Atelier « Utilisation de l'IA pour les corpus iconographiques »

Journée d'étude "Ce que l'intelligence artificielle change à l'université"
1er février 2024

Bibliothèque universitaire de Nantes

Emmanuelle Bermès (Ecole nationale des Chartes)
Jean-Philippe Moreux (BnF)

Plan

Introduction (10 mn)

Enjeux de l'analyse d'image (*computer vision)* dans les bibliothèques, pour les collections iconographiques : principalement la recherche par le contenu. Différentes approches possibles : contenu visuel, contenu sémantique, similarité. Aujourd'hui, nous vous proposons de découvrir quatre types de traitement technique considérés dans le contexte du patrimoine :

- La classification
- La détection d'objet
- La segmentation de documents
- La similarité visuelle

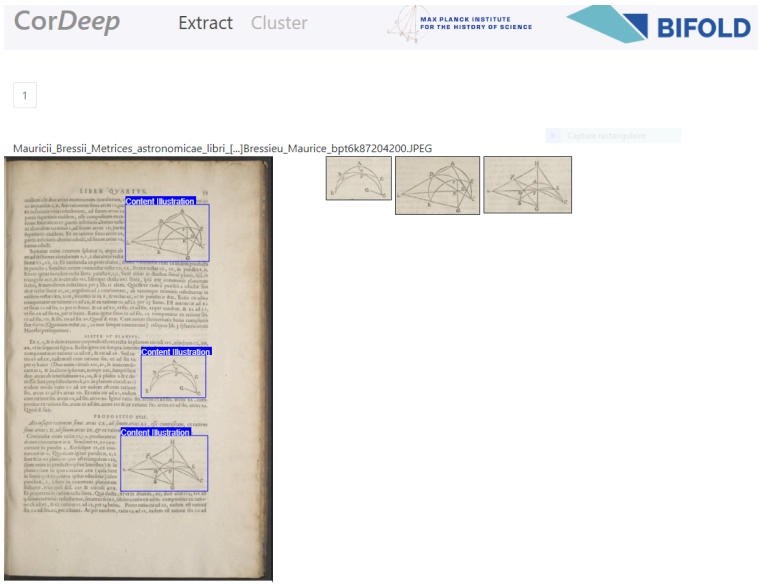
Rapide tour de table : d'où viennent les participants ? quel est leur niveau de connaissance de l'IA ; de la CV ? Qu'attendent-ils de l'atelier ?

Démonstration introductive d'une IA de détection d'illustrations dans des imprimés (15e-17e siècles)

- Tester l'outil CorDeep sur un imprimé du 16e s.
    - https://cordeep.mpiwg-berlin.mpg.de/
    - Choisir une image de document patrimonial :

https://gallica.bnf.fr/ark:/12148/bpt6k87204200/f1.planchecontact
https://gallica.bnf.fr/ark:/12148/bpt6k87204200/f44.medres
https://gallica.bnf.fr/ark:/12148/bpt6k87204200/f129.medres

- Ouvrir l'image dans CorDeep



- … ou sur un manuscrit enluminé (14e siècle) :

https://gallica.bnf.fr/ark:/12148/btv1b84497026/f233.medres
https://gallica.bnf.fr/ark:/12148/btv1b84497026/f10.medres

Bilan : que s'est-il passé ?

1. Classification d'images (20 mn)

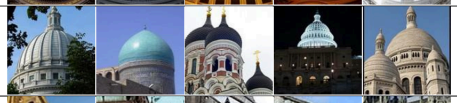
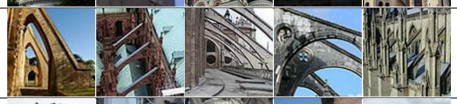
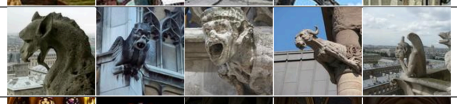
Objectif : *classer des images selon leur contenu*

À noter :

- *On ne cherche pas à localiser le contenu dans l'image.*
- *La classification peut être :*
  - binaire *: une image ne peut appartenir qu'à une classe parmi deux possibles*
  - multiclasse *: on cherche à ranger l'image dans une classe parmi* n
  - multilabel *: une image peut se voir associer plusieurs classes*

