RAPPORT INDIVIDUEL

SAE - CREATION D'UN SYSTEME DE CHATBOT CONVERSATIONNEL BASE SUR GPT-2

RESUME

Rapport individuel présentant les tâches effectuées dans le cadre du projet « Création d'un système de chatbot conversationnel basé sur GPT-2 » et le déroulement général du projet.

Bastien Hottelet

B.U.T 3 Apprentissage | 2023 - 2024

Table des matières

Introduction

Dans un monde où la technologie et la santé se croisent de plus en plus, l'importance de l'innovation dans le domaine des soins de santé est indéniable. Notre projet s'inscrit dans cette dynamique, cherchant à exploiter les capacités avancées de l'intelligence artificielle pour donner un premier diagnostic. Bien sûr, cela n'est pas voué à remplacer les professionnels de santé qui sont bien plus qualifiés. L'objectif principal de ce travail a été de développer un chatbot conversationnel intelligent, basé sur le modèle GPT-2, capable de comprendre et de répondre aux interrogations des utilisateurs avec une précision et une pertinence remarquable par rapport au modèle original de GPT-2.

Face à ce défi, notre équipe, composée de sept membres, a uni ses efforts et a réparti les tâches pour couvrir l'ensemble des étapes nécessaires au développement du projet. Le sujet qui nous a été donné, nous a permis de nous immerger dans des domaines variés, allant de la compréhension technique du modèle GPT-2 à la collecte et l'analyse de données spécifiques au secteur de la santé, en passant par l'affinement et l'optimisation de notre modèle pour qu'il réponde au mieux aux attentes des utilisateurs.

Ce rapport s'occupe de retracer le parcours accompli, mettant en lumière les défis rencontrés, les solutions apportées, et surtout, les avancées que nous avons pu faire, à notre échelle, sur le modèle GPT-2 de OpenAI. À travers ce travail, nous avons l'occasion d'en apprendre beaucoup sur l'intelligence artificielle et d'approfondir et mettre en pratique le cours de développement avancé de M. FAYE que nous avons eu plus tôt dans l'année. Nous nous sommes également heurtés à des sujets comme l'éthique, que ce soit pour les réponses apportés par l'IA ou les données utilisés pour son entrainement.

En nous appuyant sur un ensemble de données centré sur le secteur de la santé, nous avons pu affiner notre modèle pour qu'il réponde spécifiquement aux besoins et aux préoccupations des utilisateurs dans ce domaine crucial. Ce rapport détaille non seulement les aspects techniques de notre projet mais explore également les implications éthiques, les défis de la collecte de données dans le domaine de la santé, et la manière dont notre travail nous à permis d'évoluer en tant qu'informaticiens et en tant qu'humain.

Compréhension et Utilisation de GPT-2

Au cœur de notre projet se trouve le modèle GPT-2, un LLM (Large Language Model) crée par OpenAI en 2019. En effet, ce modèle n'est pas le plus récent mais il est open source et assez mauvais pour qu'il soit intéressant de voir ce qu'en font des BUT3 Informatique débutants en IA.

Nous aurions en effet pu utiliser une IA de Mistral ou Meta comme « Mistral 8x7B » ou « LLaMA 2 70B » mais non seulement, utiliser de tels IA n'aurait pas démontré nos compétences dans l'IA car ces IA sont déjà performantes dans le domaine. Mais en plus il aurait été compliqué pour des étudiants d'avoir la puissance nécessaire pour les utiliser.

Exploration Théorique

Initialement, notre démarche a commencé par une phase d'exploration théorique, visant à saisir les subtilités du modèle GPT-2. Dans cette phase, je me suis principalement consacré (en vue des prochaines tâches) à la compréhension des méthodes de fine tuning qui existent sur ce modèle. C'est durant cette période que j'ai pu découvrir les travaux de recherche relatifs à GPT-2, notamment la méthode d'entrainement appelée LoRA (Low Rank Adaptation). Nous y reviendrons plus tard.

