**Recherches sur la fraude à l’assurance**

Définition de la fraude à l’assurance :

C’est un acte intentionnel, réalisé par une personne morale ou physique, afin d’obtenir indûment un profit du contrat d’assurance.

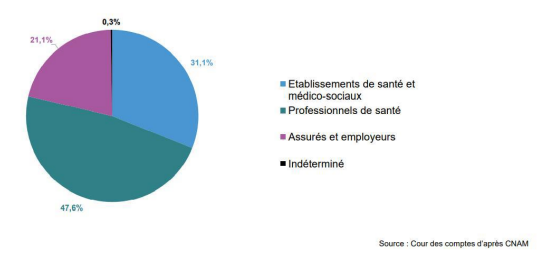
D’après l’étude Trustpair X Accenture X SAP, 95% des entreprises ont fait l’objet d’une tentative de fraude en 2021. Pour les 3/4 de ces tentatives, des pertes financières ont été à déplorer. Sans surprise, ce sont les cyberfraudes qui causent le plus de dommages – et parmi elles, les fraudes à la fausse facture sont en tête de file.

**Deux types d’activités :**

***Les activités fautives*** ce sont des « *activités qui rassemblent de manière réitérée des faits irréguliers au regard de textes juridiques (notamment les nomenclatures), commis par une ou plusieurs personnes physiques ou morales en dehors des circonstances définissant la fraude ou les activités abusives.* » (Selon la Caisse Nationale de l’Assurance Maladie (CNAM)). Par exemple : le non-respect des règles de facturation des majorations prévues par la nomenclature.

***Les activités abusives*** sont des activités « qui rassemblent de manière réitérée des faits caractéristiques d’une utilisation d’un bien, d’un service ou d’une prérogative outrepassant des niveaux acceptables par rapport à des références ou des comparatifs non juridiques. » (Selon la Caisse Nationale de l’Assurance Maladie (CNAM)). Notamment l’exécution d’actes médicaux au-delà du besoin de soins du patient.

La fraude apparait dans plusieurs domaines comme présenter ci-dessous.



Selon le rapport de la cour des comptes française au sujet de la lutte contre la fraude sur les prestations sociales publiée en 2020, il apparait que la fraude se situe très majoritairement du côté des professionnels de santé et des établissements de soins (près de 80% du montant des fraudes selon les données du rapport de la Cour des comptes contre 20% du côté des assurés).

**Facture fictive :**

Elle ne correspond à aucune prestation de service ou livraison de biens. Souvent éditée par des sociétés en redressement ou en liquidation judiciaire, elle vise à alimenter une « caisse noire » ou atterrit directement dans les poches des dirigeants. En d’autres termes, cette facture est purement imaginaire et n’a aucune assise dans une réalité économique concrète, un faux médecin traitant, un faux numéro matricule, un faux nom, la structure qui génère la facture (le nom de la structure) n’existe pas et les dates de réclamation de factures médicales sont antérieures au début de la période d'assurance.

Différents procédés émergent en voici quelques exemples :

### ****Procédé primaire****

Ce processus vise à émettre des factures avec de faux noms ou adresses et des factures pour des quantités ne correspondant pas à la réalité.

### ****Procédé intermédiaire : fausse facture travaux****

Ici, on parle de la transformation d'une opération non déductible en une opération déductible. Prestation pas soumise au régime obligatoire

**Exemple**

Camouflage d'une construction de court de tennis pour le dirigeant en travaux immobiliers de la société.

### ****Procédé plus élaboré : fausse signature****

Cela vise l’utilisation de la **fausse facture taxi (facture classique)** où la facture couvre une opération réelle mais qu’elle est établie par un tiers qui n’a pas effectué la prestation.

### ****Facture fictive d’achat ou de vente****

Ici, on note l’absence de livraison de biens ou de prestation réelle, l’alimentation d'une caisse noire ou le détournement par les dirigeants.

Elle est souvent émise par une société en défaillance fiscale.

**Facture de complaisance :**

Dans ce cas-là, la transaction (ou prestation de service) a eu lieu, mais pas dans les conditions indiquées sur la facture.

Exemple :

1. La quantité d’article livrés inscrite sur la facture ne correspond pas à la réalité. Ou le commanditaire de la prestation n’est pas le bon. Ce type de fraudes généralement employé pour dissimuler du travail non déclaré, ou pour s’octroyer une déduction fiscale sur une opération non éligible.
2. La facture frauduleuse classique qu’on appelle la facture taxi couvre une opération réelle, qui va être établie par une autre personne que celle qui a réellement réalisé l’opération (le plus souvent pour cacher une sous-traitance).
3. Une entreprise peut aussi faire une fausse facture afin de tromper sa banque en vue d’obtenir le déblocage d’un prêt. Notamment le cas des fausses factures de travaux, ou les escrocs vont facturer des prestations inexistantes, ou gonfler le prix des travaux pour disposer de fonds.

## **Comment vérifier si une facture est fausse ?**

L'administration se concentre d'abord sur les **mouvements financiers** pour identifier des factures fictives ou de complaisance. Ces factures permettent de :

* Réduire les bénéfices imposables ;
* Récupérer la TVA facturée ;
* Et de dissimuler une activité non déclarée.
* Récupérer de l’argent (grosse somme)

**Différentes méthode existe pour contrôler une facture, en voici quelques-unes :**

1. Identité du prestataire
2. Numéro SIREN/SIRET
3. Numéro TVA
4. Les coordonnées bancaires de l’entreprise

**Mentions obligatoires factures**

Pour être considérée comme vraie, une facture doit comporter les éléments suivants :

* La date de la facture (la date à laquelle la facture a été émise),
* Le numéro de facture (numéro unique de chaque facture obtenue à partir d’une séquence chronologique et continue), (optionnel)
* La **date de la prestation de service ou de la vente**,
* L’identité du prestataire de service ou du vendeur,
* L’identité du client ou de l’acheteur,
* Le numéro d’identification à la TVA, (optionnel)
* Le numéro du bon de commande, (ça dépend des factures)
* Le **décompte et la désignation des produits et services rendus**,
* Le prix catalogue, (montant de la prestation)
* Le taux de TVA légalement applicable, (pas sur toutes les factures)
* Le prix HT,
* Le montant de la TVA, (optionnel)
* Le montant TTC.

**A savoir :**

Dès le 1er juillet 2024, il ne sera plus possible de demander à vos fournisseurs des factures au format papier ou en PDF simple. Toutes les entreprises devront :  
Accepter la réception des factures électroniques de leurs fournisseurs  
Être en mesure de les intégrer dans leur système de comptabilité ou de gestion commerciale.

**Recherche sur les réseaux de neurones :**

La reconnaissance de caractères manuscrits revêt une importance particulière, du fait que comme la reconnaissance de la parole, elle s’inscrit dans un schéma général de communication homme machine. On peut très bien imaginer des systèmes fiables, permettant de donner des ordres à une machine au moyen de l’´écrit. De tels systèmes seraient aussi capables d’intégrer des bases de données manuscrites, de reconnaitre un code postal, une somme sur un chèque bancaire, voire pourquoi pas, de reconnaitre le scripteur. Selon le type d’application envisage, on peut opérer plusieurs distinctions sur l’approche à employer ainsi que sur les performances à obtenir.

