Université virtuelle du Burkina Faso (UV-BF)

université virtuelle Burkina Faso

Option : Master Fouille de données et Intelligence Artificielle

Projet Réseaux de Neurones

Auteur : ILBOUDO Issoufou ; MASTER II FDIA

Introduction

Contexte

Ce travail s'inscrit dans le cadre du Master 2 en Fouille des données et intelligence artificielle (MFDIA) première promotion 2021_2023. Il fait suite au cours de Réseaux de Neurones dispensé aux étudiants de cette promotion. Des travaux ont été donnés aux étudiants en vue mettre en pratique les connaissances théoriques acquises sur les réseaux de neurones.

Objectif général du travail : L'objectif de ce mini projet est de mettre en pratique les connaissances théoriques et pratiques sur les réseaux de neurones en répliquant un papier de recherche issue du site web paperswithcode.com.

Objectifs spécifiques :

- Répliquer un papier de recherche sur des "sequence to sequence". "
- Écrire un petit rapport de deux pages sur la reproduction (i.e., tenter d'exécuter le code sur le dataset fourni ou un autre dataset) pour démontrer/justifier ce qui n'a pas marché ;
- déposer le devoir directement sur moodle.

Pour atteindre les objectifs notre travail se déclinera autour des points suivants : compréhension de l'article et le modèle original, Collecte des ressources, Configuration de l'environnement et de l'implémentation du modèle

1. Compréhension du document original à répliquer

L'article Sequence to Sequence Learning with Neural Networks de Sutskever IIya et al. (2014) a été l'un des articles pionniers à montrer que les réseaux neuronaux profonds peuvent être utilisés pour effectuer une traduction « de bout en bout ». L'article démontre que LSTM peut être utilisé avec un minimum d'hypothèses, proposant une architecture 2 LSTM (un « Encodeur » - « Décodeur ») pour faire du Language Translation de l'anglais vers le français, montrant la promesse de la traduction automatique neuronale (NMT) par rapport à la traduction automatique statistique (SMT). La tâche consiste à effectuer la traduction d'une « séquence » de phrases / mots de l'anglais vers le français.Le document propose d'utiliser 2 réseaux LSTM profonds : un agit comme un encodeur : prend votre entrée et la mappe en un vecteur de dimension fixe et le second agit comme un décodeur : prend le vecteur fixe et le mappe à une séquence de sortie. Cela permet d'entraîner le LSTM sur plusieurs paires de langues simultanément. Le document met vraiment en évidence l'astuce consistant à inverser la séquence d'entrée lors du mappage à la séquence de sortie, toute chose qui facilite « l'établissement de la communication » entre l'entrée et la sortie tout en améliorant également les prévisions à court et à long terme du LSTM grace à un « décalage temporel minimal » où la distance entre les mots générés et les mots sources est minimisée en inversant l'ordre. Le LSTM est chargé de prédire la probabilité conditionnelle d'une séquence cible étant donné une séquence d'entrée générée à partir de la dernière couche. La séquence générée à l'aide de cette probabilité peut avoir une longueur différente du texte source. Pour l'entrainement du modèle, Il a été utilisé 160 000 des mots les plus fréquents pour la langue source et 80 000 des mots les plus fréquents pour la langue cible. Chaque mot hors vocabulaire a été remplacé par un jeton spécial « UNK ». Il a consisté à maximiser des log probabilités et à retenir la probabilité la plus élevée :une recherche de faisceau est arrêtée lorsqu'elle atteint un « <EOS> » (caractère de fin de chaîne). Tous les paramètres du LSTM sont initialisés avec la distribution uniforme entre -0,08 et 0,08. Pour faire face aux gradients explosifs, les auteurs mettent à l'échelle les gradients de chaque lot

Les meilleurs résultats sont obtenus avec un ensemble de LSTM qui diffèrent par leurs initialisations aléatoires et par l'ordre aléatoire des mini-lots. Les traductions produites par la

LSTM atteignent en effet un score BLEU de 34,8 et un système SMT basé sur les phrases obtient un score BLEU de 33,3 sur le même ensemble de données. Pour les phrases longues, la traduction a montré des résultats surprenants de longue durée. Les représentations effectuées par projection PCA des LSTM sont sensibles à l'ordre des mots tout en étant assez insensibles au remplacement d'une voix active par une voix passive.

2) Collecte des ressources, Configuration de l'environnement et réplication du modèle

Elle s'est faite selon les étapes ci-dessous :

- **Installation des librairies Pytorch et torchtex avec pip install dans jupyter notebbok et conda** :

  Les modèles ont été codés dans PyTorch en utilisant torchtext. Nous utiliserons également spaCy pour aider à la tokenisation des données. La principale difficulté a été de trouver une version de Pytorch et torchtext compatibes simultanement pour l'exécution des codes dans notebook.

- **Création du repository seq2seq sur le compte** :

  https://github.com/IssoufouILBOUDO/SeqtoSeqModeleIMPL

- **Importation des données et le code source dans jupyther.**
- **Adaptation et réplication.**