SAÉ 5.A.01 - Rapport sur le travail réalisé n°4

Rapport de comparaison, début d'optimisation | 05/02 - 16/02/2024

Pour cette quatrième phase, j'ai commencé à étudier des pistes d'optimisation du modèle avec mon collègue Samuel, et j'ai apporté une assistance sur la rédaction du rapport de comparaison entre les modèles GPT-2 générique et spécialisé. En tant que manager de cette période, j'ai également supervisé la réalisation des tâches de ces semaines par les membres de mon équipe.

Dans un premier temps, j'ai apporté une aide sur la rédaction des rapports de comparaison entre le modèle GPT-2 générique et celui entraîné par fine-tuning que nous avons réalisé. J'ai établi la structure du rapport sur un notebook Python et expliqué à mes collègues Lucas et Quentin les principes et objectifs de la comparaison à effectuer, consistant essentiellement à une comparaison de réponses généras par les modèles générique et spécialisé, ainsi qu'un comparatif de performance via des tests d'évaluations des modèles basé sur la précision et le taux d'échec / perte de ces derniers.

Dans un deuxième temps, je me suis focalisé sur un début de processus d'optimisation de notre modèle GPT-2 en collaboration avec Samuel, membre de mon équipe. Nous avons étudié plusieurs pistes d'optimisation, commençant par la réalisation de nouveaux entraînements par fine-tuning sur différents modèles GPT-2. Celui-ci étant disponible sous diverses variantes (gpt2 base, medium, large…) avec des performances différentes (nb de couches, de paramètres), il était question d'expérimenter et d'entraîner ces autres modèles sur le même jeu de données, afin de chercher des résultats de génération plus cohérents et corrects. Individuellement, nous avons tenté plusieurs entraînements avec des paramètres différents : modèle GPT2 base avec 1 seul epochs, base avec 3 epochs, medium avec 3 epochs, entraînement sur le jeu de données scientific papers d'Hugging Face plutôt que celle téléchargée et divisée manuellement…

Ayant l'expérience de fine-tuning avec les phases précédentes, nous connaissions déjà les actions à effectuer et nous nécessitons seulement d'expérimenter avec différents paramètres, ce qui a pris le plus de temps à travers les entraînements. Les principales contraintes et problèmes rencontrés ont été les limites en termes de mémoire vive et GPU, étant limités sur Google Collaboratory. Pour remédier en partie à cela, des tentatives ont été réalisées en environnement local sur nos propres machines, mais ont été compliquées compte tenu de certaines limites techniques et matérielles.

En parallèle, Samuel et moi avons étudié et expérimenté avec une piste d'optimisation du modèle via l'utilisation de la technique LoRA (Low-Rank Adaptation), qui est à approfondir avec plus de recherches à l'avenir.

D'autre part, j'ai consacré du temps à l'organisation de mon équipe, en répartissant les tâches à effectuer au cours de ces semaines, et en donnant instructions et conseils sur ces derniers. Je me suis également chargé brièvement du déploiement de l'application Gradio sur Hugging Face Spaces en début de période, et tout au long de cette session à l'organisation des répertoire GitHub, Hugging Face et Google Collaboratory.