

# Détecteur de signatures sur documents commerciaux

Cédric Soares • 16/10/2021



# **Le contexte**



# Porteurs du projet

Deux équipes de l'entité Orange  
France / DTSI / DS / DEVRAP / DS  
DO :

- 3MU
- CATOP

Périmètre d'action :

- Implémentation rapide de solutions
- Développement d'automates logiciels (RPA / RDA)
- Centre d'optimisation de coûts

Au total 16 personnes :

- 2 Managers
- 9 Développeur.r.se.s d'automates
- 2 Développeurs d'intelligences artificielles
- 1 Administrateur systèmes
- 1 DevOps
- 1 Product Owner / Chef de projet



# Pourquoi des développeurs d'intelligences artificielles ?

Depuis  
novembre 2019



Vitali Shchutski

## Les motivations :

- Des automates limités par des règles métier et des champs pré-existants
- Favoriser la montée en compétence des équipes.
- Réaliser d'expérimentations / POC issus de besoins identifiés dans des business units



Cédric Soares

## Les travaux :

- Api FastText de détection de langages
- DéTECTeur d'appels frauduleux



# Les besoins du projet

## Les réalités du marché Pro PME :

- Un process de ventes moins direct que le segment particuliers
- Un souscripteur n'est pas forcément l'utilisateur
- Un volume de document plus importants
- Des contraintes réglementaires plus importantes
- In fine un traitement qui doit être fait manuellement

## Les enjeux de la DC2P :

- Automatiser le processus
- Réduire des coûts

**MANDAT DE PRELEVEMENT SEPA - comptes FE**

La présente demande est valable jusqu'à annulation de ma part à notifier en temps voulu au créancier.

Business Services																									
Raison Sociale (*)																									
Adresse (*)																									
Complément Adresse																									
Code postal (*)																									
Ville (*)																									
Pays (*) FR																									
L'information obligatoire																									
DESIGNATION DE L'ETABLISSEMENT TENANT DU COMPTE A DEBITER (Nom et adresse de votre banque)																									
Designation Etablissement (*)																									
Adresse (*)																									
Complement Adresse																									
Code postal (*)																									
Ville (*)																									
Pays (*) FR																									
L'information obligatoire																									
CODE INTERNATIONAL BANQUE DEBiteur (BIC) - Obligatoire																									
COMPTE A DEBITER (Identification internationale du compte bancaire - IBAN - Obligatoire)																									
code pays clé	code banque	code guichet	numéro de compte	code RIB																					
F	R	7	6	3	0	0	0	3	0	1	9	1	0	0	0	2	0	1	5	3	5	8	5	3	5
NOM ET ADRESSE DU CONTRACTANT (à l'effacement si débiteur)																									
Raison Sociale (*)																									
Adresse (*)																									
Complement																									
Code postal (*)																									
Ville (*)																									
Pays (*)																									
L'information obligatoire																									
COMPTES DE FACTURATION ORANGE A PRELEVER (*)																									
<input type="checkbox"/> Comptes de facturation à prélever cocher si nécessaire																									
[Redacted]																									
Adresse retour de la demande																									
Orange Business Services TSA 60816 82008 Montauban Cedex																									
Identité de l'entreprise																									
Identifiant Orange BSN (ICB)																									
Type de paiement : récurrent																									
Référence Unique Mandat (RUM) - (seules exclusivement au créancier)																									
Identifiant Orange BSN (ICB)																									
Prise de renvoi est imprime complète à l'adresse de retour de la demande mentionnée ci-dessus, et y joignent obligatoirement un Relevé d'Identité Bancaire (R.I.B.).																									
A (Ind.) : [Redacted] 3 2019																									
Le : 1																									
Nom du signataire																									
Fonction du signataire B.G.																									
Signature et cachet du débiteur																									
Vis-dire concernant le prélèvement SEPA sont reproduits dans un document joint à ce mandat, à l'effacement duquel il sera suivi par la demande de remboursement. Les informations contenues dans le présent mandat, ou dont il comporte, sont destinées à être utilisées par le créancier pour la gestion de sa relation avec son client. Elles sont destinées à être conservées par le créancier, ou à son débiteur, ou à ses deux échelons, de remise en disponibilité tel que prévu par la loi « Le droit d'Intimité et de Liberté » du 1er juillet 1978 modifiée.																									
Référence unique mandat (RUM)																									
Orange, S.A. au capital de 10 640 226 396 EUR - 79 rue Olivier de Serres, 75015 Paris - 360 129 866 RCS Paris - Code APE 6110Z - N° TVA intra-communautaire FR 360 129 866																									