1. Les techniques dites online, utilisent un capteur spécifique (tablette graphique, stylo ´électronique) pour acquérir les informations. Ces informations reflètent la dynamique de l’´écriture, dynamique pouvant se révéler fort utile pour segmenter des caractères (i.e. isoler chaque caractère), voire pour le reconnaitre explicitement (ainsi l’ordre du dessin d’un caractère chinois influence son sens).
2. Les techniques dites offline, permettent de travailler sur une image fournie par un scanner. Cette approche est adaptée au traitement de documents existants (chèques, textes, ...). Cependant le manque d’information dynamique y rend la segmentation plus difficile. On pourra tolérer des performances moyennes dans le cas de la reconnaissance de texte. Les erreurs pourront en effet ensuite être traitées à l’aide d’un dictionnaire. Dans le cas de l’identification d’un scripteur (vérification d’une signature, autorisation d’accès, ...), de hautes performances sont cependant nécessaires. Les applications de notre travail sont plutôt à envisager dans le cadre d’un système offline n´nécessitant de bonnes performances.

**Le système OCR :**

Il consiste en général en un ensemble de modules de prétraitement, segmentation, extraction de traits, classification et traitement contextuel. Un document papier est dans un premier temps numérise pour obtenir une image en niveaux de gris (´éventuellement en noir et blanc). Durant le prétraitement, on filtre le signal pour ´éliminer le bruit, les zones contenant du texte sont d´détectées et converties en images binaires à l’aide d’une m´méthode de seuillage, soit globale, soit localement adaptative. Durant la phase de segmentation, les images texte sont séparées en caractères distincts. Ce qui est extrêmement d´délicat dans le cadre de textes manuscrits du fait de la connexité des caractères. Une des techniques souvent mise en œuvre consiste à découper les patrons en unités plus petites (sur-segmentation), puis à trouver les points de segmentation corrects à l’aide des sorties d’un classificateur de patrons. Du fait de la grande variation des ´écritures et d’un haut niveau de bruit, reconnaitre des caractères segmentés n’est pas chose aisée. Le lecteur pourra en juger sur la figure 3.1 qui montrent des exemplaires normalises en taille d’un ´échantillon de la base de caractères manuscrits du National Institute of Standards and Technology (NIST).

La figure 3.2 présente deux manières d’utiliser un RNA (réseau neuronal artificiel) dans un système OCR. La première utilise un extracteur de traits (pas nécessairement de type RNA). Par exemple, les traits caractéristiques peuvent ˆêtre les contours des lettres, qui sont donnés en entrée au réseau non-boucle. Cette approche est extrêmement flexible puisqu’elle permet d’incorporer une grande variété de caractéristiques. L’autre approche n’utilise pas de manière explicite une extraction des traits `à partir des données brutes. L’extraction se fait de façon implicite dans les couches cachées du réseau. L’avantage de cette m´méthode est que l’extraction et la classification sont intégrées et entrainées simultanément afin de produire une classification optimale. Il n’est pas ´établi que l’extraction implicite soit plus efficace dans le cadre de la reconnaissance de caractères et qui plus est, cette approche n´nécessite un réseau de taille beaucoup plus conséquente que la première méthode. Le Cun a mis en œuvre cette seconde technique pour la reconnaissance de codes postaux (LeNet). On présente en entrée au réseau une image normalisée en niveaux de gris `a un réseau possédant trois couches cachées. La première couche réalise une cartographie des traits localisés. Les cellules de la seconde couche combinent les informations de la première couche. Le niveau d’activation en sortie pouvant s’interpréter comme une approximation de la probabilité `a posteriori qu’un caractère appartienne `a une classe donnée. Les catégories en sortie sont ordonnées en fonction des niveaux d’activation et sont fournies à l´étape de post-processing. A cette ´étape, l’information contextuelle est utilisée pour mettre `a jour les classifications. Ceci peut mettre en jeu, par exemple, un dictionnaire des mots admissibles ou utiliser des contraintes syntaxiques présentes dans un numéro de téléphone ou de sécurité sociale. Les RNA sont performants dans le cas des systèmes OCR. Cependant, aucun résultat ne permet d’´établir leur supériorité par rapport `a des techniques classiques comme les classificateurs statistiques. Lors de la première conférence en 1992 “Census Optical Character Recognition System Conférence”, plus de quarante systèmes de reconnaissance de caractères manuscrits furent ´évalués sur une base commune. Les dix premiers utilisaient soit un type particulier de réseaux non bouclés, soit un classificateur basé sur les plus proches voisins.

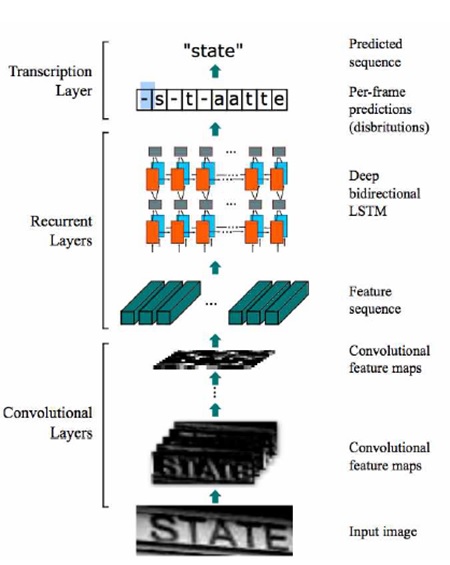
CRNN

Le CRNN est un réseau qui combine CNN et RNN pour traiter des images contenant des informations de séquence telles que des lettres.

Il est principalement utilisé pour la technologie OCR et présente les avantages suivants.

* 1. Un apprentissage de bout en bout est possible
  2. Des données de séquence de longueur arbitraire peuvent être traitées grâce au LSTM qui est libre en termes de taille de séquence d’entrée et de sortie.
  3. Il n’est pas nécessaire d’avoir recours à un détecteur ou à une technique de recadrage pour retrouver chaque caractère

Il est possible d’utiliser un CRNN pour de l’OCR, pour détecter des plaques d’immatriculation ou encore du texte ... tout dépend des données d’entrainements



CNN

Les **CNN** sont une sous-catégorie de réseaux de neurones et sont à ce jour un des modèles de classification d’images réputés être plus performant. Le principal cas d’usage est la **reconnaissance d’image**. Les réseaux convolutifs apprennent plus rapidement et ont un meilleur taux d’erreur. Dans une moindre mesure, on les utilise aussi pour analyse vidéo.

**Mode de fonctionnement :**

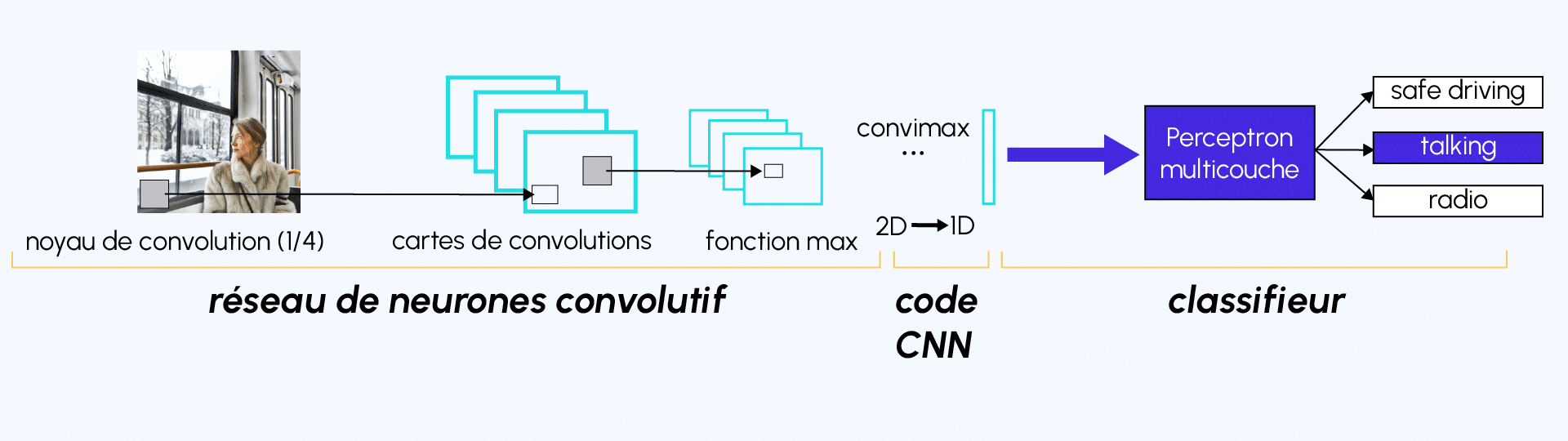
L’utilisateur fournit en entrée une image sous la forme d’une matrice de pixels en 3 dimensions :