Expérimentation Pratique

À la suite de cette immersion théorique, j'ai procédé à des expérimentations pratiques avec l'aide de Pascal et Kevin, sur le modèle GPT-2. Nous avons pour cela utilisé la librairie KerasNLP pour obtenir le modèle pré-entraîné GPT2CausalLM et son pre-processor qui nous dispense d'utiliser un Tokenizer. Ces tests préliminaires nous ont permis de mieux comprendre comment GPT-2 génère du texte et réagit à différents types de requêtes, une compréhension cruciale pour l'ajustement (fine-tuning) du modèle à notre contexte spécifique de santé.

Défis et Solutions

C'est également à ce moment que nous avons réalisé le grand travail que nous avions à faire non seulement sur le fond mais également sur la forme. En effet GPT-2 avait tendance à beaucoup se répéter et parfois dire des informations fausses ou hors sujet.

La manipulation de GPT-2 a donc révélé plusieurs défis, notamment en ce qui concerne la personnalisation du modèle pour qu'il réponde avec précision aux questions de santé des utilisateurs. En effet, sur un chatbot qui concerne le domaine médical, on doit à tout prix limiter les erreurs commises.

Conclusion

Au cours de cette partie du projet, j'ai eu l'occasion de prendre en main de nouveaux outils et concepts, notamment GPT-2 et la librairie KerasNLP avec Kevin et Pascal. Mais j'ai également eu l'occasion d'explorer seul les concepts de LoRA et de comprendre comment d'ingénieux chercheurs avaient trouver un moyen de réduire le coût d'entrainement en gardant des performances similaires. Ce savoir nous sera d'ailleurs très utile dans la suite du projet pour le fine tuning.

Grâce à cette exploration approfondie et à ces expérimentations avec GPT-2, nous avons posé les bases nécessaires pour développer un chatbot capable de dialogues naturels et informatifs.

Collecte et Analyse de données

La collecte et l'analyse des données ont constitué une phase cruciale de notre projet, définie par l'objectif de développer un chatbot conversationnel dans le secteur spécifique de la santé. Bien que la focalisation sur ce domaine ne fût pas de notre initiative, elle a néanmoins orienté notre recherche vers des ensembles de données répondant à des critères précis : pertinence médicale et format question-réponse, principalement, avec quelques exceptions notables qui ont enrichi le processus de fine-tuning de notre modèle.

Sources et Sélection des Données

Nos recherches nous ont menés vers des sources de données publiques, choisies pour leur accessibilité et leur conformité aux exigences légales et éthiques. Cette démarche s'est cependant heurtée à divers défis, notamment la diversité des formats (JSON, CSV) et la qualité souvent inégale des données, marquée par des incohérences et des lacunes dans le formatage. L'une des difficultés majeures résidait dans la rareté des datasets médicaux adaptés à nos besoins, à savoir des ensembles de données suffisamment simples pour une manipulation directe, sans la complexité d'une base de données SQL multi-tables (Comme MIMIC-III par exemple).

Traitement et Analyse

Face à ces obstacles, nous avons adopté une approche méthodique pour le nettoyage et l'analyse des données. L'utilisation des librairies Python json et pandas a facilité l'importation et le traitement préliminaire des données, nous permettant de normaliser les formats et de pallier les lacunes et incohérences détectées. Pour évaluer la pertinence des datasets par rapport à notre objectif, nous avons généré des nuages de mots (wordclouds), une technique visuelle simple mais efficace pour appréhender le contenu thématique des données et ajuster notre stratégie de sélection.

Pour le dataset obligatoire (MedQuAD), je me suis chargé de l'importation avec un peu d'aide de la part de Fatih, qui s'est ensuite chargé de réaliser le code pour l'analyse via nuage de mots de ce même dataset. Plus tard, Pascal à réutiliser ce code pour réaliser le nuage de mots d'autres datasets que nous avons utilisé.

