Publication autour de l'intelligence artificielle

Chapitre 1 : Quelles sont les applications de l'IA dans le monde du WEB ?

Auteur : ÉTIENNE COLLET
Co-auteur : MEHDI ZAÏDI
Relecteur : CLÉMENCE ARNOULD

Résumé

Cet article a pour but informatif et a vocation à ne pas plonger pleinement dans les aspects obscurs du sujet. Nous allons évoquer deux catégories d'IA majeures dans le domaine du WEB (Générative et Classificative) et s'intéresser à comment intégrer et aborder ces éléments au sein d'un projet, tout en respectant les contraintes du WEB.

Résumé

Comment intégrer et aborder l'IA au sein d'un projet ou d'une application WEB, tout en respectant les contraintes de notre domaine



Table des matières

1 Les types majeurs de l'IA **2**
1.1 L'IA générative . . . 2
1.1.1 Fonctionnement . . . 2
1.1.2 Les utilisations possibles . . . 2
1.2 L'IA classificative . . . 3
1.2.1 Fonctionnement . . . 3
1.2.2 Les utilisations possibles . . . 3
1.3 Importance de la supervision humaine . . . 3

2 Exemple et aspects techniques **3**
2.1 Outils . . . 4
2.1.1 Python et ses librairies . . . 4
2.1.2 Google Collab . . . 4
2.1.3 Exemple de code . . . 5
2.2 Tokenisation . . . 7
2.3 Apprentissage . . . 8
2.4 Fine-Tuning . . . 8

3 Les contraintes des domaines **8**
3.1 Impact environnemental . . . 8
3.2 Coût . . . 8
3.3 RGPD/Confidentialité . . . 9

4 Conclusion **9**

L'intelligence artificielle (IA) est un domaine fondamental de l'informatique visant à créer des systèmes complexes capables d'exécuter ou de résoudre des tâches qui, initialement, requièrent une intelligence humaine. On appellera le système résultant un **modèle**, les cas pratiques sont nombreux et terriblement efficaces. Les plus connus actuellement sont des modèles comme Mistral, ChatGPT, Llama ou Gemini pour la gestions des langages naturels. Ce sont tous des Large Language Models (LLM). Les IA sont utilisées dans divers secteurs professionnels pour améliorer l'efficacité et l'innovation. Par exemple, dans le secteur de la santé, elle aide au diagnostic, aux traitements personnalisés, etc. En finance, elle détecte les fraudes, gère les risques et fournit des conseils. Du coté des ESN et du WEB, l'IA est très utilisée en interne pour assister, aider et automatiser de nombreuses tâches. Cependant, les implémentations proposées par les entreprises se limitent encore à la mise en place de chatbots et d'assistants virtuels qui automatisent le support client.

1 Les types majeurs de l'IA

Il existe un très grand nombre de modèles différents comprenant chacun leurs spécificités techniques. Chaque modèle d'IA peut être combiner à un autre, ce qui leur permet d'être complémentaires. Chez Globalis, nous nous intéressons à deux modèles, qui sont ceux qui ont le plus d'applications avec le métier : l'IA générative et l'IA classificative.

1.1 L'IA générative

Une IA générative est conçue pour créer du contenu nouveau et original en s'inspirant d'un grand volume de données existantes qui lui ont été transmises lors de sa création. En se référent à du contenu créé par les humains, les résultats produits par ce type de modèle semblent, eux aussi, avoir été créés par un humain.

1.1.1 Fonctionnement

Une IA, quelle qu'elle soit, est fondée sur un grand jeu de donné que le modèle utilise pour s'entraîner à la tâche qu'il doit réaliser. Par exemple, pour une IA générative de texte, elle est alimentée par des milliers de livres, articles, et autres documents écrits. Ces données doivent être nettoyées et organisées pour permettre à l'IA de les interpréter. Les modèles sont conçus pour imiter la façon dont le cerveau humain manipule l'information. Dès lors que le modèle est entraîné, celui-ci utilise les informations apprises pour prédire ce qui vient ensuite dans une séquence de données et ainsi générer du contenu. Par exemple, si nous demandons la rédaction d'un texte, l'IA écrit des phrases qui suivent une certaine logique et un style appris durant l'entraînement afin de le générer. Par la suite, il est tout à fait possible d'ajuster les données fournies pour obtenir des résultats différents, pour assurer la cohérence des donnés générées.