Deux dimensions pour une image en niveaux de gris

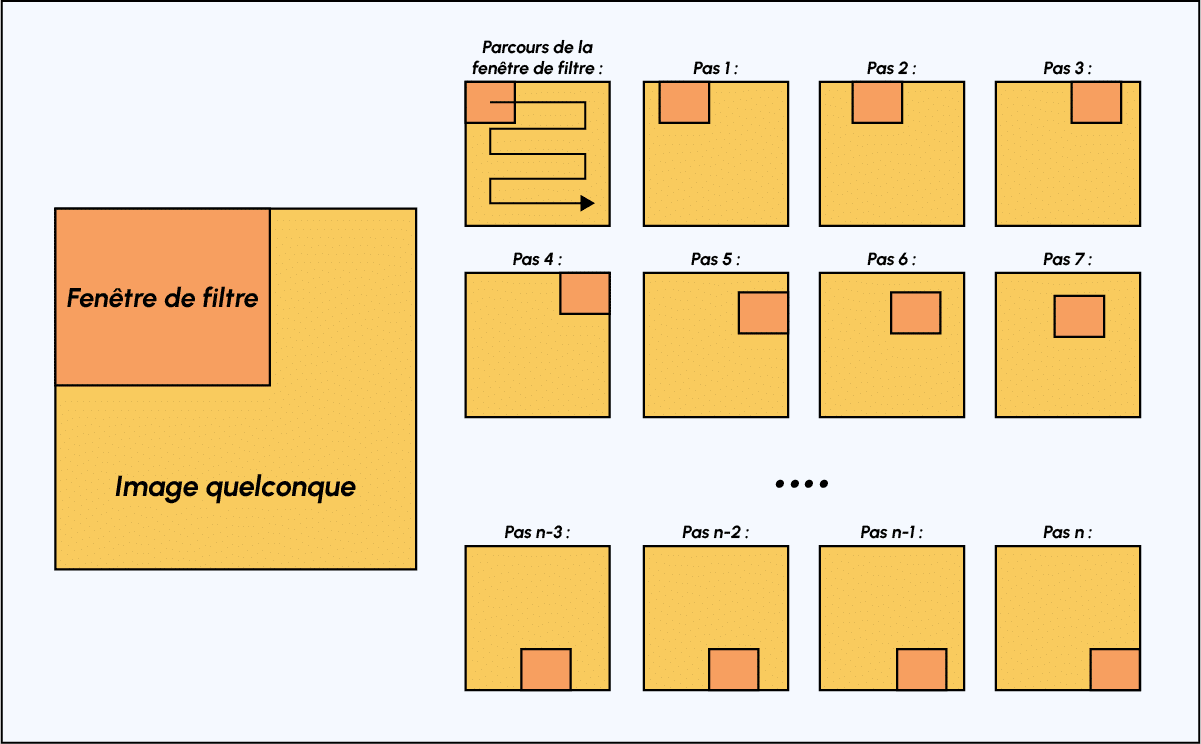
Une troisième dimension, de profondeur pour les couleur fondamentales (RGB)

L’architecture du Convolutional Neural Network dispose en amont d’une partie convolutive et comporte par conséquent deux parties bien distinctes :

* **Une partie convolutive**: Son objectif final est d’**extraire des caractéristiques propres à chaque image** en les **compressant**de façon à **réduire leur taille initiale**. En résumé, l’image fournie en entrée passe à travers une **succession de filtres**, créant par la même occasion de nouvelles images appelées **cartes de convolutions**. Enfin, les **cartes de convolutions obtenues sont concaténées** dans un vecteur de caractéristiques appelé **code CNN**.
* **Une partie classification** : Le **code CNN obtenu en sortie** de la partie convolutive est **fourni en entrée dans une deuxième partie**, constituée de couches entièrement connectées appelées perceptron multicouche (MLP pour Multi Layers Perceptron). Le rôle de cette partie est de **combiner les caractéristiques du code CNN afin de classer l’image**



La première étape est l’étape de convolution, c’est une opération mathématique simple généralement utilisée pour le **traitement et la reconnaissance d’images.** Sur une image, son effet s’assimile à un filtrage dont voici le fonctionnement :



Dans un premier temps, on **définit la taille de la fenêtre de filtre**située en haut à gauche.

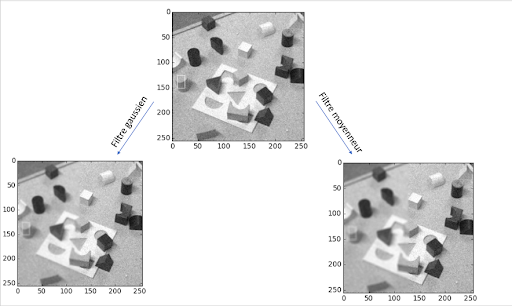
La fenêtre de filtre, représentant la feature, **se déplace progressivement** de la **gauche vers la droite** d’un certain nombre de cases défini au préalable (**le pas**) jusqu’à arriver au bout de l’image.

À chaque portion d’image rencontrée, un calcul de convolution s’effectue permettant d’obtenir en sortie une **carte d’activation**ou **feature map** qui indique où est localisées les features dans l’image : plus la feature map est élevée, plus la portion de l’image balayée ressemble à la feature

Lors de la partie convolutive d’un Convolutional Neural Network, l’image fournie en entrée passe à travers une **succession de filtres de convolution**. Par exemple, il existe des filtres de convolution fréquemment utilisés et permettant d’extraire des caractéristiques plus pertinentes que des pixels comme la détection des bords (**filtre dérivateur**) ou des formes géométriques. Le choix et l’application des filtres se fait **automatiquement** par le modèle.

Parmi les filtres les plus connus, on retrouve notamment le **filtre moyenneur** (calcule pour chaque pixel la moyenne du pixel avec ses 8 proches voisins) ou encore le **filtre gaussien** permettant de réduire le bruit d’une image fournie en entrée :

Voici un exemple des effets de ces deux différents filtres sur une image comportant un bruit important (on peut penser à une photographie prise avec une faible luminosité par exemple). Toutefois, un des inconvénients de la réduction du bruit est qu’elle s’accompagne généralement d’une réduction de la netteté :



