EBOOK

Discuter pour faire avancer les choses

Créez une infrastructure de données pour l'IA conversationnelle

NetApp

Sommaire

Trêve de bavardage, parlons sérieusement

NLP, traitement du langage naturel, IA conversationnelle, *robots parlants*...

Quel que soit le nom que vous lui donnez, un système d'IA conversationnelle parle comme un être humain, comprend le contexte d'une phrase et donne des réponses intelligentes en s'appuyant sur le deep learning, une méthode qui connaît un développement fulgurant et donne aux systèmes d'IA un ton plus naturel.

Le deep learning permet non seulement de créer des systèmes plus conviviaux, mais également d'éliminer l'intervention d'experts en linguistique et les processus basés sur les règles en arrière-plan. Avec lui, les entreprises des secteurs qui utilisent des langages particulièrement complexes (services financiers, domaine de la santé et sciences de la vie, gouvernement, secteurs automobiles et industriels, et retail) peuvent adopter des solutions de NLP.

Les données améliorent la qualité des conversations

Ces modèles d'IA peuvent être très volumineux et hautement complexes. Ils nécessitent la transmission instantanée d'importants volumes de données. Une infrastructure NLP efficace doit être en mesure de :

1. Monter le son
2. Fluidifier les pipelines
3. Répondre du tac au tac

NLP : au-delà des chatbots

Assistants connectés, moteurs de recherche, saisie intuitive, etc. Le NLP est devenu le nouveau langage universel. Il est partout, même là où on l'attend le moins.

Évaluation de la solvabilité

Le NLP peut servir à générer des cotes de crédit sur la base de données telles que la position géographique, l'activité sur les médias sociaux, les recherches en ligne, les contacts, etc.

Sélection de patients pour les essais cliniques

Il n'est pas toujours aisé de trouver des participants pour les essais cliniques, notamment parce que les candidats potentiels ne savent pas toujours que ces tests existent. Avec le NLP, les chercheurs et fabricants peuvent trouver automatiquement des patients qui correspondent aux critères de leurs essais.

Maintien de l'ordre

Les services de police utilisent le NLP pour identifier les motifs de crime. Ainsi, ils améliorent la protection des citoyens et réduisent la violence, tout en augmentant leur réactivité et compréhension de la situation.

Entretien des véhicules

Le NLP aide les automobilistes à assurer l'entretien de leur véhicule. S'ils ont une question, au lieu de chercher la réponse dans un gros manuel, ils la posent directement à leur voiture : « Quel est ce voyant qui s'allume ? » ou « Comment faire pour changer une bougie ? ».

Réparation d'aéronefs

Le NLP permet aux mécaniciens de synthétiser les informations des épais manuels d'entretien et ainsi de mieux comprendre les problèmes signalés par les pilotes.

1. Montez le son

Le NLP nécessite une quantité titanesque de données : tous les mots jamais prononcés, voire plus.

Le système de NLP doit traiter, comprendre et référencer la parole en s'appuyant sur une immense bibliothèque de données afin de générer une réponse intelligible en quelques millisecondes.

La complexité du langage humain, avec toutes ses règles et exceptions, ses nuances et ses sous-entendus, complique cette tâche déjà ardue. Les modèles propres à un secteur doivent également intégrer des informations précises sur un domaine, une entreprise ou des produits spécifiques.

Voilà pourquoi la taille des modèles d'IA n'a cessé de grossir jusqu'à contenir des millions voire des milliards de paramètres. Plus il y a de données, plus le modèle est précis. Toutefois, l'entraînement d'un modèle aussi énorme peut prendre des semaines et nécessite les meilleurs frameworks de machine learning et de deep learning.