Applications :

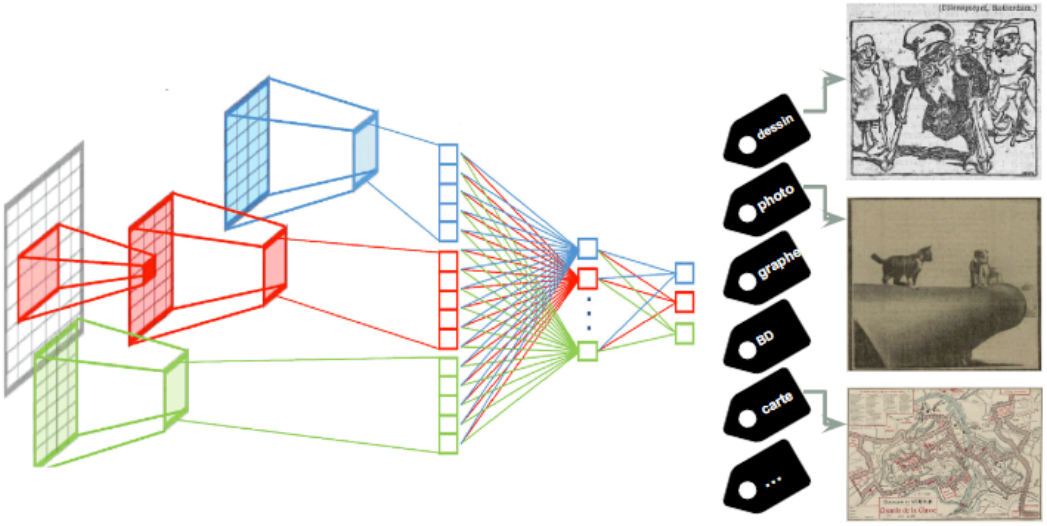
- *Classer des illustrations selon leur contenu (principal)*

Category	Examples
Altar (829 images)	
Apse (514 images)	
Bell tower (1059 images)	
Column (1919 images)	
Dome (inner) (616 images)	
Dome (outer) (1177 images)	
Flying buttress (407 images)	
Gargoyle (and Chimera) (1571 images)	
Stained glass (1033 images)	
Vault (1110 images)	

Classification of Architectural Heritage Images Using Deep Learning Techniques
J. Llamas, P. Lerones, +2 authors Jaime Gómez-García-Bermejo, 26 September 2017

GallicaPix : motifs de papiers-peints avec des lignes

- *Classer des types d'illustration (technique de production, fonction, genre…)*



GallicaPix : classer des types d'illustrations selon leur technique, leur fonction

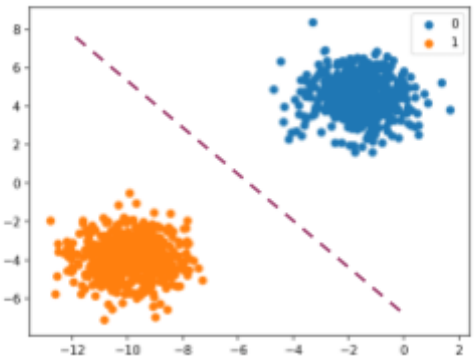
Résultats **:** une "étiquette" ("label") par image (plusieurs dans le cas multilabel). Ces données peuvent nourrir un catalogue, un portail documentaire.

Approches

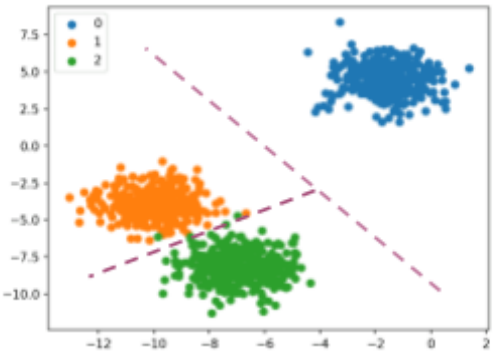
Supervisées

- On calcule les caractéristiques visuelles de chaque image d'un jeu d'images "annotées" (rangées manuellement par classe) :
    - approche "classiques" (abandonnées) : par exemple avec des algorithmes de détection de contour, de texture… (OpenCV).