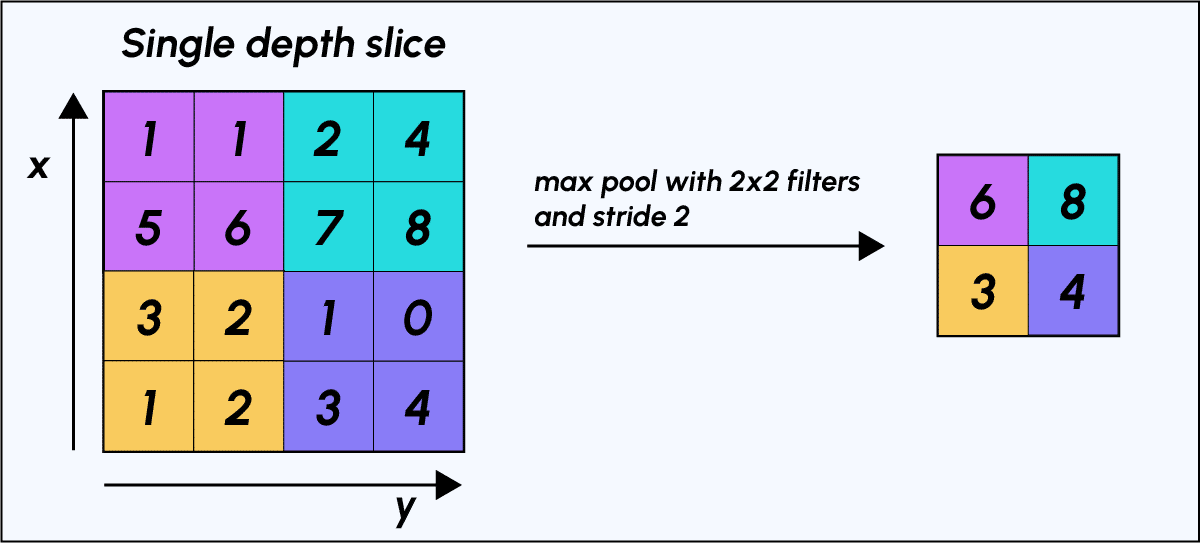
Comme on peut l’observer, contrairement au filtre moyenneur, le filtre gaussien réduit le bruit sans pour autant réduire significativement la netteté.

**MAX-POOLING**

L’intérêt de la **partie convolutive d’un CNN**est qu’elle permet d’**extraire des caractéristiques propres à chaque image**en les compressant de façon à réduire leur taille initiale, via des méthodes de sous-échantillonnage tel que le Max-Pooling.

Le Max-Pooling est un processus de discrétisation basé sur des échantillons. Son objectif est de sous-échantillonner une représentation d’entrée (image, matrice de sortie de couche cachée, etc.) en réduisant sa dimension. De plus, son intérêt est qu’il réduit le coût de calcul en **réduisant le nombre de paramètres à apprendre** et fournit une **invariance par petites translations** (si une petite translation ne modifie pas le maximum de la région balayée, le maximum de chaque région restera le même et donc la nouvelle matrice créée restera identique).

Pour rendre plus concret l’action du Max-Pooling, voici un exemple : imaginons que nous avons une matrice 4×4 représentants notre entrée initiale et un filtre d’une fenêtre de taille 2×2 que nous appliquerons sur notre entrée. Pour chacune des régions balayées par le filtre, le max-pooling prendra le maximum, créant ainsi par la même occasion une nouvelle matrice de sortie où chaque élément correspondra aux maximums de chaque région rencontrée.



L’ajout en amont de la partie convolutive permet d’**obtenir en sortie une « carte de caractéristiques » ou « code CNN » dont les dimensions** sont **plus petites**que celles de l’image initiale ce qui va avoir l’avantage de **diminuer**grandement le nombre de **paramètres à calculer dans le modèle.**

**Paramètre important :**

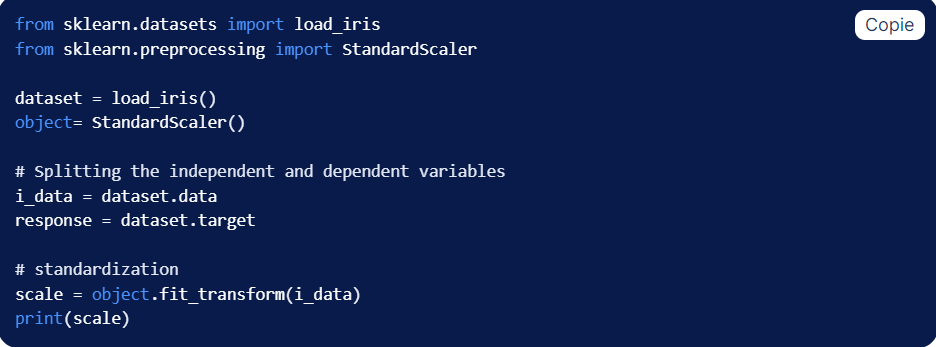
**Padding :** permet de sortir du bord de l’image afin de prendre les pixels extérieurset d’avoir la même taille d’image en sortie.

**Strides :** le pat avec le qu’elle le kernel va se déplacer sur les pixels.

**StandardScaler**

Cette dernière permet de standardiser les valeurs des données dans un format standard (une mise à l’échelle).

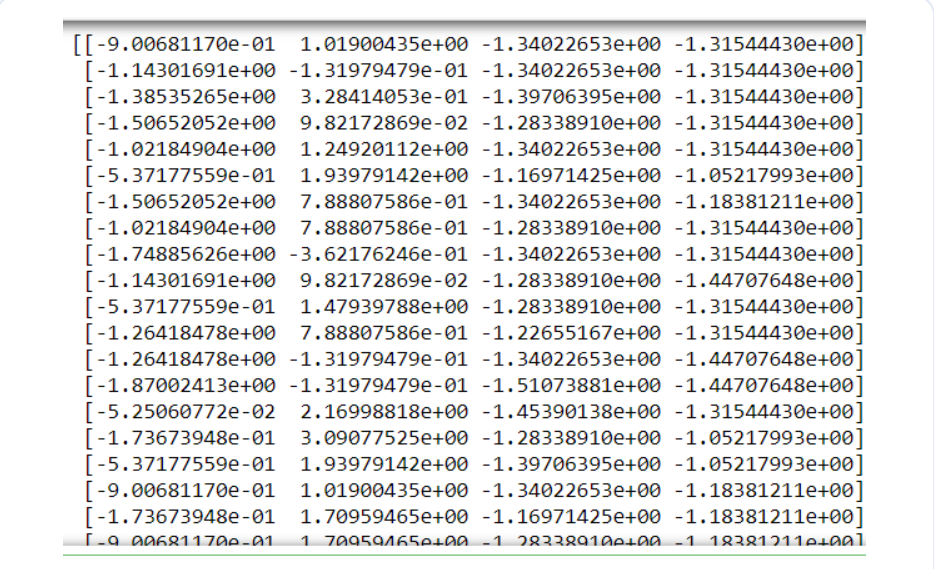
Exemple de code :



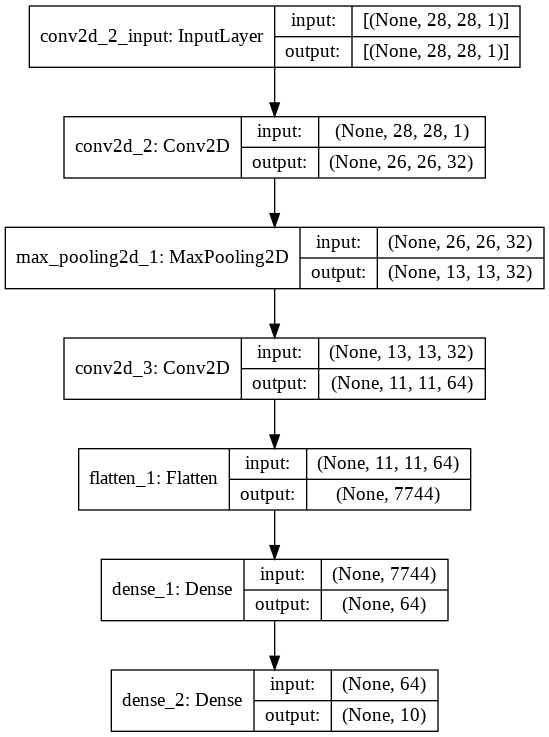
Explication :