Cette phase d'analyse préliminaire a révélé l'importance d'une sélection rigoureuse des données, soulignant le défi de trouver un équilibre entre la richesse informative nécessaire pour l'entraînement du modèle et la spécificité requise pour un chatbot de santé efficace. La diversité des formats et la qualité variable des datasets ont exigé un effort considérable de nettoyage et de prétraitement, mais ont également offert une opportunité d'approfondissement technique et de créativité dans l'optimisation de notre chatbot.

D’amateur à expert | Fine-tuning de GPT-2

Le processus de fine-tuning de notre modèle GPT-2 s'est avéré être un véritable parcours d'apprentissage, ponctué de défis techniques et de réussites significatives. Cette étape cruciale a été divisée en trois phases distinctes, marquées par une évolution constante de notre maîtrise technique et une compréhension approfondie des subtilités du modèle.

Première Phase : Initiatives Précoces

Kevin, l'un de mes camarades, et moi-même, avons initié le fine-tuning en exploitant LoRA mais nous avons rencontré des difficultés lors de la sauvegarde sous le format ".keras" lorsque nous suivions la documentation de KerasNLP.

Pour contourner ces difficultés, Kevin a trouvé une méthode qui consiste à sauvegarder seulement les poids du modèle sous forme de checkpoint. Bien que cette première tentative ait posé les bases, les résultats n'étaient pas à la hauteur de nos attentes, nous poussant à réévaluer notre approche.

Seconde Phase : Redéfinition et Correction

Face aux limitations observées, j'ai décidé de prendre en main le fine-tuning, armé de ma connaissance acquise lors des étapes précédentes et d'une analyse approfondie de la documentation de Keras.

C'est dans ce contexte que j'ai identifié et corrigé une erreur critique dans l'implémentation LoRA recommandée par Keras. En ajustant le code pour éliminer correctement les LoRA layers après l'entraînement, j'ai non seulement optimisé le processus mais également rendu possible la sauvegarde efficace du modèle. Cette correction technique a été un tournant, me permettant d'exploiter pleinement les capacités de Google Colab Pro+ pour entraîner le modèle avec une sélection enrichie de datasets.

```
570  570  for layer_idx in range(lora_model.backbone.num_layers):
571  571      self_attention_layer = lora_model.backbone.get_layer(
572  572          f"transformer_layer_{layer_idx}"
573  573      )._self_attention_layer
574  574
575  575      # Merge query dense layer.
576  576      query_lora_layer = self_attention_layer._query_dense
577  577
578  578      A_weights = query_lora_layer.A.kernel  # (768, 1) (a, b)
579  579      B_weights = query_lora_layer.B.kernel  # (1, 12, 64) (b, c, d)
580  580      increment_weights = tf.einsum("ab,bcd->acd", A_weights, B_weights) * (ALPHA / RANK)
581  581      query_lora_layer.original_layer.kernel.assign_add(increment_weights)
582  582
583  583      # Merge value dense layer.
584  584      value_lora_layer = self_attention_layer._value_dense
585  585
586  586      A_weights = value_lora_layer.A.kernel  # (768, 1) (a, b)
587  587      B_weights = value_lora_layer.B.kernel  # (1, 12, 64) (b, c, d)
588  588      increment_weights = tf.einsum("ab,bcd->acd", A_weights, B_weights) * (ALPHA / RANK)
589  589      value_lora_layer.original_layer.kernel.assign_add(increment_weights)
     590  +
     591  +    # Put back in place the original layers with updated weights
     592  +    self_attention_layer._query_dense = query_lora_layer.original_layer
     593  +    self_attention_layer._value_dense = value_lora_layer.original_layer
```

Figure 1 - Pull Request KerasNLP

Troisième Phase : Expansion et Affinement

Conscient de l'importance de la diversité des connaissances pour notre chatbot, j'ai élargi notre corpus d'entraînement pour inclure non seulement des données spécifiquement médicales mais aussi des ensembles de questions-réponses générales. Cette stratégie visait à doter notre modèle d'une base solide en conversation, tout en affinant sa spécialisation dans le domaine de la santé grâce à un troisième dataset médical. L'équilibre dans la sélection et l'ordonnancement des epochs s'est avéré déterminant, en particulier lorsque nous avons dû ajuster notre approche à la suite d'une sur-spécialisation accidentelle du modèle dans le style des articles de recherche médicale.

