ネットアップの AI による地合い分析

NetApp Solutions

NetApp
April 10, 2024

目次

ネットアップの AI による地合い分析

TR-4910 ：『 NetApp AI と顧客コミュニケーションを組み合わせた感情分析』

Sathish Thyagarajan 、 Rick Huang 氏、および SFL Scientific 、 Diego Sosa-coba 、 David Arnette 氏

このテクニカルレポートでは、転送学習と会話型 AI を使用して、ネットアップのデータ管理テクノロジと NVIDIA ソフトウェアフレームワークを使用して、エンタープライズレベルのグローバルサポートセンターで感情分析を行うための設計ガイダンスを提供します。この解決策は、チャットログ、 E メール、およびその他のテキストまたは音声通信を表す録音された音声ファイルやテキストファイルから顧客の洞察を得たいと考えているあらゆる業界に適用されます。ネットアップはエンドツーエンドのパイプラインを実装して、ネットアップのクラウド対応オールフラッシュストレージを使用した GPU アクセラレーションコンピューティングクラスタで、自動音声認識、リアルタイムの感情分析、ディープラーニングの自然言語処理モデル再トレーニング機能をデモンストレーションしました。大規模で最先端の言語モデルのトレーニングと最適化により、世界規模のサポートセンターで推論を迅速に実行できるようになり、優れたカスタマーエクスペリエンスと目標を達成し、長期的な従業員パフォーマンス評価を実施できます。

感情分析は、正、負、または中性感情がテキストから抽出される Natural Language Processing （ NLP ）内の研究分野です。会話型 AI システムは、より多くの人がコミュニケーションを行うようになったため、ほぼグローバルレベルの統合にまで成長しました。感情分析には、サポートセンターの従業員のパフォーマンスを発信者との会話で決定し、適切な自動チャットボット応答を提供し、四半期ごとの収益呼における企業の代表者と対象者間のやり取りに基づいて会社の株価を予測するなど、さまざまなユースケースがあります。さらに、感情分析を使用して、ブランドが提供する製品、サービス、サポートに関するお客様の見解を判断できます。

このエンドツーエンドの解決策は、 NLP モデルを使用して、サポートセンター分析フレームワークを可能にする高度なセンチメント分析を実行します。音声録音は文書化されたテキストに処理され、会話の各文から感情が抽出されます。結果はダッシュボードに集約され、会話の感情を分析するために、従来とリアルタイムの両方で巧妙に細工することができます。この解決策は、データモダリティと出力ニーズが似ている他のソリューションに汎用化できます。適切なデータを使用することで、他のユースケースにも対応できます。たとえば、企業収益の問い合わせを、同じエンドツーエンドパイプラインを使用して、センチメントについて分析することができます。また、パイプラインの柔軟性が高いため、トピックモデリングや Named Entity Recognition （ NER ）などの他の形式の NLP 解析も可能です。

これらの AI 実装は、 NVIDIA Rivea 、 NVIDIA TAO Toolkit 、 NetApp DataOps ツールキットが連携して実現しました。NVIDIA のツールを使用すると、あらかじめ組み込まれたモデルとパイプラインを使用して、ハイパフォーマンスな AI ソリューションを迅速に導入できます。NetApp DataOps ツールキットにより、さまざまなデータ管理タスクが簡易化され、開発期間が短縮されます。

お客様にもたらされる価値

企業は、感情分析のためのテキスト、音声、ビデオの会話について、従業員評価および顧客対応ツールから価値を得ています。マネージャーは、ダッシュボードに表示される情報を活用して、会話の両側に基づいて従業員と顧客満足度を評価できます。

さらに、 NetApp DataOps ツールキットは、お客様のインフラストラクチャ内でのデータのバージョン管理と割り当てを管理します。その結果、ダッシュボードに表示される分析情報が頻繁に更新されるため、データストレージのコストを抑えることができません。

ユースケース

これらのサポートセンターで処理されるコールの数が原因で、手動で実行した場合はコールパフォーマンスの評価にかなりの時間がかかる可能性があります。BAG of Words カウンティングなどの従来のメソッドは、いくつかの自動化を実現できますが、これらのメソッドは、ダイナミック言語の微妙な側面や意味をキャプチャしません。AI モデリング手法を使用すると、このように詳細な分析を自動化された方法で実行できます。さらに、 NVIDIA 、 AWS 、 Google などが公開している最新のトレーニング済みモデリングツールを使用することで、複雑なモデルを含むエンドツーエンドのパイプラインを容易に構築し、カスタマイズできるようになりました。

サポートセンターの感情分析のためのエンドツーエンドのパイプラインは、従業員が発信者と会話するときに、音声ファイルをリアルタイムで取り込みます。次に、これらのオーディオファイルは音声テキストコンポーネントで使用するために処理され、テキスト形式に変換されます。会話中の各文は、感情（肯定的、否定的、または中立的）を示すラベルを受け取ります。

感情分析は、コールパフォーマンスを評価するための会話の重要な側面を提供することができます。これらの感情は、従業員と発信者間のやり取りにさらに深いレベルを追加します。AI を活用した感情ダッシュボードは、マネージャーが会話内の感情をリアルタイムに追跡し、従業員の過去の問い合わせを過去に分析します。

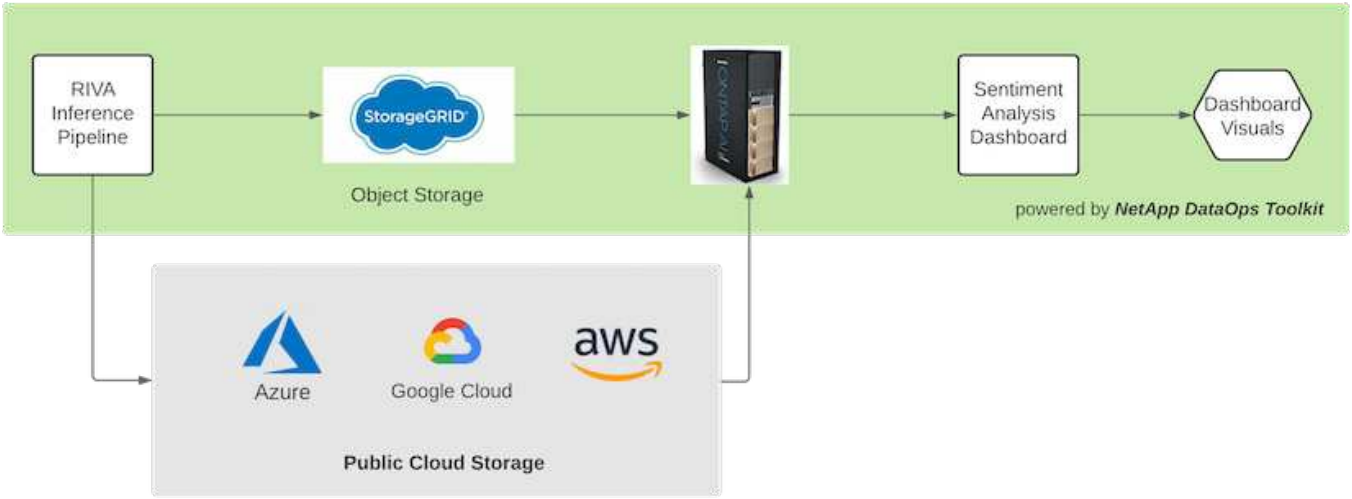
この問題を解決するエンドツーエンドの AI パイプラインを迅速に構築するための強力な方法が用意されています。この場合、 NVIDIA Riva ライブラリを使用して、音声変換と感情分析の 2 つの直列タスクを実行できます。1 つ目は教師あり学習信号処理アルゴリズムで、 2 つ目は教師あり学習 NLP 分類アルゴリズムです。NVIDIA TAO Toolkit を使用すれば、ビジネス関連のデータを使用して、関連するあらゆるユースケースに合わせてアルゴリズムを微調整できます。その結果、コストとリソースの数分の 1 に過ぎず、より正確で強力なソリューションを構築できます。お客様はを組み込むことができます "NVIDIA Maxine の 2 つのポートが" サポートセンター設計における GPU アクセラレーションビデオ会議アプリケーションのフレームワーク。

