SCHOEN Adrien

# Rapport de réunion n°18

*Date de la réunion : 28/05/2021*

*Date de la prochaine réunion : 04/06/2021*

## Ordre du jour : Génération de séquence

**1. Ce que le stagiaire a dit :**

Cette semaine je me suis concentré sur la génération de la payload. Nous avions décidé de modéliser ce problème sous la forme de génération de données séquentielles. Je me suis donc rapproché de ce qui se faisait du côté de la génération textuelle.

J'ai commencé par travailler sur nos données de payload en essayant de les représenter les octets dans un espace de dimension finie et continue. Pour cela j'ai effectué un encodage. Dans la payload d'un paquet, j'ai considéré des bigrams de deux octets. J'ai ensuite encodé la séquences des bigrams représentant la payload en séquence d'entiers. Un entier était égal à la somme des valeurs numériques du bigram considéré. Ma séquence était ainsi une séquence d'entier qui allaient être les index pour des one-hot vector. Exemple, le bigram (86, AF) sera représenté par un vecteur de 0 ou le 309ieme indice est égale à 1. J'ai donc converti une séquence d'octets en une sequence de vecteurs.

J'ai ensuite entraîné un Word2Vec mais l'entraînement est beaucoup trop long sur Google Collab. J'ai donc mis cette partie en pause et considéré directement la séquence de one-hot vector sans faire d'embedding.

Ayant du mal à appliquer directement SeqGAN, j'ai décidé d'approfondir mes connaissances en NLP notamment en m'intéressant au Générateur de SeqGAN, les Model Langage.

Un Model Langage est un modèle entraîné pour comprendre les spécificités d'un langage. Ce modèle peut ensuite être transposé sur différentes taches relatives à ce langage (génération de phrase, détection d'émotion dans un texte....). Un Model Langage peut être composé de plusieurs parties, être entraîné de différentes manières mais si l'on veut générer de nouvelles étapes à partir d'un Langage Model, il faut toujours réaliser quatre étapes:

1) Pré traitement des données, C'est là où l'on segmente les séquences en mot. On peut faire ici la réunion en bigram par exemple

2) Tokenisation : Ici on regroupe les mots pour former des tokens, ça peut être des one hot vector par exemple, ou des vecteurs d'embedding

3) La définition et l'entraînement du modèle : Il y a ici plusieurs méthodes pour entraîner un réseau. Certaines méthodes consistent à prédire certain tokens cachés d'une séquence, d'autres consistent à prédire une séquence cible depuis une séquence source. D'une certaine façon, on veut que le modèle comprenne un langage.

4) Tirage: Ici, on tire les tokens de la séquence à générer. Il y a plusieurs méthodes de tirage.

Cette semaine j'ai entraîné un langage model très simple avec juste une couche d'embedding et une LSTM. C'est ce modèle de langage qui peut être utilisé comme acteur dans SeqGAN.

**2. Ce que les encadrants ont ajouté :**

Pour l'encodage, il faut s'assurer que les différents mots aient tous des tokens différents. Si l'on décide de faire la somme des valeurs numériques pour obtenir le token d'un bigram, il faut s'assurer que la somme des deux valeurs d'un bigram soit unique à ce bigram. C'est très improbable.

Word2Vec peut être entraîné de deux façons. Soit skip-gram soit Continuous Bag of Word. Il faut que je me renseigne sur la différence entre ces deux méthodes. Une peut être plus rapide qu'une autre

On peut aussi essayer de voir si l'on pourrait pas directement générer des séquences depuis Word2Vec, en entraînant le modèle à prédire le mot suivant d'une séquence donné. Par exemple lui donner x2, X3 comme échantillon pour lui faire prédire x4

Il faudrait que je commence un résumé de ce que j'ai appris. Le mieux serait de commencer par résumé mes découvertes en NLP, puisque c'est ce que j'ai fait de plus récent. Ca m'aidera à voir les défis de ce domaine et de prendre du recul sur mes recherches.

Il faut aussi préparer la réunion de Jeudi.

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

Continuer d'acquérir des connaissances sur les LM, essayer de faire un LM avec des transformer

Faire le point sur les connaissances acquises en NLP.

Préparer la réunion de jeudi.