1.1.2 Les utilisations possibles

La génération de textes, d'images, de sons ou d'autres données numériques est devenue très simple grâce aux modèles d'IA génératifs. Il est désormais facile de créer une IA capable de générer des gabarits spécifiques (lettres d'avocat, articles de presses, etc). Il est également possible de réaliser de la complétion de texte ou de document, en modifiant légèrement les formats de données. Les modèles peuvent générer des graphiques complexes à partir de données et inversement, déterminer les données potentielles à partir d'un graphique. La capacité de génération est pratiquement illimitée, tant la base d'entraînement est vaste. Il est aussi possible de créer des mails de relance personnalisés pour des plate-formes, voire des images plus attractives pour la vente de produits.

L'accessibilité est un domaine majeur du Web, comme le montre cet article publié par Globalis `https://www.globalis-ms.com/actualites/accessibilite-introduction/`. Dans ce domaine, les modèles d'IA peuvent apporter des contributions significatives avec, par exemple, de la transcriptions audio qui peut aider les malvoyants à mieux naviguer sur une page internet. De même, le remplissage de formulaires par la parole devient facilement accessible, simplifiant l'interaction pour les utilisateurs ayant des difficultés avec les interfaces traditionnelles.

1.2 L'IA classificative

Une IA classificative est un modèle d'intelligence artificielle conçu pour classifier des données dans des catégories définies. Le modèle apprend, à partir d'un ensemble de données d'entraînement étiquetées, à prédire la classe d'appartenance d'une nouvelle observation non étiquetée. Ce type d'IA est largement utilisé dans divers domaines tels que la reconnaissance d'images ou la prédiction médicale.

1.2.1 Fonctionnement

Pour entraîner un modèle d'IA classificative, il est nécessaire de créer un ensemble de données où chaque exemple est associé à une étiquette de classe connue. Par exemple, pour la reconnaissance d'images d'animaux, chaque image est étiquetée par l'animal qu'elle contient. Il est important d'avoir des données homogènes et normalisées afin d'avoir un format adapté à l'entraînement du modèle. Les modèles couramment utilisés incluent les réseaux de neurones, les machines à vecteurs de support (SVM), et les arbres de décision. Chaque modèle a ses avantages selon le type de données et la complexité du problème. Le modèle est entraîné de sorte à minimiser une fonction de perte, qui mesure la différence entre les prédictions du modèle et les véritables étiquettes de classe.

Une fois entraînée, l'IA classificative peut prédire la classe d'appartenance d'une nouvelle observation non étiquetée. Par exemple, elle peut identifier si une image contient un chat ou un chien. La performance de l'IA est évaluée à l'aide de différentes mesures qui permettent de quantifier à quel point le modèle est capable de classer correctement les données. Le modèle peut toujours être affiné en ajustant ses hyperparamètres, en augmentant la taille de l'ensemble de données d'entraînement, ou en utilisant des techniques plus avancées.

1.2.2 Les utilisations possibles

Dans l'univers du web, la classification est tout aussi, voire même plus, importante que la génération de contenu, comme le montre clairement le SEO.

Les modèles utilisés par les moteurs de recherche peuvent être réduits pour se limiter à un seul site, permettant ainsi d'orienter rapidement l'utilisateur vers la ou les pages souhaitées. Nous pouvons facilement imaginer que des commerciaux souhaitant classifier des CVs selon leur qualité puissent créer une base d'entraînement via les IA classificatives. Une autre application possible est la détection de messages injurieux ou non autorisés dans une messagerie ou un courriel. Il est également envisageable de reconnaître et de trier automatiquement les documents téléchargés sur un site, tels que des factures, des contrats ou des formulaires, facilitant ainsi leur gestion. Enfin, dans le domaine du e-commerce, il est possible de classifier les utilisateurs en différents groupes, en fonction de leur comportement, de leurs intérêts et de leurs interactions, afin de mener des campagnes marketing plus ciblées.

1.3 Importance de la supervision humaine

Même si les IA génératives et classificatives peuvent produire des résultats impressionnants, elles nécessitent toujours une supervision humaine pour garantir que le contenu est correct, approprié, précis, et éthique. De plus, seul l'évaluation humaine peut permettre de faire évoluer un modèle vers l'objectif qui lui est fixé.

2 Exemple et aspects techniques

Quelque soit le niveau de compréhension de l'univers de l'IA, il y a toujours des parts d'ombre et d'incompréhension. Cependant, lorsque l'on ne sait pas par où débuter, il y a deux méthodes pour acquérir des compétences : Utiliser les modèles déjà existant comme Mistral ou bien créer et entraîner soi-même ses modèles. Nous allons nous intéresser uniquement à cette seconde option.

2.1 Outils

2.1.1 Python et ses librairies

Python est l'un des langages de programmation les plus populaires pour le développement d'intelligence artificielle, en raison de sa simplicité, sa lisibilité et sa vaste collection de bibliothèques spécialisées. Des outils comme TensorFlow, PyTorch et Keras facilitent la création, l'entraînement et le déploiement de modèles d'IA. TensorFlow et PyTorch sont particulièrement puissants pour les réseaux de neurones et l'apprentissage profond. Keras, qui peut s'appuyer sur TensorFlow en backend, offre une interface haut niveau pour construire et entraîner des modèles de réseaux de neurones de manière plus intuitive. Ces bibliothèques permettent aux développeurs de réaliser des tâches variées telles que la reconnaissance d'image, la compréhension du langage naturel, et la prévision de séries temporelles avec une efficacité accrue.

2.1.2 Google Collab

Google Colab est une plateforme de développement basée sur le cloud qui permet aux utilisateurs d'écrire, d'exécuter et de partager des notebooks interactifs contenant du code Python. Elle offre un accès gratuit à des ressources de calcul telles que des GPU et des TPU. Les notebooks créés dans Google Colab peuvent intégrer du texte, du code exécutable, des visualisations et des commentaires, facilitant ainsi la collaboration et le partage de projets entre utilisateurs. Ces points en font la plateforme gratuite la plus utilisée pour l'IA.

2.1.3 Exemple de code

Dans un premier temps, voici un exemple de code pour entraîner votre premier modèle exécutable avec Google Collab.

```
# Importer les bibliothèques nécessaires
import tensorflow as tf
from tensorflow.keras.datasets import mnist
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Flatten
from tensorflow.keras.utils import to_categorical

# Charger les données MNIST ( Une base d'image )
(x_train, y_train), (x_test, y_test) = mnist.load_data()

# Préparer les données
x_train = x_train / 255.0
x_test = x_test / 255.0
y_train = to_categorical(y_train)
y_test = to_categorical(y_test)

# Construire le modèle
model = Sequential([
    Flatten(input_shape=(28, 28)),
    Dense(128, activation='relu'),
    Dense(10, activation='softmax')
])

# Compiler le modèle
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Entraîner le modèle ( le nombre d'epochs est le nombre d'itération d'entrainement )
model.fit(x_train, y_train, epochs=5, validation_data=(x_test, y_test))

# Évaluer le modèle
test_loss, test_acc = model.evaluate(x_test, y_test)
print('Test accuracy:', test_acc)

# sauvegarde le modèle dans un fichier
model.save("number_classification.h5")
```