Google Translate

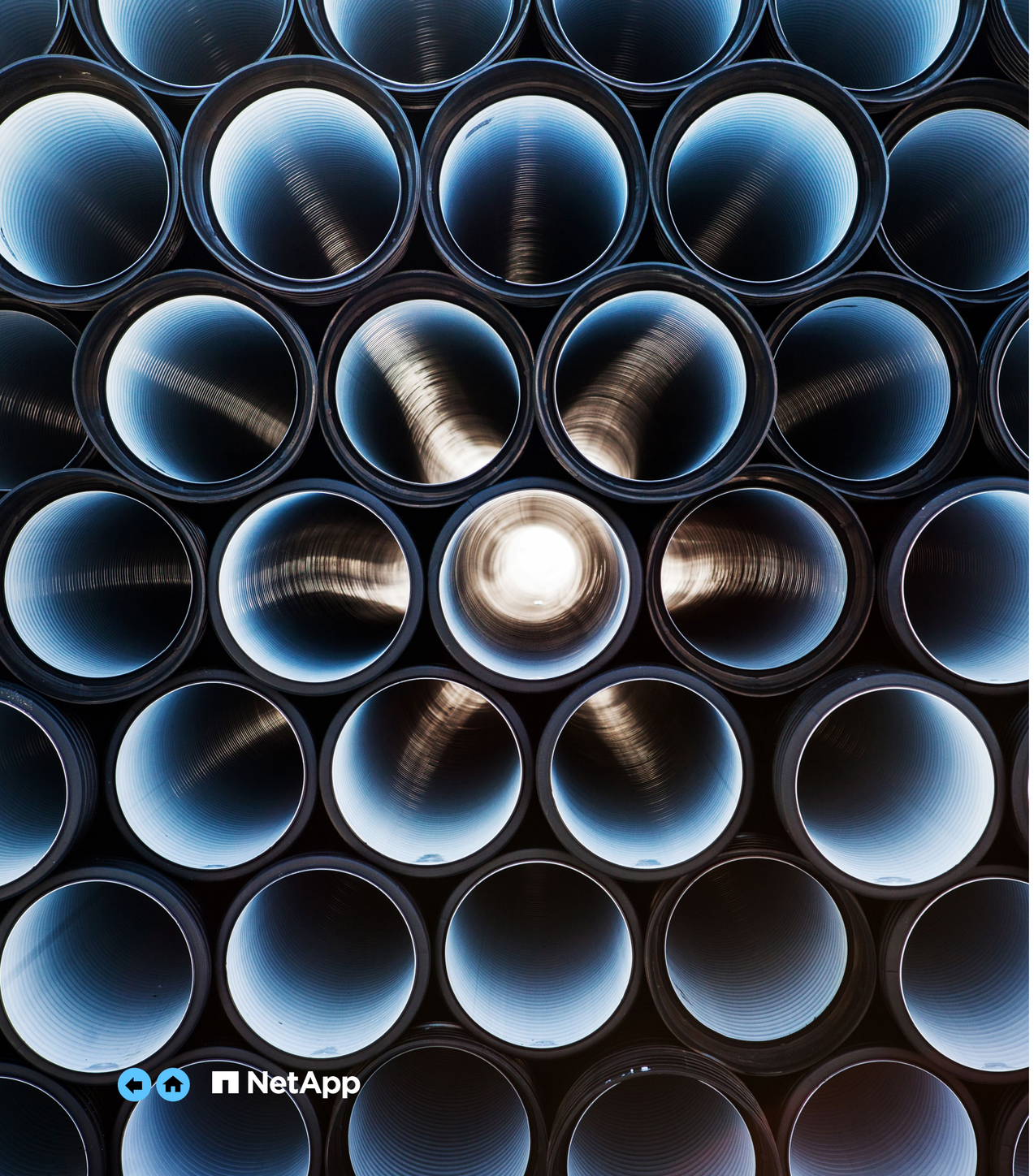
Google Translate prend en charge plus de 100 langues et recourt au crowdsourcing pour tester et améliorer les traductions ainsi que l'entraînement des modèles pour les langues peu représentées. Google Translate traite 140 milliards de mots chaque jour, ce qui équivaut au travail de 70 millions de traducteurs humains. *Tous les jours*.

