kaggle_review_2017

June 26, 2023

1 Revue de compétitions Kaggle (2017)

Les gagnants des compétitions Kaggle décrivent parfois leurs solutions sur le blog de Kaggle No Free Hunch. Il y a toujours de bonnes idées à glaner.

[1]:
```
%matplotlib inline
from jyquickhelper import add_notebook_menu
add_notebook_menu()
```

[1]:
```
<IPython.core.display.HTML object>
```

Le blog Kaggle publie régulièrement des interviews des gagnants des compétitions. C'est l'occasion de découvrir la meilleur solution et les outils qui ont permis de la mettre en place. Certains sujets sont des compétitions académiques et les gagnants mettent parfois leur code à disposition sous Github.

1.1 The Nature Conservancy Fisheries Monitoring

kaggle

- **Objectif :** classification d'images de poissons
- **données :** images de poisssons et l'espèces à reconnaître

The Nature Conservancy Fisheries Monitoring Competition, 1st Place Winner's Interview: Team 'Towards Robust-Optimal Learning of Learning'

1.1.1 Eléments clés de la solution

- La solution s'inspire de **transfer learning** : VGG-16, ResNet-101.
- Un modèle de **Fast-RNN** a été utilisée pour extraire les boîtes englobantes des poissons.
- Les images étaient de **tailles différentes**, grandes et petites. Plutôt que de les normaliser toutes à la même taille, **deux modèles** pré-entraînés sur des tailles différentes ont été mis en concurrence.
- Les images ont été réannotées avec des **boîtes englobantes polygonales**. Les poissons sur le pont d'un bateau sont différentes des images de poissons habituelles. Il fallait faire attention à ce que le modèle de classification n'utilise pas le bateau comme features pour classer un type de poisson que ce bateau pêche souvent.
- La base d'images d'apprentissage a été augmentée avec des opérations classiques (**rotations, flous, ...**)
- Il y avait des **images prises de nuits** mais sous-représentées. Une des tâches à consister à créer plus d'images de nuits à partir des images prises de jours en redressant les distributions des couleurs.
- Carte NVIDIA : 2x NVIDIA GTX 1080, 1x NVIDIA TITAN X
- Le meilleur modèle a nécessité 4h d'apprentissage et nécessitait 0.5 secondes pour la prédiction.

Sur GitHub : amsqr/Allen_AI_Kaggle

1.1.2 Idées à récupérer

- Le bateau et l'espèce de poisson étaient corrélés. Il fallait réduire cette corrélation.
- Certaines images étaient très types (de nuit) et sous-représentées. Il a fallu redresser la base d'apprentissage de ces petits biais.
- Transfer Learning en première approche.

1.2 Data Science Bowl 2017: Can you improve lung cancer detection?

kaggle

- **objectif :** détecter la présence d'un cancer du poumon
- **données :** scan 3D des poumons haute résolution

2017 Data Science Bowl, Predicting Lung Cancer: 2nd Place Solution Write-up, Daniel Hammack and Julian de Wit, 2nd place solution for the 2017 national datascience bowl

1.2.1 Eléments de solutions

- Les gagnants ont **augmentés leurs données** avec les résultats du challenge Lung Nodule Andalysis 2016 (de cet article Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge), et des tutoriels Full Preprocessing Tutorial ainsi que les autres comme celui-ci Applying a 3D convolutional neural network to the data
- La première étape a été de **reconstituer les images 3D** (une image 3D = séquence d'images 2D prise à intervalle régulier)
- Le problèmes **a été modifié** en prédictions de la taille des nodules (à partir d'autres jeux de données LIDC) sur des bouts de scans puis prédiction de la probabilités de cancer.
- Utilisation de la méthode Curriculum Learning qui consiste à apprendre un modèle **d'abord sur des données propres** puis à ajouter petit à petit des exemples plus difficiles à classer.
- Pas d'utilisation de modèles connus U-net (modèles de deep learning spécialisés dans la segmentation d'images médicales) mais utilisation de **CNN sur des fenêtres glissantes**.
- Le modèle C3D (**Convolution 3D**) n'a pas marché tel quel mais le modèle final est très proche de la même architecture (github/C3D).
- Deux auteurs, deux codes : code 1, code 2, rapport et autres références.
- L'équipe arrivée 9ème a décrit sa solution Predicting lung cancer - un peu plus clair et didactique.

