10/03/2024

Rapport individuel

SAE – Création d'un système de chatbot conversationnel base sur GPT-2

Pascal Zhan

DRACOLIA

Table des matières

Introduction

Dans un contexte où technologie et santé se fondent de plus en plus, l'innovation dans le domaine médical devient cruciale. Notre projet s'insère dans cette tendance en exploitant les capacités avancées de l'intelligence artificielle pour offrir des diagnostics préliminaires. Notre but n'est pas de remplacer les experts de la santé, mais plutôt de développer un chatbot intelligent, basé sur le modèle GPT-2, capable de répondre de manière précise et pertinente aux questions des utilisateurs, en améliorant les performances du modèle GPT-2 initial.

Notre équipe de sept a collaboré en répartissant les tâches pour englober toutes les phases de développement du projet. Ce travail nous a permis d'aborder différents domaines, de la compréhension technique de GPT-2 à l'analyse de données spécifiques au secteur de la santé, ainsi que l'amélioration et l'optimisation de notre modèle selon les besoins des utilisateurs.

Ce rapport détaille notre parcours, soulignant les obstacles rencontrés, les solutions adoptées, et les avancées que nous avons réalisées sur le modèle GPT-2 d'OpenAI. Ce projet nous a permis de mieux comprendre l'intelligence artificielle et de mettre en pratique les connaissances acquises dans le cadre du cours de développement avancé donné par M. FAYE. Nous avons également considéré les enjeux éthiques, tant au niveau des réponses de l'IA que de l'utilisation des données pour son entraînement.

En utilisant un ensemble de données centré sur le domaine de la santé, nous avons pu adapter notre modèle aux besoins spécifiques des utilisateurs de ce secteur. Ce rapport couvre les aspects techniques, les questions éthiques, les défis de collecte de données en santé, et comment ce travail nous a fait évoluer tant sur le plan professionnel qu'humain.

Prise en main de GPT-2 pour le Développement d'un Chatbot en Santé

Au cœur de notre initiative se situe le modèle GPT-2, un modèle de langage de grande taille développé par OpenAI en 2019. Bien que non le plus récent et open source, il offre une base intéressante pour les débutants en intelligence artificielle, comme notre groupe d'étudiants en troisième année d'informatique. Nous avons opté pour GPT-2 plutôt que pour des modèles plus avancés comme Mistral 8x7B ou LLaMA 2 70B de Mistral ou Meta, car cela nous permet de démontrer nos compétences en IA sans nécessiter une puissance de calcul hors de notre portée.

Notre première étape fut l'exploration théorique, centrée sur la compréhension des subtilités de GPT-2 et des méthodes de fine tuning associées, dont la méthode LoRA (Low Rank Adaptation). Cette phase théorique nous a préparés pour la suite du projet et enrichi notre compréhension du modèle.

La seconde phase fut l'expérimentation pratique, menée avec mes coéquipiers Bastien et Kevin. Nous avons utilisé KerasNLP pour manipuler le modèle pré-entraîné GPT2CausalLM, ce qui nous a dispensés de l'usage d'un tokenizer et nous a aidés à comprendre le processus de génération de texte de GPT-2, essentiel pour son ajustement à des contextes spécifiques comme celui de la santé.

Nous avons rencontré plusieurs défis, notamment la répétitivité et l'exactitude des informations fournies par GPT-2. Ceux-ci étaient particulièrement critiques puisque notre chatbot cible le domaine sensible de la santé. Nous avons donc travaillé sur la personnalisation du modèle pour minimiser les erreurs et fournir des réponses plus précises et pertinentes.

En conclusion, cette partie du projet a été enrichissante, me permettant de maîtriser de nouveaux outils et concepts tels que GPT-2, KerasNLP, et LoRA. Cette exploration et ces expérimentations ont jeté les bases nécessaires au développement d'un chatbot capable de fournir des dialogues naturels et informatifs dans le domaine de la santé.

Optimisation de Données pour le Chatbot de Santé : Collecte, Analyse et Application

Dans la phase de collecte et d'analyse des données de notre projet, l'accent a été mis sur le développement d'un chatbot spécialisé dans le domaine de la santé. Cette orientation, bien que prédéfinie, a guidé notre quête de bases de données adaptées, privilégiant celles qui sont médicalement pertinentes et qui suivent un format de questions-réponses, tout en incorporant certaines exceptions qui ont contribué à l'enrichissement du processus de fine-tuning de notre modèle.