Google BERT

Google BERT est un modèle de NLP très utilisé qui compte 340 millions de paramètres. BERT représente une avancée importante dans le domaine du NLP, car il va au-delà des interfaces voix transactionnelles, telles que les algorithmes de chaîne téléphonique, pour devenir réellement conversationnel. Il peut lire des textes et répondre aux questions de manière extrêmement précise.

BioMegatron

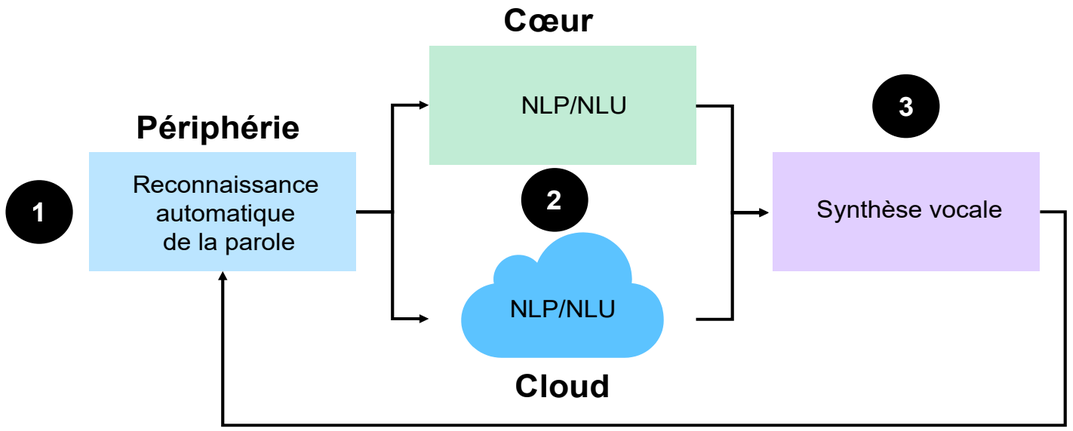
BioMegatron est le plus grand modèle linguistique biomédical de type Transformer jamais entraîné. Il compte jusqu'à 1,2 milliard de paramètres. Il a été entraîné à partir de 6,1 milliards de mots issus de PubMed, un référentiel de résumés et d'articles médicaux complets.



2. Fluidifiez les pipelines

Pour un NLP rapide et efficace, vous avez besoin d'un pipeline qui s'étend à votre écosystème tout entier, de l'ingestion à la reconnaissance, jusqu'à la synthèse vocale. Les données doivent circuler rapidement et librement d'une étape du pipeline à l'autre pour que le traitement du langage s'effectue en temps réel.

En général, un pipeline de NLP comprend trois étapes :



Dans une infrastructure de NLP moderne, des milliers d'emplacements en périphérie du réseau recueillent des téraoctets de données chaque jour. Malheureusement, quand l'accès à ces données est limité par une infrastructure en silo, le deep learning ne peut qu'effleurer la surface.

3. Répondez du tac au tac

Pour qu'une IA puisse parler comme un humain, elle doit fonctionner à la même vitesse qu'un cerveau humain, voire plus vite. Mais plus le modèle est vaste, plus le délai entre la question de l'utilisateur et la réponse de la machine augmente. Pour que la conversation semble naturelle, tous les calculs doivent s'effectuer en 300 millisecondes au maximum.

Ce processus se découpe en plusieurs étapes :

1. Conversion des paroles de l'utilisateur en texte
2. Compréhension du sens
3. Recherche de la meilleure réponse en contexte
4. Formulation de la réponse à l'oral