1.2.2 Idées à récupérer

- Utilisation de modèles existants pour prédire des résultats intermédiaires sur les nodules pour augmenter les données.
- Exemple de code avec des modèles de convolution 3D.
- Architecture gagnante : U-net](https://arxiv.org/abs/1505.04597), C3D
- Base de données médicales : Lung Nodule Andalysis 2016, LIDC

1.3 March Machine Learning Mania 2017: Predict the 2017 NCAA Basketball Tournament

kaggle

- **objectif :** prédire des résultats de basket-ball
- **données :** données des matchs passés

March Machine Learning Mania, 5th Place Winner's Interview: David Scott

1.3.1 Elements de solution

- Le participant arrivé 5ème est un **passionné de basket et de régression logistique** et c'est tout ce qu'il avouera.
- Le premier a réutilisé la solution d'ancien vainqueur Building an NCAA mens basketball predictive model and quantifying its success.
- Utilisation de Linear Mixed Effects Models qui s'appliquent sur des **données non indépendantes** : utilisation d'une variable de groupe.
- Le challenge était évalué avec la formule suivante : $LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i \log \hat{y_i} + (1 - y_i) \log(1 - \hat{y_i})$ avec $\hat{y_i}$ la probabilité pour l'équipe 1 de battre l'équipe 2. Ce type d'erreur marche plutôt bien pour la régression logistique puisqu'elle optimise cette erreur.
- Le gagnant a construit **15 features** issues de la connaissance du basket (degré d'offensivité…).
- Il fallait aussi **prédire le gagnant d'un tournoi** : cette prédiction a été réalisée à partir de simulation faite à partir du modèle appris.

1.3.2 Idées à récupérer

- Il vaut connaître le basket.

1.4 Dstl Satellite Imagery Feature Detection

kaggle

- **objectif :** reconnaître des objets (voitures, bâtiments, routes, arbres, cours d'eau, lacs…) dans de grandes images satellites
- **données :** grandes images satellistes, la même zone est recouverte avec différents types de capteurs (noir et blanc, infra…)

Dstl Satellite Imagery Competition, 3rd Place Winners' Interview: Vladimir & Sergey

1.4.1 Eléments de la solution

- Pas d'utilisation de Fast-RNN ou SSD trop complexes à mettre en oeuvre.
- **Distributions des classes** : il y a plus d'arbres que de voitures, les prédictions doivent refléter cette distribution. Il est possible de modifier les prédictions afin de respecter cette distribution.
- Utilisation des **détecteurs d'images simples** tels que NDWI, cela marche assez bien pour les cours d'eau
- Utilisation de **réseaux de neurones** : u-net, Tiramisu, un optimiseur différent Incorporating Nesterov Momentum into Adam
- **Choisir les features** pour le réseau de neurones : RBG, M, P images, on laisse tomber ECI, SAVI, NDWI
- **Découper les images** plutôt que les normaliser : un réseau de neurones est appliqué sur chaque image. Problème : les prédictions sont moins bonnes sur les bords. On résoud cela avec un quadrillage qui se superposent (**overlapping** grid) et on ajoute une couche dans le réseau de neurones qui enlève les bords Cropping2D.
- Les images 3600x3600 ont été découpées en zones 112x112 : cela laisse des zones incomplètes de l'autre côté de l'image. Ces zones sont complétées avec une image symétrique (comme dans un miroir).
- Prédire sur les images symétriques et réconcilier les prédictions.

1.4.2 Idées à retenir

- Détecteurs d'images pour les arbres, les cours d'eau ECI, SAVI, NDWI
- Corriger les sorties avec des informations sur une distribution a priori des objets à détecter

1.5 Autres compétitions

- Two Sigma Financial Modeling Code Competition, 5th Place Winners' Interview: Team Best Fitting | Bestfitting, Zero, & CircleCircle - algorithme de trading, la solution est assez classique et introduit des caractéristiques que les modèles ont du mal à capturer comme des courtes périodes de fortes volatilités, des périodes assez calmes…

- Santander Product Recommendation Competition: 3rd Place Winner's Interview, Ryuji Sakata - recommandation de produits, ce participant a isolé deux catégories de produits (tendances inhabituelles pour l'un des groupes) pour lesquels deux modèles et deux jeux de features distincts ont été construits.

- Seizure Prediction Competition, 3rd Place Winner's Interview: Gareth Jones - prédire les crises (épilepsie, …). Les features s'appuient sur des électro-encéphalogramme (EEG). A noter l'utlisation de RUSboost pour les problèmes de données mal balancées.

[2]: