Analisi del sentimento con NetApp ai

NetApp Solutions

NetApp
April 26, 2024

Sommario

Analisi del sentimento con NetApp ai

TR-4910: Analisi del sentimento da Customer Communications con NetApp ai

Rick Huang, Sathish Thyagarajan e David Arnette, NetApp Diego Sosa-Coba, SFL Scientific

Questo report tecnico fornisce ai clienti indicazioni di progettazione per eseguire l'analisi del sentimento in un centro di supporto globale di livello Enterprise utilizzando le tecnologie di gestione dei dati NetApp con un framework software NVIDIA che utilizza l'apprendimento del trasferimento e l'intelligenza artificiale conversazionale. Questa soluzione è applicabile a qualsiasi settore che desideri ottenere informazioni sui clienti da file di testo o vocali registrati che rappresentano registri di chat, e-mail e altre comunicazioni di testo o audio. Abbiamo implementato una pipeline end-to-end per dimostrare il riconoscimento vocale automatico, l'analisi del sentimento in tempo reale e le funzionalità di riqualificazione del modello di elaborazione del linguaggio naturale di apprendimento approfondito su un cluster di calcolo accelerato dalla GPU con storage all-flash NetApp connesso al cloud. È possibile formare e ottimizzare enormi modelli linguistici all'avanguardia per eseguire rapidamente l'inferenza con il centro di supporto globale, al fine di creare un'esperienza del cliente eccezionale e valutazioni obiettive e a lungo termine delle performance dei dipendenti.

L'analisi del sentimento è un campo di studio all'interno di Natural Language Processing (NLP) attraverso il quale i sentimenti positivi, negativi o neutri vengono estratti dal testo. I sistemi di intelligenza artificiale convergente sono saliti a un livello di integrazione quasi globale man mano che sempre più persone vengono a interagire con loro. L'analisi del sentimento ha una varietà di casi di utilizzo, dalla determinazione delle performance dei dipendenti del centro di supporto nelle conversazioni con i chiamanti e la fornitura di risposte dei chatbot automatizzate appropriate alla previsione del prezzo delle azioni di un'azienda in base alle interazioni tra i rappresentanti dell'azienda e il pubblico alle chiamate trimestrali sui guadagni. Inoltre, l'analisi del sentimento può essere utilizzata per determinare la posizione del cliente sui prodotti, servizi o supporto forniti dal marchio.

Questa soluzione end-to-end utilizza modelli NLP per eseguire un'analisi del sentimento di alto livello che abilita i framework analitici del centro di supporto. Le registrazioni audio vengono elaborate in testo scritto e il sentimento viene estratto da ciascuna frase della conversazione. I risultati, aggregati in una dashboard, possono essere creati per analizzare i sentimenti delle conversazioni, sia storicamente che in tempo reale. Questa soluzione può essere generalizzata ad altre soluzioni con modalità dati e esigenze di output simili. Con i dati appropriati, è possibile eseguire altri casi di utilizzo. Ad esempio, è possibile analizzare le richieste di guadagno dell'azienda per verificare il sentimento utilizzando la stessa pipeline end-to-end. Altre forme di analisi NLP, come la modellazione topica e il riconoscimento di entità nominate (NER), sono possibili anche grazie alla natura flessibile della pipeline.

Queste implementazioni di ai sono state rese possibili da NVIDIA RIVA, NVIDIA TAO Toolkit e NetApp DataOps Toolkit che hanno collaborato. I tool NVIDIA vengono utilizzati per implementare rapidamente soluzioni ai dalle performance elevate utilizzando modelli e pipeline precostruiti. Il NetApp DataOps Toolkit semplifica varie attività di gestione dei dati per accelerare lo sviluppo.

Valore per il cliente

Le aziende vedono il valore di uno strumento di valutazione dei dipendenti e reazione dei clienti per conversazioni testuali, audio e video per l'analisi del sentimento. I manager traggono vantaggio dalle informazioni presentate nella dashboard, consentendo una valutazione dei dipendenti e della soddisfazione dei clienti in base a entrambe le parti della conversazione.

Inoltre, il NetApp DataOps Toolkit gestisce la versione e l'allocazione dei dati all'interno dell'infrastruttura del cliente. Questo porta ad aggiornamenti frequenti delle analisi presentate all'interno della dashboard senza creare costi di storage dei dati ingombranti.

Casi di utilizzo

A causa del numero di chiamate che questi centri di supporto elaborano, la valutazione delle performance delle chiamate potrebbe richiedere molto tempo se eseguita manualmente. I metodi tradizionali, come il conteggio delle parole e altri metodi, possono ottenere una certa automazione, ma questi metodi non acquisiscono aspetti più sfumati e contesto semantico del linguaggio dinamico. È possibile utilizzare tecniche di modellazione ai per eseguire alcune di queste analisi più sfumate in modo automatizzato. Inoltre, con gli attuali tool di modellazione pre-addestrati e all'avanguardia pubblicati da NVIDIA, AWS, Google e altri, una pipeline end-to-end con modelli complessi può ora essere messa in piedi e personalizzata con relativa facilità.

Una pipeline end-to-end per l'analisi del sentimento del centro di supporto consente di acquisire file audio in tempo reale mentre i dipendenti conversano con i chiamanti. Quindi, questi file audio vengono elaborati per l'utilizzo nel componente voce-testo che li converte in un formato di testo. Ogni frase della conversazione riceve un'etichetta indicante il sentimento (positivo, negativo o neutro).

L'analisi del sentimento può fornire un aspetto essenziale delle conversazioni per la valutazione delle performance delle chiamate. Questi sentimenti aggiungono un ulteriore livello di profondità alle interazioni tra dipendenti e chiamanti. Il dashboard dedicato al sentimento assistito dall'ai offre ai manager un monitoraggio in tempo reale del sentimento all'interno di una conversazione, oltre a un'analisi retrospettiva delle chiamate passate del dipendente.

Esistono tool precostruiti che possono essere combinati in modi potenti per creare rapidamente una pipeline di ai end-to-end per risolvere questo problema. In questo caso, la libreria NVIDIA RIVA può essere utilizzata per eseguire le due attività in-series: Trascrizione audio e analisi del sentimento. Il primo è un algoritmo di elaborazione del segnale di apprendimento supervisionato e il secondo è un algoritmo di classificazione NLP di apprendimento supervisionato. Questi algoritmi pronti all'uso possono essere ottimizzati per qualsiasi caso di utilizzo pertinente con dati rilevanti per l'azienda utilizzando NVIDIA TAO Toolkit. Questo porta a soluzioni più accurate e potenti che vengono costruite solo per una frazione dei costi e delle risorse. I clienti possono incorporare "NVIDIA Maxine" Framework per applicazioni di videoconferenza accelerate dalla GPU nella progettazione del centro di supporto.