Dans mon travail individuel, je me suis concentré sur l'examen et l'analyse des différents ensembles de données disponibles. J'ai employé des wordclouds pour identifier et visualiser les termes les plus fréquents et pertinents au sein de ces datasets, ce qui a facilité la compréhension des thèmes récurrents et des questions importantes dans le contexte médical. Cet outil a été particulièrement utile pour distiller les informations essentielles des datasets et pour orienter efficacement notre stratégie de fine-tuning.

Un ensemble de données a été spécifiquement fourni pour le sujet de notre projet, tandis que les autres ont été découverts et intégrés grâce aux efforts de Bastien. Cette diversité de sources a enrichi notre base de connaissances et a permis une analyse plus nuancée et approfondie, essentielle pour la conception d'un chatbot répondant avec précision aux besoins des utilisateurs en matière de santé. La combinaison de ces datasets a non seulement élargi notre horizon de données mais a également renforcé la robustesse de notre modèle en lui permettant de traiter un éventail plus large de requêtes médicales.

Optimisation de GPT-2 : Stratégies et Résultats du Fine-Tuning

Le fine-tuning de notre modèle GPT-2 s'est transformé en une véritable expérience d'apprentissage, jalonnée par des défis techniques et couronnée par des succès remarquables. Cette étape fondamentale a témoigné de notre développement technique constant et de notre compréhension de plus en plus nuancée des particularités du modèle.

Dans cette dynamique, mes coéquipiers Kévin et Bastien se sont concentrés sur l'application de LoRA pour le fine-tuning de GPT-2, tandis que de mon côté, j'ai opté pour une approche de fine-tuning sans LoRA. L'objectif de cette division était de permettre une comparaison éclairée entre ces deux méthodes de fine-tuning distinctes.

Suite à nos expériences de fine-tuning sur le dataset spécifique fourni, nous avons observé que le modèle ajusté avec LoRA démontrait une plus grande précision. De plus, il s'est avéré que le temps nécessaire pour le fine-tuning sans LoRA était presque le double de celui requis avec LoRA, mettant en lumière les avantages en termes d'efficacité et de performance offerts par LoRA.

Bastien a joué un rôle clé en élargissant notre corpus d'entraînement pour inclure un mélange de données médicales et de questions-réponses générales. Cela visait à équiper notre modèle d'une capacité conversationnelle versatile tout en affinant son expertise dans le domaine médical. Cette démarche a nécessité une attention particulière pour maintenir un équilibre adéquat pendant l'entraînement, surtout pour éviter que le modèle ne se sur-spécialise dans un style trop académique.

En conclusion, cette phase de fine-tuning de GPT-2 a non seulement renforcé notre compréhension technique et notre capacité à manipuler des modèles d'intelligence artificielle complexes, mais elle a également mis en évidence l'importance des méthodes d'optimisation telles que LoRA dans l'amélioration de l'efficacité et de l'exactitude des modèles. Cette expérience a souligné l'importance de choisir la bonne stratégie de fine-tuning pour répondre aux besoins spécifiques d'un projet, tout en optimisant les ressources et le temps disponibles. Notre travail a contribué à une meilleure adaptation de GPT-2 aux exigences du domaine de la santé, en fournissant une base solide pour le développement futur de notre chatbot conversationnel.

Hugging Face et la mise en place de Gradio

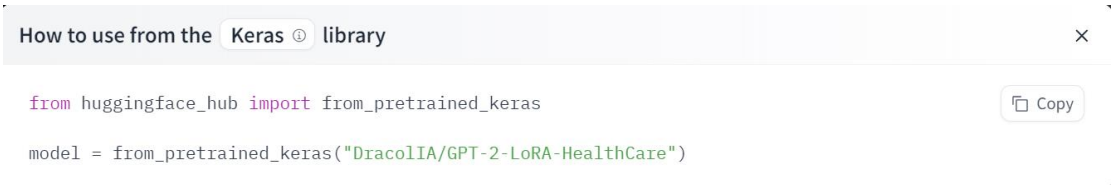
Le développement de l'interface utilisateur de notre chatbot avec Gradio était essentiel pour assurer la compatibilité avec la plateforme d'hébergement Hugging Face. Malgré certaines restrictions initiales, Gradio s'est avéré être un outil précieux, nous permettant de concevoir une interface conviviale adaptée à notre modèle conversationnel basé sur GPT-2, tout en naviguant à travers des défis techniques importants.

Collaboration en Continu : Le Pont entre GitHub et Hugging Face

La synchronisation entre notre espace Hugging Face et notre dépôt GitHub, orchestrée habilement par notre collègue Fatih, a joué un rôle crucial dans l'évolution et la collaboration de notre projet. Cette intégration clé garantit que toute mise à jour effectuée sur la branche principale de notre dépôt GitHub se répercute automatiquement dans notre espace Hugging Face, facilitant ainsi un développement fluide et une collaboration efficace parmi les membres de notre équipe.

Défis de Synchronisation : Naviguer entre GitHub et Hugging Face

En outre, la synchronisation entre notre dépôt GitHub et Hugging Face a posé un problème spécifique : le modèle, trop lourd pour être stocké sur GitHub, devait être manuellement ré-uploadé sur Hugging Face après chaque synchronisation. Pour résoudre ce problème, j'ai décidé d'utiliser huggingface_hub, qui promet une intégration plus fluide. Cependant, le code fourni pour charger le modèle via huggingface_hub s'est révélé défectueux, car il passait le chemin d'un dossier au lieu de celui du modèle spécifique à la fonction tf.keras.models.load_model. J'ai corrigé ce problème pour assurer un chargement efficace et correct du modèle, améliorant ainsi la fiabilité et la facilité d'utilisation de notre chatbot.



Premiers Pas Interactifs : L'Aventure Gradio Commence

La première version de notre interface, créée par Tamij, a établi les fondements en exploitant les fonctionnalités de Gradio et de GPT-2 pour simuler une interaction fluide, à l'image de celles gérées par des IA plus sophistiquées telles que GPT-3.5 et Gemini. Cette base initiale a été essentielle pour orienter le développement de notre interface.

Adaptation Stratégique : Surmonter les Limites de GPT-2

Face aux limitations de GPT-2, en particulier la restriction de taille d'entrée, Bastien a adapté notre chatbot vers un modèle de questions et réponses. Cette transition a permis de simplifier les interactions tout en restant dans les capacités de notre modèle, offrant ainsi une expérience utilisateur optimisée malgré ces contraintes.

Mémoire Numérique : Ajout de la Fonctionnalité d'Historique

J'ai ensuite ajouté une fonctionnalité permettant de suivre l'historique des interactions pour améliorer davantage l'expérience utilisateur, permettant ainsi une certaine continuité et une meilleure immersion malgré le format question-réponse imposé par les limitations techniques.

Conclusion

Malgré les défis initiaux, le développement de notre interface utilisateur avec Gradio et l'intégration avec Hugging Face se sont avérés bénéfiques. L'approche initiale de Tamij a posé des bases solides, et l'ajustement vers un format question-réponse par Bastien a aligné notre projet avec les capacités de GPT-2. L'ajout d'une fonctionnalité d'historique et l'adaptation réussie aux contraintes de synchronisation ont marqué des avancées significatives, soulignant notre capacité à répondre de manière flexible et innovante aux besoins spécifiques du domaine de la santé. Ce processus a mis en évidence notre engagement à améliorer continuellement notre produit pour répondre aux attentes des utilisateurs, tout en soulignant la valeur de la collaboration et de la résolution créative de problèmes dans le développement technologique.

Réflexions Finales : Le Parcours du Développement de Chatbot

En conclusion, notre projet a représenté un voyage remarquable à travers les défis et les innovations dans le développement d'un chatbot conversationnel dans le domaine de la santé, en s'appuyant sur les capacités de l'intelligence artificielle, en particulier le modèle GPT-2. Tout au long de ce parcours, nous avons rencontré des obstacles techniques et conceptuels, depuis la compréhension initiale et le fine-tuning du modèle d'IA, jusqu'à la collecte et l'analyse des données spécifiques au secteur de la santé, et la création d'une interface utilisateur intuitive avec Gradio.

