Génération de jeux de données de synthèse

Datalchemy

Génération de données pour faciliter l'entraînement d'une intelligence artificielle.

Pour quoi faire ?

Pour obtenir de meilleurs résultats d'IA en élargissant les jeux de données existants.

- Pour diminuer les coûts de constitution de datasets labélisés
- Pour une meilleure représentation des cas d'échec dans les données générées
- Pour pallier le manque de données en utilisant la majorité des données réelles pour la validation
- Pour l'initialisation d'une architecture data/IA et son démarrage
- Pour l'entraînement ou le réentraînement d'algorithmes
- Pour l'usage de données sensibles et difficiles à exporter à l'extérieur d'un SI pour apprentissage

En entrée

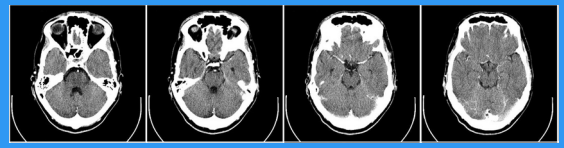
- Les jeux de données existants
- Les formats et les métadonnées
- Toutes les contraintes connues des objets et des processus, notamment d'acquisition et de labélisation

En sortie

- Un jeu de données de synthèse
- Un outil de génération de données
- Une architecture de données
- Un modèle ré-entraîné
- Une documentation et une évaluation de résultats

CAS D'USAGE TYPES

Un acteur industriel utilisant des robots pilotés par vision souhaite améliorer les performances des algorithmes utilisés mais l'acquisition de données nécessite d'arrêter la production. Nous générons des datasets d'images à partir de modélisations 3D, utilisées pour l'apprentissage du système de vision.



Un groupe leader de l'imagerie médicale souhaite développer de nouvelles approches d'analyse d'imagerie médicale mais manque de données d'entraînement. A partir d'un premier jeu de données labélisées, nous générons un volume suffisant à l'entraînement, permettant d'utiliser l'intégralité de la donnée réelle pour valider la preuve de concept.

Un acteur industriel ne pouvant diffuser ses données sur le cloud souhaite pouvoir entraîner ses modèles à l'extérieur à partir de données de synthèse sans enjeu de confidentialité.

Accélération de labélisation de datasets d'apprentissage

Datalchemy

Pré-traitement de données pour une réduction drastique du temps de labélisation.

Pour quoi faire ?

Pour pouvoir mettre en œuvre de l'IA dans des cas où la labélisation est un frein.

- Pour des gains de productivité majeurs (x10) sur le temps d'annotation
- Pour pouvoir atteindre rapidement un volume d'apprentissage suffisant
- Pour gagner du temps pour lancer un projet de recherche
- Pour mettre en œuvre de l'IA sur beaucoup plus de cas d'usage en valorisant/labélisant des jeux de données existants
- Pour disposer d'une analyse statistique sur des données existantes
- Pour mieux comprendre des processus réels et les données associées

En entrée

- Tous les jeux de données utiles, annotés ou pas
- Une description du processus de labélisation, incluant les formats d'entrée et de sortie
- Toutes les contraintes connues des objets et des processus

En sortie

- Une analyse des données étudiées et des recommandations
- Un algorithme de pré-labélisation de données
- Un outil d'annotation optimisant le processus
- Une documentation et une évaluation de résultats

CAS D'USAGE TYPES

Une entreprise dans le domaine du vivant analyse des images réalisées en pleine nature afin d'identifier des zones spécifiques. La labélisation d'une image prend 30 minutes en moyenne. Nos algorithmes permettent de diminuer ce temps à 3 minutes en moyenne.

Un grand groupe ferroviaire dispose d'un flux d'images réalisées à hauteur de rails et souhaite accélérer voire automatiser la labélisation des images en vue de la mise en place d'un modèle deep learning pour une détection de défauts automatisée.

Un fabricant de matériel industriel utilisant une caméra haute fréquence afin d'analyser les défauts de fonctionnement de ses appareils cherche à traiter ces vidéos avec de l'IA pour pré-sélection des images présentant des défauts potentiels.