Une fois le modèle créé et déployé, celui-ci peut être utilisé en lui envoyant une requête. Par exemple, si l'on veux appeler le modèle de MistralAI depuis un script PHP, le code minimal suivant fonctionne :

```php
<?php
// Donné initiales
$api_url = 'https://api.mistral.ai/v1/chat/completions';
$api_key = 'your_api_key_here'; // Remplacez par votre clé API
$data = array( // Données à envoyer à l'API
    'model' => 'mistral-large-latest',
    'messages' => [['role' => 'user', 'content' => 'Bonjour, comment ça va?']],
    'max_tokens' => 50,
);

$ch = curl_init($api_url);

// Configuration des options CURL
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_HTTPHEADER, array(
    'Content-Type: application/json',
    'Authorization: Bearer ' . $api_key
));
curl_setopt($ch, CURLOPT_POST, true);
curl_setopt($ch, CURLOPT_POSTFIELDS, json_encode($data));

// Exécute la requête CURL et capture la réponse
$response = curl_exec($ch);

// Vérifie les erreurs CURL
if ($response === false) {
    $error = curl_error($ch);
    echo 'Erreur cURL: ' . $error;
} else {
    // Traite la réponse
    $response_data = json_decode($response, true);
    echo 'Réponse de l\'API MistralAI: ';
    print_r($response_data);
}

// Ferme la session cURL
curl_close($ch);
?>
```

Enfin, si cette fois-ci on veux accéder a l'API de chatGPT depuis la partie client d'une application, le script TypeScript suivant fonctionne. Cependant, celui-ci requière la librairie d'OpenAI, que l'on peut installer avec la commande suivante :

```
npm install --save openai
```

De plus, il est possible de retrouver la documentation ainsi que des exemples semblables à celui ci-dessous sur le lien suivant : OpenAI Platform

```
const apiKey = 'YOUR_API_KEY';
const openai = new OpenAI({
    apiKey: apiKey,
    dangerouslyAllowBrowser:true
    // Cette option est importante lorsque le scipte est utiliser sur un site internet
});
async function main() {
    const completion = await openai.chat.completions.create({
      messages: [{ role: "system", content: "You are a helpful assistant." }],
      model: "gpt-4o-mini",
    });

    console.log(completion.choices[0]);
}

main()
```

2.2 Tokenisation

La tokenisation est un processus clé dans le traitement du langage naturel (NLP) et l'analyse de texte. Elle consiste à diviser un texte en unités plus petites, appelées "tokens". Un token peut être un mot, une phrase ou même un caractère, selon le contexte et les besoins de l'analyse. Par exemple, dans la phrase "Le chat dort", la tokenisation pourrait produire les tokens "Le chat" et "dort".

Le processus de tokenisation commence par la segmentation du texte, généralement en supprimant les espaces et les ponctuations pour isoler les mots ou les phrases. Cette étape peut inclure la gestion de cas spéciaux, comme les contractions ou les noms composés. Ensuite, des techniques plus sophistiquées peuvent être appliquées.

Ce traitement est essentiel car il permet de transformer un texte brut en une forme manipulable et analysable par les modèles d'IA. Cela facilite la recherche de mots-clés, l'analyse de la fréquence des termes, et la compréhension contextuelle des textes.

2.3 Apprentissage

L'apprentissage d'un modèle est le processus par lequel un algorithme est entraîné, à partir de données, à effectuer une tâche spécifique, telle que la classification, la génération, ou la reconnaissance de motifs.

Pendant l'entraînement, l'algorithme ajuste pas à pas ses paramètres internes en minimisant une fonction de coût, souvent à l'aide de techniques d'optimisation, telle que la descente de gradient, cela dans l'objectif de réduire l'erreur entre les prédictions du modèle et les valeurs réelles.

Une fois l'entraînement terminé, le modèle est évalué sur l'ensemble de test pour vérifier sa capacité à généraliser ses prédictions à de nouvelles données. Si les performances sont satisfaisantes, le modèle peut être déployé pour faire des prédictions sur de nouvelles données. Sinon, des ajustements peuvent être nécessaires, tels que la modification de l'architecture du modèle, l'ajout de plus de données, ou l'utilisation de techniques de régularisation pour éviter le sur-apprentissage. Chacune de ces modifications implique de relancer un apprentissage complet du modèle.

En somme, l'apprentissage d'un modèle est un processus itératif et méthodique qui vise à créer des algorithmes capables de prendre des décisions ou de faire des prédictions basées sur des données.

