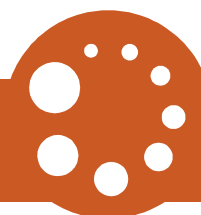




ADELE: Reconnaissance d'images aériennes pour la statistique nationale de la superficie

Projet pilote de la stratégie d'innovation des données:
Application de l'apprentissage automatique profond (Deep Learning)

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

Éditeur: Office fédéral de la statistique (OFS)
Rédaction: Claudio Facchinetti
Domaine: 00 Bases statistiques des généralités
Langue du texte original: Français
Traduction: Services linguistiques de l'OFS

Concept de mise en page: Section DIAM
Téléchargement: www.statistique.ch
Copyright: OFS, Neuchâtel 2019
La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée

Sommaire

1	Objet de ce document	4
2	Résumé	4
3	Situation initiale	4
4	Objectifs	4
5	Résultats	5
5.1	«Business Understanding Overview»	5
5.2	«Data Understanding Overview».....	5
5.3	«Data Preprocessing Overview»	5
5.4	«Modeling Overview»	5
5.5	«Evaluation Overview».....	6
5.6	«Deployment Overview»	7
5.7	Direction du projet.....	7
6	Conséquences/Lessons learned	8
6.1	Lessons learned relatives à la gestion du projet, de la qualité, des processus et des risques.....	8
6.2	D'autres recommandations / expériences utiles pour des projets futurs	8

1 Objet de ce document

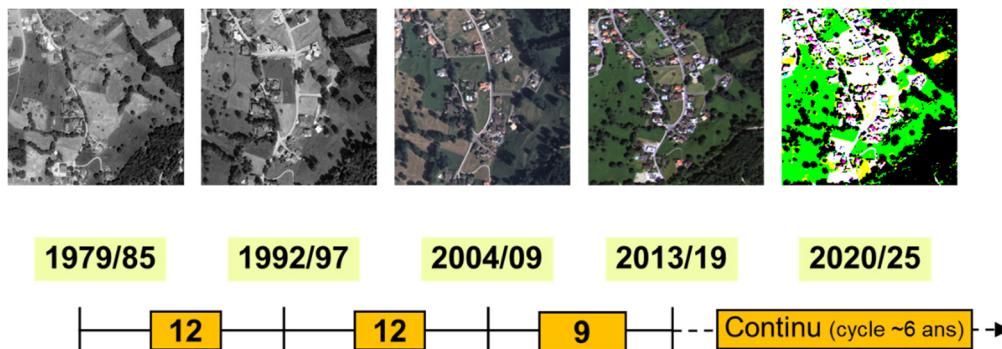
Ce document a pour objet de résumer les événements et résultats du projet-pilote « Arealstatistik Deep Learning » (ADELE). Il contient les conclusions qui peuvent être tirées du travail et propose de clôturer le projet conduit dans le cadre de la stratégie d'innovation sur les données de l'OFS.

2 Résumé

Quatre workshops conduits d'août 2018 à février 2019 ont permis de familiariser le groupe de projet avec les principes de l'intelligence artificielle (dénommée IA ci-après) en exécutant des cycles d'entraînement de *Deep Learning*, tout en variant les paramètres sensibles de l'algorithme, ainsi qu'en testant les modèles obtenus en fonction de différents critères de qualité. Les résultats obtenus ont été intégrés dans le projet de révision de la statistique de la superficie 2020 (*Methodenrevision Arealstatistik* - MA 2020) durant sa phase d'initialisation, conduite en parallèle. La phase de conception, libérée par le comité de direction le 22.05.2019, poursuit les travaux en IA dans le cadre du sous-projet « MA 2020 - ADELE ».

3 Situation initiale

La statistique de la superficie du sol de l'OFS constitue un outil indispensable de l'observation à long terme du territoire. Ses résultats permettent d'évaluer dans quelle mesure l'évolution de l'utilisation du sol est en adéquation avec les objectifs du développement territorial et d'une gestion mesurée de la ressource « sol ». La méthodologie a subi différentes modifications techniques au cours des près de 40 années de production. Depuis 2013, l'enquête est menée de manière continue et une surface équivalente à un sixième de la Suisse est analysée chaque année (cf. figure ci-dessous).



Les ressources sont actuellement allouées presque entièrement à l'interprétation visuelle, au détriment des activités d'analyse et de diffusion. Dans ce contexte, certains algorithmes issus de l'IA peuvent servir à la classification de l'utilisation et de la couverture des sols, ainsi qu'à la détection des changements.