Une autre chose importante sur laquelle j'ai travaillé pour rendre notre modèle plus efficace à sa tâche de question réponse est le prompt engineering. En effet GPT-2 n'est pas fait pour répondre à des questions mais pour générer du texte. C'est donc pour cela que j'ai choisi d'introduire dans les données d'entrainement la notion de Tokens sous la forme de « [QUESTION] » et « [ANSWER] » de sorte à ce que quand GPT-2 voit ces marques, il sait ce qui lui reste à faire.

Datasets utilisés :

Nom	Lien	Objectif
SQuAD	https://rajpurkar.github.io/SQuAD-explorer/	Question réponse – Généraliste
QuAC	https://quac.ai	Question réponse – Généraliste
PubMedQA	https://pubmedqa.github.io	Question réponse - Article de recherche Domaine médicale
BioASQ – Task B	http://participants-area.bioasq.org/datasets/	Question réponse – Type Examen Domaine médicale
MedQuAD	https://www.kaggle.com/datasets/jpmiller/layoutlm	Question réponse – Domaine médicale

Leçons Apprises et Impacts

Le fine-tuning avec LoRA a marqué une avancée notable dans la spécialisation de notre modèle, offrant une efficacité accrue en évitant des mises à jour inutiles de l'ensemble des poids du modèle. Cette phase a non seulement démontré l'importance cruciale des choix techniques comme le nombre d'epochs et la gestion des datasets, mais elle a également illustré notre capacité à naviguer à travers des défis complexes, à les résoudre avec créativité et à pousser les limites de ce que notre projet pouvait atteindre.

Implémentation et mise en place de Gradio

Dans le cadre de notre projet, l'utilisation de Gradio pour développer l'interface utilisateur de notre chatbot a été une contrainte imposée par le besoin de compatibilité avec la plateforme d'hébergement Hugging Face. Malgré cette contrainte, l'outil s'est révélé être un atout précieux, nous permettant de créer une interface adaptée à notre modèle conversationnel basé sur GPT-2, bien que nous ayons dû naviguer à travers certaines limitations techniques significatives.

Première Implémentation par Tamij

La première version de l'interface, conçue par mon camarade Tamij, visait à exploiter au mieux les possibilités offertes par Gradio et GPT-2. Cette étape initiale a posé les fondations de notre interface, et avait pour but de créer un format de conversation comme le font les IA actuelles (GPT-3.5, Gemini, Claude, …).

Confrontation aux Limitations de GPT-2

Cependant, l'ambition initiale de créer un chatbot avec un historique conversationnel s'est heurtée aux limites de la taille d'entrée de GPT-2, qui restreint sévèrement la quantité de contexte pouvant être prise en compte (1024 caractères). Cette contrainte nous a forcés à reconsidérer notre approche, nous orientant vers un modèle d'interaction basé sur le style question-réponse, plus adapté aux capacités de notre modèle.

Transition vers un Modèle Question-Réponse

Face à la nécessité d'adapter notre chatbot aux contraintes de GPT-2, j'ai pris l'initiative de transformer l'interface pour adopter un format question-réponse. Cette modification a nécessité un remaniement significatif de l'interface utilisateur pour simplifier les interactions, en s'assurant que chaque question soit indépendante, afin de contourner la limitation de la taille d'entrée du modèle.

Gestion de l'Historique par Pascal

Pour enrichir l'expérience utilisateur malgré le format question-réponse, mon camarade Pascal a intégré une fonctionnalité d'historique des interactions. Cette innovation permet aux utilisateurs de visualiser leurs questions précédentes ainsi que les réponses du chatbot, offrant un semblant de continuité et une meilleure immersion dans la conversation, malgré les contraintes techniques.