Indicateur d'incertitude dans des résultats de deep learning

Datalchemy

Il est essentiel pour certaines activités d'assortir un résultat issu de deep learning d'une probabilité d'incertitude. Les modèles bayésiens aident à cela.

Pour quoi faire ?

Pour ajouter une notion d'incertitude (ou un niveau de certitude) à un résultat deep learning.

- Pour permettre à des utilisateurs d'aide à la décision IA de disposer d'un indice de confiance.
- Pour sécuriser l'automatisation de processus industriels sensibles.
- Pour minimiser les recours à un expert métier en ne portant à sa connaissance que les cas où l'incertitude dépasse un certain seuil.
- Pour favoriser l'acceptation d'aide au diagnostic par l'IA, notamment dans le domaine médical, et plus généralement rassurer les utilisateurs sur le fait que l'humain garde la mainmise sur les cas complexes.

En entrée

- Tous les jeux de données utiles
- Les modèles existants
- Disponibilité des experts métier

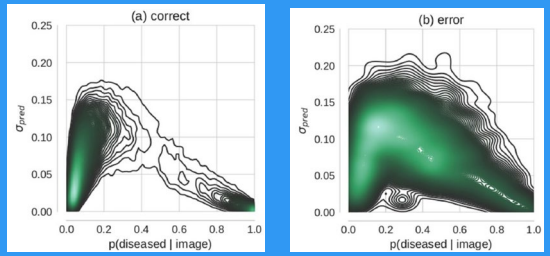
- Et en plus : une bonne analyse des enjeux décisionnels

En sortie

- Un prototype de test et son évaluation
- Une analyse des données étudiées

- Et aussi un état de l'art sur le sujet

CAS D'USAGE TYPES



Un client veut mettre en oeuvre du deep learning dans un domaine sensible où les faux négatifs représentent un danger à éviter absolument, et les faux positifs peuvent représenter une charge de travail importante de discrimination pour les experts. Réalisation d'un état de l'art et d'une preuve de concept aux résultats concluants (incertitude plus haute quand la donnée est corrompue ou en présence de distribution drift).

ENJEUX TECHNIQUES

- Etat de l'art sur les modèles bayésiens
- Comparaison des résultats vs. la performance des solutions possibles
- Etablissement d'une stratégie de développement et de tests incrémentaux pour obtenir les meilleurs résultats dans un budget défini
- Analyse des implémentations MC Dropout et frameworks Deepmind Laplace et tests
- Développement d'une preuve de concept et tests
- Confirmation de l'incertitude à partir de modifications contrôlées des données d'entrée

Transcription de voix vers texte pour les contrôleurs du ciel – DGAC

Datalchemy

Entraîner des modèles pour transcrire automatiquement les conversations entre des pilotes et la tour de contrôle dans l'aviation civile.

Pour quoi faire ?

Pour automatiser la transcription de conversations bruitées et éloignées des standards du langage naturel.

- Pour des besoins de formation à la DGAC
- Pour des besoins d'analyse post-OPS
- Pour une amélioration de la qualité des transcriptions
- Pour ouvrir un champ d'analyse en disposant de nouvelles données cruciales mais inédites dans ce format

En entrée

- Définition de la problématique
- Contraintes spécifiques
- Les jeux de données déjà existants

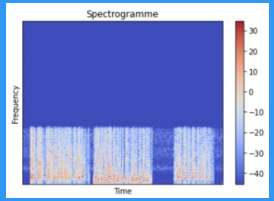
En sortie

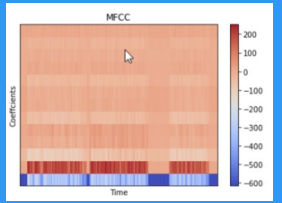
- Architecture data
- Modèles et algorithmes
- Résultats d'entraînement
- Evaluation des données de sortie

AUTRE CAS D'USAGES

Un acteur du transport ferroviaire souhaite retranscrire de manière automatisée les dialogues entre les superviseurs et les équipes d'intervention en cas de problème technique, ceci dans le but de disposer de retour d'expériences consolidés en bases de données.