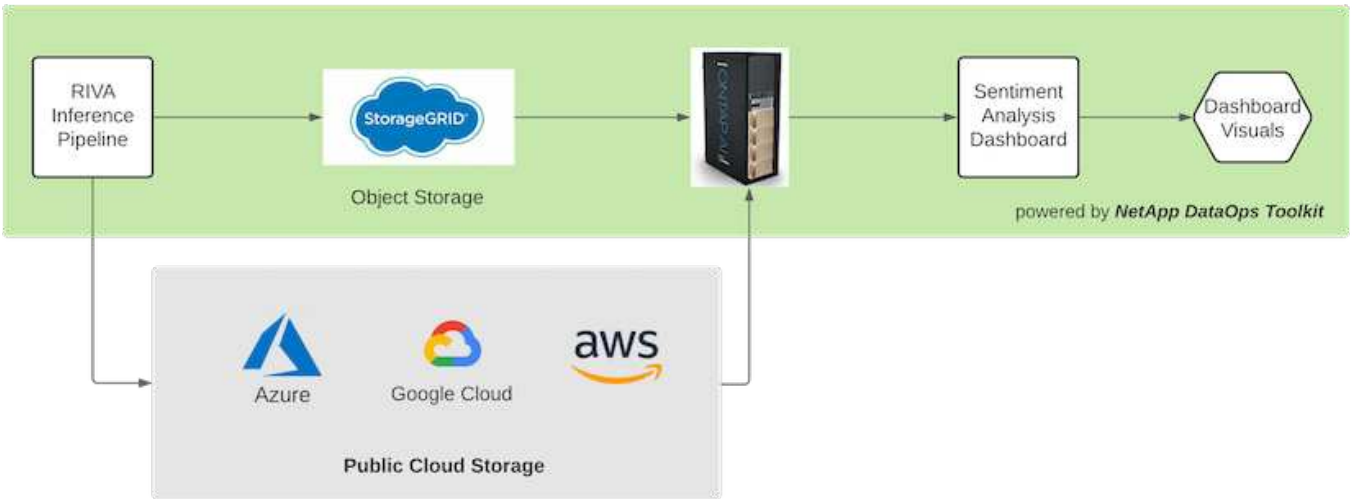
I seguenti casi di utilizzo sono alla base di questa soluzione. Entrambi i casi di utilizzo utilizzano il toolkit TAO per la messa a punto del modello e RIVA per l'implementazione del modello.

- Voce-testo
- Analisi del sentimento

Per analizzare le interazioni del centro di supporto tra dipendenti e clienti, ogni conversazione con il cliente sotto forma di chiamate audio può essere eseguita attraverso la pipeline per estrarre sentimenti a livello di

frase. Tali sentimenti possono quindi essere verificati da un essere umano per giustificare i sentimenti o modificarli in base alle necessità. I dati etichettati vengono quindi trasferiti nella fase di messa a punto per migliorare le previsioni del sentimento. Se esistono già dati di sentimento etichettati, è possibile accelerare la messa a punto del modello. In entrambi i casi, la pipeline è generalizzabile con altre soluzioni che richiedono l'acquisizione di audio e la classificazione delle frasi.

Gli output di ai sentiment vengono caricati su un database cloud esterno o su un sistema storage gestito dall'azienda. Gli output del sentimento vengono trasferiti da questo database più grande allo storage locale per l'utilizzo all'interno della dashboard che visualizza l'analisi del sentimento per i manager. La funzionalità principale del dashboard consiste nell'interfacciarsi con il dipendente del servizio clienti in tempo reale. I manager possono valutare e fornire un feedback sui dipendenti durante le loro chiamate con aggiornamenti in tempo reale del sentimento di ciascuna frase, nonché una revisione storica delle performance del dipendente o delle reazioni dei clienti.



Il "NetApp DataOps Toolkit" Può continuare a gestire i sistemi di storage dei dati anche dopo che la pipeline di inferenza RIVA ha generato etichette di sentimento. Questi risultati ai possono essere caricati su un sistema storage gestito dal NetApp DataOps Toolkit. I sistemi di storage dei dati devono essere in grado di gestire centinaia di inserti e selezionare ogni minuto. Il sistema di storage dei dispositivi locali esegue query in tempo reale sull'ampio storage dei dati per l'estrazione. È inoltre possibile eseguire query sull'istanza di storage dei dati più grande per ottenere dati storici per migliorare ulteriormente l'esperienza del dashboard. Il NetApp DataOps Toolkit facilita entrambi questi utilizzi clonando rapidamente i dati e distribuirli in tutte le dashboard che li utilizzano.

Pubblico di destinazione

Il pubblico di riferimento per la soluzione comprende i seguenti gruppi:

- Responsabili dei dipendenti
- Data engineer/data scientist
- Amministratori IT (on-premise, cloud o ibridi)

Tenere traccia dei sentimenti durante le conversazioni è uno strumento prezioso per valutare le performance dei dipendenti. Utilizzando la dashboard di ai, i manager possono vedere come dipendenti e chiamanti cambiano le proprie sensazioni in tempo reale, consentendo valutazioni live e sessioni di guida. Inoltre, le aziende possono ottenere preziose informazioni sui clienti dai clienti impegnati in conversazioni vocali, chat di testo e videoconferenze. Tali analisi dei clienti utilizzano le funzionalità di elaborazione multimodale su larga scala con modelli e flussi di lavoro ai moderni e all'avanguardia.

Dal punto di vista dei dati, un gran numero di file audio viene elaborato quotidianamente dal centro di supporto. Il NetApp DataOps Toolkit facilita questa attività di gestione dei dati per la messa a punto periodica di modelli e dashboard di analisi del sentimento.

Gli amministratori IT traggono vantaggio anche dal NetApp DataOps Toolkit, che consente loro di spostare rapidamente i dati tra ambienti di implementazione e produzione. Anche gli ambienti e i server NVIDIA devono essere gestiti e distribuiti per consentire l'inferenza in tempo reale.

Architettura

L'architettura di questa soluzione di Support Center si basa sugli strumenti predefiniti di NVIDIA e sul NetApp DataOps Toolkit. I tool NVIDIA vengono utilizzati per implementare rapidamente soluzioni ai ad alte performance utilizzando modelli e pipeline precostruiti. Il NetApp DataOps Toolkit semplifica varie attività di gestione dei dati per accelerare lo sviluppo.

Tecnologia della soluzione

"NVIDIA RIVA" È un SDK con accelerazione GPU per la creazione di applicazioni ai converazionali multimodali che offrono performance in tempo reale sulle GPU. Il toolkit NVIDIA Train, Adapt, and Optimize (TAO) offre un modo più rapido e semplice per accelerare la formazione e creare rapidamente modelli di ai altamente accurati e performanti, specifici per il dominio.