この解決策の中核をなすのは、次のユースケースです。どちらのユースケースでも、 TAIO ツールキットを使用してモデルの微調整を行い、 Rivea を使用してモデルを展開します。

- 音声テキスト
- 感情分析

従業員と顧客の間のサポートセンターのやり取りを分析するために、音声コールの形式で各顧客との会話をパイプラインを通じて実行し、文レベルの感情を抽出できます。そのような感情は、人間が感情を正当化するか、必要に応じて調整することができます。次に、ラベル付けされたデータが微調整ステップに渡され、感情の予測が改善されます。ラベル付きの感情データがすでに存在する場合は、モデルの微調整を迅速に行うことができます。どちらの場合も、パイプラインは音声の取り込みと文章の分類を必要とする他のソリューションに対して一般化可能です。

AI の感情に関するアウトプットは、外部クラウドデータベースまたは企業が管理するストレージシステムにアップロードされます。この大規模なデータベースからローカルストレージにセンチメント出力が転送され ' 管理者のセンチメント分析を表示するダッシュボード内で使用されますダッシュボードの主な機能は、カスタマーサービスのスタッフとリアルタイムで連携することです。マネージャは、コール中の従業員に関する評価やフィードバックを行い、各文章の感情を最新の状態に更新したり、従業員の過去のパフォーマンスや顧客からの反応を履歴的に確認したりすることができます。



。 "NetApp DataOps ツールキット" Riva 推論パイプラインが感情ラベルを生成した後も、データストレージシステムの管理を継続できます。これらの AI 分析結果は、 NetApp DataOps ツールキットで管理するデータストレージシステムにアップロードできます。データストレージシステムは、数百ものインサートを管理し、毎分選択できる能力を備えている必要があります。ローカルデバイスストレージシステムは、大容量のデータストレージをリアルタイムで照会して抽出します。大規模なデータストレージインスタンスを照会して履歴データを照会することで、ダッシュボードのエクスペリエンスを強化することもできます。NetApp DataOps ツールキットを使用すると、データを迅速にクローニングし、データを使用するすべてのダッシュボードにデー

タを分散できるため、この 2 つの方法を簡単に使用できます。

対象読者

解決策の対象となるグループは次のとおりです。

- 従業員のマネージャー
- データエンジニア / データサイエンティスト
- IT 管理者（オンプレミス、クラウド、ハイブリッド）

会話中に感情を追跡することは、従業員のパフォーマンスを評価するための貴重なツールです。AI ダッシュボードを使用することで、マネージャーは従業員と発信者が自分の感情をリアルタイムでどのように変化させるかを確認できるため、ライブ評価やガイダンスセッションが可能になります。さらに、音声会話、テキストチャットボット、ビデオ会議に参加しているお客様から、価値ある顧客インサイトを得ることができます。このような顧客分析では、最新の AI モデルとワークフローを使用して、大規模なマルチモーダル処理の機能を活用しています。

データ側では、多数のオーディオファイルがサポートセンターによって毎日処理されます。NetApp DataOps ツールキットを使用すると、モデルの定期的な微調整と感情分析用ダッシュボードの両方で、このデータ処理タスクを容易に行うことができます。

IT 管理者は、 NetApp DataOps ツールキットを利用して、導入環境と本番環境の間でデータを迅速に移動することもできます。また、リアルタイム推論のためには、 NVIDIA 環境とサーバも管理、分散する必要があります。

アーキテクチャ

このサポートセンターの解決策のアーキテクチャは、 NVIDIA が構築したツールと NetApp DataOps ツールキットを中心にしています。NVIDIA のツールを使用すると、構築済みのモデルとパイプラインを使用して、ハイパフォーマンスな AI ソリューションを迅速に導入できます。NetApp DataOps ツールキットにより、さまざまなデータ管理タスクが簡易化され、開発期間が短縮されます。

解決策テクノロジ

"NVIDIA RIVA" GPU でリアルタイムのパフォーマンスを実現する、マルチモーダルな会話型 AI アプリケーションを構築するための GPU アクセラレーション対応 SDK です。NVIDIA Train 、 Adapt 、 Optimize （ TAO ）ツールキットは、トレーニングを高速化し、高精度で高性能なドメイン固有の AI モデルをすばやく簡単に作成する方法を提供します。

NetApp DataOps ツールキットは Python ライブラリで、開発者、データサイエンティスト、 DevOps エンジニア、データエンジニアはさまざまなデータ管理タスクを簡単に実行できます。これには、新しいデータボリュームまたは JupyterLab ワークスペースのほぼ瞬時のプロビジョニング、データボリュームまたは JupyterLab ワークスペースのほぼ瞬時のクローニング、データボリュームまたは JupyterLab ワークスペースのほぼ瞬時の Snapshot コピーによるトレーサビリティとベースライン設定が含まれます。

