SCHOEN Adrien

Rapport de réunion n°19

Date de la réunion : 04/06/2021

Date de la prochaine réunion : 11/06/2021

Ordre du jour : Transformers

1. Ce que le stagiaire a dit :

J'ai simplifié l'entraînement de Word2Vec et implémenté l'entraînement Continuous Bag of Word. Je n'ai pas encore essayé la génération simple avec W2V CBOW.

J'ai mené des recherches exhaustives sur les tranformers. J'ai expliqué leurs avantages et leurs inconvénients, comment les concevoir. Je me suis notamment attardé sur le positionnal encoding, le mécanisme d'attention et le masking.

J'ai ensuite décrit les problèmes que je rencontrais pour adapter mon Langage Model assez simple sur nos données de payload. Premièrement, l'entraînement devait être changé, car nous entraînions notre modèle non pas sur un grand corpus de mot, mais sur une liste de séquences. Ensuite, ces séquences n'ont pas toutes la même longueur, certains payload sont plus longs que d'autres ; et je ne sais pas comment faire un langage model qui me permette de tirer des séquences de longueurs variables. Dernier problème : pour l'instant nous générons les différents tokens d'une séquence itérativement, cad que l'on génère les mots 1 par un. Le problème avec cette forme de génération est qu'il faut fournir un mot ou une séquence initial (ou seed dans la littérature), et je ne sais pas quel séquence/mot mettre dans le cadre de notre génération de payload.

Actuellement, j'essaie de compiler mes connaissances en NLP pour prendre du recul sur notre problème.

2. Ce que les encadrants ont ajouté :

Il faut maintenant vraiment se poser et faire un récapitulatif de ce qui a été fait

On commence à entrer dans un domaine vraiment technique alors que nous ne somme même pas sûr que la génération à l'echelle des octets soit quelquechose de viable. Il faut vraiment se poser pour prendre du recul sur ce qui a été généré.

On peut aussi imaginer de la génération de comportement, ou l'on génére un ensemble d'action réalisé par un utilisateur que l'on pourrait effectuer dans une VM ou autre. Il faudrait à ce moment-là réussir à conceptualiser ce que l'on voudrait faire et trouver des données en conséquences.

3. Ce qu'il faut faire pour la prochaine séance :

Commencer le résumé de ce que l'on a appris en NLP. Mettre en pause la partie code.

Mettre à jour le github pour pouvoir échanger le code avec les encadrants.