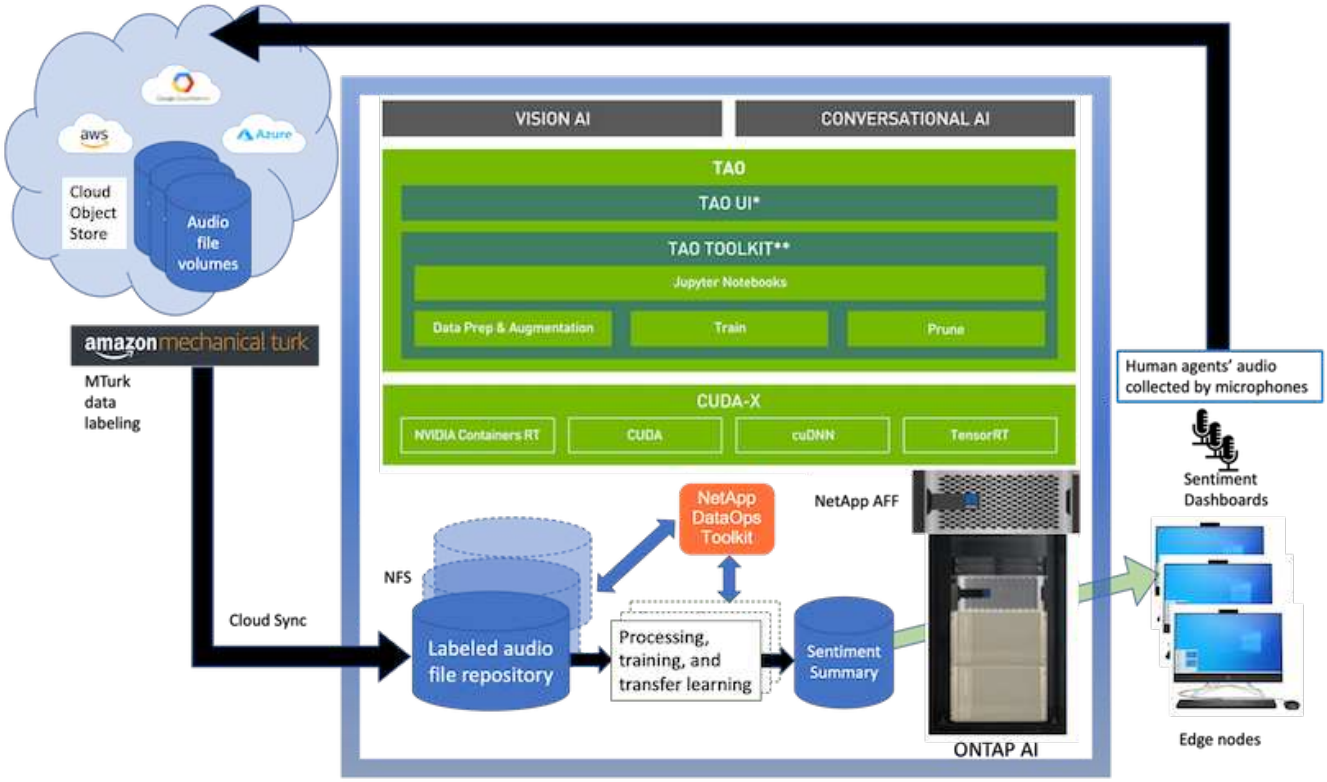
Il NetApp DataOps Toolkit è una libreria Python che semplifica l'esecuzione di varie attività di gestione dei dati da parte di sviluppatori, data scientist, ingegneri DevOps e data engineer. Ciò include il provisioning quasi istantaneo di un nuovo volume di dati o di uno spazio di lavoro JupyterLab, la clonazione quasi istantanea di un volume di dati o di uno spazio di lavoro JupyterLab e lo snap-shoting quasi istantaneo di un volume di dati o di uno spazio di lavoro JupyterLab per la tracciabilità e la baselining.

Diagramma architetturale

Il seguente diagramma illustra l'architettura della soluzione. Esistono tre categorie di ambiente principali: Cloud, core e edge. Ciascuna delle categorie può essere distribuita geograficamente. Ad esempio, il cloud contiene archivi di oggetti con file audio in bucket in regioni diverse, mentre il core potrebbe contenere data center collegati tramite una rete ad alta velocità o Copia e sincronizzazione di NetApp BlueXP. I nodi edge denotano le piattaforme di lavoro quotidiane dei singoli agenti umani, in cui sono disponibili strumenti di dashboard interattivi e microfoni per visualizzare il sentimento e raccogliere dati audio dalle conversazioni con i clienti.

Nei data center con accelerazione GPU, le aziende possono utilizzare NVIDIA "RIVA" Framework per la creazione di applicazioni ai conversazionali, alle quali il "Toolkit Tao" Si connette per il finetuning e la riqualificazione dei modelli utilizzando tecniche di trasferimento L-learning. Queste applicazioni di calcolo e i flussi di lavoro sono basati su "NetApp DataOps Toolkit", Che offre le migliori funzionalità di gestione dei dati offerte da ONTAP. Il toolkit consente ai team di dati aziendali di prototipare rapidamente i propri modelli con dati strutturati e non strutturati associati tramite snapshot e cloni per tracciabilità, controllo delle versioni, test A/B, fornendo così sicurezza, governance, e conformità alle normative. Vedere la sezione "Progettazione dello storage" per ulteriori dettagli.

Questa soluzione dimostra l'elaborazione dei file audio, il training sul modello NLP, l'apprendimento del trasferimento e le fasi dettagliate della gestione dei dati. La pipeline end-to-end risultante genera un riepilogo dei sentimenti che viene visualizzato in tempo reale sui dashboard degli agenti di supporto umano.



Requisiti hardware

La seguente tabella elenca i componenti hardware necessari per implementare la soluzione. I componenti hardware utilizzati in una particolare implementazione della soluzione possono variare in base ai requisiti del cliente.

Test di latenza della risposta	Tempo (millisecondi)
Elaborazione dei dati	10
Deduzione	10

Questi test dei tempi di risposta sono stati eseguiti su oltre 50,000 file audio in 560 conversazioni. Ogni file audio era di ~100 KB come MP3 e ~1 MB quando convertito in WAV. La fase di elaborazione dei dati converte gli MP3 in file WAV. I passaggi di inferenza convertono i file audio in testo ed estraggono un sentimento dal testo. Questi passaggi sono tutti indipendenti l'uno dall'altro e possono essere parallelizzati per accelerare il processo.

Tenendo conto della latenza del trasferimento dei dati tra gli archivi, i manager dovrebbero essere in grado di visualizzare gli aggiornamenti dell'analisi del sentimento in tempo reale entro un secondo dalla fine della frase.