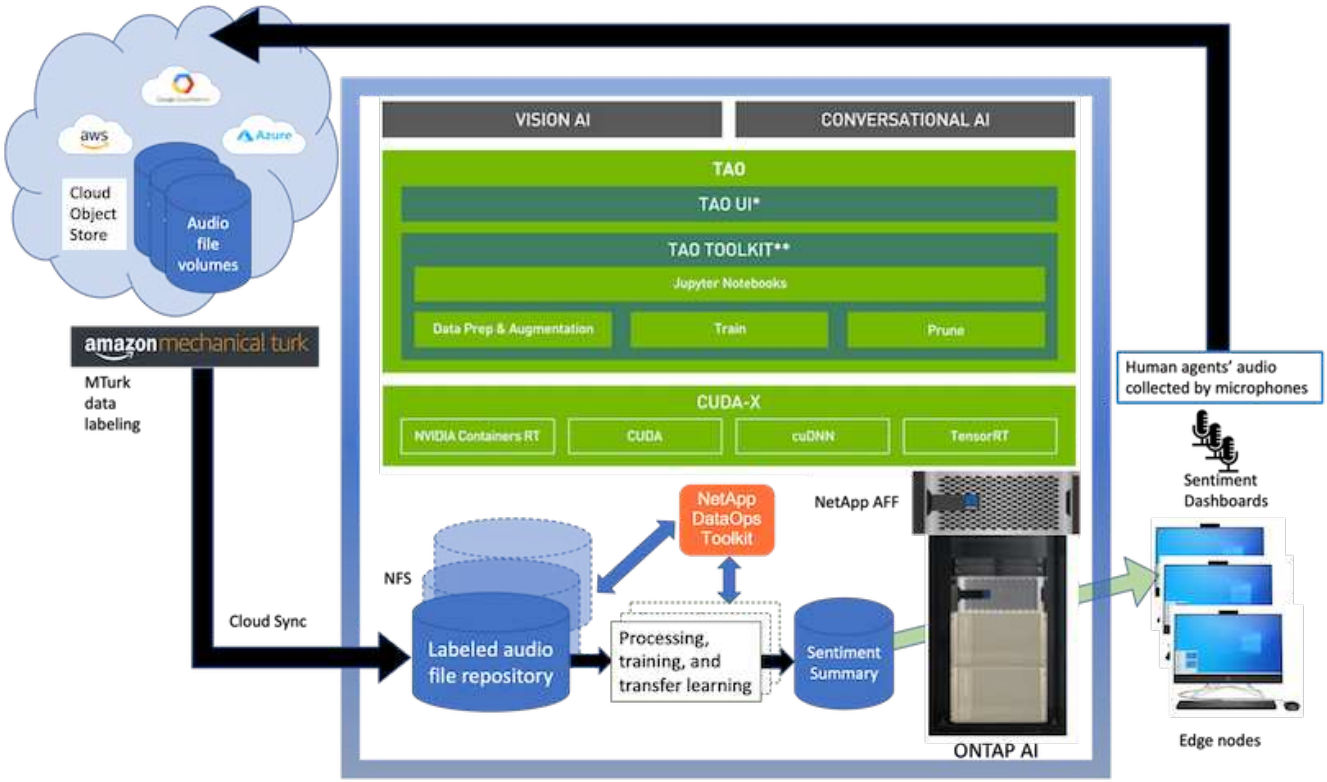
アーキテクチャ図

次の図は、解決策のアーキテクチャを示しています。環境には、クラウド、コア、エッジの 3 つのカテゴリ

があります。各カテゴリは地理的に分散させることができます。たとえば、クラウドにはさまざまなリージョンのバケットにオーディオファイルが格納されたオブジェクトストアが含まれていますが、コアには高速ネットワークやNetApp BlueXPのコピーと同期でリンクされたデータセンターが含まれている場合があります。エッジノードは、ヒューマンエージェントの日常的な作業プラットフォームを表しています。このプラットフォームでは、対話型ダッシュボードツールとマイクを使用して感情を視覚化したり、顧客との会話から音声データを収集したりできます。

GPU によって高速化されたデータセンターでは、 NVIDIA を使用できます "リバ" 会話型 AI アプリケーションを構築するためのフレームワーク。それには、があります "Tao ツールキット" Transfer L ラーニング技術を使用して、モデルのフィニッチニングと再トレーニングを接続します。これらのコンピューティングアプリケーションとワークフローは、を基盤としています "NetApp DataOps ツールキット"ONTAP が提供する最高のデータ管理機能を実現します。このツールキットを使用すると、企業のデータチームは、スナップショットやクローンを使用して、構造化データと非構造化データでモデルのプロトタイプを迅速に作成できるため、トレーサビリティ、バージョン管理、 A/B テストを実現し、セキュリティ、ガバナンス、 コンプライアンスを実現できます。を参照してください "ストレージ設計" 詳細：

この解決策では、オーディオファイル処理、 NLP モデルトレーニング、トランスファーラーニング、およびデータ管理の詳細な手順について説明します。最終的なパイプラインが生成され、ヒューマンサポートエージェントのダッシュボードにリアルタイムで表示されるセンチメントの概要が生成されます。



ハードウェア要件

次の表に、解決策の実装に必要なハードウェアコンポーネントを示します。解決策の特定の実装で使用されるハードウェアコンポーネントは、お客様の要件に応じて異なる場合があります。

応答遅延テスト	時間（ミリ秒）
データ処理	10.
推論	10.

この応答時間テストは、 560 の会話で 50,000 以上のオーディオファイルで実行されました。各オーディオファイルのサイズは、 MP3 の場合は約 100 KB 、 WAV の場合は約 1 MB でした。データ処理手順では、 MP3 を WAV ファイルに変換します。推論の手順では、オーディオファイルをテキストに変換し、テキストから感情を抽出します。これらのステップは互いに独立しており、並列化することでプロセスを高速化できます。

ストア間でのデータ転送の遅延を考慮すると、マネージャは、文章の最後の 2 番目の時間内にリアルタイムの感情分析の更新を確認できるようになります。

NVIDIA Riva ハードウェア

ハードウェア	要件
OS	Linux x86_64
GPU メモリ（ASR）	ストリーミングモデル：最大 5600 MB の非ストリーミングモデル：約 3100 MB
GPU メモリ（NLP）	1 つの BERT モデルで最大 500MB

NVIDIA TAO ツールキットハードウェア

ハードウェア	要件
システム RAM	32 GB
GPU RAM	32 GB
CPU	8 コア
GPU	NVIDIA （A100 、 V100 、 RTX 30x0）
SSD の場合	100GB

フラッシュストレージシステム

NetApp ONTAP 9.

ネットアップの最新世代のストレージ管理ソフトウェア ONTAP 9.9 は、インフラの刷新とクラウド対応データセンターへの移行を可能にします。ONTAP は、業界をリードするデータ管理機能を活用して、データの格納場所に関係なく、単一のツールセットでデータの管理と保護を実現します。エッジ、コア、クラウドなど、必要な場所に自由にデータを移動することもできます。ONTAP 9.9 には、データ管理を簡素化し、重要なデータを高速化および保護し、ハイブリッドクラウドアーキテクチャ全体で次世代のインフラ機能を実現する、多数の機能が含まれています。

NetApp BlueXPのコピーと同期

"BlueXPのコピーと同期" は、高速でセキュアなデータ同期を実現するネットアップのサービスです。オンプレミスの NFS または SMB ファイル共有間で、次のいずれかのターゲットにファイルを転送できます。

- NetApp StorageGRID
- NetApp ONTAP S3
- NetApp Cloud Volumes Service の略
- Azure NetApp Files の特長
- Amazon Simple Storage Service （Amazon S3）

- Amazon Elastic File System （Amazon EFS ）
- Azure Blob の略
- Google クラウドストレージ
- IBM クラウドオブジェクトストレージ

BlueXPのCopy and Syncは、必要な場所に迅速かつ安全にファイルを移動します。転送されたデータは、ソースとターゲットの両方で完全に使用できます。BlueXPのCopy and Syncは、事前定義されたスケジュールに基づいてデータを継続的に同期し、差分のみを移動するため、データレプリケーションにかかる時間とコストを最小限に抑えることができます。BlueXPのCopy and Syncは、セットアップと使用が簡単なソフトウェアサービス（SaaS）ツールです。BlueXPのCopyとSyncによってトリガーされるデータ転送は、データブローカーによって実行されます。BlueXPのCopy and Syncデータブローカーは、AWS、Azure、Google Cloud Platform、オンプレミスに導入できます。