2.4 Fine-Tuning

Le fine-tuning est une technique qui vise à ajuster un modèle pré-entraîné sur corpus de données pour qu'il s'adapte à une tâche spécifique. Par exemple, un modèle de langage pré-entraîné sur des milliards de mots peut être affiné pour une application spécifique comme la classification de sentiments dans des critiques de films.

Le processus de fine-tuning commence par la sélection d'un modèle qui a déjà appris des représentations générales du langage, tel que Mistral. Ensuite, ce modèle est ré-entraîné (ou affiné) sur un jeu de données plus petit et spécifique à la tâche cible. Il a récemment été prouvé que le formatage des données impacte les résultats du modèle résultant[1]. Pendant cette phase, les paramètres du modèle sont ajustés pour minimiser l'erreur sur la nouvelle tâche tout en conservant les connaissances générales acquises lors de l'entraînement initial.

Le fine-tuning est crucial car il permet d'obtenir des performances élevées avec relativement peu de données spécifiques. Au lieu de partir de zéro, nous utilisons la puissance des modèles génériques et nous les adaptons à des cas d'utilisation particuliers, ce qui économise du temps et des ressources de calcul.

3 Les contraintes des domaines

L'IA est certes un sujet qui révolutionne de nombreux domaines, mais cette puissance de résolution implique des éléments importants à connaître et à prendre en compte avant de vouloir l'intégrer.

3.1 Impact environnemental

L'entraînement d'un modèle est très coûteux et, en terme de temps, est tout autant coûteux d'un point de vue énergétique. La base d'entraînement et les ressources de calculs nécessaires pour l'entraînement du modèle affectent énormément les résultats finaux de l'IA. Par exemple les grands modèles tels que mistral-large ont nécessité environ deux mois d'entraînement sans interruption sur une infrastructure de pointe. Ainsi, quelques soient les dimensions du modèle, si l'on prend en compte l'entraînement, le fonctionnement de l'IA et les requêtes qui lui sont envoyées, il est actuellement impossible d'avoir un bilan énergétique sobre/propre.

3.2 Coût

Ce lourd coût énergétique apporte inévitablement un coût financier. Que ce soit pour l'électricité, l'achat des installations et la main-d'oeuvre, en plus des connaissances de pointe qui sont nécessaires pour avoir des résultats performants. Pour les applications les plus courantes, l'utilisation d'outils existants comme Mistral permet un coup de quelques dizaines d'euros par mois, tandis que la création des énormes modèles que l'on connaît coûte plusieurs milliard d'euros. Enfin la création de petits modèles spécialisés dans la réalisation de petites tâches pointues vont plutôt avoisiner le millier d'euro, en comptant la main-d'oeuvre, l'entraînement et la validation des performances.

3.3 RGPD/Confidentialité

Enfin, la partie précédente amène une question évidente : si un modèle large coûte un milliard et que chaque requête ne coûte que quelques centimes, comment une entreprise peut-elle être rentable ? Une partie de la réponse est que les données qui sont transmises et générées par ces modèles sont utilisées par l'entreprise, que ce soit pour mieux connaître leurs clients, pour former leur future base d'entraînement ou encore, simple hypothèse, pour être revendu. De nombreuses polémiques émergent sur ce sujet et les limites ne sont pas encore clairement définies. Ainsi, pour des données sensibles, il est souvent préférable d'opter pour la création de sa propre IA, garantissant la source et l'utilisation des données.

4 Conclusion

L'IA révolutionne le monde et celui du WEB n'en fait pas exception. Malgré cela, tous les modèles qui existent n'y sont pas adaptés et seuls quelques-uns ont actuellement du sens. Aujourd'hui, il est important de comprendre comment fonctionnent les différents modèles d'IA afin que leur utilisation soit la plus pertinente possible, en fonction des secteurs et des besoins.

Références

[1] Kuniaki Saito, Kihyuk Sohn, Chen-Yu Lee, and Yoshitaka Ushiku. Where is the answer ? Investigating Positional Bias in Language Model Knowledge Extraction. 2024.