Hardware NVIDIA RIVA

Hardware	Requisiti
SISTEMA OPERATIVO	Linux x86_64
Memoria GPU (ASR)	Modelli di streaming: ~5600 MB modelli senza streaming: ~3100 MB
Memoria GPU (NLP)	~500 MB per modello BERT

Hardware NVIDIA TAO Toolkit

Hardware	Requisiti
RAM di sistema	32 GB
RAM GPU	32 GB
CPU	8 core
GPU	NVIDIA (A100, V100 e RTX 30x0)
SSD	100 GB

Sistema storage flash

NetApp ONTAP 9

ONTAP 9.9, l'ultima generazione di software per la gestione dello storage NetApp, consente alle aziende di modernizzare l'infrastruttura e passare a un data center predisposto per il cloud. Sfruttando le funzionalità di gestione dei dati leader del settore, ONTAP consente la gestione e la protezione dei dati con un singolo set di strumenti, indipendentemente dalla posizione dei dati. Puoi anche spostare liberamente i dati ovunque siano necessari: Edge, core o cloud. ONTAP 9.9 include numerose funzionalità che semplificano la gestione dei dati, accelerano e proteggono i dati critici e abilitano le funzionalità dell'infrastruttura di nuova generazione nelle architetture di cloud ibrido.

Copia e sincronizzazione di NetApp BlueXP

"Copia e sincronizzazione di BlueXP" È un servizio NetApp per una sincronizzazione dei dati rapida e sicura che consente di trasferire file tra condivisioni di file NFS o SMB on-premise a una delle seguenti destinazioni:

- NetApp StorageGRID
- NetApp ONTAP S3
- NetApp Cloud Volumes Service
- Azure NetApp Files
- Amazon Simple Storage Service (Amazon S3)
- Amazon Elastic file System (Amazon EFS)
- Azure Blob
- Storage Google Cloud
- Storage a oggetti IBM Cloud

BlueXP Copy e Sync sposta i file dove ne hai bisogno in modo rapido e sicuro. Una volta trasferiti, i dati sono completamente disponibili per l'utilizzo sia sull'origine che sulla destinazione. BlueXP Copy e Sync sincronizza costantemente i dati in base alla pianificazione predefinita, spostando solo i delta, in modo da ridurre al minimo il tempo e il denaro necessari per la replica. BlueXP Copy and Sync è un tool software as a service (SaaS) semplice da configurare e utilizzare. I trasferimenti dei dati attivati da BlueXP Copy e Sync sono effettuati dai broker di dati. Puoi implementare i broker di dati BlueXP Copy e Sync in AWS, Azure, Google Cloud Platform o on-premise.

NetApp StorageGRID

La suite di storage a oggetti software-defined di StorageGRID supporta un'ampia gamma di casi di utilizzo in ambienti multi-cloud pubblici, privati e ibridi. Grazie alle innovazioni leader del settore, NetApp StorageGRID memorizza, protegge, protegge e preserva i dati non strutturati per un utilizzo multiuso, inclusa la gestione automatica del ciclo di vita per lunghi periodi di tempo. Per ulteriori informazioni, consultare "NetApp StorageGRID" sito.

Requisiti software

La seguente tabella elenca i componenti software necessari per implementare questa soluzione. I componenti software utilizzati in una particolare implementazione della soluzione possono variare in base ai requisiti del cliente.

Computer host	Requisiti
RIVA (in precedenza JARVIS)	1.4.0
TAO Toolkit (in precedenza Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
SISTEMA OPERATIVO DGX	5.1
DOTK	2.0.0

Software NVIDIA RIVA

Software	Requisiti
Docker	>19.02 (con nvidia-docker installato)>=19.03 se non si utilizza DGX
Driver NVIDIA	465.19.01+ 418.40+, 440.33+, 450.51+, 460.27+ per GPU Data Center

Software	Requisiti
Sistema operativo container	Ubuntu 20.04
CUDA	11.3.0
CuBLAS	11.5.1.101
CuDNN	8.2.0.41
NCCL	2.9.6
TensorRT	7.2.3.4
Server di inferenza Triton	2.9.0