    - ou avec des réseaux de neurones artificiels.
- On entraîne un modèle à séparer les classes d'image à partir des caractéristiques calculées (*embeddings*) :



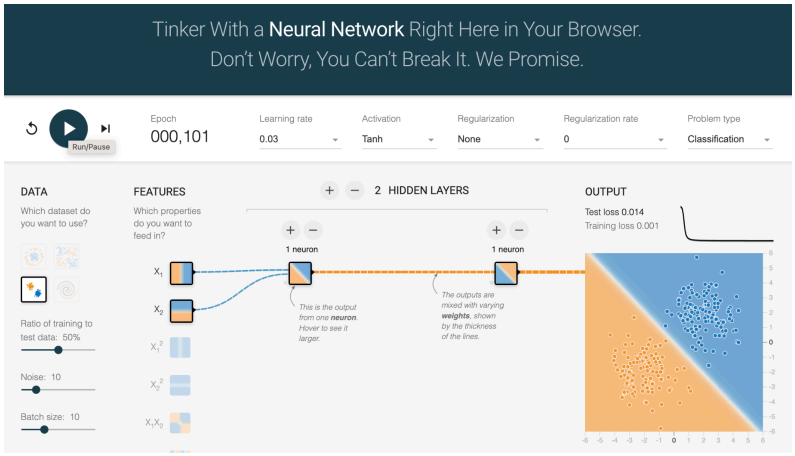
Classification binaire (ex. : chat-chien)



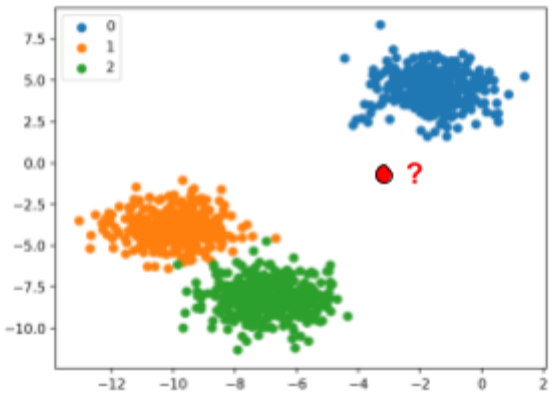
Classification multiclasse (chat-chien-plante)

playground.tensorflow.org

- Choisir un jeu de données gaussien avec du bruit
- Un modèle à 2 neurones
- Entraîner le modèle

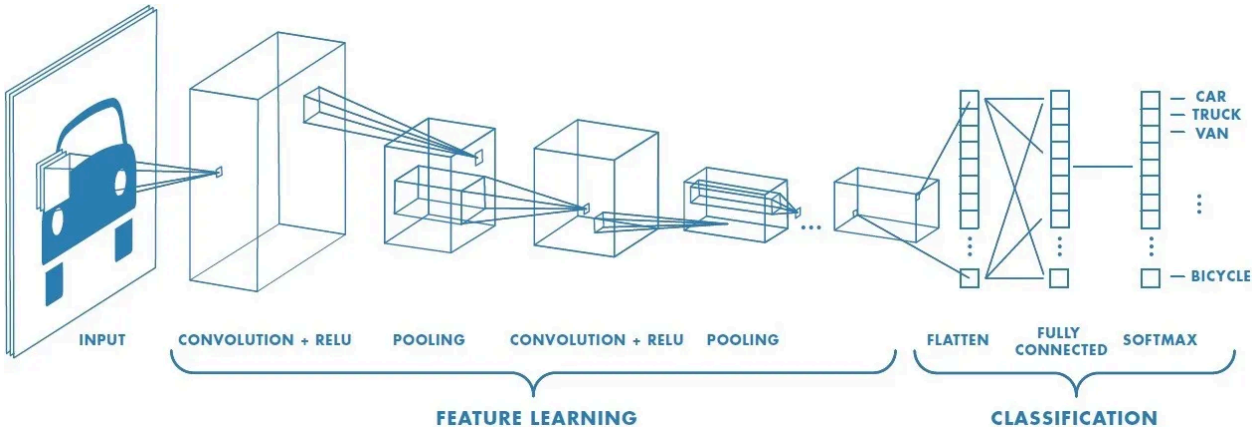


- On applique ensuite ce modèle aux images que l'on veut classifier :



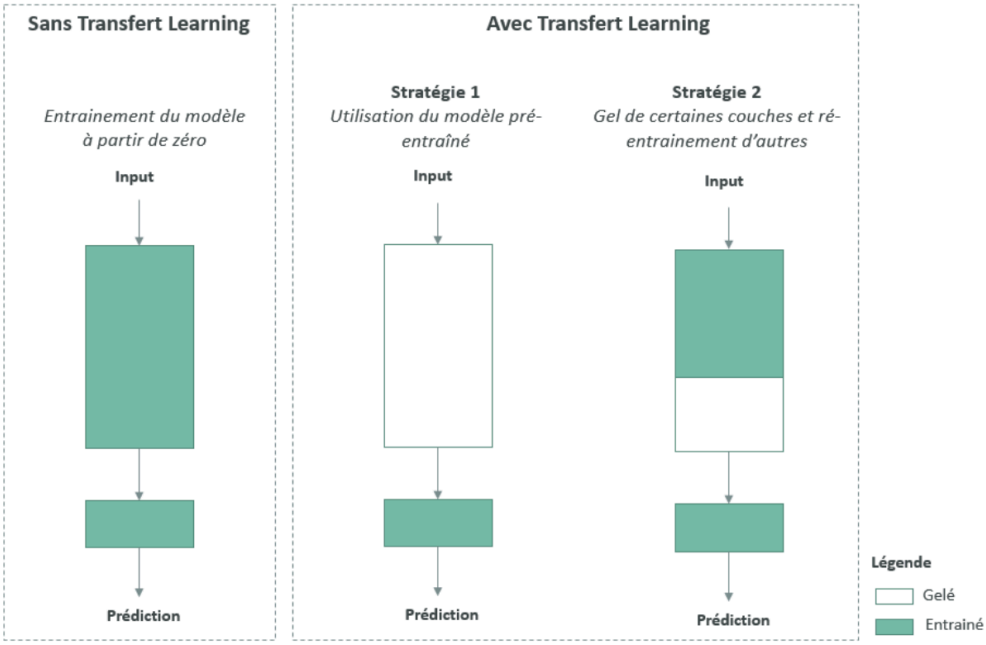
Avantage des modèles neuronaux : le modèle apprend seul les caractéristiques qui vont permettre de bien distinguer les classes visuelles, à l'aide d'exemples.

En pratique, des architectures CNN (réseaux neuronaux convolutifs) sont utilisées : une fenêtre glissante parcourt les pixels de l'image, extrait et synthétise les caractéristiques visuelles pertinentes. La dernière couche de neurones prédit la classe.



Approfondissement hors atelier
Démo CNN : entraînement d'un réseau de neurones / Démo CNN *:* détail du fonctionnement de l'architecture d'un réseau de neurones

Ces architectures sont entraînées sur de grands corpus d'apprentissage (par ex. ImageNet), et les modèles résultants sont soit appliqués tel que, soit adaptés à des cas spécifiques par apprentissage par transfert (*transfert learning*) :



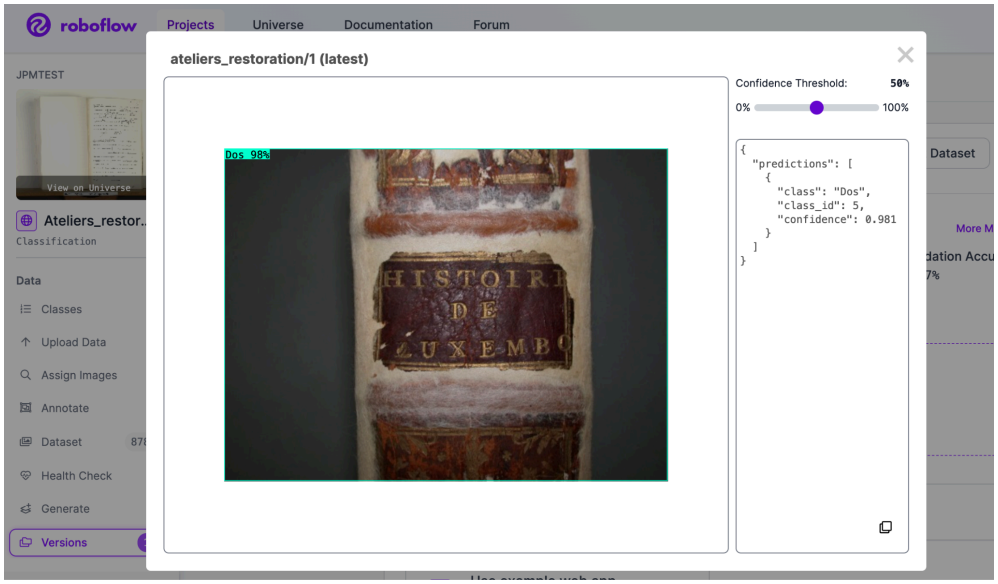
Gallica Pix (BnF, 2017) : classification de types d'image (modèle CNN Inception, apprentissage par transfert)

