SCHOEN Adrien

Rapport de réunion n°17

Date de la réunion : 21/05/2021

Date de la prochaine réunion : 28/05/2021

Ordre du jour : Génération Header UDP

1. Ce que le stagiaire a dit :

J'ai expliqué les changements que j'avais apporté aux deux modèles : J'ai changé les couches d'activation de sortie du réseau pour les faire correspondre au mieux avec le format de nos données en sortie. En effet pour la plupart des colonnes, nous avons en sortie des données binaires, il faut donc adapter les sorties des fonctions d'activation pour qu'elles nous renvoient des valeurs comprises entre 0 et 1, qu'on arrondira à 0 ou à 1 (sigmoid peut être utilisé ici). J'ai aussi essayé de remplacer les 0 par des -1 pour utiliser tanh et profiter de son gradient plus important en 0, mais les résultats sont un peu pres équivalent..

J'ai réduit le nombre de neurones dans les couches denses des deux modèles afin de diminuer la complexité du réseau et d'accélérer sa convergence. J'ai remarqué que les GAN étaient plus longs à converger que les réseaux que j'avais l'habitude de manipuler jusqu'à maintenant (vae, lstm, cnn...). Pour accélérer et améliorer la convergence j'ai aussi changé des paramètres d'entraînement (learning rate, poid de la pénalité de gradient..)

On avait des problèmes avec la valeur de durée la semaine dernière, ce problème était dû au fait que la durée était la seule valeur à être traité de manière continue et que les couches d'activation avaient été conditionnées pour générer des valeurs binaires entre 0 et 1. Donc le réseau générait toujours des 0 pour la durée. J'ai donc discrétisé la valeur de durée en la transformant d'abord en valeur entière, puis en valeur binaire comme ce que j'avais fait pour le nombre d'octet (une durée de 18 prend alors la valeur 10010).

J'ai ensuite défini une fonction d'évaluation de la génération d'un modèle. J'ai d'abord utilisé une forme de différence d'aires sous les courbes de distribution des valeurs générées et réelles, mais je suis ensuite passé à une fonction de distance JS. Plus la distance entre la distribution générée et la distribution réelle est faible, meilleur est le modèle. Nos meilleures modèles ont un score d'environ 3.

J'ai ensuite résumé les recherches bibliographiques de cette semaine et de la semaine dernière. Pour rappel, on cherchait à répondre à deux questions :

En ce moment j'essaye d'implémenter un encodage des header à base d'embedding et d'un autre côté j'essaie de générer les payload en fonction des header générés en regardant les applications de SeqGAN.

2. Ce que les encadrants ont ajouté :

Le choix d'une fonction d'activation du générateur dans notre cas doit se faire en accord avec les données d'entrée, en effet les données de sortie du modèle devront avoir le même encodage que les données d'entrée . Il faudrait peut-être rajouter une fonction entre le discriminateur et le générateur

Pour améliorer la distribution des variables de port, il faudrait songer à diviser les variables en deux, car on remarque qu'il y a à chaque fois un port déterminé et un port aléatoire.

Il faut que je tienne un rapport des expériences qui ont été tentées et de leurs résultats et il faut que je puisse un maximum justifier ces résultats.

Continuer les rapports quotidiens si cela m'aide

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

Continuer la génération de payload avec des hypothèses fortes pour pouvoir avoir des premiers résultats.