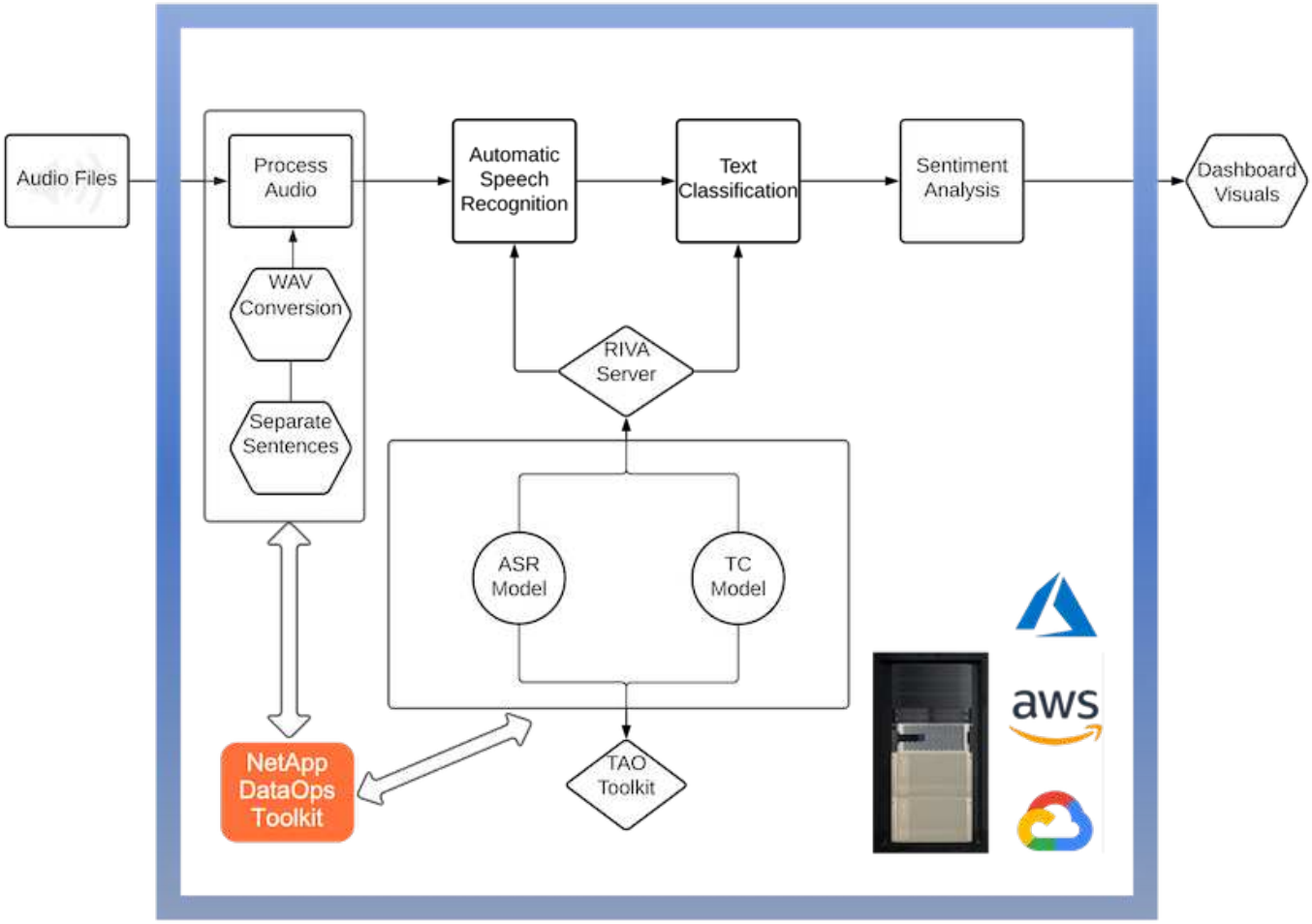
Software NVIDIA TAO Toolkit

Software	Requisiti
Ubuntu 18.04 LTS	18.04
python	>=3.6.9
docker-ce	>19.03.5
API docker	1.40
nvidia-container-toolkit	>1.3.0-1
nvidia-container-runtime	3.4.0-1
nvidia-docker2	2.5.0-1
driver nvidia	>455
python-pip	>21.06
nvidia-pyindex	Ultima versione

Utilizza i dettagli del caso

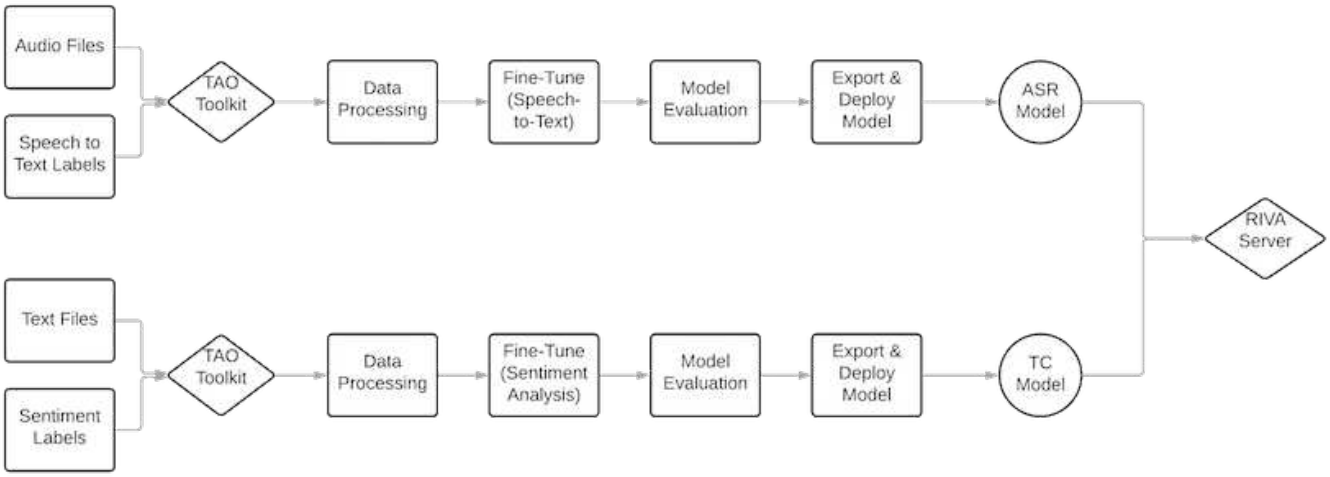
Questa soluzione si applica ai seguenti casi di utilizzo:

- Voce-testo
- Analisi del sentimento



Il caso d'utilizzo del parlato-to-text inizia con l'acquisizione di file audio per i centri di supporto. Questo audio viene quindi elaborato per adattarsi alla struttura richiesta DA RIVA. Se i file audio non sono già stati suddivisi nelle unità di analisi, è necessario eseguire questa operazione prima di passare l'audio a RIVA. Una volta elaborato, il file audio viene trasmesso al server RIVA come chiamata API. Il server utilizza uno dei numerosi modelli che ospita e restituisce una risposta. Questa voce-testo (parte del riconoscimento vocale automatico) restituisce una rappresentazione testuale dell'audio. Da qui, la pipeline passa alla parte di analisi del sentimento.

Per l'analisi del sentimento, l'output di testo del riconoscimento vocale automatico funge da input per la classificazione del testo. Text Classification è il componente NVIDIA per la classificazione del testo in un numero qualsiasi di categorie. Le categorie di sentimento variano da positivo a negativo per le conversazioni del centro di supporto. Le performance dei modelli possono essere valutate utilizzando un set di holdout per determinare il successo della fase di fine tuning.



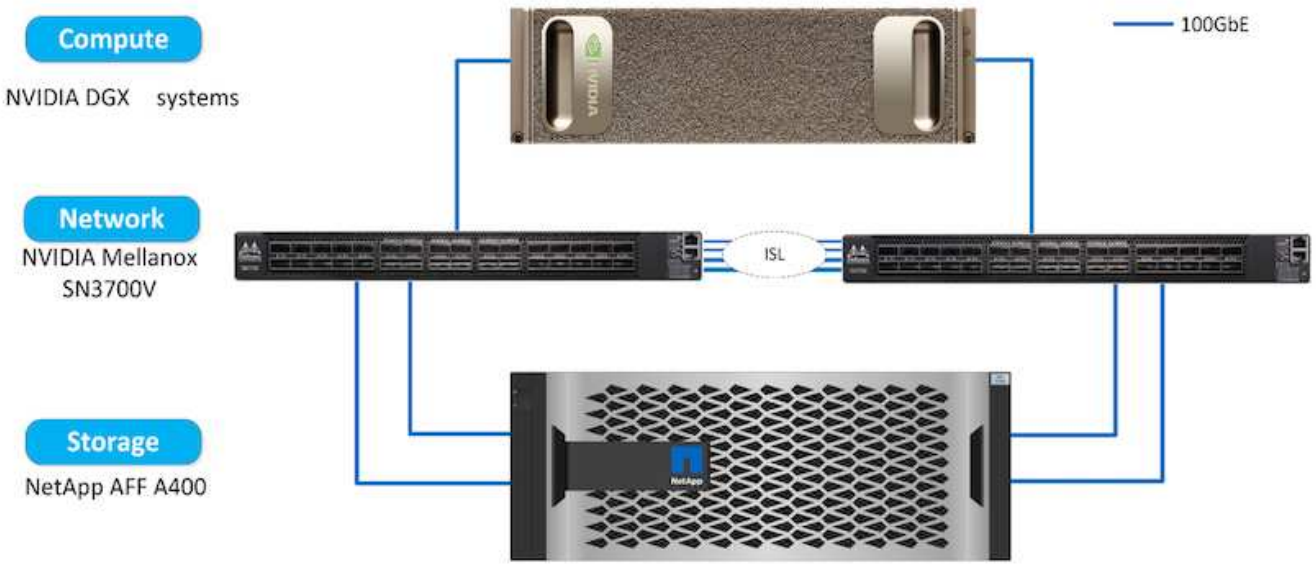
Una pipeline simile viene utilizzata sia per l'analisi del parlato-to-text che per l'analisi del sentimento all'interno del toolkit TAO. La differenza principale è l'utilizzo di etichette necessarie per la messa a punto dei modelli. La pipeline TAO Toolkit inizia con l'elaborazione dei file di dati. Poi i modelli preformati (provenienti da "Catalogo NGC NVIDIA") vengono perfezionati utilizzando i dati del centro di supporto. I modelli perfezionati vengono valutati in base alle metriche di performance corrispondenti e, se sono più performanti dei modelli preformati, vengono implementati sul server RIVA.

