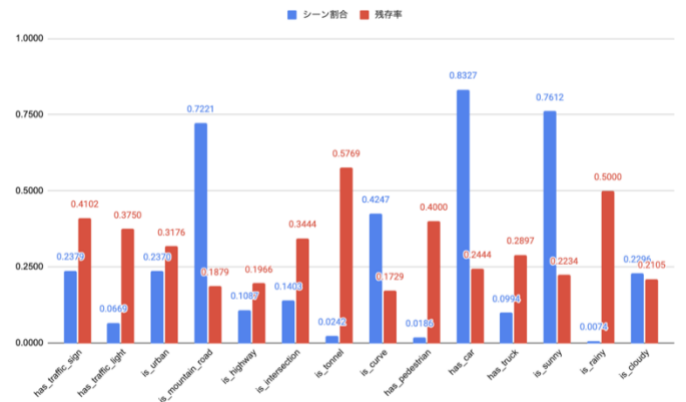
検証内容	データの状態	期待される結果
検証①	特徴 A を持つ画像が多い	重複排除で特徴 A を持つ画像の数が減りやすい
検証①	特徴 A を持つ画像が少ない	重複排除で特徴 A を持つ画像の数が減りにくい
検証②	停車シーンの 242 枚	重複排除の後で、どの程度数が減るかを確認する

【検証の結果】

重複排除の実施後、検証①と検証②で以下の結果が得られました。

＜検証①の結果＞

結果をみると、シーン割合が大きいもののほど残存率が低く、シーン割合が小さいもののほど残存率が高くなる傾向になったことから、キャプションに対し重複判定に用いたい特徴を盛り込むことで、その特徴に注目して重複排除できていることがうかがえます。

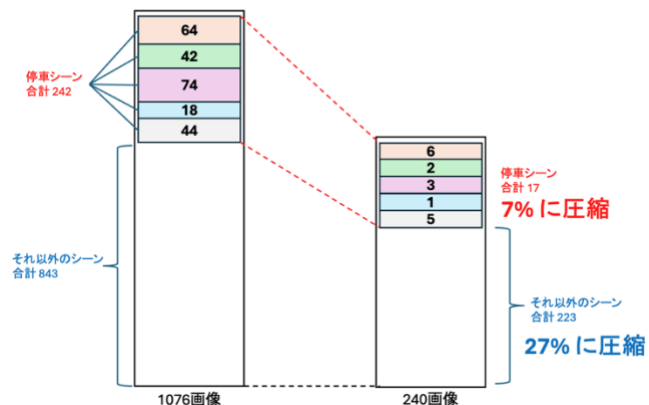


シーン割合（青）：metadata の各項目に対して、重複排除前に画像全体に対して、その項目が含まれる割合

残存率（赤）：重複排除前の各シーンの数に対して、重複排除後に残った画像の割合

＜検証②の結果＞

停車シーンのように、見た目がほぼ同じ画像において、重複排除後は元のデータ数の7%まで9割以上も減少し、停車シーン以外の画像は27%まで減少しました。なお、停車シーンの242枚に限ってシーン割合と残存率を確認したところ、シーンによらず一貫して低残存率となりました。画素的な重複画像に対し同様のキャプションが付与された結果、その類似度が高いため、重複排除がうまくいったと考えられます。

**【AI モデル開発を効率化するソリューション「FastLabel Data Curation」を提供開始】**

Nemo Curator を活用しデータキュレーションを行う今回の検証を通じ、VLM を用いたキャプションングと SemDeDup を組み合わせることで、画像内において指定した特徴に注目した重複排除が可能であることが確認できました。本検証結果を踏まえ、当社では AD/ADAS 開発向けにデータキュレーションの効率化を推進するために Nemo Curator と VLM を活用した AI モデル開発を効率化するソリューション「FastLabel Data Curation」を提供開始し、下記のようなメリットを創出してまいります。

本プレスリリースに関するお問い合わせ

FastLabel 株式会社 広報担当 pr@fastlabel.ai

お問い合わせフォーム URL <https://fastlabel.ai/contact-pr>

<「FastLabel Data Curation」のメリット>

必要なデータがすぐに手に入る、AI データパイプラインの構築を実現します。

- ・ VLM による自動タグ付け：画像やデータの内容から AI が自動タグ付け。面倒なタグ付けや目視による選別作業を撤廃。
- ・ Nemo Curator を活用した高速な重複排除/類似検索：データの意味を理解した高速な重複排除により、学習データ品質と網羅性の担保作業を大幅に効率化。

なお、当社は本日開催された NVIDIA の開発者向けのイベント、NVIDIA AI Day Tokyo のセッションに登壇し、本ソリューションについて初公開しています。ご興味がある方はぜひご相談ください。

【当社事業について】

当社は「データセット提供」、「アノテーション代行」、「モデル開発支援」、「FastLabel Data Factory」など、AI 開発を行うお客様に向けたトータルソリューションを提供しています。

データセット提供：権利クリアかつ高品質な各種データの収集、販売を行います。ストックの提供や新規撮り下ろしにも対応しています。

アノテーション代行：あらゆる非構造化データに対応し、弊社独自の品質管理によりデリバリー品質を担保しています。ドメイン知識が必要な仕様にも対応可能です。

モデル開発支援：画像や動画の撮影条件や正確性／統一性を読み解きながらモデルの学習・評価を実施し、評価結果を精度向上につなげていきます。

FastLabel Data Factory：データ収集・管理、アノテーション、モデル開発までワンストップで提供する SaaS です。DataOps 構築を実現し教師データ準備を大幅に効率化できます。

当社は各業界リーディング企業への豊富な支援実績を有し、権利クリアかつ高品質なデータ作成に強みを置いております。豊富な経験を通して培ったアノテーション仕様作成をはじめとした AI 開発のノウハウで、今後も多くのお客様を支援してまいります。

【当社の概要について】

社 名：FastLabel 株式会社

代 表 者：代表取締役 CEO 鈴木健史

事 業 内 容：Data-centric AI 開発を支援するプロフェッショナルサービスとプロダクトの提供

設 立：2020 年 1 月 23 日

本 社 所 在 地：〒163-0224 東京都新宿区西新宿 2-6-1 新宿住友ビル 24 階

U R L：<https://fastlabel.ai/>

本プレスリリースに関するお問い合わせ

FastLabel 株式会社 広報担当 pr@fastlabel.ai

お問い合わせフォーム URL <https://fastlabel.ai/contact-pr>