1. Importation des bibliothèques ici load\_iris est un jeu de donnée d’images de fleur, l’importation de Standarcaler une méthode pré-processing.
2. On instancie le dataset iris a la variable dataset et instancie un object qui est la méthode standardscaler ()
3. Qu’est-ce que dans le dataset une variable dépendante et indépendante ?
4. On standardise en appliquant la méthode StandardScaler grâce à fit\_transform () sur i\_data

Et voici la sortie :

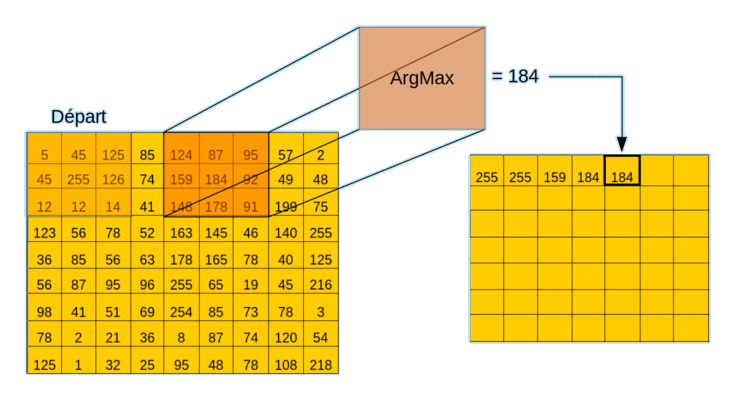


**Evolution des dimensions de l’image**



**Différentes couches**

**Couche de convolution :**



Cette couche permet de récupérer les informations de l’image d’entrée comme la bordure, les trait, les motifs etc...

Elle applique des filtres ou kernel pour effectuer une opération de convolution sur l’image, cette couche produit une nouvelle carte de caractéristiques.

Cette couche a un ensemble de paramètres au niveaux des poids ou paramètres, en voici quelques-uns :

Filtre : nombre d’exécution de l’action si filters = 32 alors l’action sera exécutée 32fois. Soit 32 features map possédant chacune des caractéristiques différentes de l’image.

Taille du filtre : dans notre exemple la taille du filtre est de 3x3

Le pas ou le stride : nombre de pixel ou le filtres se déplace si stride=1 le filtres se déplace de 1pixel à la fois.

Fonction d’activation :

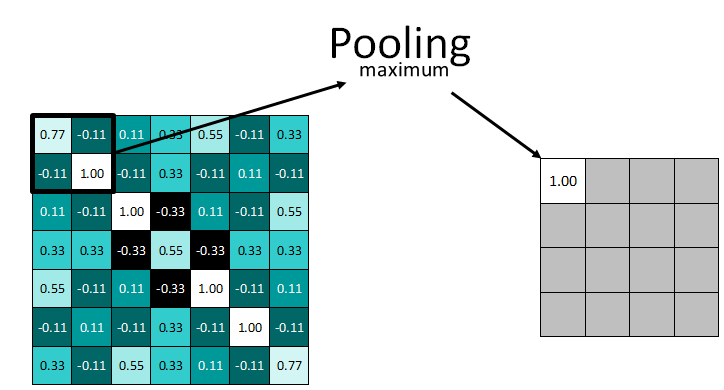
**ReLU** : si la sortie de convolution est -1 elle a transformé en 0 et si c’est 2 elle reste 2.

**Sigmoïde** : si la sortie de convolution est 2 la fonction la transforme en une valeur entre 0 et 1.

**SoftMax** : souvent utiliser dans la couche de sortie pour classifier.

La fonction prend un vecteur de scores en entrée et renvoie un vecteur de probabilité, ou la somme de toutes les probabilités est égale à 1.

**Couche de pooling :**

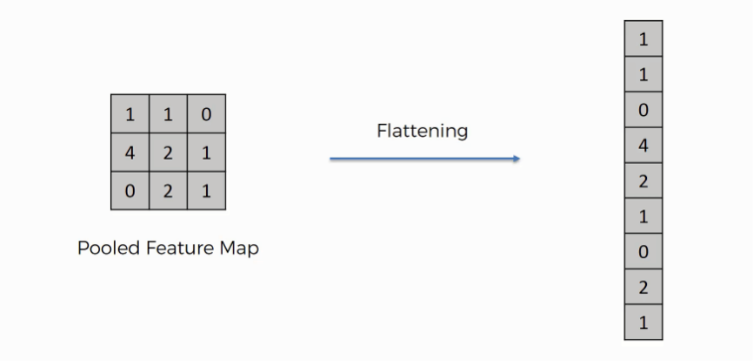


L’objectif de cette couche est de réduire la dimension spatiale (largeur et hauteur) des cartes de caractéristiques extraites par les couches de convolution. Elle garde les caractéristiques les plus important tout en réduisant la quantité de données.

Elle effectue une opération de sous-échantillonnage en prenant la valeur maximal (max pooling) ou la moyenne (average-pooling) d’une région, cela permet de réduire l’image dans perdre les caractéristiques importantes.

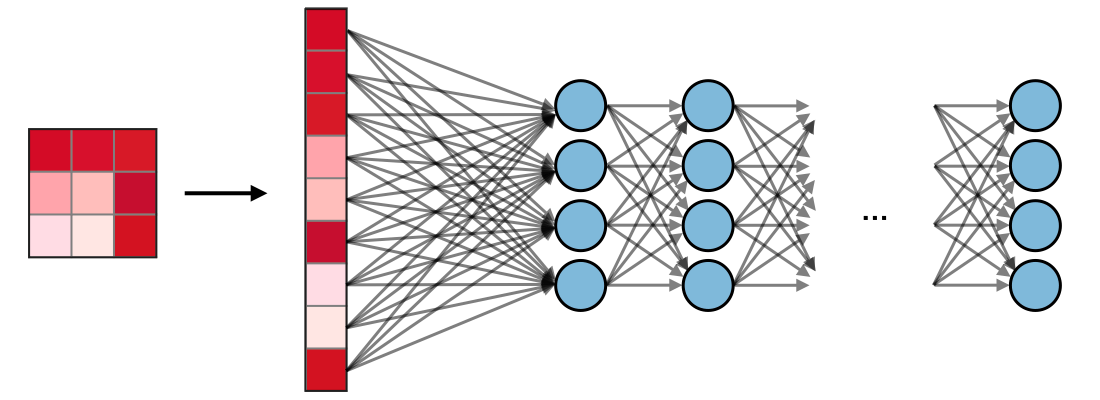
Elle n’a pas réellement de paramètres mise à part la taille du kernel.

**Couche flatten :**



Permet d’aplatir le tenseur, de réduire sa dimension. Elle prend en entrée un 3D-tensor et retourne un 1D-tensor, cette opération est essentielle car elle permet d’établir une connexion entres les couches de base du deep learning.

**Couche fully-connected (Dense) :**



Cette couche a pour but d’effectuer une classification en utilisant les caractéristiques extrait des couches de convolution précédentes.

La couche fully connected est responsable de la prise de décision finale dans le réseau. Elle combine les informations extraites des couches précédentes pour produire une sortie qui est utilisée pour la classification.

Les CNNs sont généralement formés de plusieurs couches de convolution et de pooling, suivies par une couche fully-connected qui combine les features extraites par les couches précédentes pour classifier l'image, elle renvoie un vecteur de taille N, où N est **le nombre de classes dans notre problème de classification** d'images. Chaque élément du vecteur indique la probabilité pour l'image en entrée d'appartenir à une classe.