Considerazioni di progettazione

In questa sezione vengono descritte le considerazioni di progettazione relative ai diversi componenti di questa soluzione.

Progettazione di rete e calcolo

A seconda delle restrizioni sulla sicurezza dei dati, tutti i dati devono rimanere all'interno dell'infrastruttura del cliente o in un ambiente sicuro.



Progettazione dello storage

Il NetApp DataOps Toolkit funge da servizio principale per la gestione dei sistemi storage. DataOps Toolkit è una libreria Python che consente a sviluppatori, data scientist, ingegneri DevOps e data engineer di eseguire diverse attività di gestione dei dati, come il provisioning quasi istantaneo di un nuovo volume di dati o di un'area di lavoro JupyterLab, la clonazione quasi istantanea di un volume di dati o di un'area di lavoro JupyterLab, E lo snap-shoting quasi istantaneo di un volume di dati o di uno spazio di lavoro JupyterLab per la tracciabilità o il baselining. Questa libreria Python può funzionare come un'utility a riga di comando o una libreria di funzioni che possono essere importate in qualsiasi programma Python o Jupyter notebook.

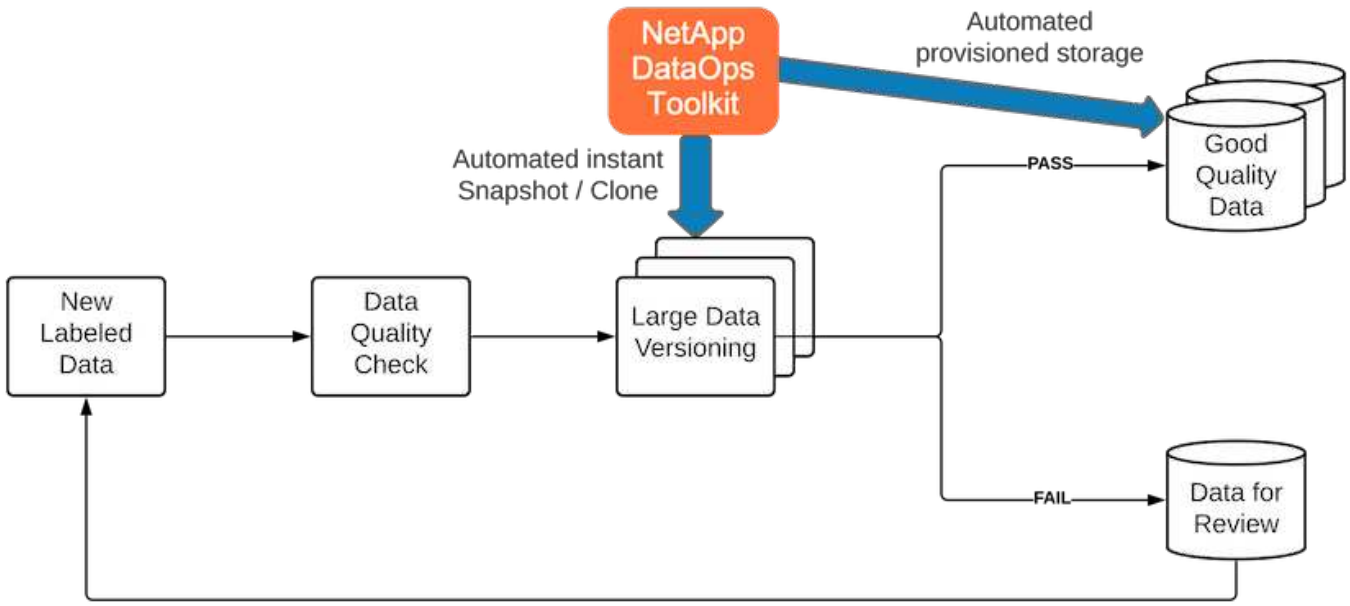
Best practice RIVA

NVIDIA offre diverse funzionalità generali "best practice per i dati" Per utilizzare RIVA:

- **Se possibile, utilizzare formati audio senza perdita di dati.** l'utilizzo di codec con perdita di dati come MP3 può ridurre la qualità.
- **Aumentare i dati di training.** l'aggiunta di rumore di fondo ai dati di training audio può inizialmente ridurre la precisione e aumentare la robustezza.
- **Limitare la dimensione del vocabolario se si utilizza il testo scartato.** molte fonti online contengono messaggi o voci accessorie e parole non comuni. La rimozione di questi elementi può migliorare il modello linguistico.
- **Se possibile, utilizzare una frequenza di campionamento minima di 16 kHz.** tuttavia, provare a non ricampionare, perché in questo modo si riduce la qualità audio.

Oltre a queste Best practice, i clienti devono dare la priorità alla raccolta di un set di dati campione rappresentativo con etichette accurate per ogni fase della pipeline. In altre parole, il set di dati di esempio dovrebbe riflettere in modo proporzionale le caratteristiche specificate esemplificate in un set di dati di destinazione. Allo stesso modo, gli annotatori dei set di dati hanno la responsabilità di bilanciare la precisione e la velocità dell'etichettatura in modo da massimizzare la qualità e la quantità dei dati. Ad esempio, questa soluzione di Support Center richiede file audio, testo etichettato ed etichette di sentimento. La natura sequenziale di questa soluzione significa che gli errori dall'inizio della pipeline vengono propagati fino alla fine Se i file audio sono di scarsa qualità, anche le trascrizioni di testo e le etichette di sentimento saranno.

Questa propagazione degli errori si applica allo stesso modo ai modelli addestrati su questi dati. Se le previsioni del sentimento sono accurate al 100% ma il modello da voce a testo non funziona correttamente, la pipeline finale è limitata dalle trascrizioni audio-testo iniziali. È essenziale che gli sviluppatori considerino le performance di ciascun modello singolarmente e come un componente di una pipeline più ampia. In questo caso specifico, l'obiettivo finale è sviluppare una pipeline in grado di prevedere con precisione il sentimento. Pertanto, la metrica generale su cui valutare la pipeline è la precisione dei sentimenti, che la trascrizione vocale-testuale influisce direttamente.



