PROPOSITION DE STAGE RECHERCHE – MASTER 1

Année 2017

Sujet de stage :

Apprentissage profond pour la correction post-OCR

Résumé du travail proposé :

Tester des approches d'apprentissage profond (Deeplearning) pour la détection et la correction d'erreurs OCR (Optical Character Recognition). Ce travail exploratoire se déroulera à Hanoi dans le laboratoire ICTLab dans le cadre d'un co-encadrement L3i/ICTLab. Les résultats obtenus donneront lieu à une participation à la « ICDAR2017 Competition on Post-OCR Text Correction ».

Mots clés : Optical Character Recognition, Deeplearning, Recurrent neural network, Word2Vec

Informations complémentaires :

Encadrant(s) : Antoine DOUCET / Guillaume CHIRON / Ton LE HUU / Mickaël COUSTATY

Equipe/Thème : Images et Contenus / Valorisation de contenus numériques

Cadre de coopération : France ULR L3i / Vietnam ICTLab

Date de début du stage : mi-avril 2017

Durée du stage : 4 mois minimum

Financement : selon la durée, à discuter

Lieu du stage : ICTLab, Hanoi, Vietnam

Contexte du stage :

Ce stage s'inscrit dans l'initiative « ICTLab Challenge : Deep Learning for OCR errors correction », fruit d'une collaboration entre le laboratoire L3i de la Rochelle et le laboratoire ICTLab de Hanoi, Vietnam. Deux stages de type « échange franco-vietnamiens » y sont proposés, l'un au L3i avec un séjour de 4 mois au Vietnam, l'autre au Vietnam avec un séjour de 2 mois en France. Les deux stagiaires seront invités à travailler ensembles, localement et à distance, accompagné des chercheurs du L3i (Guillaume CHIRON, Antoine DOUCET), et du ICTLab (Huu Ton LE, Hien Phuong LAI, Hoang Tung TRAN) qui travaillent sur la thématique. Ces stages sont en quelques sortes des études de faisabilités pionnières, qui vont au-delà des approches classiques testées et mises en œuvre dans le cadre d'un projet antérieur «AméliOCR » initié en 2016 entre la BnF (Bibliothèque nationale de France) et le Laboratoire L3i.

Description du sujet :

Nous proposons au stagiaire, dans un premier temps d'explorer et de tester des approches d'apprentissage profondes sur des documents textuels (p. ex. https://github.com/karpathy/char-rnn). Les frameworks récents de deeplearning (p. ex. Tensorflow, Keras) sont de plus en plus accessibles et permettent un prototypage rapide. Dans un second temps, nous proposons d'adapter des modèles de langages existants au problème de détection et de correction d'erreurs textuelles présentes sur des documents OCRisées. Un corpus de plusieurs millions de mots composé de textes bruités (résultant d'un pipeline d'OCRisation) accompagnés leur version corrigée sera mis à disposition pour permettre l'entrainement de modèles de langages. Concernant ce volet lié aux données, le projet AméliOCR mentionné ci-dessus fournira un socle de

travail solide grâce au partage de l'expertise accumulé sur le problème depuis un an. Pour finir, nous proposons que les résultats obtenus donnent lieu à une participation à la « ICDAR2017 Competition on Post-OCR Text Correction ».

Prérequis et contraintes particulières :

- Autonomie importante
- Bonne maitrise de l'anglais
- Forte adaptabilité
- Curiosité pour la recherche
- Compétences en programmation (linux bash, python, lua)

Lien / Références :

1. Projet AméliOCR : http://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/A2016000030_post-correction-d-ocr-pour-les-ouvrages-anciens-en-exploitant-les-associations-lexicales-de-l-ocr-bruite
2. Compétition ICDAR2017 : https://sites.google.com/view/icdar2017-postcorrectionocr
3. Exemple de modèle de langages : http://karpathy.github.io/2015/05/21/rnn-effectiveness/, ainsi que

Contacts :

Email : guillaume.chiron@univ-lr.fr / antoine.doucet@univ-lr.fr