# Les utilisateurs

- Opérateur : soumet les images
- Mainteneur : entraîne le modèle
- Administrateur : supervise le bon fonctionnement



# Les contraintes

- Une interface graphique pour soumettre les images
- Une api REST interrogeable indépendamment de l'interface
- Un script d'entraînement du modèle
- Permet de réaliser des détections successives
- Des composants conteneurisés

# La base de données analytique



# Le dataset

## Tobacco-800 : 789 documents

- 1290 documents dont 789 annotés
- Issus de l'industrie du tabac
- Créée par Illinois Institute of Technology en 2006

## Nanonets : 174 documents

- Issus des moteurs de recherche Bing et Google
- Mis à disposition pour tester une API en 2018

## GSA Lease: 772 documents

- Baux commerciaux contractés par l'administration Américaine
- Classés par états

*Lorillard*

LORILLARD, INC. Research Center, 420 English Street, P.O. Box 21688, Greensboro, North Carolina 27420-1688

February 11, 1987

Dr. Peter T. Thomas  
Senior Toxicologist  
IIT Research Institute  
10 West 35th Street  
Chicago, IL 60616

Dear Dr. Thomas:

We have completed our review of four Immunomodulatory Screening Tests, recently submitted to Lorillard, and have the following comment.

SN45LOR(A76)

Page 9, Paragraph 1, last sentence: This statement is too vague as written. We would agree that an increase in spleen weight could indicate immunomodulation; however, the relationship between toxicity and the observed immune response is unclear, as written. Please clarify this statement.

We have also reviewed SN42LOR(B109), SN42LOR(A166), and SN41LOR(A16). These reports are acceptable to us in their present form and can be considered final reports.

Thank you for your attention to these reports. If you have any questions, please feel free to call.

Sincerely,



Thomas A. Vollmuth, Ph.D.  
Toxicologist, Life Sciences

/tb:1

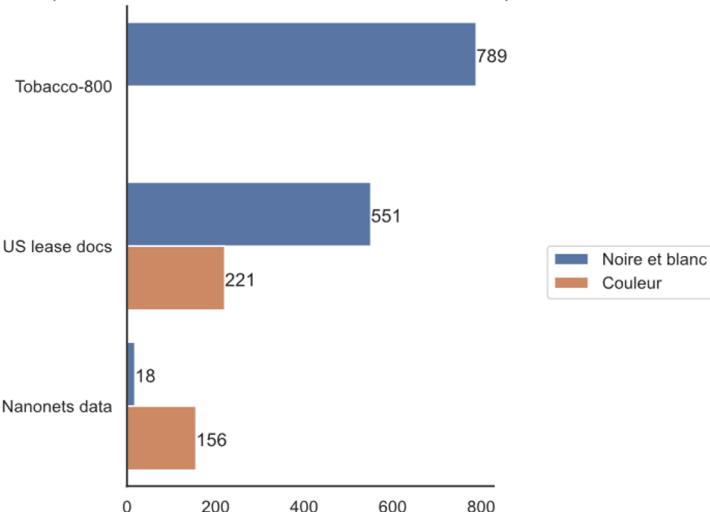
8595251CS

Soit un total, à date, de 1735 documents scannés

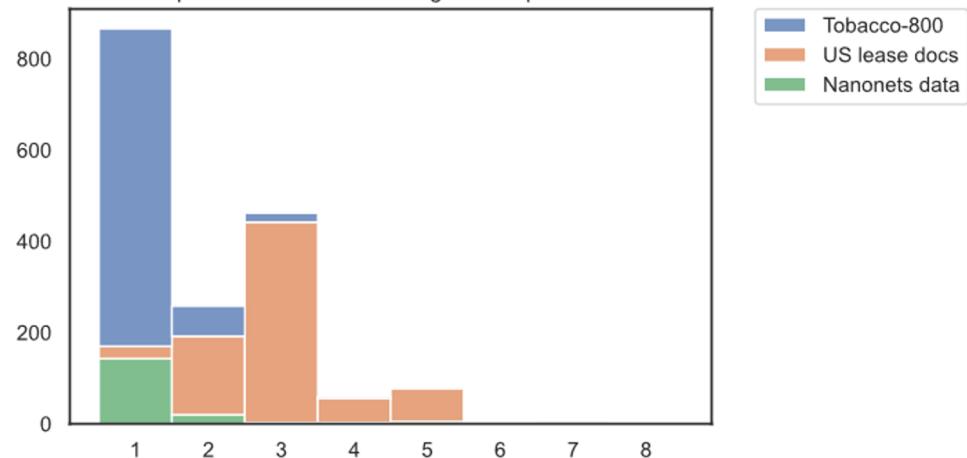


# Analyse des contenus des documents

Répartition des documents noirs et blancs vs en couleurs par source

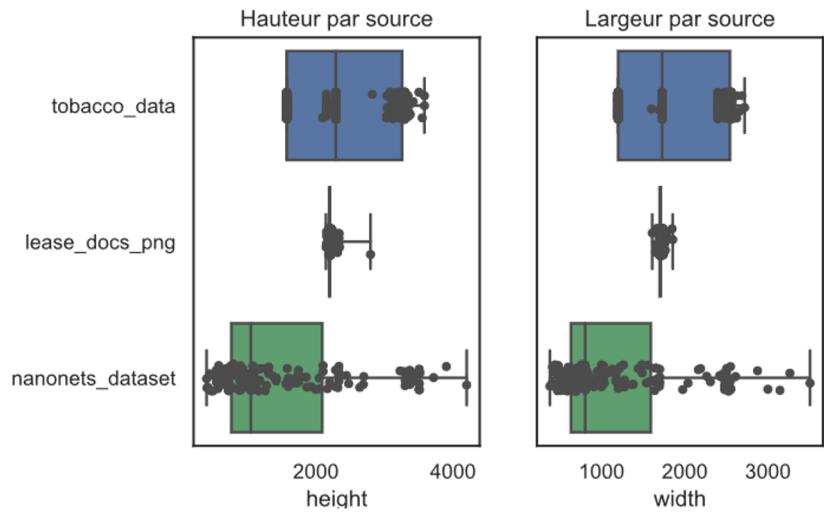
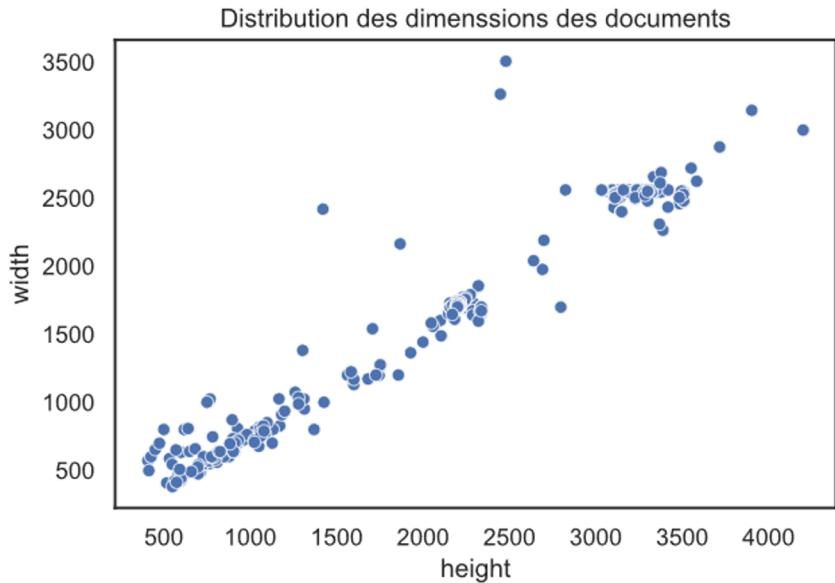


Répartition du nombre de signatures par source





# Analyse de la forme des documents



# Le modèle



# Les critères de choix

## Rapport Rapidité / Précision

- Permet la détection d'images scannées à la volée sans détériorer les performances

## Réactivité

- Doit pouvoir réaliser des détections en série

## Maturité

- Un modèle reconnu et largement déployé

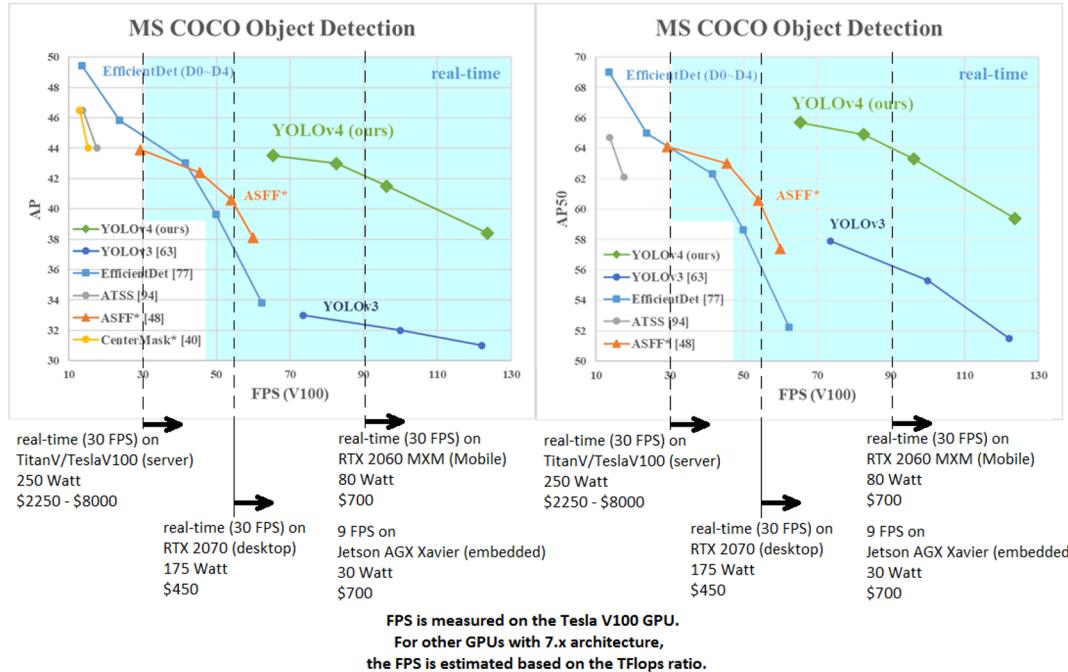


# Le classement paperswithcode (mai 2020)

	Score mAP (Position)	Score FPS (Position)	Score Inference Time (Position)
EfficientNet-D7x	55.1 (1e)	6.5 (18e)	Pas de mesure dans le benchmark
CSPResneXt	33.4 (17e)	58 (3e)	17 (3e)
YolovoV4 - 512	43 (7e)	83 (1e)	12 (1e)

Si Yolo V4 n'est pas le modèle offrant le meilleur rapport précision sur recall mais il est de loin le plus réactif

# L'article de recherche



Par contre les performances sont plus robustes face à l'augmentation du nombre d'images par secondes



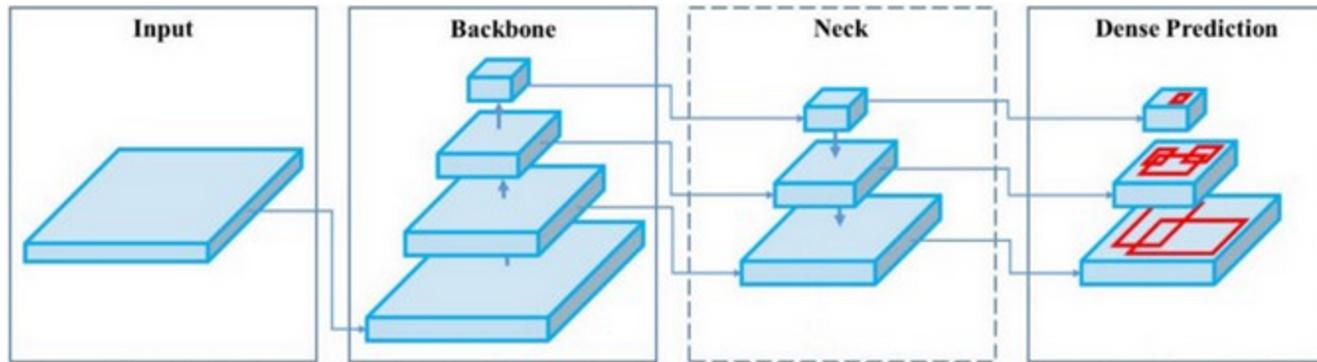
# Un modèle reconnu

- Une 4ème version d'un projet qui a débuté en 2016
- Utilisé dans 145 articles de recherche
- 13 600 Stars et 14 900 forks sur le repository
- Utilisé par Andrew Ng pour ses formations