ENJEUX TECHNIQUES





- Etat de l'art sur la modélisation du bruit et le débruitage
- Définition et modélisation de l'architecture data nécessaire
- Analyse des sources de données aux regard des algorithmes envisagés
- Modèles acoustiques
- Adaptation de modèles NLP (pour du langage technique contraint)
- Choix et paramétrage des modèles de langage et gestion du réentraînement

Anonymisation de données

Datalchemy

Détection précise d'informations permettant l'identification d'individus, et détection de références à des textes juridiques en vue d'une anonymisation et de la mise en place de liens automatiques vers Legifrance

Pour quoi faire et comment ?

Fournir à notre client un outil leur permettant de repérer toutes les informations personnelles et références juridiques dans un document donné.

- Etablir d'une analyse l'état de l'art
- Disposer d'une architecture de données adaptée
- Compléter les données en volume trop faible avec des données synthétiques
- Récupérer et labéliser par webscrapping de centaines de décisions du conseil d'état
- Sélectionnet les meilleurs modèles de langage
- Mettre en œuvre, entraîner et évaluer les modèles de langage sélectionnés
- Livrer d'un outil opérationnel (développé et documenté par nos soins)

En entrée

- Données type à traiter (contrats et autres documents juridiques)

En sortie

- Des jeux de données d'apprentissage générés automatiquement
- Outil de détection des informations personnelles et références juridiques
- Documentation complète

CAS D'USAGE TYPES

Un éditeur juridique qui souhaite utiliser une base de contrats réels, et de décisions juridiques pour constituer une base de données de modèles génériques, et donc sans référence pouvant permettre l'identification des parties prenantes.

Un groupe d'écoles supérieures cherche à mettre en œuvre des algorithmes d'intelligence artificielle sur les données concernant les élèves et doit procéder à une identification et une anonymisation automatisée de l'ensemble de leurs données disponibles.

Un opérateur télécoms et fibre optique souhaite intégrer les données de compte-rendu d'intervention chez les clients pour entraîner des modèles de machine learning, mais doit s'assurer que les données d'identification des clients sont supprimées.

Débruitage et super résolution Imagerie médicale – GE Healthcare

Datalchemy

La manipulation d'images médicales nécessite de s'assurer que les prétraitements qui peuvent être réalisés ne font pas disparaître des informations importantes.

Pour quoi faire ?

Analyser l'état de l'art pour mettre à jour les connaissances des experts d'imagerie médicale, à partir de l'état de l'art sur les algorithmes de débruitage et de super résolution (images médicales ou non), en mesurant risques et défis pour chacun.

- Pour disposer d'une analyse de l'état de l'art sur un sujet pointu
- Pour réaliser une analyse de risque des différents modèles possibles, après contrôle du code quand cela est possible, et mesurer le risque de perte d'information
- Pour évaluer la robustesse
- Pour être en mesure de dispenser du meilleur niveau de conseil sur un sujet clé

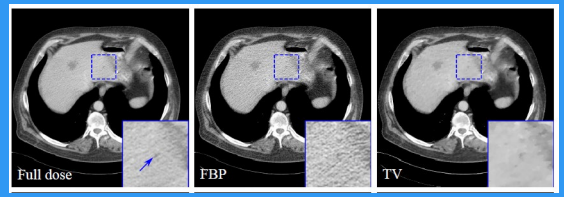
En entrée

- Données scientifiques liées aux cas d'usage spécifiques
- Contraintes d'exploitation
- Description des outils connus ou utilisés

En sortie

- Analyse de l'état de l'art
- Formation d'un public d'experts
- Documentation de comparaison de modèles et d'algorithmes disponibles

CAS D'USAGE TYPES



Un grand groupe spécialisé dans la fabrication de matériels d'imagerie médicale et de solutions logicielles de traitement d'images souhaite faire progresser en permanence les technologies qui les sous-tendent.