Au niveau de la classification, cette couche est souvent associée à une fonction d’activation softmax qui attribue des probabilités à une classe. Ainsi elle permet de déterminer à quelle classe appartient l’image.

Cette couche admet certains paramètres :

**Le nombre de neurones :** c’est le nombre de neurones dans la couche fully-connected, chaque neurone est connecté à tous les neurones de la couche précédentes.

**Fonction d’activation :** comme au-dessus ReLU, sigmoid…

**Biais(bias) :** chaque neurone de la couche a un biais qui permettent au réseau d’apprendre des décalages et des constantes dans les données.

**Fonction de perte :** ou aussi appeler loss function mesure l’erreur entre la sortie du réseau et la valeur cible (exemple dans le cas de la classification, la vraie classe de l’image). La fonction de perte est utilisée pour ajuster les poids et les biais du réseau pendant l’entrainement.

**Couche de Normalisation (Batch Normalization) :** Cette couche est utilisée pour normaliser les activations des neurones dans une couche, ce qui peut améliorer la stabilité et la vitesse de convergence du modèle.

**Couche de Dénormalisation (Instance Normalization) :** Elle normalise les activations de chaque exemple (instance) plutôt que sur l'ensemble des données d'entraînement. Elle est souvent utilisée dans des applications telles que la génération d'images.

**Couche de Goutteout (Dropout) :** La couche de dropout désactive aléatoirement un pourcentage des neurones lors de l'entraînement pour prévenir le surajustement (overfitting).

**Couche de Concaténation (Concatenation)** : Cette couche combine les sorties de plusieurs couches pour former une seule couche fusionnée. Elle est utile lorsque vous avez plusieurs sources d'informations à combiner.

**Couche de Global Average Pooling (GAP) :** Plutôt que d'utiliser une couche fully connected traditionnelle pour réduire les caractéristiques en un vecteur, la couche GAP calcule la moyenne des caractéristiques spatiales, réduisant ainsi la dimension sans perte d'informations importantes.

**Couche de Remplissage (Padding)** : Les couches de remplissage ajoutent des zéros ou d'autres valeurs autour des bords des images ou des caractéristiques pour maintenir la taille spatiale après la convolution.

**Couche de Convolution Transposée (Transposed Convolution ou Deconvolution) :** Cette couche est utilisée dans les réseaux génératifs, comme les GAN, pour augmenter la résolution spatiale des caractéristiques.

**Couche d'Unpooling :** Cette couche est parfois utilisée pour inverser l'opération de pooling et restaurer la résolution spatiale.

**Couche de Réduction Dimensionnelle (Dimensionality Reduction)** : Les couches telles que la couche de réduction dimensionnelle (PCA, Principal Component Analysis) sont utilisées pour réduire la dimensionnalité des caractéristiques.

**Couche d'Embedding (Incorporation) :** Cette couche est couramment utilisée dans le traitement du langage naturel pour convertir des mots en vecteurs numériques.

**Vocabulaire**

1. Neurone (ou unité) : L'unité de base d'un réseau de neurones artificiels, qui effectue des opérations mathématiques sur ses entrées et produit une sortie.

(rajouter image)

2. Couche (ou layer) : Une couche est un groupe de neurones qui effectuent des opérations en parallèle. On distingue généralement les couches d'entrée, les couches cachées et les couches de sortie.

3. Réseau de neurones (ou neural network) : Un réseau composé de couches de neurones interconnectées qui sont utilisées pour effectuer des tâches d'apprentissage automatique.

4. Apprentissage automatique (ou machine learning) : Une approche informatique qui permet aux systèmes d'apprendre à partir de données et d'améliorer leur performance avec l'expérience.

5. Apprentissage profond (ou deep learning) : Une sous-discipline de l'apprentissage automatique qui implique des réseaux de neurones avec plusieurs couches (d'où le terme "profond").

6. Fonction d'activation : Une fonction mathématique appliquée à la sortie d'un neurone pour introduire de la non-linéarité dans le réseau, généralement ReLU (Rectified Linear Unit) ou sigmoïde.

7. Propagation avant (ou forward propagation) : Le processus de calcul de la sortie d'un réseau neuronal pour une entrée donnée.

8. Propagation arrière (ou backpropagation) : Le processus de mise à jour des poids des neurones dans un réseau en fonction de l'erreur commise lors de la propagation avant, pour permettre l'apprentissage.

9. Poids (ou weights) : Les coefficients qui déterminent l'importance des connexions entre les neurones dans un réseau.

10. Fonction de coût (ou loss function) : Une mesure qui quantifie l'écart entre les prédictions du réseau et les valeurs attendues, utilisée pour guider l'apprentissage.

11. Optimisation : Le processus d'ajustement des poids du réseau pour minimiser la fonction de coût.

12. Époch (ou epoch) : Une passe complète à travers l'ensemble de données d'apprentissage lors de l'entraînement du réseau.(par défault)

13. Mini-lot (ou mini-batch) : Un sous-ensemble des données d'apprentissage utilisé pour mettre à jour les poids du réseau de manière plus efficace. (par défault)

14. Surapprentissage (ou overfitting) : Un phénomène où le réseau apprend trop bien les données d'entraînement, ce qui le rend moins capable de généraliser sur de nouvelles données.

15. Taux d'apprentissage (ou learning rate) : Un paramètre qui contrôle la taille des pas effectués lors de l'optimisation des poids du réseau.(inclut dans la fonction de poids)

16. Réseau de neurones convolutifs (CNN) : Un type de réseau de neurones spécialement conçu pour le traitement d'images, utilisant des couches de convolution.

17. Réseau de neurones récurrents (RNN) : Un type de réseau de neurones conçu pour traiter des données séquentielles, en maintenant une mémoire interne.

18. Réseau de neurones artificiels (ANN) : Un terme générique pour tout réseau de neurones non spécialisé, généralement avec des couches denses.

19. Tensorflow et PyTorch : Deux des cadres de développement populaires pour la création de réseaux de neurones.

20. Prétraitement de données : Les techniques utilisées pour préparer les données avant de les utiliser pour l'entraînement d'un réseau, y compris la normalisation, la réduction de dimension, etc.

Dataset : <https://adamharley.com/rvl-cdip/>

Plusieurs liens pour dataset : https://www.stat4decision.com/fr/10-sites-pour-trouver-des-donnees-pour-modeles/

Lien : <https://datascientest.com/convolutional-neural-network#:~:text=Les%20CNN%20d%C3%A9signent%20une%20sous,d'une%20matrice%20de%20pixels>.

<https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python>

<https://datascientest.com/tout-savoir-sur-scikit-learn>

https://www.youtube.com/watch?v=581X9wsnWJs

Recherche sur comment on peut avoir les méthodes de fraudes à l’assurance :

Ce lien emmène vers une mutuelle "M comme mutuelle" et eux on, sur 756 dossiers investigués ils ont 286 fraudes avérées et si on essayait de les contacter ?

<https://www.mcommemutuelle.com/nos-offres/>

Se rendre sur le darkweb et accéder à un réseau de fraudeur.

Interroger un fraudeur mais aussi une personne qui s’est fait frauder.

Se rendre sur un réseau télégramme.

Pas mal de source devrais êtres trouver sur X aussi

<https://twitter.com/CharlesPrats/status/1554709763363147776/photo/3>

<https://www.fdesouche.com/2022/08/03/le-senat-a-vote-20-millions-de-credits-pour-lancer-la-mise-en-place-dune-carte-vitale-biometrique-afin-de-lutter-contre-la-fraude-qui-atteindrait-6-milliards-deuros-sel/>

Une personne qui profite de l’aah (aide handicapé) sans rien faire et donc perçois des sous, il a manipulé son médecin.

https://twitter.com/CharlesPrats/status/1703804081737703485

Zip toutes les données que j’ai conçues + script pour les envoyer par mail

Remplacer l’entrainement du texte blanc sur fond blanc par les factures avec du texte blanc sur fond blanc

Faire recherche sur la détection de flou

Recherches sur la détection de la provenance d’un document

Ajouter deux autres critères sur l’API numéro finess, adhérant suspicieux.

Détection provenance d’un document PDF

**Qu’est-ce qu’un fichier PDF ?**

PDF est un format utilisé pour afficher des documents dans un format électronique indépendant du logiciel, matériel ou système d’exploitation dans lequel ils sont visualisé. Développé initialement par Adobe comme un format de fichier universellement compatible basé sur PostScript, le Portable Document Format devenu de facto un standard international pour l'échange de documents et d'informations.

**Normes des fichier PDF**

Ce qu’il faut comprendre avec les fichiers PDF, c’est que chaque nouvelle version n’élimine pas les précédentes. Chaque nouvelle version accroit les capacités au format, mais cela ne signifie pas que toutes doivent être utilisées lors de la création d’un PDF.

En voici quelques-unes :

**PDF/UA**

UA signifie « Universal Accessibility » (accessibilité universelle) et le PDF/UA est une spécification qui définit comment rendre un PDF lisible par des technologies d’assistance (logiciels, voire appareils spéciaux) afin qu’un ordinateur puisse lire à voix haute le contenu d’un tel document à quiconque dépendant de ces technologies. Alors que les PDF deviennent si courants dans notre vie, en particulier dans des domaines tels que les services publics, la banque, les services à la personne, le recrutement, le médical et bien d’autres types de services, l’éducation, etc., garantir une accessibilité à ces documents, facile et égale pour tous, est vraiment essentiel.

Un PDF/UA a une structure logique clairement et correctement définie, bien décrite. Grâce à la description de cette structure, les technologies d’assistance savent et sont capables de dire quel est le titre du document, dans quel ordre lire les paragraphes du document et les colonnes de texte, quelles sont les listes, où sont les images et ce qu’elles décrivent, ignorer les titres répétitifs et pieds de page contenant les numéros de page, etc.

**PDF/A**

Le format PDF a été conçu comme un format permettant l’interopérabilité entre différents logiciels, ordinateurs et plateformes. Le format PDF/A reprend cette idée en l’étalant dans le temps : il garantit que les documents PDF pourront être ouverts à l’avenir.

PDF/A est un format PDF conçu pour l’archivage, la conservation à long terme et l’échange de documents électroniques. L’aspect visuel des documents électroniques est préservé au fil du temps, quels que soient les outils et systèmes utilisés pour leur production, stockage et reproduction. Les documents source peuvent être des documents papier, des emails, des documents PDF « habituels », des pages-web, et bien d’autres. Le format PDF/A offre un cliché numérique fiable de tout document, ce dernier restant interrogeable et exploitable pour :

* La gestion des documents
* Les flux automatisés
* La gestion des dossiers juridiques
* L’archivage du courrier : papier, fax, email
* L’archivage de documents, la migration d’archives (e-gouvernementales, légales, etc.)

En raison de sa capacité à fournir une représentation stable et uniforme des documents au fil du temps et à travers plusieurs plateformes, le format PDF/A est également utilisé comme le format de la gestion des documents et de l’échange continu de documents. Il s’agit d’un « papier numérique », tout aussi fiable, inaltéré et permanent que les documents papier auxquels nous sommes habitués à faire confiance.

**L’outil OCR ABBYY FineReader :**

ABBYY FineReader est un système de reconnaissance optique de caractères (OCR). Il permet de convertir des documents numérisés, des documents PDF et des fichiers d'images, y compris les photos numériques, en formats modifiables.

**Avantages d'ABBYY FineReader**

• Le système OCR utilisé par ABBYY FineReader offre aux utilisateurs une reconnaissance rapide et précise et préserve le format source de tout document (y compris le texte sur images d'arrière-plan, le texte en couleur sur des arrières plans colorés et le texte autour d'une image, etc.).

• Grâce à la technologie de reconnaissance de document adaptative d'ABBYY (ADRT®), ABBYY FineReader peut analyser et traiter un document en entier, au lieu de procéder page par page. Cette approche conserve la structure du document source, y compris le format, les hyperliens, les courriels, les en–têtes et les bas de page, les légendes d'images et de tableaux, les numéros de page et les annotations en bas de page.

• ABBYY FineReader peut reconnaître des documents écrits dans une ou plusieurs des 189 langues prises en charge, y compris l'arabe, le vietnamien, le coréen, le chinois, le japonais, le thaï et l'hébreu. ABBYY FineReader propose aussi une détection automatique des langues du document.

• De plus, ABBYY FineReader est quasiment infaillible quant aux défauts d'impression et peut reconnaître des textes imprimés dans pratiquement toutes les polices. • Ce programme contient également une large gamme d'options pour les données de sortie : les documents peuvent être enregistrés sous une grande variété de formats, envoyés par courriel ou transférés vers d'autres applications pour un traitement complémentaire.

**Nouveautés d'ABBYY FineReader 11**

• Augmentation de l'efficacité du traitement Grâce au mode Noir et blanc, FineReader 11 propose un traitement de document plus rapide de 30% lorsque vous n'avez pas besoin de couleur. De plus, le programme exploite efficacement les processeurs multicœurs permettant ainsi une conversion encore plus rapide.

• Création flexible de livres électroniques Numérisez des livres puis convertissez–les aux formats EPUB et FB2 pour les lire en déplacement sur votre iPad, sur votre tablette ou sur votre appareil portable préféré. Vous pouvez aussi les envoyer directement vers votre compte Kindle. Convertissez des livres ou des articles papier aux formats de livres électroniques appropriés pour les ajouter à votre bibliothèque électronique ou à votre archive.

• Support natif pour OpenOffice.org Writer FineReader 11 reconnaît et convertit les images de documents et les fichiers PDF directement au format OpenOffice.org Writer (ODT). Il préserve la présentation et le formatage d'origine avec précision. À présent, vous pouvez facilement ajouter des documents à votre archive \*.odt en quelques clics.

• Interface utilisateur améliorée

• L'éditeur de styles amélioré vous permet de configurer les paramètres de style dans une seule fenêtre conviviale. Toutes les modifications sont appliquées à l'intégralité du document et ce, en une fois.

• Organisez les pages des documents FineReader pour préserver au mieux la présentation.

• Lancez la conversion de documents dès que le programme démarre grâce à un accès encore plus facile à toutes les tâches de conversion courantes ou avancées.

• OCR pour appareil photo nouvelle génération avec un large éventail d'outils efficaces d'édition d'images FineReader 11 offre une gamme complète de nouveaux outils d'édition d'images efficaces, notamment les curseurs de luminosité et de contraste et les outils de niveau vous permettant d'obtenir des résultats plus précis et d'améliorer les paramètres de l'image.

• Amélioration de la précision de l'OCR et de la conservation de la présentation grâce à la technologie ADRT 2.0 Minimisez le temps nécessaire à l'édition des documents convertis grâce à une meilleure détection des styles, des notes en bas de page, des en–têtes et bas de page des documents.

• Sortie optimale de PDF Les trois paramètres d'images prédéfinis pour les fichiers PDF fournissent des résultats optimaux selon vos besoins : mode Meilleure qualité, Taille compacte ou Équilibré.