NetApp StorageGRID

StorageGRID の Software-Defined オブジェクトストレージスイートは、パブリッククラウド、プライベートクラウド、ハイブリッドマルチクラウド環境のすべてをシームレスにサポートし、幅広いユースケースに対応しています。業界をリードするイノベーションにより、 NetApp StorageGRID は、非構造化データを長期にわたって自動化されたライフサイクル管理などの多目的に保管、保護、保管します。詳細については、を参照してください "NetApp StorageGRID" サイト

ソフトウェア要件

次の表に、この解決策を実装するために必要なソフトウェアコンポーネントを示します。解決策の特定の実装で使用されるソフトウェアコンポーネントは、お客様の要件に応じて異なる場合があります。

ホストマシン	要件
Riva (以前の開発コード名 Jarv)	1.4.0
Tao ツールキット (以前の Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
DGX OS	5.1
DTK	2.0.0

NVIDIA Riva ソフトウェア

ソフトウェア	要件
Docker です	>19.02 （ NVIDIA - Docker をインストール済み） >=19.03 （ DGX を使用していない場合
NVIDIA ドライバ	465.19.01 + 418.40 + 、 440.33 + 、 450.51 + 、 460.27 + （データセンターの GPU の場合
コンテナ OS	Ubuntu 20.04
CUDA （ CUDA	11.3.0
cuBLAS	11.5.1.101
cuDNN	8.2.0.41

ソフトウェア	要件
NCCL	2.9.6
TensorRT	7.2.3.4.
Triton Inference サーバ	2.9.0

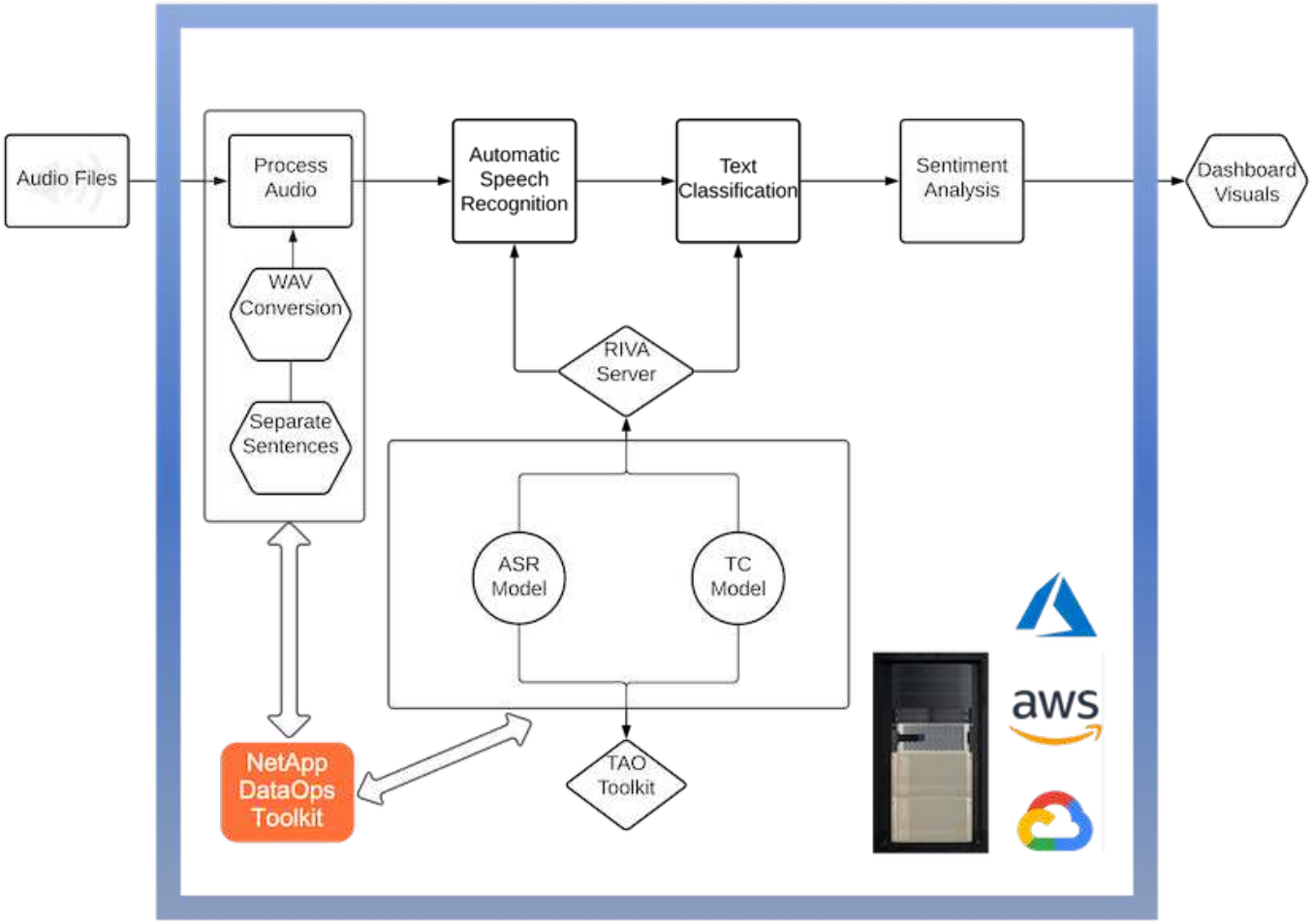
NVIDIA TAO ツールキットソフトウェア

ソフトウェア	要件
Ubuntu 18.04 LTS	18.04
Python	3.6.9 以上
Docker - CE	19.03.5
Docker - API	1.40
nvidia -container-toolkit	>1.3.0-1
nvidia Container - ランタイム	3.4.0 -1
nvidia - docker2	2.5.0-1
nVidia ドライバ	> 455
python-pip	>21.06
nvidia -pyindex	最新バージョン

ユースケースの詳細

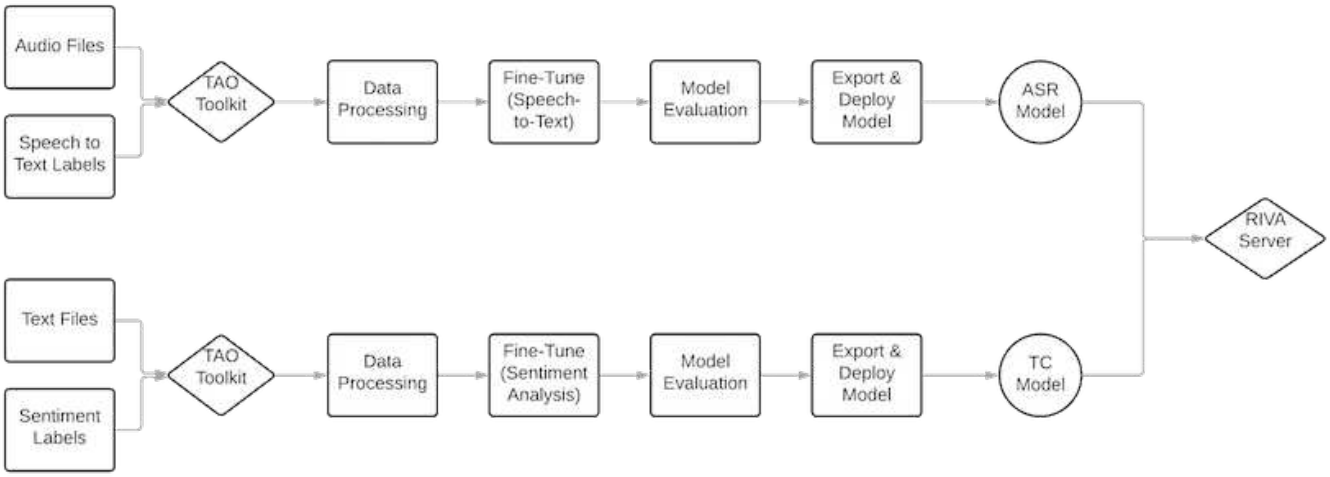
この解決策環境のユースケースは次のとおりです。

- 音声テキスト
- 感情分析



音声テキスト変換のユースケースは、まずサポートセンターの音声ファイルを取り込むことから始まります。このオーディオは、 Riva が必要とする構造に合わせて処理されます。オーディオファイルが解析単位に分割されていない場合は、オーディオを Riva に渡す前にこれを行う必要があります。オーディオファイルが処理されると、 API 呼び出しとして Riva サーバーに渡されます。サーバは、ホスティングしている多くのモデルの 1 つを採用し、応答を返します。この音声 / テキスト（自動音声認識の一部）は、音声のテキスト表現を返します。そこから、パイプラインはセンチメント分析部分に切り替わります。

感情分析では、自動音声認識からのテキスト出力がテキスト分類への入力として機能します。Text Classification は、任意の数のカテゴリにテキストを分類するための NVIDIA コンポーネントです。サポートセンターとの会話では、感情のカテゴリがプラスからマイナスになります。モデルのパフォーマンスは、ホールドアウトセットを使用して、微調整ステップの成功を判断することができます。



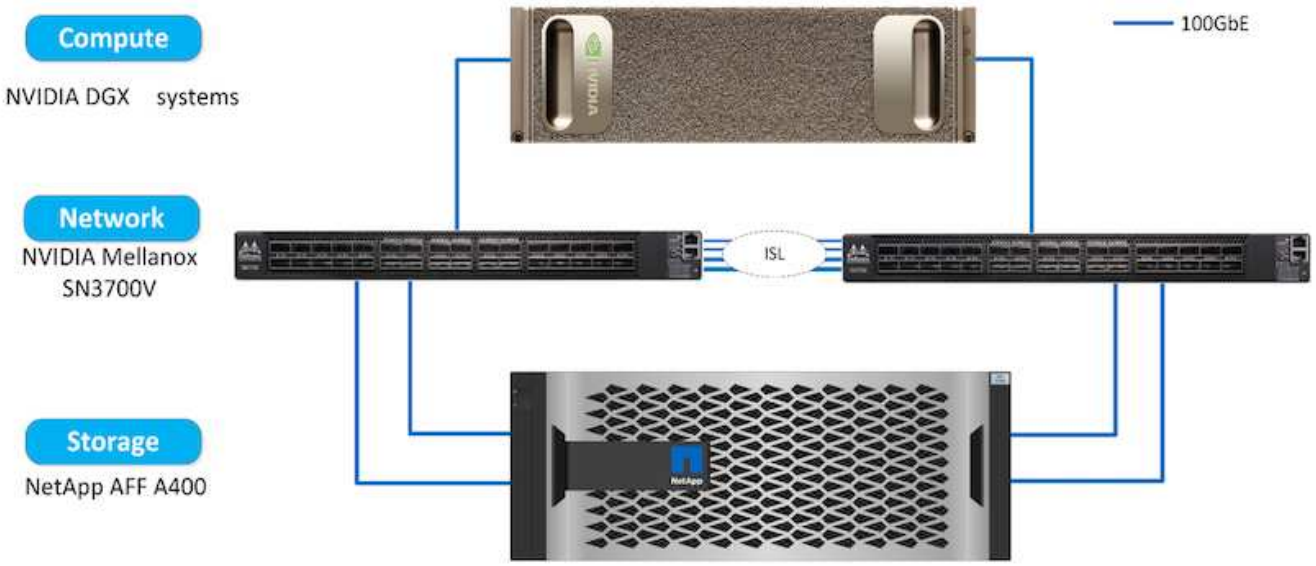
TAO ツールキット内の音声テキスト分析と感情分析にも、同様のパイプラインが使用されています。主な違いは、モデルの微調整に必要なラベルの使用です。TAO ツールキットパイプラインは、データファイルの処理から始まります。次に、事前にトレーニングされたモデル（から入手可能 "NVIDIA NGC カタログ"）は、サポートセンターのデータを使用して微調整されます。微調整されたモデルは、対応するパフォーマンス指標に基づいて評価され、事前トレーニングされたモデルよりもパフォーマンスが高い場合は、 Riva サーバに導入されます。

設計上の考慮事項

このセクションでは、この解決策のさまざまなコンポーネントの設計上の考慮事項について説明します。

ネットワークとコンピューティングの設計

データセキュリティの制限に応じて、すべてのデータはお客様のインフラストラクチャまたはセキュアな環境内に保持されている必要があります。



ストレージ設計

NetApp DataOps ツールキットは、ストレージシステムを管理するための主要なサービスです。DataOps ツールキットは Python ライブラリで、開発者、データサイエンティスト、 DevOps エンジニア、データエンジニアは、新しいデータボリュームや JupyterLab ワークスペースのほぼ瞬時のプロビジョニング、データボリュームや JupyterLab ワークスペースのほぼ瞬時のクローニングなど、さまざまなデータ管理タスクを簡単に実行できます。 トレーサビリティやベースライン設定のためのデータボリュームまたは JupyterLab ワークスペースのほぼ瞬時のスナップショット作成。この Python ライブラリは、任意の Python プログラムまたは Jupyter Notebook にインポートできるコマンドラインユーティリティまたは関数ライブラリとして機能します。

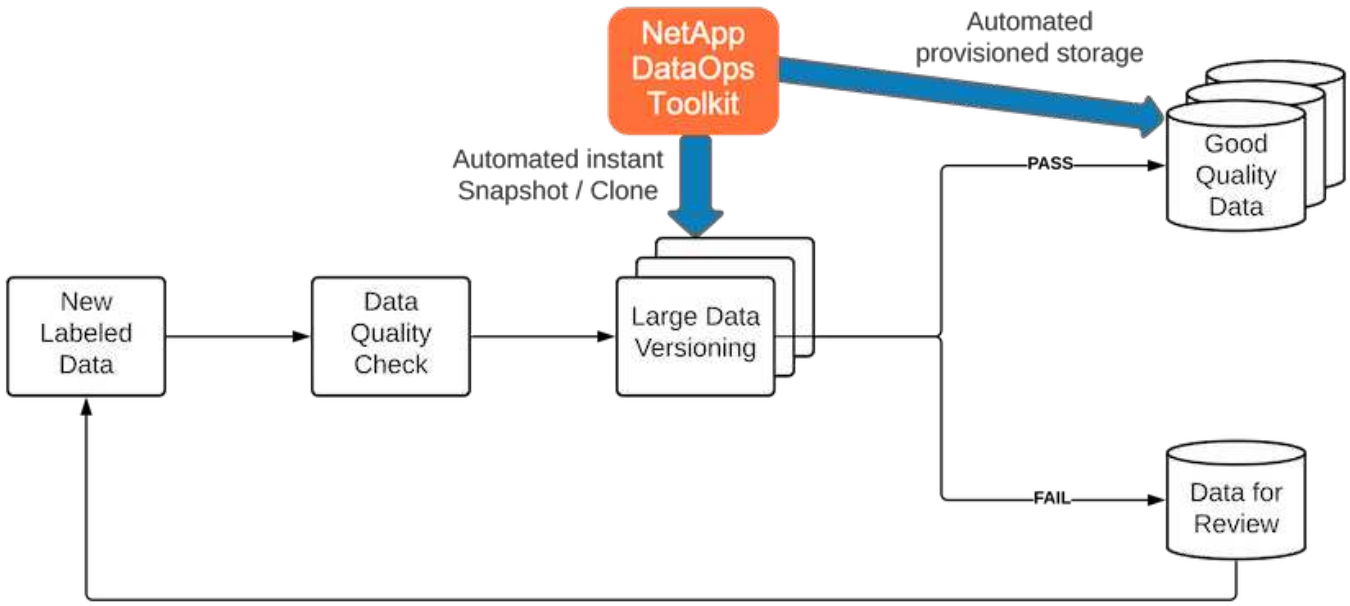
RIVA のベストプラクティス

NVIDIA はいくつかの一般的な機能を提供 "ベストプラクティスに基づくデータ保護" リベットを使用する場合：

- * 可能であれば、ロスレスのオーディオフォーマットを使用します。 * MP3 などの損失のあるコーデックを使用すると、品質が低下する可能性があります。
- * トレーニングデータの増加。 * 音声トレーニングデータにバックグラウンドノイズを追加することで、当初は精度を低下させながら堅牢性を高めることができます。
- * スクラップテキストを使用すれば語彙のサイズを制限しなさい。 * 多くのオンライン源にタイプミスまたは補助発音および珍しい単語を含んでいる。これらを削除すると、言語モデルが改善されます。
- * 可能であれば、最小サンプリングレート 16kHz を使用します。 * ただし、オーディオ品質が低下するため、リサンプルしないようにしてください。

これらのベストプラクティスに加えて、パイプラインの各ステップで正確なラベルを持つ代表的なサンプルデータセットの収集に優先順位を付ける必要があります。つまり、サンプルデータセットには、ターゲットデータセットに典型的な指定された特性を比例的に反映させる必要があります。同様に、データセットの注釈には、データの品質と量を最大化するために、正確性とラベル付けの速度のバランスをとる責任があります。たとえば、このサポートセンターの解決策には、音声ファイル、ラベル付きテキスト、および感情ラベルが必要です。この解決策は、シーケンシャルなので、パイプラインの開始時に発生したエラーが最後まで伝播されます音声ファイルの品質が悪い場合は、テキスト文字変換と感情ラベルも同様になります。

このエラーの伝播も同様に、環境 the models Trained on this data です。感情の予測が 100% 正確であるにもかかわらず、音声テキスト変換モデルのパフォーマンスが低い場合、最終的なパイプラインは最初の音声テキスト変換によって制限されます。開発者は、各モデルのパフォーマンスを個別に、また大きなパイプラインのコンポーネントとして考慮する必要があります。この場合、最終目標は、感情を正確に予測できるパイプラインを開発することです。そのため、パイプラインを評価する全体的な指標は感情の精度であり、音声からテキストへの変換は直接影響を与えます。



NetApp DataOps ツールキットは、ほぼ瞬時のデータクローニングテクノロジを使用して、データ品質チェックパイプラインを補完します。各ラベル付きファイルを評価し、既存のラベル付きファイルと比較する必要があります。これらの品質チェックをさまざまなデータストレージシステムに分散させることで、これらのチェックを迅速かつ効率的に実行できます。

サポートセンターのセンチメント分析の導入

解決策の導入には、次のコンポーネントが含まれます。

1. NetApp DataOps ツールキット
2. NGC の設定
3. NVIDIA Rivea サーバ
4. NVIDIA TAO ツールキット
5. TAO モデルを Riva にエクスポートします

導入を実行するには、次の手順を実行します。

NetApp DataOps ツールキット：センターのセンチメント分析をサポート

を使用します "NetApp DataOps ツールキット"、次の手順を実行します。

1. PIP でツールキットをインストールします。

```
python3 -m pip install netapp-dataops-traditional
```

2. データ管理を設定

```
netapp_dataops_cli.py config
```

NGC 構成：センターの感情分析をサポート

セットアップするには "NVIDIA NGC"、次の手順を実行します。

1. NGC をダウンロード

```
wget -O ngccli_linux.zip
https://ngc.nvidia.com/downloads/ngccli_linux.zip && unzip -o
ngccli_linux.zip && chmod u+x ngc
```

2. 現在のディレクトリをパスに追加します。

```
echo "export PATH=\"\$PATH:$(pwd)\"" >> ~/.bash_profile && source
~/.bash_profile
```

3. コマンドを実行できるように、 NGC CLI を設定する必要があります。次のコマンドを入力します。プロンプトが表示されたら、 API キーも入力します。

```
ngc config set
```

Linux ベースではないオペレーティングシステムについては、を参照してください "こちらをご覧ください"。

NVIDIA Rivea サーバ：センタ心理分析をサポートします

セットアップするには "NVIDIA RIVA"、次の手順を実行します。

1. NGC から Riva ファイルをダウンロード

```
ngc registry resource download-version
nvidia/riva/riva_quickstart:1.4.0-beta
```

2. Riva セットアップを初期化します (`Riva_init.sh`)
3. Riva サーバ (`Riva_start.sh`) を起動します
4. Riva クライアント (`Riva_start_client.sh`) を起動します
5. Riva クライアント内で、オーディオ処理ライブラリをインストールします（ "FFmpeg"）

```
apt-get install ffmpeg
```

6. を起動します "Jupyter" サーバ
7. Riva Inference Pipeline Notebook を実行します。

NVIDIA TAO Toolkit ：センターの感情分析をサポートします

NVIDIA TAO Toolkit をセットアップするには、次の手順を実行します。

1. を準備してアクティブ化します "仮想環境" TAO ツールキット用。
2. をインストールします "必須パッケージ"。
3. トレーニング中および微調整中に使用したイメージを手動で引き出します。

```
docker pull nvcr.io/nvidia/tao/tao-toolkit-pyt:v3.21.08-py3
```