Déploiement et Optimisation

La mise en place de notre chatbot sur la plateforme Hugging Face via Gradio a finalement représenté une étape cruciale, nous permettant de partager notre travail avec une audience plus large. Cela nous a surtout permis d'enfin voir les fruits de notre travail, avec une IA utilisable avec une interface graphique et pas en ligne de code sur un notebook. C'était selon moi une petite étape techniquement mais à l'échelle du projet, cela marquait la conclusion d'une période.

Conclusion

Au terme de ce projet ambitieux, il est temps de réfléchir aux accomplissements réalisés, aux défis surmontés et aux leçons apprises tout au long de notre parcours. De l'exploration théorique de GPT-2 à l'implémentation d'une interface utilisateur intuitive via Gradio, chaque étape nous a rapprochés de notre objectif : créer un chatbot conversationnel en santé capable de répondre aux interrogations des utilisateurs avec une pertinence et une précision notable.

Ce projet ne fut pas seulement une aventure technique, mais également une exploration des implications éthiques et sociales liées à l'utilisation de l'intelligence artificielle dans le secteur de la santé. Les défis rencontrés, qu'ils soient liés à la collecte et à l'analyse de données ou au fine-tuning du modèle, nous ont enseigné l'importance de l'adaptabilité, de la collaboration et de l'innovation.

Notre capacité à transformer les limitations en opportunités de croissance, comme le passage d'un modèle de conversation à un modèle question-réponse à cause de la contrainte de taille d'entrée de GPT-2, témoigne de notre engagement et de notre flexibilité. La collaboration entre les membres de l'équipe a été cruciale.

L'implémentation finale sur la plateforme Hugging Face a pour nous été un grand moment qui nous a permis de constater tout le travail effectué et le parcours réalisé depuis le début du projet.

En conclusion, ce projet a été une étape significative dans notre parcours académique et professionnel. Il a non seulement renforcé notre compréhension des modèles d'intelligence artificielle comme GPT-2 mais nous a également préparés à aborder avec confiance les défis futurs dans le domaine de l'informatique et de l'intelligence artificielle. Nous partons de ce projet enrichis, prêts à explorer de nouvelles frontières technologiques et à continuer de contribuer à l'innovation dans le domaine de la santé et au-delà.

Ce rapport clôt un chapitre de notre formation, mais ouvre la voie à de futures explorations, avec la certitude que les compétences et les connaissances acquises nous accompagneront dans nos prochaines aventures.

Disclaimer

Ce rapport a été rédigé avec l'assistance d'une intelligence artificielle (IA), dans le souci d'efficacité, de productivité et de qualité rédactionnelle. Toutefois, il est important de noter que l'IA n'a pas été simplement sollicitée en lui donnant le sujet du projet pour générer automatiquement le contenu. Au contraire, chaque étape du projet, chaque difficulté rencontrée, a été soigneusement expliquée à l'IA afin qu'elle puisse aider à structurer et à formuler le récit de manière plus cohérente et détaillée que je n'aurais pu le faire seul.

Le recours à cette technologie ne signifie pas que le contenu du rapport a été généré sans supervision. Après la création de chaque section, j'ai attentivement vérifié, corrigé et validé les informations pour m'assurer qu'elles correspondent parfaitement à la réalité de notre travail et aux résultats obtenus. Cette démarche garantit que le rapport reste fidèle à nos expériences et réflexions tout au long du projet.

L'utilisation de l'IA dans la rédaction de ce rapport est un choix réfléchi, visant à tirer parti des avancées technologiques pour améliorer la qualité de notre travail tout en restant rigoureusement attaché à l'exactitude et à l'intégrité académique. Ce choix illustre également notre engagement à explorer de nouvelles méthodologies de travail, en phase avec les compétences et les défis de l'ère numérique.

En somme, bien que l'IA ait joué un rôle d'assistant dans la composition de ce document, chaque mot a été examiné et approuvé par moi-même, reflétant ainsi une synthèse entre innovation technologique et responsabilité humaine.