L'objectif du projet pilote NDS est de se familiariser avec ce domaine en vue de l'implémentation d'un prototype informatique, développé en parallèle par un mandat externe. Ce dernier a comme but d'appliquer des algorithmes de l'IA, par étapes, à l'entier du territoire national. Il est intégré à la révision méthodologique de la statistique de la superficie (MA 2020) en prévision de la prochaine enquête.

4 Objectifs

Le but du projet est d'initier les collaborateurs de la section GEO aux nouvelles technologies de l'IA, en vue de leur intégration en production courant 2021.

Les objectifs suivants sont à satisfaire:

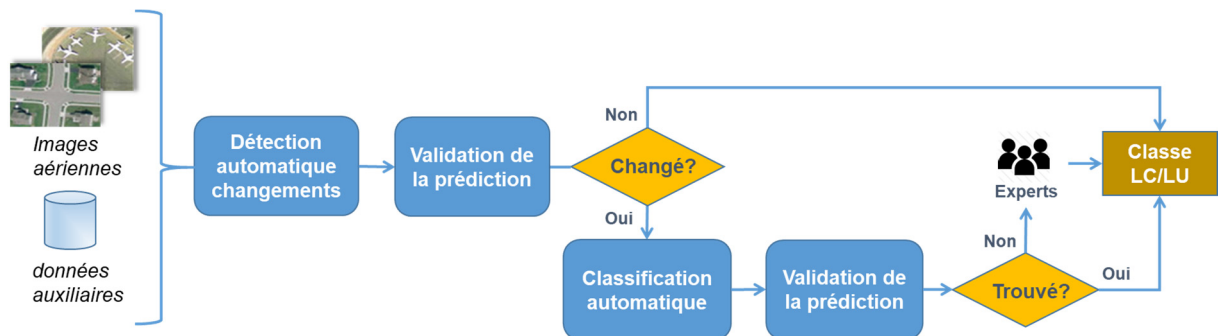
- Comprendre le métier, les enjeux et critères de succès sous la forme d'un problème à résoudre en termes de données et de méthodes issues de l'IA
- Préparer un ensemble partiel de données et d'outils (framework) et l'utiliser pour la formation, l'analyse et l'exploration

5 Résultats

5.1 «Business Understanding Overview»

Une étude de faisabilité¹ a été réalisée en 2017 dans le but de comprendre le métier, formaliser le problème à résoudre en termes de données, comprendre les enjeux, connaître les critères de réussite et mettre en place un plan initial pour réaliser cet objectif. Elle a démontré que, parmi les algorithmes d'apprentissage issus de l'IA pour la reconnaissance d'images, le *Deep Learning* peut contribuer de manière significative à la classification de l'utilisation et de la couverture des sols, ainsi qu'à la détection des changements.

Les workshops conduits pour le projet-pilote ADELE ont ensuite permis de mieux comprendre les enjeux de l'étude de faisabilité et de définir la vision d'un nouveau processus, intégrant l'IA dans l'interprétation visuelle d'images aériennes par des experts. Une attention particulière est attribuée à l'évaluation systématique de la qualité de la prédiction, comme présenté dans le diagramme suivant.



Les étapes partiellement automatisées du processus consisteront en une détection automatique des changements, suivi d'une classification de l'utilisation et de la couverture des sols. Un indicateur de qualité permettra de juger si le résultat nécessite une décision par des experts ou non.

5.2 «Data Understanding Overview»

L'interprétation visuelle est réalisée à partir de photographies aériennes stéréoscopiques, qui forment la base de données primaires. Comme il n'est pas toujours possible d'identifier clairement les points d'échantillonnage, des jeux de données supplémentaires sont utilisés pour appuyer l'interprétation visuelle. Une étude réalisée mi-2018 présente un inventaire actualisé des données secondaires (par ex. classifications précédentes modèles de terrain) et des données auxiliaires (par ex. cadastre, biotopes, constructions), formant l'essentiel des données supplémentaires.

5.3 «Data Preprocessing Overview»

Données primaires

Un échantillon de données primaires issu des enquêtes 2004-2009 et 2013-2019 a été préparé durant le 3^{ème} trimestre 2018 et transféré dans le *cloud Atlantica NDS*. Le traitement consiste en une transformation en tuiles (ou *tiles*) adaptées à la modélisation par apprentissage et leur classification en 4 catégories, selon les données de référence des enquêtes précédentes : zones boisée, agricole, habitat et infrastructure, ou improductive.

Données supplémentaires

Cette phase de manipulation des données a démontré qu'une partie importante des données secondaires et auxiliaires nécessite un prétraitement spécifique (par ex. alignement temporel) ou sont inutilisable (par ex. ombres ou effets de bordures). Elles ont été écartées au vu de l'effort trop important et des ressources à disposition dans le cadre du projet-pilote.