La question du pré-processing des images issus des appareils permet de pouvoir envisager une amélioration de la taille ou de la qualité des images brutes, mais alors la question de la perte d'information cruciales devient essentielle à maîtriser.

Si une tumeur est potentiellement détectable dans une image, il faut qu'elle reste aussi détectable quels que soient les traitements que l'image brute pourrait avoir subis.

L'étude de la littérature scientifique et l'analyse détaillée des meilleurs algorithmes candidats a permis de déterminer exactement les avantages et inconvénients de chacun d'entre eux afin de déterminer aussi bien des axes d'évolution des solutions, que des programmes de formations pour les experts en charge des évolutions des produits et services manipulant les images.

Consolidation juridique automatisée – DILA

Datalchemy

Les Services Généraux du Premier Ministre sont en charge de la publication du journal officiel (JO) et de la tenue à jour du site Légifrance qui diffuse les textes de loi et les décrets.

Pour quoi faire ?

La modification des textes de loi à partir des textes modificateurs parus au Journal Officiel est une tâche fastidieuse. Elle est automatisable grâce à l'intelligence artificielle.

- Pour identifier les meilleurs algorithmes de traitement du langage naturel
- Pour disposer une base d'apprentissage adaptée au vocabulaire juridique
- Pour maîtriser tous les cas complexes qui rendent la consolidation difficile
- Pour construire une architecture de données adaptées
- Pour faire gagner un temps précieux aux opérateurs en éliminant la détection dans 90% des cas pour ne garder que la validation

En entrée

- La totalité des textes de loi français
- Les règles de rédaction propres aux textes juridiques
- Les règles et processus de consolidation

En sortie

- Une architecture de données adaptée
- On modèle de langage adapté
- Le code de traitement
- Un déploiement sur un serveur de test
- Un rapport d'analyse et de résultats
- Une évaluation des résultats sur trois jours du JO

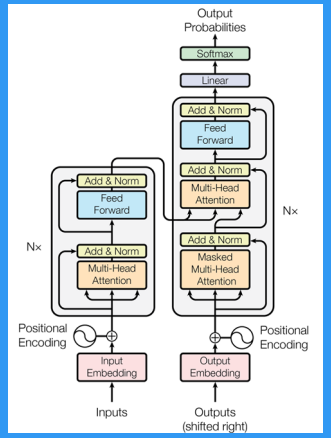
LES ENJEUX SPÉCIFIQUES

Plusieurs enjeux sont à prendre en compte :

- identifier ce qui est une modification dans l'annonce du JO
- Identifier ce qui doit changer (article, alinea, groupe de mots) et comment (suppression, modification, ajout)
- Identifier la référence du texte qui doit être changé
- Identifier la partie à changer dans le texte cible

Nous avons spécialisé le modèle BERT (RoBERTa) sur les textes juridiques

Un enjeu crucial de ce projet : obtenir les meilleurs résultats possibles en seulement 4 mois.



Entraînement et mise en œuvre du réseau PointNet

Datalchemy

A partir des recommandations d'un article de recherche, entraîner le réseau PointNet pour une entreprise majeure du secteur médical.

Pour quoi faire ?

Entraîner le réseau PointNet afin d'évaluer ses usages potentiels pour des besoins d'imagerie médicale. Le but était d'entraîner le réseau sur des données médicales publiques en travaillant sur les hyperparamètres, pour permettre au client d'évaluer les résultats du modèle sur leurs propres données.

