SAÉ 5.A.01 - Rapport sur le travail réalisé

Étude du modèle GPT-2 | 06/11/2023 - 17/11/2023

Pour cette première phase, je me suis chargé d'effectuer une étude du modèle GPT-2 et de réaliser des tests sur Jupyter Notebook pour découvrir son fonctionnement, son utilisation, ses capacités et limitations, ainsi que les résultats fournis par ce dernier. J'ai pu collaborer en binôme avec mon collègue et camarade Lucas AGUETAÏ afin de travailler sur cette mission.

J'ai débuté cette période en individuel en consultant la documentation existante sur KerasNLP et du modèle GPT-2 associé à celui-ci, dont la consultation était suggérée dans le sujet fourni sur cette SAE. Je me suis donc intéressé brièvement à ce qu'était KerasNLP, qui est donc une librairie pour le Natural Language Processing, étant une extension de Keras. C'est avec cette librairie que nous pouvons accéder et utiliser le modèle GPT-2.

J'ai pris connaissance de la documentation de celui-ci et en particulier à la section "GPT2 Text Generation with KerasNLP", où il a été question pour moi de découvrir le code Python et en particulier GPT2 Causal LM, le modèle permettant d'initialiser GPT-2 et d'utiliser les fonctions associées pour générer du texte. Par la même occasion, j'ai consulté en détail les différents composants mis à disposition par l'API Keras + Tensorflow, avec donc les modèles/modules GPT-2 tels que GPT2Backbone, GPT2CausalLM, les tokenizers et préprocesseurs respectifs à ces derniers.

Afin d'expérimenter avec le modèle, j'ai donc créé un notebook .ipynb sur Google Collab afin de tester le code mis à disposition sur la documentation de GPT-2 sur keras.io, concernant la génération de texte avec ce modèle. Cela a donc été l'opportunité pour moi de réaliser des premières générations de résultats et d'observer la complétion du texte fournie en entrée.

Suite à cela, j'ai fait part de cette courte expérimentation avec les autres membres de mon groupe, durant lequel nous avons pu échanger sur un aperçu des capacités de GPT-2 en ce qui concerne la complétion de texte. Quelques tests mineurs ont été réalisés collectivement sur des idées qui ont été proposées lors de nos échanges, durant lesquelles nous avons pu commencer à relever quelques défauts sur le modèle, tels que la répétition excessive de termes, ou encore des cas de contradictions dans le sens des phrases.

Pendant le déroulement de cette période, j'ai échangé régulièrement avec mon collègue Lucas AGUETAÏ afin d'approfondir la découverte de GPT-2, notamment en élaborant de nouveaux tests de génération de texte. Nous avons entretenu plusieurs réflexions sur comment il serait possible d'exposer les forces, faiblesses et les limites du modèle, en interrogeant le modèle sur plusieurs sujets et domaines différents, techniques ou plus générales, et cela sous forme de questions-réponses.

Nous avons donc pensé à plusieurs cas de tests : un test de complétion de texte classique, un test de génération traitant des calculs mathématiques, une génération de réponse à une question de sujet général, ainsi qu'à une question concernant un sujet technique tel que

ceux contenus dans les articles scientifiques du set de données qui nous a été fourni. La diversité de ces tests nous permet d'identifier plusieurs points, comme la limitation en connaissance avancée dans un domaine technique malgré pré-entraînement du modèle, du format question-réponse moins adapté pour le modèle GPT-2.

Progressivement, nous avons eu aussi l'idée d'essayer les différentes variantes de modèles de GPT-2, chacun proposant un nombre de couches et de paramètres différents, plus ou moins importants selon les besoins. Ayant réalisé jusqu'à présent principalement des tests sur le modèle *gpt2_base_en* avec un réseau de neurones à 12 couches, il pouvait donc être intéressant de répéter les mêmes tests sur les autres modèles disponibles proposant des réseaux à 24 (*gpt2_medium_en*), 36 (*gpt2_large_en*) et plus de couches. Cela nous a permis de voir si des réponses plus élaborées, précises et cohérentes peuvent être générées avec un modèle GPT-2 constitué d'un réseau de neurones plus complexe.

D'autre part, j'ai consacré du temps à travailler sur le notebook constituant le rapport d'étude du modèle GPT-2, en me focalisant sur son format et sa structure, ainsi que le contenu textuel de ce dernier.

En globalité, j'ai trouvé cette phase plutôt intéressante par le fait que j'ai été assez impressionné par le modèle GPT-2 et de ses capacités de génération de texte, en constatant comment son pré entraînement lui permet de générer un très grand panel de texte tout en restant grammaticalement structuré et en partie cohérent, malgré les nombreux défauts rencontrés.
Néanmoins, nous pouvons constater l'évolution importante qu'a entrepris OpenAI avec son modèle GPT-2 comparé à ses futures modèles GPT-3 et 4 mobilisés dans son application ChatGPT, avec les améliorations qui ont été réalisées en prenant en considération les défauts que l'on a pu relevés.