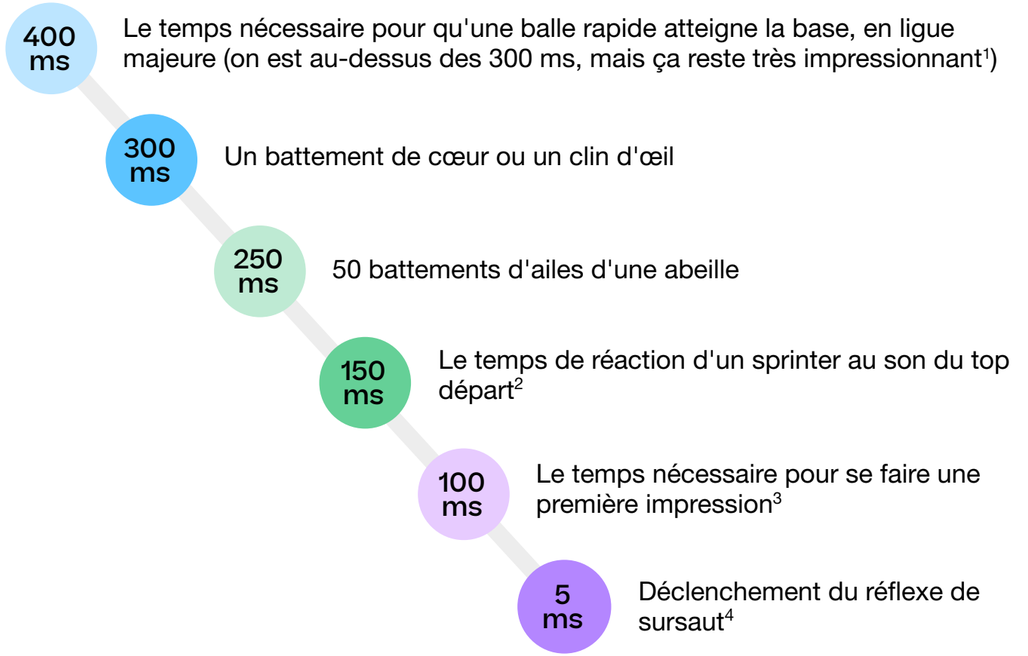
Au vu des exigences rigoureuses en matière de latence, les développeurs de système d'IA doivent souvent se résoudre à faire des compromis. Un modèle de traitement du langage complexe et de très haute qualité sera légèrement plus lent qu'un modèle plus léger qui répondra plus rapidement, mais de manière moins nuancée.

Comme nous, les assistants vocaux ont recours à des petites diversions pour gagner un peu de temps lorsqu'ils cherchent une réponse, « Laissez-moi chercher cela », ou émettent des bips ou autres sons pour combler un silence gênant. En revanche, l'IA conversationnelle idéale, le Graal du NLP, est suffisamment sophistiquée pour comprendre précisément les questions qu'on lui pose et assez rapide pour y répondre de façon parfaitement naturelle.

Quelques chiffres qui donnent le tournis

En général, un système de NLP peut fournir une réponse en moins de 300 millisecondes, soit 0,3 seconde. Rapide non ? Très !

Voici quelques exemples d'événements qui se déroulent en 300 millisecondes ou moins :



NetApp parle votre langue

Avec l'IA NetApp® ONTAP® optimisée par les systèmes NVIDIA DGX et les systèmes de stockage 100 % Flash connectés au cloud NetApp, il est possible d'entraîner et d'optimiser des modèles linguistiques de pointe pour accélérer l'inférence. Une Data Fabric optimisée par NetApp simplifie la gestion des données tout au long du pipeline de données d'IA, de la périphérie au cœur, et jusqu'au cloud.

- **Les solutions NetApp pour l'IA** éliminent les goulots d'étranglement pour optimiser la collecte de données, accélérer les workloads d'IA et faciliter l'intégration au cloud.
- **Les solutions de gestion unifiée des données de NetApp** permettent un déplacement fluide et économique des données dans l'environnement multicloud hybride.
- De renommée internationale, **l'écosystème de partenaires NetApp** propose des intégrations techniques complètes avec des leaders de l'IA, des partenaires channel et des intégrateurs système, des fournisseurs de logiciels et de matériel, et des partenaires cloud. Ils assemblent des solutions d'IA intelligentes, puissantes et fiables qui vous aident à atteindre vos objectifs commerciaux.
- **Les services professionnels de NetApp** fournissent l'expertise dont vous avez besoin pour réduire la complexité et tirer le meilleur parti de l'IA pour développer votre activité.