Le fine-tuning du modèle a été un parcours d'apprentissage en soi, mettant en évidence l'importance de la personnalisation dans le domaine de l'IA. Les approches divergentes adoptées par l'équipe, comparant les méthodes avec et sans LoRA, ont mis en exergue les compromis entre précision et efficacité, soulignant l'impact significatif de techniques d'optimisation avancées sur les performances et les temps de traitement.

La synchronisation entre notre espace Hugging Face et notre dépôt GitHub, réalisée grâce à l'ingéniosité de notre camarade Fatih, a posé les fondements d'un travail collaboratif efficace et d'un processus de mise à jour fluide. L'initiative et la créativité de chaque membre de l'équipe, notamment les contributions de Tamij, Bastien et les miennes, ont été cruciales pour naviguer à travers les limitations de GPT-2 et pour adapter notre chatbot aux exigences spécifiques de la communication dans le domaine de la santé.

Les défis rencontrés nous ont poussés à explorer des solutions innovantes et à développer des compétences techniques et analytiques avancées. Cette expérience a non seulement renforcé notre compréhension de l'intelligence artificielle et de ses applications potentielles mais nous a également permis de réfléchir sur les enjeux éthiques et les responsabilités liées au développement de technologies de santé.

Ce projet a souligné l'importance de la flexibilité, de la persévérance et de la collaboration dans le développement de solutions technologiques répondant aux besoins réels des utilisateurs. Les enseignements tirés de cette expérience enrichiront indubitablement notre approche des projets futurs, et la réussite de ce chatbot dans le secteur de la santé sert de témoignage à notre engagement envers l'amélioration de la qualité de vie et du bien-être des individus à travers la technologie.

Alors que nous clôturons ce chapitre de notre parcours éducatif, nous restons inspirés par les possibilités infinies que l'avenir de l'intelligence artificielle détient pour le domaine de la santé et au-delà. Ce projet n'est pas seulement le fruit de notre apprentissage académique ; il représente également un pas en avant dans notre développement en tant qu'informaticiens et citoyens conscients des impacts de la technologie sur la société.

Disclaimer

Ce rapport a été produit avec l'aide d'une intelligence artificielle (IA) pour affiner la rédaction et l'organisation du contenu. Chaque aspect et défi du projet a été méthodiquement détaillé à l'IA pour un accompagnement rédactionnel adapté. Tous les éléments produits par l'IA ont été scrupuleusement révisés, modifiés et confirmés pour garantir qu'ils reflètent fidèlement le travail réalisé. L'utilisation attentive de l'IA montre notre désir d'intégrer des avancées technologiques tout en respectant l'intégrité académique et la précision.

Tandis que l'IA a aidé à formuler ce rapport, chaque section a fait l'objet d'une vérification exhaustive et a reçu l'approbation de l'auteur, créant ainsi une combinaison d'innovation technologique et de rigueur humaine.

Ce document est le fruit d'un projet étudiant et sert des objectifs pédagogiques et de recherche. Les points de vue, analyses et conclusions ici exprimés sont propres aux auteurs et ne représentent pas nécessairement ceux de leur institution éducative. Bien que nous ayons veillé à la justesse et à la pertinence de l'information fournie, ce rapport ne vise pas à offrir des conseils médicaux ou professionnels.

Le chatbot développé durant ce projet utilise le modèle GPT-2 d'OpenAI et est conçu comme une ressource informative supplémentaire, pas comme une alternative à l'expertise médicale professionnelle. Il est recommandé de consulter des professionnels de santé pour des conseils médicaux adaptés.

Les technologies, méthodes et concepts discutés dans ce document sont en constante évolution et peuvent être mis à jour post-publication. Les auteurs ne peuvent donc pas assurer la pertinence ou l'exactitude continues des contenus et méthodologies décrits.

Toutes les données et informations utilisées ont été créées à des fins de démonstration ou obtenues de sources publiques, et toutes les mesures appropriées ont été prises pour honorer les droits de propriété intellectuelle et la confidentialité des données.