- Pour disposer d'une analyse des données publiques disponibles
- Pour disposer d'un modèle entraîné
- Pour construire une architecture data adaptée
- Pour sélectionner des jeux de données publics pertinents
- Pour développer un code réutilisable documenté, avec un rapport de résultats

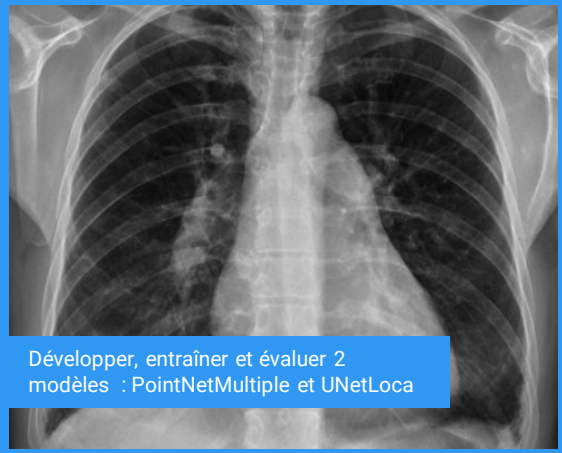
En entrée

- Une liste des jeux de données publics d'imagerie
- Le code développé par le client, non fonctionnel

En sortie

- Des jeux de données sélectionnés
- Une architecture de données
- Une documentation et une évaluation de résultats
- Un code de traitement opérationnel
- Un rapport d'évaluation complet

CAS D'USAGE TYPE



Développer, entraîner et évaluer 2 modèles : PointNetMultiple et UNetLoca

Le client est un acteur majeur de l'imagerie médicale, et pense que PointNet, une avancée académique majeure de 2017 dans le traitement des images 3D, devrait résoudre un de leurs besoins actuels, mais leurs tentatives de mise en œuvre ont jusque-là échoué.
Ils font appel à Datalchemy pour entraîner et déployer le modèle sur des données publiques (pour des questions de confidentialité des données patient) et vont ensuite évaluer les résultats du modèle entraîné sur leurs propres données.

Le meilleur modèle évalué a été le réseau PointNetMultipe, le modèle de convergence pour UNetLoca était bien plus lent.
Les résultats sur les données client ont été concluants et le modèle part en industrialisation.

Environnement de simulation de réseau ferroviaire – SNCF

Datalchemy

Simulation complète du réseau ferré français, avec les circulations, les jauges des gares, les contraintes des conducteurs, les flux passagers et les incidents.

Pour quoi et comment faire ?

Mettre en œuvre un simulateur complet afin de permettre l'intégration d'outils de deep learning et la formation.

- Pour disposer d'un environnement modulaire de simulation permettant le test de modèles de deep learning
- Pour développer un outil de formation des agents terrain

- Intégration de plusieurs micro et macro simulateurs
- Nettoyage, factorisation et optimisation
- Industrialisation et mise en production

Projet intégré en production dans l'environnement R&D du client

CAS D'USAGE TYPES

Le gestionnaire du réseau ferré souhaite disposer d'un modèle de simulation de son réseau afin de mesurer l'impact des circulations sur différents paramètres influençant notamment la maintenance et l'asset management.

L'opérateur de transport ferroviaire souhaite disposer d'un environnement de simulation des circulations afin de pouvoir former ses contrôleurs de trafic.

Un gestionnaire de réseau autoroutier souhaite disposer d'un simulateur de fonctionnement de son réseau afin de pouvoir aisément tester des scénarios de résilience ou de réaction à des enchaînements d'événements spécifique. Il souhaite aussi en faire la première brique d'une plateforme permettant notamment la prédiction de trafic ou de survenance d'événements.

Apprentissage non-supervisé et inférence variationnelle

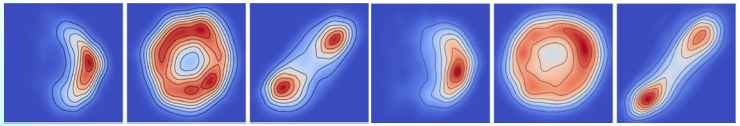
Datalchemy

Utiliser l'apprentissage non supervisé et l'inférence variationnelle pour le clustering, pour identifier des données manquantes dans les jeux de données et la détection des données aberrantes.

Pour quoi ?

Aider un client qui dispose d'un jeu de données de taille très importante à avoir une meilleure compréhension de ses données, à trouver les données aberrantes et prévoir le meilleur processus d'annotation en vue de développements deep learning.

- Pour disposer d'une vision complète de l'état de l'art à date
- Pour procéder à de l'inférence variationnelle avec des modèles génératifs (VAE/Flows) sur les vecteurs latents
- Pour détecter les valeurs aberrantes ou atypiques et estimer la proximité potentielle
- Pour pouvoir faire des recommandations sur des jeux de données déséquilibrés



L'inférence variationnelle rend possible de représenter une distribution de données et d'obtenir des résultats porteurs de sens. La normalisation des flux est une de ces méthodes.

En entrée

- Jeu de données
- Disponibilité des experts métier

En sortie

- Analyses et données transformées
- Modèles à appliquer

CAS D'USAGE TYPE

Une entreprise génère de très nombreuses données. Il devient de plus en plus difficile de prendre des décisions éclairées sur les zones manquant de données pour un bon équilibre des jeux d'apprentissage. De plus, la détection des données aberrantes est cruciale, quitte à développer un outil spécifique pour cela,
La labélisation a été fortement accélérée en concentrant le travail sur les sous-groupes manquants.

ENJEUX SPÉCIFIQUES

- Analyse statistique des données
- Entraînement de modèles supervisés pour l'extraction de vecteurs latents et leur analyse
- Entraînement en inférence variationnelle avec deux méthodes (VAE et Flux) sur les données actuelles ou les vecteurs latents
- Tests de réduction de dimension sur les les vecteurs latents (t-SNE, uMAP)
- Analyse de regroupements/clusters réalisés avec différentes méthodes
- Détection de valeurs aberrantes basé sur une approche VAE croisée avec les clusterings
- Recommandations pour un meilleur jeu de données

Prédire la détérioration de la santé des personnes en maintien à domicile – Groupe Up

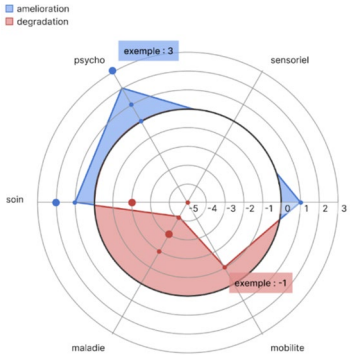
Datalchemy

Aider les établissements d'aide à domicile à générer automatiquement des alertes quand les conditions de dépendance de leurs bénéficiaires se dégradent.

Comment ?

Utiliser les champs commentaires enregistrés entre les aides à domicile et leurs entreprises pour prédire la dégradation des états de dépendance des bénéficiaires.

- En anonymisant les données
- En réalisant au préalable des analyses statistiques sur les données
- En initialisant une architecture data et en la mettant en œuvre
- En entraînant des modèles de langage et en les déployant
- En concevant et en développant une expérience utilisateur



Projet en cours de mise en production en 2023

En entrée

- Les données fournies par des sociétés d'aide à domicile
- Des ateliers avec des experts métier pour mieux comprendre les indicateurs clés

En sortie

- Architecture de données adaptées
- Modèles entraînés à utiliser
- Interfaces utilisateurs pour consulter les résultats

CAS D'USAGE TYPES

Un consortium souhaite développer un outil pour les entreprises d'aide à domicile afin qu'elles disposent d'alertes prédictives sur la dégradation de l'état des bénéficiaires de leurs services, et que leur niveau de dépendance se détériore. Ils perçoivent la valeur ajoutée de l'intelligence artificielle mais ne disposent pas de ressources en interne pour réaliser ce travail au meilleur niveau de performance et à l'état de l'art.

Une entreprise de maintenance industrielle souhaite utiliser le contenu des comptes-rendus d'intervention pour prédire la survenance de panne ou de dégradation de l'état des installations maintenues.

ENJEUX SPÉCIFIQUES

- Anonymisation des données
- Echanges avec les experts métiers pour comprendre les facteurs de risque quel que soit le secteur
- Entraîner le modèle de langage FlauBERT et l'appliquer aux données