

STAGES DEEP LEARNING 2019/2020

CLASSIFICATION, APPRENTISSAGE SEMI-SUPERVISÉ/ROBUSTE, OPTIMISATION DE MODÈLES

1. Contexte

Dans le cadre des solutions qui accompagnent notre offre de signature électronique, nous éditons une solution complète de *Smart Document Management*, qui couvre la classification et la lecture de documents, le contrôle de l'ensemble des données du dossier et la lutte contre la fraude. Le but de l'équipe Deep Learning est d'améliorer ces modules en s'appuyant sur les récentes avancées en la matière appliqués au texte et à l'image, en visant un niveau de performances élevé, tant du point de vue de la précision des algorithmes que de leur temps d'exécution.

2. Sujets proposés

Les problématiques principales abordées cette année concernent : la classification des documents, avec la volonté d'une mise à jour permanente de nos modèles pour suivre les évolutions des documents rencontrés en production (cf. sujets 1 et 2). Par ailleurs, nous souhaitons aussi travailler sur les problématiques de compression et de temps d'exécution de modèles (sujet 3).

Le détail des sujets est disponible à la fin du document.

Le stagiaire aura pour objectif de mettre en application les techniques à l'état de l'art de la recherche sur des problématiques concrètes, au sein d'une équipe R&D d'une quinzaine de personnes dont 4 en Deep Learning. L'équipe bénéficie d'une base de données conséquente pour la mise en place des apprentissages en Deep Learning, point essentiel pour obtenir des bonnes performances avec ces algorithmes.

Après un premier travail bibliographique, le stagiaire devra proposer des méthodes qu'il s'agira d'implémenter et d'évaluer. Il participera aux réflexions sur les choix algorithmiques, d'outils et d'intégration avec le reste de l'équipe R&D.

3. Profil recherché

Étudiant en M2 ou dernière année d'école d'ingénieur spécialisé en Machine Learning et éventuellement en traitement d'images, pour un stage de 5 à 6 mois.

- Curieux, autonome, passionné par le Machine Learning, bonnes bases mathématiques
- Python 3. Le stage se fera avec le framework TensorFlow 2.0.
- Apprécié : un projet ou une expérience en Deep/Machine Learning (utilisation de Scikit Learn, Caffe, Torch, Theano ou TensorFlow)
- Apprécié : une implication dans un projet open-source ou des compétitions de Data Science (type Kaggle) est très positive.

4. Process de recrutement

Un call pour faire connaissance, et un entretien technique avec travail bibliographique à faire en amont sur un sujet Deep Learning.

5. Pour en savoir plus...

Sur QuickSign : retrouvez-nous sur Welcome to the Jungle
<https://www.welcometothejungle.co/fr/companies/quicksign>

Et n'hésitez pas à envoyer votre candidature :

catherine.herold@quicksign.com

6. Détails des sujets

Sujet 1 : Compréhension (explainability), Confiance du classifieur de documents et renforcement de sa capacité de généralisation (image)

Afin de définir la pipeline de traitement d'un document, il est nécessaire en premier lieu de déterminer son type (permis de conduire, carte d'identité, avis d'imposition, ...), et ce peu importe le type d'acquisition (photo, scan, screenshot smartphone, ...). Les documents et sources d'acquisition évoluant dans le temps, le but de ce stage sera de travailler sur la capacité de généralisation d'un modèle.

Par ailleurs, afin de mieux comprendre les erreurs et limites des modèles appris, nous souhaitons investiguer la compréhension des sorties du modèle, afin de compléter de façon pertinente les datasets d'apprentissage. Pour cela, nous nous appuyerons sur les récents travaux autour de *l'explainability* des modèles Deep Learning.

Enfin, afin de déterminer la pertinence de la sortie du module de classification, nous souhaiterions pouvoir associer des notions de confiance aux prédictions du classifieur, pour ne faire intervenir un contrôle manuel que dans les cas où la prise de décision n'est pas possible automatiquement.

Bibliographie :

1. Stefano Teso – [Toward Faithful Explanatory Active Learning with Self-explainable Neural Nets](#), Interactive Adaptive Learning, ECML-PKDD 2019.
2. Riccardo Guidotti, Anna Monreale, Stan Matwin, Dino Pedreschi, [Black Box Explanation by Learning Image Exemplars in the Latent Feature Space](#), ECML-PKDD 2019.
3. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, ["Why Should I Trust You?": Explaining the Predictions of Any Classifier](#) (LIME)
4. Ancona et al, [Towards better understanding of gradient-based attribution methods for Deep Neural Networks](#) (DeepExplain), ICLR 2018.
5. Ramprasaath R. Selvaraju et al., [Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization](#)

6. Aditya Chattopadhyay et al., [Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks](#), WACV 2018
7. Alexandre Rame, [OMNIA Faster R-CNN: Detection in the wild through dataset merging and soft distillation](#)
8. Kei Akuzaw, [Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization](#), ECML 2019
9. Guo et al, "[On Calibration of Modern Neural Networks](#)" , 2017.

Sujet 2 : Apprentissage continu et robuste d'un classifieur de documents (image)

Dans le même contexte que précédemment, nous souhaitons mettre à jour les modèles régulièrement avec de nouvelles données récentes, et ce, à coût minimum: des techniques d'apprentissage semi-supervisé doivent donc être mises en place, ainsi que des approches robustes à de potentielles erreurs d'annotations.

Il s'agira donc d'explorer les méthodes d'apprentissage semi-supervisé, ainsi que des approches de sélection des données pour les apprentissages récurrents. Par ailleurs, ayant à disposition des labels, mais parfois erronés, il sera intéressant de parcourir les méthodes d'apprentissage robuste à des erreurs d'annotations.

Bibliographie :

1. Qizhe Xie et al., [Unsupervised Data Augmentation for Consistency Training](#)
2. Ahmet Iscen et al., [Label Propagation for Deep Semi-supervised Learning](#), CVPR 2019
3. Bo Han et al, [Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels](#), NIPS 2018
4. Qizhe Xie et al., [Unsupervised Data Augmentation for Consistency Training](#)
5. Ahmet Iscen et al., [Label Propagation for Deep Semi-supervised Learning](#), CVPR 2019

Sujet 3 : Compression de modèles

Les architectures des réseaux de neurones convolutifs sont de plus en plus complexes et cela provoque une augmentation de la taille mémoire et de la complexité de calcul.

L'intérêt de la compression de modèles permet non seulement de réduire les temps de traitement mais aussi de réduire les coûts de production. Le stage aura pour but d'explorer les différentes méthodes de quantification, d'élagage de réseaux mais aussi la construction de nouveaux réseaux plus efficaces afin de compresser l'architecture cible sans perte de performance. L'aspect temps de traitement sera également à approfondir, afin de proposer des modèles rapides utilisant des ressources limitées.

Bibliographie :

1. Song Han et al., [Deep compression : Compressing deep neural networks with pruning, trained quantization and Huffman coding](#). ICLR 2016
2. Yu Cheng et al., [A survey of model compression and acceleration for Deep Neural Networks](#)