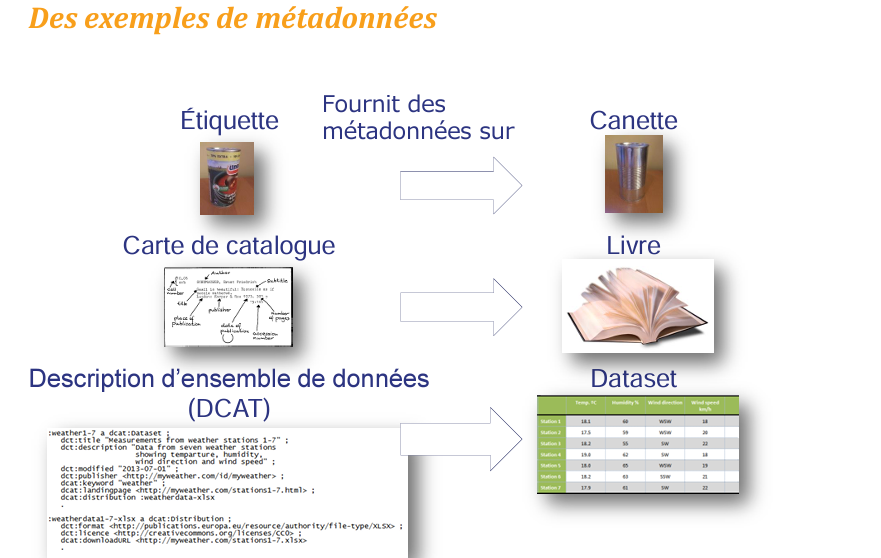
• Nouvelles langues de reconnaissance\* – Arabe, Vietnamien, et Turkmène (alphabet latin).

**Métadonnées d’un document :**

Les métadonnées d’un document sont des informations non visuelles contenues dans un document qui fournissent un contexte supplémentaire. Par exemple l’auteur du document et la date a laquelle il a été créé. Elles peuvent également aider a classer les documents ex. si le document doit être disponible en interne ou au grand public.

Il existe différents types de métadonnées :

* Métadonnées descriptives, décrivent une ressource à des fins de découverte et d’identification.
* Métadonnées structurelles, par ex. des modèles de données et des données de référence.
* Métadonnées administratives, renseignent afin d’aider à la gestion d’une ressource.



**Détection de flou**

**Méthodes:**

1. **Méthodes de Défloutage :** Explorez les différentes méthodes et algorithmes de défloutage d'images. Ces méthodes sont conçues pour restaurer la netteté des images qui ont été affectées par le flou.
2. **Caractéristiques du Flou :** Comprendre les caractéristiques du flou dans les images. Il peut s'agir de flous de mouvement, de flous optiques, ou d'autres types de flou. Chaque type de flou peut nécessiter une approche spécifique pour la détection et la correction.
3. **Traitement du Signal :** Étant donné que le flou peut être considéré comme une altération du signal dans le domaine spatial, explorez les techniques de traitement du signal qui peuvent être appliquées à la détection du flou.
4. **Apprentissage Automatique :** Des approches basées sur l'apprentissage automatique, telles que les réseaux de neurones convolutifs, sont également utilisées pour la détection du flou. Explorez les modèles qui ont été développés pour cette tâche.
5. **Évaluation de la Qualité de l'Image :** Les mesures de qualité d'image peuvent être utilisées pour évaluer la netteté d'une image. Comprenez les métriques couramment utilisées pour évaluer la qualité des images.
6. **Applications Pratiques :** Explorez les applications pratiques de la détection du flou, telles que la photographie, la vision par ordinateur, la télédétection, etc.

**Méthode de défloutage :**

**Caractéristiques du Flou :**

Pour commencer il existe différents types de flou un lien est disponible pour avoir des exemples et comment éviter si cela est possible : https://initiationphoto.com/les-differents-types-de-flou-et-comment-les-eviter/

**Traitement du Signal :**

**Apprentissage Automatique :**

En ce qui concerne l'utilisation de l'apprentissage automatique, notamment les réseaux de neurones convolutifs (CNN), pour la détection du flou, voici quelques points supplémentaires à considérer :

1. Réseaux de Défloutage : Certains modèles de CNN sont spécifiquement conçus pour le défloutage d'images. Ces modèles apprennent à estimer et à compenser le flou présent dans une image. Ils peuvent être entraînés sur des paires d'images floues et nettes pour apprendre la relation entre les deux.
2. Base de Données d'Entraînement : La qualité des résultats dépend souvent de la base de données sur laquelle le modèle est entraîné. Des bases de données contenant des images avec différents types de flou (mouvement, optique, etc.) sont utilisées pour garantir que le modèle est capable de traiter divers scénarios.

Dataset kaggle : https://www.kaggle.com/datasets/kwentar/blur-dataset

1. Architecture du Modèle : L'architecture du CNN est cruciale. Certains modèles utilisent des architectures profondes pour capturer des caractéristiques complexes du flou, tandis que d'autres peuvent être plus légers pour des applications en temps réel.
2. Fonctions de Perte : Les fonctions de perte dans l'apprentissage supervisé sont essentielles. Elles quantifient la différence entre la sortie du modèle et la vérité terrain (image nette). Des fonctions de perte adaptées sont nécessaires pour garantir que le modèle apprend correctement.
3. Transfert d'Apprentissage : Certains modèles peuvent bénéficier du transfert d'apprentissage en utilisant des pré-entraînements sur de grandes bases de données d'images avant de se spécialiser dans la détection du flou.
4. Évaluation des Performances : Les performances du modèle doivent être évaluées sur des ensembles de données de test distincts pour s'assurer de sa généralisation à de nouvelles données. Des mesures telles que la PSNR (Peak Signal-to-Noise Ratio) et le SSIM (Structural Similarity Index) sont couramment utilisées pour évaluer la qualité des images défloutées.
5. Applications : Les modèles de détection du flou basés sur l'apprentissage automatique sont largement utilisés dans divers domaines tels que la photographie numérique, la vidéographie, la surveillance vidéo, et d'autres domaines où la qualité visuelle est cruciale.

**Évaluation de la Qualité de l'Image :**

Il existe deux méthodes pour évaluer la qualité d’une image il y’a PSNR et l’erreur quadratique moyenne EQM

**PSNR** (sigle de Peak Signal to Noise Ratio) est une mesure de distorsion utilisée en image numérique, tout particulièrement en compression d'image. Elle permet de quantifier la performance des codeurs en mesurant la qualité de reconstruction de l'image compressée par rapport à l'image originale.

Il est aussi possible d’évaluer la qualité d’une image en connaissant le nombre de pixel présent sur une image, plus de pixels sont présent sur l’image plus l’image est nette, bien-sur ce n’est pas le seul paramètre possible il faut prendre aussi en compte la clarté, la netteté…, mais pour les pixels nous pourrions se fixer quelques paramètres généraux :

1. **Très Basse Qualité :** Moins de 100 000 pixels. Ces images peuvent paraître floues ou pixélisées, en particulier lorsqu'elles sont agrandies.
2. **Basse Qualité :** Entre 100 000 et 500 000 pixels. Bien que ces images puissent être acceptables pour des utilisations en ligne ou sur des appareils mobiles, elles peuvent ne pas convenir à des impressions de grande taille.
3. **Qualité Standard :** Entre 500 000 et 1 million de pixels. Cela convient généralement pour la plupart des utilisations courantes, y compris l'affichage en ligne, les impressions de taille modérée, etc.
4. **Haute Qualité :** Entre 1 million et plusieurs millions de pixels. Ces images sont adaptées à des impressions de grande taille ou à des utilisations professionnelles nécessitant une qualité exceptionnelle.
5. **Très Haute Qualité :** Plusieurs millions de pixels et au-delà. Ce niveau de résolution est généralement utilisé dans des domaines tels que la photographie professionnelle, la conception graphique avancée, etc.