4. を起動します "Jupyter" サーバ
5. TAO 微調整ノートブックを実行します。

TAO モデルを Riva にエクスポート：センターの感情分析をサポートします

を使用してください "Rivea の Tao ツールキットモデル"、次の手順を実行します。

1. TAO 微調整ノートブックにモデルを保存します。
2. TAO トレーニング済みモデルを Riva モデルディレクトリにコピーします。
3. Riva サーバ (`Riva_start.sh`) を起動します

導入の障害です

独自の解決策を開発する際に留意すべき点をいくつかご紹介します。

- 最初に NetApp DataOps ツールキットをインストールし、データストレージシステムが最適に動作するようにします。
- NVIDIA NGC は、イメージとモデルのダウンロードを認証するため、それ以外のコンポーネントよりも先にインストールする必要があります。
- Rivea は、 TAO ツールキットの前にインストールする必要があります。Riva インストールでは、必要に応じてイメージをプルするように Docker デーモンが設定されます。
- DGX および Docker でモデルをダウンロードするには、インターネットアクセスが必要です。

検証結果

前のセクションで説明したように、 2 つ以上の機械学習モデルが順番に実行されている場合は常に、エラーがパイプライン全体に伝播されます。この解決策では、企業の株価

リスクレベルを測定する上で最も重要な要因は、文章の感情です。音声対テキストモデルは、パイプラインに不可欠ですが、感情を予測する前に前処理単位として機能します。実際に重要なのは、基本的な真実文と予測された文の感情の違いです。これは、ワードエラーレート（ WER ）のプロキシとして機能します。音声とテキストの正確さは重要ですが、 WER は最終的なパイプラインメトリックでは直接使用されません。

```
PIPELINE_SENTIMENT_METRIC = MEAN(DIFF(GT_sentiment, ASR_sentiment))
```

これらの感情指標は、 F1 スコア、リコール、各文章の精度について計算できます。結果は集約され、各メトリックの信頼間隔とともに混乱マトリックス内に表示されます。

転送学習を使用する利点は、データ要件、トレーニング時間、コストの数分の 1 でモデルのパフォーマンスが向上することです。また、微調整されたモデルをベースラインバージョンと比較して、転送学習がインペアリングではなくパフォーマンスを向上させるようにする必要があります。つまり、調整済みモデルの方が、サポートセンターのデータのパフォーマンスが事前トレーニング済みモデルよりも優れているはずです。

パイプラインの評価

テストケース	詳細
テスト番号	パイプラインのセンチメント指標
テストの前提条件	音声 / テキストおよび感情分析モデル向けに微調整されたモデル
予想される結果	微調整されたモデルのセンチメント・メトリックは、元の事前トレーニング済みモデルよりも優れています。

パイプラインのセンチメント指標

1. ベースラインモデルのセンチメントメトリックを計算します。
2. 微調整モデルのセンチメントメトリックを計算します。
3. これらの指標間の差異を計算します。
4. すべての文の違いを平均化します。

ビデオとデモ

センチメント分析パイプラインを含むノートブックが 2 つあります。 "「サポート - センター - モデル - 転送 - 学習と微調整 .ipynb 」" および "「サポート - センタ - センチメント - 分析 - パイプライン .ipynb 」"。これらのノートブックは、ユーザーのデータに微調整された最先端のディープラーニングモデルを使用して、サポートセンターのデータを取り込み、各文から感情を抽出するパイプラインを開発する方法を示しています。

サポートセンター - 感情分析パイプライン .ipynb

このノートブックには、オーディオの取り込み、テキストへの変換、外部ダッシュボードで使用するための感情の抽出を行う推論 Riva パイプラインが含まれています。データセットは、まだダウンロードされていない場合は自動的にダウンロードされて処理されます。ノートブックの最初のセクションは、音声ファイルからテキストへの変換を処理する Speech to Text です。続いて、各テキスト文の感情を抽出し、それらの結果を提案されたダッシュボードと同様の形式で表示する感情分析セクションが表示されます。

> MP3 データセットをダウンロードして正しい形式に変換する必要があるため、このノートブックはモデルのトレーニングや微調整の前に実行する必要があります。

Call Center - Sentiment Analysis Pipeline

This notebook demonstrates how to build a pipeline for sentiment analysis of call center conversations. The goal of this pipeline is to develop sentiment analysis for use within an external dashboard.

This tutorial will guide you through the use of NVIDIA's RIVA for automatic speech recognition and text classification. This tutorial uses NetApp cloud storage for data storage and a pre-trained RIVA model.

Channels

These are the channels on which RIVA is hosting models.

- speech: `51051`
- voice: `61051`

These channels **must** be aligned with `riva_speech_api_port` and `riva_vision_api_port` within `config.sh`

In [4]:
```
speech_channel = "localhost:51051"
voice_channel = "localhost:61051"
```

Speech-To-Text

Automatic Speech Recognition (ASR) takes as input an audio stream or audio buffer and returns one or more text transcripts, along with additional optional metadata. ASR represents a full speech recognition pipeline that is GPU accelerated with optimized performance and accuracy. ASR supports synchronous and streaming recognition modes.

For more information on NVIDIA RIVA's Automatic Speech Recognition, visit here.

Constants

Use these constants to affect different aspects of this pipeline:

- `DATA_DIR` : base folder where data is stored
- `DATASET_NAME` : name of the call center dataset
- `COMPANY_DATE` : folder name identifying the particular call center conversation

サポートセンター - モデルトレーニングと微調整 .ipynb

ノートブックを実行する前に、 TAO Toolkit 仮想環境を設定する必要があります (インストール手順については、『 Commands Overview 』の TAO Toolkit の項を参照してください) 。

このノートブックは、 TAIO ツールキットを使用して、お客様のデータに基づいてディープラーニングモデルを微調整します。前のノートブックと同様に、この 2 つのセクションに分かれて、 Speech to Text コンポーネントと、センチメント分析コンポーネントが表示されます。各セクションでは、データ処理、モデルトレー