Il NetApp DataOps Toolkit integra la pipeline per il controllo della qualità dei dati attraverso l'utilizzo della sua tecnologia di cloning dei dati quasi istantanea. Ogni file etichettato deve essere valutato e confrontato con i file etichettati esistenti. La distribuzione di questi controlli di qualità nei vari sistemi di storage dei dati garantisce che questi controlli vengano eseguiti in modo rapido ed efficiente.

Implementazione dell'analisi del sentimento del centro di supporto

L'implementazione della soluzione comporta i seguenti componenti:

1. NetApp DataOps Toolkit
2. Configurazione NGC
3. Server NVIDIA RIVA
4. Toolkit NVIDIA TAO
5. Esportare i modelli TAO in RIVA

Per eseguire l'implementazione, attenersi alla seguente procedura:

NetApp DataOps Toolkit: Analisi del sentimento del centro di supporto

Per utilizzare "NetApp DataOps Toolkit", completare la seguente procedura:

1. PIP installare il toolkit.

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

2. Configurare la gestione dei dati

```
netapp_dataops_cli.py config
```

Configurazione NGC: Analisi del sentimento del centro di supporto

Per configurare "NVIDIA NGC", completare la seguente procedura:

1. Scarica NGC.

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

2. Aggiungere la directory corrente al percorso.

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

3. È necessario configurare l'interfaccia CLI NGC per l'utilizzo in modo da poter eseguire i comandi. Immettere il seguente comando, inclusa la chiave API quando richiesto.

```
ngc config set
```

Per i sistemi operativi che non sono basati su Linux, visitare il sito "qui".

Server NVIDIA RIVA: Analisi del sentimento del centro di supporto

Per configurare "NVIDIA RIVA", completare la seguente procedura:

1. Scarica I file RIVA da NGC.

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

2. Inizializzare la configurazione DI RIVA (`riva_init.sh`).
3. Avviare IL server RIVA (`riva_start.sh`).
4. Avviare IL client RIVA (`riva_start_client.sh`).
5. All'interno del client RIVA, installare la libreria di elaborazione audio ("FFMPEG")

```
apt-get install ffmpeg
```

6. Avviare "Jupyter" server.
7. Eseguire il notebook RIVA Inference Pipeline.

NVIDIA TAO Toolkit: Analisi del sentimento del centro di supporto

Per configurare NVIDIA TAO Toolkit, attenersi alla seguente procedura:

1. Preparare e attivare un "ambiente virtuale" Per TAO Toolkit.
2. Installare "pacchetti richiesti".
3. Estrarre manualmente l'immagine utilizzata durante l'addestramento e la messa a punto.

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

4. Avviare "Jupyter" server.
5. Eseguire il notebook TAO fine-Tuning.

Esportare i modelli TAO in RIVA: Analisi del sentimento del centro di supporto

Da utilizzare "Modelli TAO Toolkit di RIVA", completare la seguente procedura:

1. Salva i modelli nel notebook TAO fine-Tuning.
2. Copiare i modelli addestrati TAO nella directory dei modelli RIVA.
3. Avviare IL server RIVA (`riva_start.sh`).

Blocchi stradali per l'implementazione

Ecco alcuni elementi da tenere a mente durante lo sviluppo della soluzione:

- Il NetApp DataOps Toolkit viene installato per primo per garantire il funzionamento ottimale del sistema di storage dei dati.
- NVIDIA NGC deve essere installata prima di ogni altra cosa perché autentica il download di immagini e modelli.
- RIVA deve essere installato prima del toolkit TAO. L'installazione DI RIVA configura il daemon del docker per estrarre le immagini in base alle necessità.
- DGX e Docker devono disporre dell'accesso a Internet per scaricare i modelli.

Risultati della convalida

Come indicato nella sezione precedente, gli errori vengono propagati in tutta la pipeline ogni volta che vi sono due o più modelli di apprendimento automatico in esecuzione in sequenza. Per questa soluzione, il sentimento della frase è il fattore più importante nella misurazione del livello di rischio azionario dell'azienda. Il modello da voce a testo, sebbene essenziale per la pipeline, funge da unità di pre-elaborazione prima che i sentimenti possano essere previsti. Ciò che conta realmente è la differenza di sentimento tra le frasi di verità e le frasi previste. Questo serve come proxy per il tasso di errore di

parola (WER). La precisione del parlato-to-text è importante, ma il WER non viene utilizzato direttamente nella metrica finale della pipeline.

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

Queste metriche di sentimento possono essere calcolate per il punteggio F1, il richiamo e la precisione di ciascuna frase. I risultati possono quindi essere aggregati e visualizzati all'interno di una matrice di confusione, insieme agli intervalli di confidenza per ciascuna metrica.

Il vantaggio dell'utilizzo del transfer learning è un aumento delle performance del modello per una frazione dei requisiti dei dati, dei tempi di formazione e dei costi. I modelli perfezionati devono anche essere confrontati con le versioni di riferimento per garantire che l'apprendimento del trasferimento migliori le performance invece di comprometterle. In altre parole, il modello ottimizzato dovrebbe funzionare meglio sui dati del centro di supporto rispetto al modello preaddestrato.

Valutazione della pipeline

Caso di test	Dettagli
Numero del test	Metrica del sentimento della pipeline
Prerequisiti del test	Modelli ottimizzati per modelli di analisi del parlato-to-text e del sentimento
Risultato previsto	La metrica del sentimento del modello ottimizzato offre prestazioni migliori rispetto al modello originale preaddestrato.

Metrica del sentimento della pipeline

1. Calcola la metrica del sentimento per il modello di riferimento.
2. Calcola la metrica del sentimento per il modello ottimizzato.
3. Calcola la differenza tra queste metriche.
4. Calcolare la media delle differenze tra tutte le frasi.

Video e demo

Esistono due notebook che contengono la pipeline di analisi del sentimento: ""Support-Center-Model-Transfer-Learning-and-fine-Tuning.ipynb"" e. ""Support-Center-Sentiment-Analysis-pipeline.ipynb"". Insieme, questi notebook dimostrano come sviluppare una pipeline per acquisire i dati del centro di supporto ed estrarre sentimenti da ogni frase utilizzando modelli di deep learning all'avanguardia e ottimizzati sui dati dell'utente.

Support Center - Sentiment Analysis Pipeline.ipynb

Questo notebook contiene la pipeline Inference RIVA per l'acquisizione di audio, la conversione in testo e l'estrazione di sentimenti da utilizzare in una dashboard esterna. I set di dati vengono scaricati ed elaborati automaticamente, se non è già stato fatto. La prima sezione del notebook è Speech-to-Text, che gestisce la conversione dei file audio in testo. Segue la sezione analisi del sentimento che estrae i sentimenti per

ciascuna frase di testo e visualizza i risultati in un formato simile alla dashboard proposta.

> Questo notebook deve essere eseguito prima del training e della messa a punto del modello, in quanto il set di dati MP3 deve essere scaricato e convertito nel formato corretto.