D'ailleurs, NetApp fait partie des leaders dans l'étude MarketScape d'IDC sur les fournisseurs de stockage fichier scale-out[5]. C'est important, car nos workloads de vision par ordinateur sont scale-out et basés sur des fichiers.

Simplifiez la vie de vos data scientists

5x

Traitez 5 fois plus de données dans votre pipeline d'IA

< 60 quelques secondes

Copiez des datasets en quelques secondes au lieu de plusieurs heures voire jours

~ 20 minutes

Configurez votre infrastructure d'IA en 20 minutes environ avec l'intégration Ansible

NetApp Retail Assistant : un projet prometteur

À l'aide de NVIDIA Jarvis, un framework complet pour la création de services d'IA conversationnelle, NetApp et NVIDIA ont développé NARA (NetApp Retail Assistant), un assistant virtuel pour la vente au détail. Capable de traiter les demandes orales comme écrites, l'assistant répond aussi aux questions sur la météo, des points d'intérêt et des tarifs en se connectant aux API weatherstack et Yelp Fusion et au kit de développement logiciel Python d'eBay. Découvrez-le maintenant.

L'assistant de vente NARA de NetApp s'appuie sur les éléments suivants :

- **NVIDIA Jarvis.** Jarvis fournit des services accélérés par processeur graphique pour l'IA conversationnelle via un pipeline de deep learning de bout en bout pour maintenir une faible latence.
- **NetApp ONTAP AI** Cette architecture reconnue combine les systèmes NVIDIA DGX et le stockage 100 % Flash de NetApp. ONTAP AI rationalise le flux de données de façon fiable, pour que vous puissiez entraîner et exécuter des modèles conversationnels complexes en respectant les exigences en matière de latence.
- **NVIDIA NeMo.** NeMo, un kit pour la création, l'entraînement et le réglage des modèles d'IA conversationnelle accélérés par processeur graphique, vous permet de construire vos modèles à l'aide d'API faciles à utiliser, avec des applications de reconnaissance automatique de la parole, de traitement du langage naturel et de synthèse vocale.

Le NLP, ça vous tente ?

Vous vous demandez peut-être ce qu'on va bien pouvoir inventer demain. Un outil de conversation avec les animaux ? Nous ne pouvons pas apprendre à parler à un écureuil, non. Par contre, nous pouvons vous montrer comment créer l'infrastructure d'IA idéale pour le NLP.

En savoir plus sur les solutions d'IA de NetApp :

- NetApp AI
- ONTAP AI
- Solutions NetApp pour le NLP

Des questions ? N'hésitez pas à contacter nos spécialistes en solutions d'IA.

1. O'Neill, Shane. « Real-time bidding: What happens in 200 milliseconds? », Nanigans.
2. Welsh, Tim. « Exactly how long does it take to think a thought? », The Christian Science Monitor, 1er juillet 2015.
3. Wargo, Eric. « How Many Seconds to a First Impression? », Association for Psychological Science, 1er juillet 2006.
4. Wise, Jeff. « What Is the Speed of Thought? », New York Magazine, 19 décembre 2016.
5. Potnis, Amita. IDC. MarketScape : Étude sur les fournisseurs de stockage fichier scale-out 2019. IDC. Décembre 2019.

À propos de NetApp
NetApp est un spécialiste dans un monde de généralistes. Nous nous fixons un seul objectif : aider votre entreprise à valoriser ses données. NetApp migre vers le cloud les services de données haute performance que vous utilisez, et apporte à votre data center la flexibilité du cloud. Nos solutions leaders du secteur fonctionnent dans de nombreux environnements clients et les principaux clouds publics.

En tant qu'entreprise spécialisée dans les logiciels et axée sur le cloud et les données, seul NetApp peut vous aider à créer votre propre Data Fabric, à simplifier et connecter votre cloud, et à fournir les données, les applications et les services adaptés aux personnes appropriées, en tout lieu et à tout moment.

Pour en savoir plus, consultez le site **www.netapp.com/fr**.