SAÉ 5.A.01 - Rapport sur le travail réalisé n°3

Travail sur le fine-tuning | 01/01/2024 - 19/01/2024

Pour cette troisième phase, j'ai principalement travaillé sur le fine-tuning du modèle GPT-2 avec le jeu de données concernant les articles scientifiques.

En collaborant avec mes collègues Quentin et Samuel, nous avons étudié la possibilité de mobiliser le GPU pour l'entraînement du modèle GPT-2 sur notre jeu de données. En effet, les capacités limitées de mémoire vive (sur Google Collaboratory) consistaient un blocage au niveau de l'entraînement, mobilisant beaucoup de ressources en plus du chargement du jeu de données de taille importante.

Pour cela, nous avons trouvé un moyen de forcer l'utilisation du GPU sur une partie du code (spécifiquement, pour la fonction .fit), en employant la mémoire du GPU T4 de Google Collaboratory. Il était nécessaire de préciser que le code doit être exécuté avec le GPU en plaçant le code dans un structure "with tf.device('/device:GPU:0'):".

Ensuite, l'entraînement pouvant être effectué, nous avons effectué le fine-tuning du modèle GPT-2 de base en l'entrainant sur le jeu de données d'articles scientifiques. J'ai réfléchi avec Samuel et Quentin sur une méthode pour fournir parmi un grand nombre d'articles, uniquement le contenu textuel de ces derniers, dans une même liste Python.

Nous avons procédé par un entraînement sur un premier échantillon de 10 articles, puis par un échantillon de 10.000 articles, et ensuite sur la quasi-totalité du jeu de données avec 80.000 articles scientifiques. L'objectif étant de fine-tuner le modèle sur la totalité des articles d'entraînement, nous avons réussi à entraîner le modèle - récemment entraîné sur les 80.000 premiers articles - sur les articles restants (article 80.001 à 119 924 ème), nécessitant toutefois certaines manipulations sur Google Colaboratory (copie des fichiers des articles correspondants dans un répertoire séparé, afin de charger en mémoire seulement celles-ci).

Suite à cela, j'ai mené des recherches sur mon côté afin de trouver un moyen de sauvegarder l'entraînement réalisé sur notre modèle GPT-2 fine-tuné. J'ai trouvé - en me basant sur les enseignements de M. Faye - qu'il existe plusieurs moyens de réaliser cela, le premier en sauvegardant le modèle via un fichier .keras, le deuxième en enregistrant les paramètres du modèle via un fichier de points de contrôle .ckpt, la troisième en persistant le modèle et les paramètres via un fichier .h5.

En suivant les recommandations de notre enseignant, nous avons fait usage de la sauvegarde par points de contrôle (checkpoint .ckpt), permettant de sauvegarder durant les étapes d'entraînement fine-tuning les paramètres mobilisés dans le modèle. Ainsi, il suffit de charger les paramètres utilisés afin d'employer la configuration entraînée sur des articles scientifiques, et derrière de mobiliser le modèle pour la génération de texte. C'est le procédé que nous avons suivi, en initialisant sur Gradio un modèle GPT-2 de base générique sans entraînement, puis en chargeant les checkpoints sur ce modèle afin de le spécialiser.

Enfin, je me suis chargé d'effectuer quelques tests d'évaluation basiques du modèle, sous forme d'évaluation TensorFlow avec la méthode .evaluate. En évaluant ce dernier sur le jeu de données d'évaluation et de tests d'articles scientifiques, nous avons trouvé une précision d'environ 43% en se basant sur un algorithme de catégorisation (SparseCategoricalCrossentropy).