PHASE PROJET - PROJET LEYENDA

SUJET

Consignes de travail

Présentation

L'entreprise *TouNum* travaille sur la numérisation de documents (textes, images...). Leurs services sont souvent requis par des entreprises numérisant leur base de documents papier. Ils souhaitent étendre leur gamme de services pour inclure des outils de Machine Learning. En effet, certains de leurs clients ont une grande quantité de données à numériser, et un service de catégorisation automatique serait plus que valorisable.

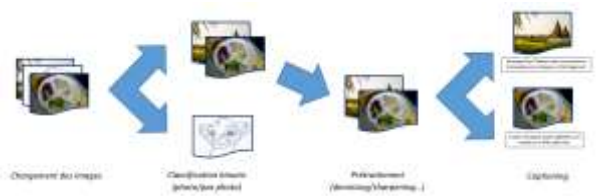
TouNum n'a pas dans son personnel de spécialiste du Machine Learning. L'entreprise fait appel à vous, les spécialistes en Data Science de CESI. On vous propose un premier contrat pour travailler sur une solution visant à analyser des photographies pour en déterminer une légende descriptive de manière automatique (du *captioning*).

Cela parait ambitieux, mais les techniques existent, et elles fonctionnent plutôt bien. Ceci dit, il y a deux challenges à relever, en plus de cet étiquetage proprement dit. Tout d'abord, la numérisation se faisant à la chaîne et sur des images de qualité variable (parfois floues, ou bruitées), il faudra tout d'abord voir ce qu'on peut faire pour nettoyer ces images. Et puis, Tounum a déjà numérisé beaucoup de documents sur lesquels ils souhaitent faire tourner les algorithmes d'apprentissage que vous allez concevoir. Or, Certaines de ces images ne sont pas des photos, mais parfois des images de documents composés, ou des schémas, voire des dessins ou des peintures. Il faudrait donc qu'on puisse, en amont de l'analyse de contenu, faire le tri entre les photos et le reste.

Heureusement, Tounum a déjà quelques milliers d'images catégorisées et étiquetées. Voilà qui devrait être utile pour effectuer de l'apprentissage supervisé.

Objectifs et contraintes

Le workflow que vous devrez concevoir aura la forme suivante :



L'implémentation des algorithmes s'appuiera sur Python et les librairies SciKit et TensorFlow. Par ailleurs, la librairie Pandas sera utilisée dès qu'il s'agira de manipuler des dataset. ImageIO sera utile pour charger des images. Enfin, vous réutiliserez des bibliothèques de calcul avec lesquelles vous avez déjà fait connaissance, comme NumPy et MatPlotLib

La classification binaire s'appuiera sur des réseaux de neurones. Elle permettra de trier les images en deux catégorie : photo, et autre type d'image. A minima, votre algorithme sera capable de différencier entre une photo et un schéma ou un texte scanné. Idéalement, l'algorithme arrivera à discriminer entre des photos et des peintures (ce qui sera plus difficile, une peinture risquant d'être plus proche visuellement d'une photo). Tounum a déjà classé un certain nombre d'images, vous aurez donc un dataset d'images catégorisées pour entraîner votre réseau de neurones.

Le prétraitement s'appuiera sur des notions assez simples autour des filtres de convolution, et les appliquera pour améliorer la qualité de l'image.

Le Captioning, c'est-à-dire la génération automatique des légendes, utilisera deux techniques avancées de Machine Learning : les réseaux de neurones récurrents (RNN), et les réseaux de neurones convolutifs (CNN). Si les RNN permettent de générer les étiquettes, il sera nécessaire de passer avant par des CNN pour prétraiter les images. Cela va nous permettre d'identifier les zones d'intérêt dans les images, et de représenter les images en question de manière plus compacte (parce que si vous devez charger des milliers d'images dans votre RAM, ça va vite devenir coûteux en performances). Vous vous appuierez sur des dataset d'étiquetage classiques pour effectuer l'apprentissage supervisé.

TouNum souhaite être en mesure de produire un logiciel dans quelques mois. Elle attend donc un prototype de votre part d'ici 5 semaines.

La solution que vous proposerez doit être entièrement automatisée (de la donnée à l'étiquetage) pour pouvoir s'adapter rapidement et simplement à tout type de données d'images. TouNum attend une réponse et une solution qui puisse se lire sous la forme d'un notebook Jupyter. À la lecture de vos documents, TouNum doit être en mesure de mettre votre solution en production ainsi que d'en assurer la maintenance. Votre notebook devra aussi présenter le contexte métier dans lequel elle s'inscrit, et démontrer la pertinence de votre modèle de manière rigoureuse et pédagogique.

TouNum vous demandera, après étude de votre livrable (notebook et workflow) de présenter votre solution devant le responsable R&D et le directeur de TouNum. Vous aurez 20 minutes pour expliquer et justifier votre worksflow, faire une démonstration de la solution, et apporter des éléments de réflexion quant aux objectifs finaux de TouNum. Vous devrez faire preuve de recul et d'esprit critique quant au travail présenté. Prévoyez de leur consacrer aussi 20 minutes en cas de questions.

Vos livrables seront donc :

- Un notebook pour chaque module présentant de manière didactique le prototypage de la solution, et la démonstration de la pertinence du modèle, statistiques à l'appui :
    - Le livrable 1 Classification binaire. Ce livrable est à fournir en fin de semaine 3
    - Le livrable 2 Traitement d'images. Ce livrable est à fournir en fin de semaine 4
    - Le livrable 3 Captioning d'images. Ce livrable est à fournir en fin de semaine 5, et reprendra les deux premiers, afin de présenter l'ensemble de la solution, sous forme d'un workflow reproductible
- Une présentation reprenant en 20 minutes l'ensemble des éléments du rapport, et une démonstration sur un dataset qui sera fourni le jour de la soutenance.

Les différents datasets nécessaires à ces livrables vous seront diffusés par l'intermédiaire de votre pilote.

Attention, certains fichiers sont assez volumineux, il est conseillé de les télécharger à l'avance