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

Support Center - Model Training e fine-Tuning.ipynb

L'ambiente virtuale TAO Toolkit deve essere configurato prima di eseguire il notebook (per istruzioni sull'installazione, consultare la sezione relativa al toolkit TAO nella Panoramica dei comandi).

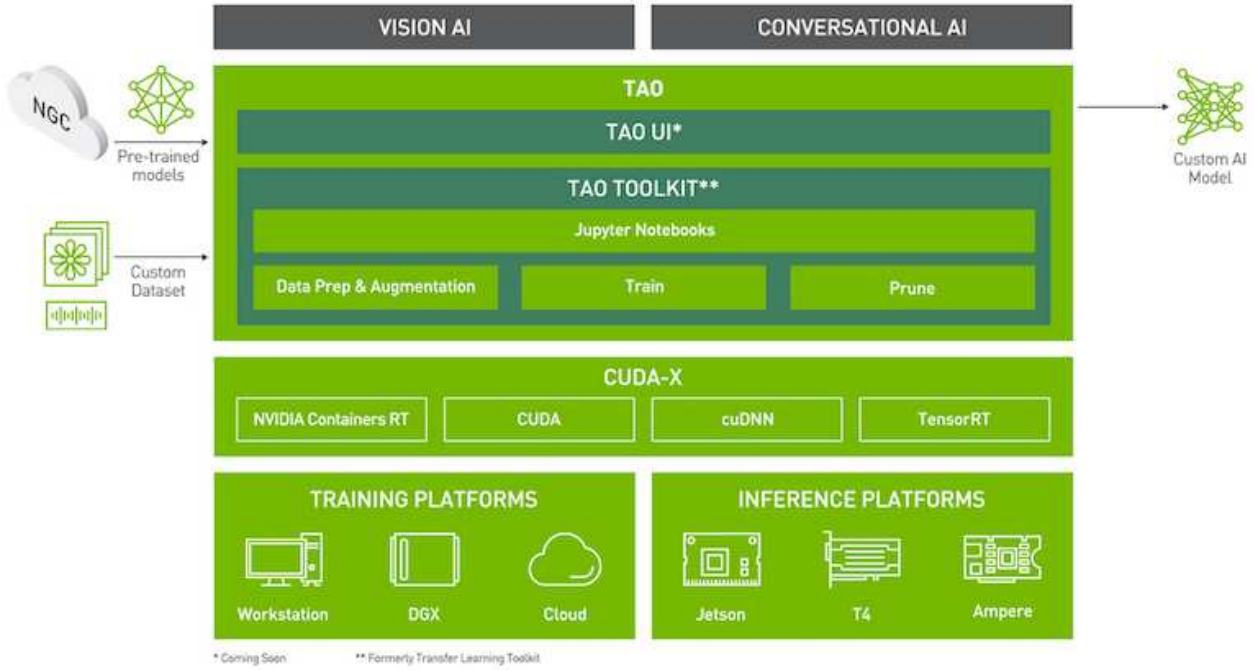
Questo notebook si affida al toolkit TAO per mettere a punto modelli di apprendimento approfondito sui dati dei clienti. Come per il notebook precedente, questo è separato in due sezioni per i componenti Speech-to-Text e analisi del sentimento. Ogni sezione passa attraverso l'elaborazione dei dati, la formazione sui modelli e la messa a punto, la valutazione dei risultati e l'esportazione dei modelli. Infine, è disponibile una sezione finale per l'implementazione di entrambi i modelli ottimizzati per L'utilizzo in RIVA.

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.

Conclusione

Poiché l'esperienza dei clienti è diventata sempre più considerata come un terreno di battaglia competitivo fondamentale, un centro di supporto globale con intelligenza artificiale diventa un componente critico che le aziende di quasi tutti i settori non possono permettersi di trascurare. La soluzione proposta in questo report tecnico è stata dimostrata per supportare l'offerta di esperienze clienti eccezionali di questo tipo e la sfida è ora quella di garantire che le aziende stiano adottando misure per modernizzare l'infrastruttura e i flussi di lavoro dell'ai.

Le migliori implementazioni dell'ai nel servizio clienti non devono sostituire gli agenti umani. Piuttosto, l'ai può consentire loro di creare esperienze eccezionali con i clienti attraverso analisi del sentimento in tempo reale,

escalation delle controversie e calcolo affettivo multimodale per rilevare indizi verbali, non verbali e facciali con i quali modelli ai completi possono fornire consigli su larga scala e integrare ciò che un singolo agente umano potrebbe non essere presente. L'ai può anche fornire una migliore corrispondenza tra un particolare cliente e gli agenti attualmente disponibili. Utilizzando l'ai, le aziende possono estrarre il prezioso sentimento dei clienti riguardo alle loro opinioni e impressioni sui prodotti, i servizi e l'immagine del marchio del provider.

La soluzione può anche essere utilizzata per costruire dati Time-Series per gli agenti di supporto che fungono da metriche di valutazione obiettiva delle performance. I sondaggi convenzionali sulla soddisfazione dei clienti spesso non hanno risposte sufficienti. Raccogliendo il sentimento a lungo termine dei dipendenti e dei clienti, i datori di lavoro possono prendere decisioni informate in merito alle performance degli agenti di supporto.

La combinazione di NetApp, SFL Scientific, framework di orchestrazione open-source e NVIDIA riunisce le più recenti tecnologie come servizi gestiti con grande flessibilità per accelerare l'adozione della tecnologia e migliorare il time-to-market per le nuove applicazioni ai/ML. Questi servizi avanzati vengono forniti on-premise e possono essere facilmente trasferiti per ambienti cloud-native e architetture di implementazione ibride.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e/o siti Web:

- Demo interattive 3D

  "www.netapp.com/ai"

- Connettiti direttamente con uno specialista ai di NetApp

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

- NVDIA base Command Platform with NetApp Solution Brief

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

- Infografica sulle buone ragioni di NetApp per ai 10

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

- Ai in Healthcare: White paper sull'apprendimento approfondito per identificare le lesioni COVID-19 nelle scansioni TC polmonari

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

- Ai in Healthcare: White paper sul monitoraggio dell'utilizzo della maschera facciale in ambito sanitario

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

- Ai in Healthcare: Report tecnico di imaging diagnostico

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

- Ai per il retail: Ai di NetApp Conversational con NVIDIA RIVA

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

- Analisi della soluzione ai di NetApp ONTAP

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

- Descrizione della soluzione NetApp DataOps Toolkit

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

- Analisi della soluzione NetApp ai Control Plane

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

- EBook trasformare il settore con Data Drive ai

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

- Analisi della soluzione ai NetApp EF-Series

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

- Analisi della soluzione NetApp ai e Lenovo ThinkSystem for ai Inferencing

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

- Analisi della soluzione NetApp ai e Lenovo ThinkSystem per ai e ML Enterprise

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

- NetApp e NVIDIA – ridefinire le possibilità con i video ai

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