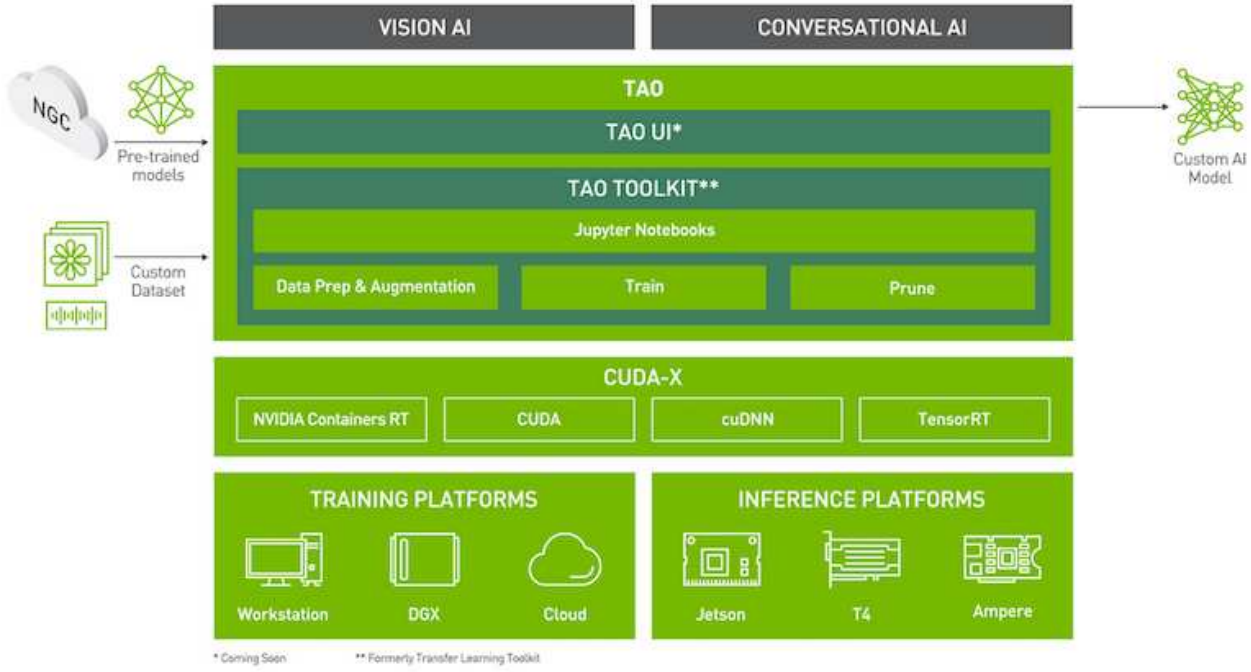
ニング、微調整、結果の評価、およびモデルのエクスポートについて説明します。最後に、 Riva で使用するために、両方の微調整済みモデルを導入するための最終セクションがあります。

Call Center - Model Transfer Learning and Fine-Tuning

TAO Toolkit is a python based AI toolkit for taking purpose-built pre-trained AI models and customizing them with your own data. Transfer learning extracts learned features from an existing neural network to a new one. Transfer learning is often used when creating a large training dataset is not feasible in order to enhance the base performance of state-of-the-art models.

For this call center solution, the speech-to-text and sentiment analysis models are fine-tuned on call center data to augment the model performance on business specific terminology.

For more information on the TAO Toolkit, please visit here.



Installing necessary dependencies

For ease of use, please install TAO Toolkit inside a python virtual environment. We recommend performing this step first and then launching the notebook from the virtual environment. Please refer to the README for these instructions.

まとめ

顧客体験が競争上の重要な戦場と見なされるようになった今、 AI を強化したグローバルサポートセンターは、ほぼすべての業界の企業が無視することができない重要な要素となっています。このテクニカルレポートで提案している解決策は、こうした優れたカスタマーエクスペリエンスの提供を支援することを実証しています。課題は、企業が AI インフラとワークフローを最新化するためのアクションを取ることです。

顧客サービスにおける AI の最適な実装は、人事担当者の代わりになるものではありません。AI は、リアルタ

イムの感情分析、紛争のエスカレーション、マルチモーダルの感情コンピューティングを通じて、優れた顧客体験を生み出す力を発揮します。これにより、包括的な AI モデルが大規模に推奨事項を提示し、個々のヒューマンエージェントが欠けている可能性のある点を補足する、言葉、言葉以外の顔の手がかりを検出できます。AI は、特定のお客様と現在対応可能なエージェントをよりよくマッチさせることもできます。AI を活用することで、企業は、プロバイダの製品、サービス、ブランドイメージに対する顧客の考えや印象に関する価値ある感情を引き出すことができます。

解決策を使用して、客観的なパフォーマンス評価指標として機能するように、サポートエージェントの時系列データを作成することもできます。従来の顧客満足度調査では、十分な回答が得られないことが雇用者は、長期的な従業員や顧客の感情を収集することで、サポートエージェントのパフォーマンスに関して十分な情報に基づいた判断を下すことができます。

ネットアップ、 SFL Scientific 、オープンソースのオーケストレーションフレームワーク、 NVIDIA を組み合わせることで、最新テクノロジをマネージドサービスとして統合し、優れた柔軟性を提供することで、テクノロジの採用を促進し、新しい AI / ML アプリケーションの市場投入期間を短縮できます。これらの高度なサービスはオンプレミスで提供され、クラウドネイティブ環境やハイブリッド導入アーキテクチャへの移植が容易です。

追加情報の参照先

このドキュメントに記載されている情報の詳細については、以下のドキュメントや Web サイトを参照してください。

- 3D 対話型デモ

  "www.netapp.com/ai"

- ネットアップの AI スペシャリストと直接つながる

  "https://www.netapp.com/artificial-intelligence/"

- ネットアップ解決策が実現する NVDIA 基本コマンドプラットフォームの概要

  https://www.netapp.com/pdf.html?item=/media/32792-DS-4145-NVIDIA-Base-Command-Platform-with-NetApp.pdf

- AI 向けネットアップ 10 の理由を解説したインフォグラフィック

  "https://www.netapp.com/us/media/netapp-ai-10-good-reasons.pdf"

- ヘルスケア分野の AI ：『 Deep learning to identify COVID-19 nscions in lung CT scans 』ホワイトペーパーを参照してください

  https://www.netapp.com/pdf.html?item=/media/31240-WP-7342.pdf

- ヘルスケア分野の AI ：ヘルスケア設定におけるフェイスマスクの使用状況を監視するホワイトペーパーです

  https://www.netapp.com/pdf.html?item=/media/37490-NA-611-Monitoring-face-mask-usage-in-healthcare-settings.pdf

- 医療分野の AI ：診断画像診断テクニカルレポート

https://www.netapp.com/pdf.html?item=/media/7395-tr4811.pdf

- 小売業向け AI ：ネットアップの会話型 AI で NVIDIA Riva を使用

  "https://docs.netapp.com/us-en/netapp-solutions/ai/cainvidia_executive_summary.html"

- NetApp ONTAP AI 解決策の概要

  https://www.netapp.com/pdf.html?item=/media/6736-sb-3939.pdf

- NetApp DataOps ツールキットの解決策概要

  https://www.netapp.com/pdf.html?item=/media/21480-SB-4111-1220-NA-Data-Science-Toolkit.pdf

- ネットアップの AI コントロールプレーン解決策の概要

  https://www.netapp.com/pdf.html?item=/media/6737-sb-4055.pdf

- 『データの活用で業界を変革』の E ブック

  "https://www.netapp.com/us/media/na-337.pdf"

- NetApp EF シリーズ AI 解決策の概要

  https://www.netapp.com/pdf.html?item=/media/26708-SB-4136-NetApp-AI-E-Series.pdf

- AI 推論向けのネットアップの AI と Lenovo ThinkSystem 解決策の概要

  https://www.netapp.com/pdf.html?item=/media/25316-SB-4129.pdf

- エンタープライズ AI および ML 向けのネットアップ AI と Lenovo ThinkSystem の解決策概要

  https://www.netapp.com/pdf.html?item=/media/25317-SB-4128.pdf

- ネットアップと NVIDIA – AI ビデオで可能なことを再定義

  https://www.youtube.com/watch?v=38xw65SteUc