# Les principes de l'architecture du modèle



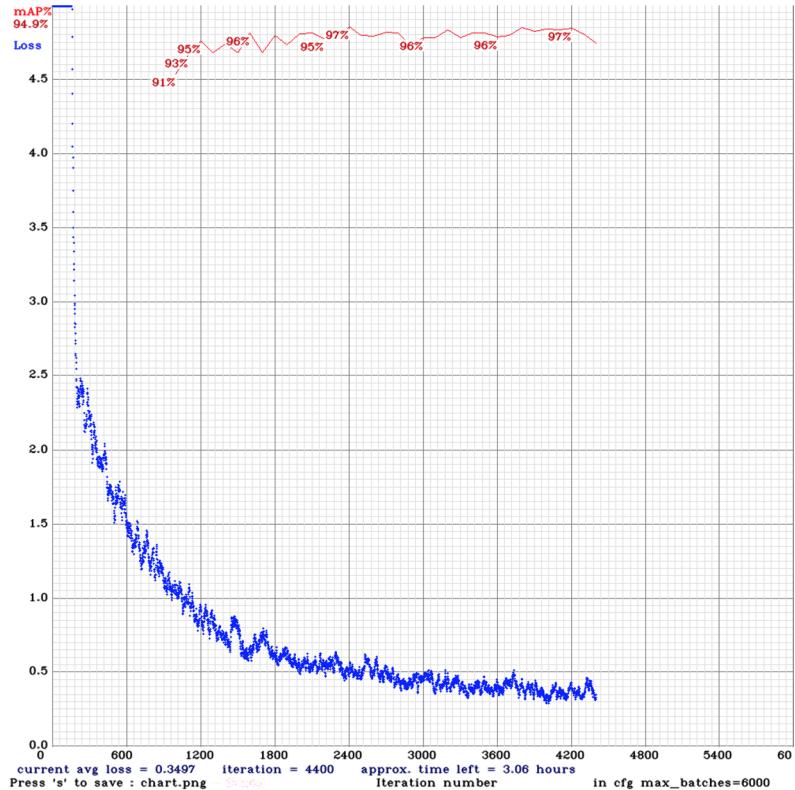
# L'entraînement du modèle



# L'entraînement

## Les hyperparamètres :

- height et width : 608
- batch-size : 64
- 6000 batches
- learning rate : 0.001





# La configuration

- CPU : Intel i9-9900k
- GPU: Nvidia RTX 2080Ti 11 Go GDDR6
- RAM : 64 Go 3600 mhz DDR4



# L'évaluation sur le jeu de test

	Nombre d'images	mAP	Precision	Recall	F1 Score
Premier entraînement	789	29.57%	44%	25%	32%
Deuxième entraînement	1113	65.06%	52%	75%	62%
Troisième entraînement	1735	68.20%	71%	69%	70%



# **Les erreurs marquantes**

- Les signatures encadrées par des annotations
- Les signatures recouvertes par un tampon

# L'intégration



# Périmètres fonctionnels

## Interface Graphique

- Mise en ligne des images et conversion
- Mise en forme des résultats
- Gestion des sessions des utilisateurs
- Enregistrement en base de données relationnelle
- Prise de traces pour le monitoring

## Api

- Appelle le modèle pour réaliser une détection
- Renvoie les données du résultat

## Entraînement

- Entraîner le modèle
- Stocker les images d'entraînements, les fichiers de configuration et de poids

Utilisation continue / usuelle

Utilisation ponctuelle



# Périmètres fonctionnels

## Interface Graphique



## Api



## Entraînement



Utilisation continue / usuelle

Utilisation ponctuelle

# Le bilan



# Synthèse

## Forces :

- Une application fonctionnelle
- Complètement conteneurisée
- Une Api déployée dans le Cloud

## Faiblesses :

- De nombreux biais importants issus du dataset
- Un niveau d'interprétabilité faible de YoloV4

Merci pour votre  
attention