5.4 «Modeling Overview»

Un 1^{er} workshop a été conduit le 29.08.2018 afin de familiariser le groupe de projet avec les principes

¹ Voir [rapport](#) sur le site web de l'OFS

de l'IA en général et du *Deep Learning* en particulier. Plusieurs cycles d'entraînement ont été effectués en démonstration, illustrant notamment la dépendance entre les images aériennes (données primaires) et les données de référence (*ground truth*).

L'environnement dans le *cloud Atlantica NDS* a été complété dans ce but avec des scripts en langage Python permettant les opérations : préparation des données, exécution de cycle d'entraînement du réseau neuronal, tests sur un set de données dédié et prédictions à partir de nouvelles images.

Un 2^{ème} workshop a été conduit le 07.12.2018 avec une présentation détaillée d'un participant expérimenté (Michael Leuenberger) permettant de tester les techniques de modélisation, varier leurs paramètres, comprendre les paramètres et explorer le domaine des solutions.

Les sujets suivants y ont été abordés, afin de préparer les participants pour des essais individuels :

- Présentations de l'algorithme fourni pour le *Deep Learning*
- Présentation des différentes mesures d'évaluation
- Mise en application avec différentes stratégies d'optimisation
- Visualisation des couches du *Deep Learning*

Un 3^{ème} workshop a été conduit le 25.02.2019 afin de comparer les résultats obtenus lors des essais individuels.

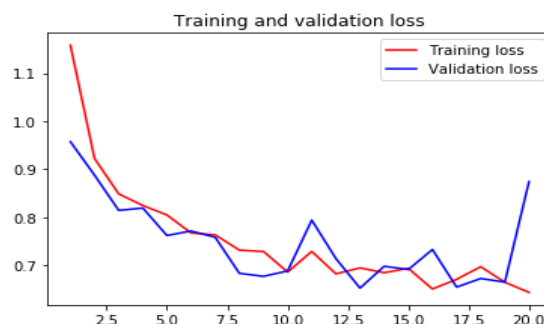
5.5 «Evaluation Overview»

Outils d'évaluation

Les résultats obtenus par les participants ont été évalués en fonction de critères d'évaluation standards, notamment : matrice de confusion, *accuracy*, *precision*, *recall*, score F1, entropie et courbes de validation. L'effort a porté principalement sur la compréhension et l'exploitation des matrices de confusion, pour des cas simples (binaires) ou à classes multiples. Elles fournissent une mesure de la qualité des systèmes de classification et contribuent à la sélection du modèle optimal.

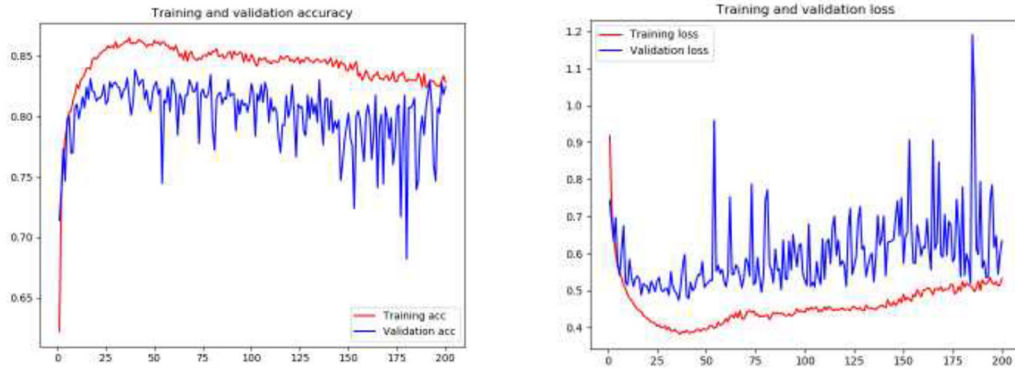
Résultats des évaluations

L'exemple dans le diagramme ci-dessous illustre à l'aide de fonctions de perte par entropie (*loss*) un surajustement observé au cours de l'entraînement initial du 1^{er} workshop, avec un réseau de 3 couches de convolution au cours de 20 itérations de l'ensemble du jeu de données (ou *epochs*). Un modèle surajusté correspond trop étroitement à l'ensemble de données d'entraînement et peut ainsi ne pas correspondre à des données supplémentaires ou ne pas prévoir de manière fiable les observations futures. On observe que la perte estimée pour le set de données de validation (courbe en bleu) augmente par rapport à celle estimée pour le set de données d'entraînement (courbe en rouge).



En considérant les données d'entraînement, de validation et les paramètres fixés, ce type de graphe a pour but de sélectionner un modèle optimal au cours de l'apprentissage du modèle. Celui-ci est identifié lorsque la courbe de perte pour les données de validation atteint un minimum (respectivement un maximum pour la valeur d'*accuracy*), soit 18 dans l'exemple.

Les deux graphes ci-dessous montrent l'évolution de l'apprentissage d'un modèle avec 4 couches de convolution au cours de 200 itérations. Le graphe de gauche affiche la valeur d'*accuracy* (proportion de résultats justes), alors que le graphe de droite indique la valeur de perte estimée par entropie. Pour chacun des graphes, ces deux indices sont calculés pour les données d'entraînement et de validation.



Dans ce cas d'étude, un modèle optimal (pour des données et paramètres déterminés) se situerait entre 25 et 30 itérations. Un modèle sélectionné avant 25 itérations sera considéré comme sous-optimal, le modèle n'a pas terminé sa phase d'apprentissage. Un modèle sélectionné après 30 itérations sera considéré comme sur-optimal (ou surajusté) le modèle ayant été trop contraint par les données d'entraînement et n'étant plus capable de généraliser, comme le démontre le résultat de prédiction pour une nouvelle image, ci-dessous (l'image à droite n'utilisant plus que 2 classes sur 4).



Modèle normal (underfitting / optimal / overfitting)

D'autres résultats ont encore été obtenus en modifiant des paramètres de l'algorithme (ou hyper-paramètres) : nombre de couches de caractéristiques (*feature map*), dimensions des couches de convolution, nombre de cycles d'entraînement (*epochs*) ou encore fonctions d'activation.

De façon générale, toutes les modifications apportées ont un effet sur les aspects suivants du *Deep Learning* :

- Limitations, effets et durée de l'apprentissage
- Complexité et dimensions de l'espace des paramètres à explorer par l'algorithme
- Relations et dépendances entre les paramètres

Cette expérience sera mise en profit dans le futur, lors de la mise en œuvre de réseaux neuronaux avec plusieurs centaines de couches de convolution et des images aériennes couvrant toute la Suisse.

5.6 «Deployment Overview»

Ce projet-pilote précède le développement d'un prototype informatique, mandaté en externe, qui a pour but de prouver courant 2019 cette approche pour l'entier du territoire national. Le projet pilote permet de préparer les collaborateurs au mieux pour la réception du prototype informatique, en particulier pour en vérifier la robustesse avec le support expert de METH.

5.7 Direction du projet

Les personnes et instances ayant été impliquées dans le projet sont les suivantes:

- Section GEO :
 - Claudio Facchinetti (chef de projet)
 - Martina Brügger
 - Romain Douard
 - Michael Bühler
 - Thierry Nippel
- Section METH :
 - Daniel Kilchmann
 - Michael Leuenberger
- Université de Neuchâtel :
 - Michael Leuenberger

Il est prévu que les mêmes personnes et instances continueront à contribuer au projet de révision de la statistique de la superficie (MA2020).

6 Conséquences/Lessons learned

6.1 Lessons learned relatives à la gestion du projet, de la qualité, des processus et des risques

Le volume important de données à traiter (notamment les images aériennes) a dû être limité en raison du volume de stockage restreint dans le *cloud Atlantica*.

Le programme de formation NDS a permis de préparer une partie des participants et leur fournir la connaissance requise, en particulier pour la compréhension du *Deep Learning* et des nombreux critères d'évaluation de la performance. L'échange avec les enseignants et les participants était bénéfique. Malgré des différences importantes dans la manière de traiter les sujets, l'équilibre entre théorie et pratique (projet-pilote) était bon et le « timing » idéal.

Le projet-pilote a permis de démontrer dans le cadre NDS le potentiel d'innovation pour l'OFS dans l'application de l'IA au traitement d'images. Il amène une plus-value au projet MA 2020 en cours, pour mieux comprendre les enjeux et les technologies. Les ressources et les moyens requis peuvent désormais être dimensionnés correctement.

Les objectifs ont pu être atteints en particulier grâce à l'expertise et au soutien régulier des participants d'instances interne (METH) et externe (Uni NE), que nous remercions.

6.2 D'autres recommandations / expériences utiles pour des projets futurs

- L'échange du savoir acquis et des expériences conduites avec l'IA, ainsi qu'une coordination des travaux effectués devrait être assurée au sein de l'OFS. La mise sur pied d'un groupe de spécialistes/experts issus des différentes divisions est souhaitée.
- La poursuite du support externe fourni par Uni NE en matière d'IA appliquée à la statistique publique est souhaitée.
- La focalisation de projets-pilotes sur des thèmes spécifiques et limités est importante, plutôt que de vouloir traiter trop de sujets simultanément.