- https://gallicapix.bnf.fr/
- Choisir le corpus par défaut 14-18
- Choisir un critère Technique ou Fonction
- Saisir un mot-clé et lancer la recherche

Approfondissement hors atelier

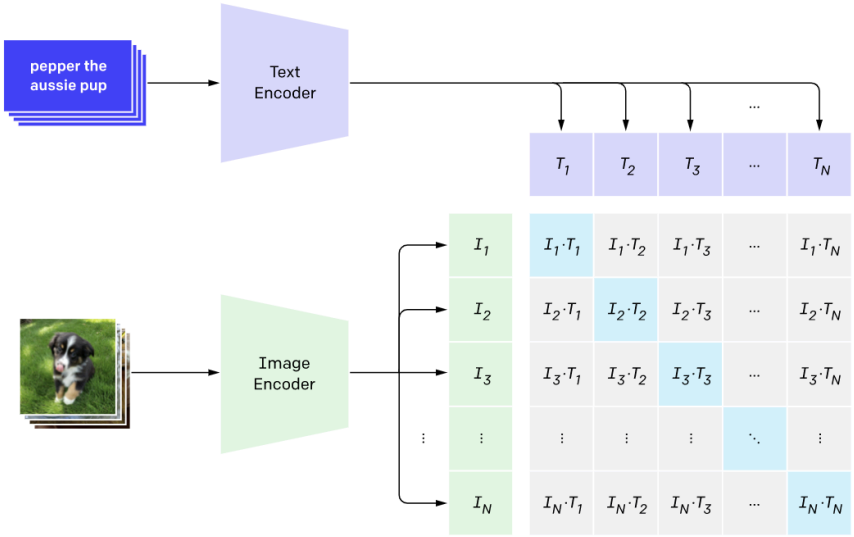
Roboflow (plateforme IA en ligne) : démonstration d'apprentissage par transfert :

- Télécharger un jeu d'images (un dossier par classe)
- Entraîner un modèle de classification
- Tester avec une image

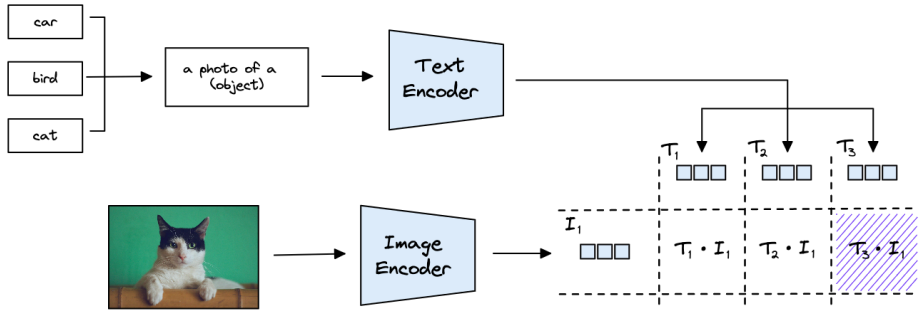


Une approche plus récente : modèles neuronaux multimodaux texte-image (CLIP - *Contrastive Language–Image Pre-training*, OpenCLIP…) : on entraîne un modèle doté des modalités texte (légendes) et image sur de grandes quantités d'images du Web (CLIP : 400 M). Les deux modalités sont agrégées dans un espace de caractéristiques commun.

1. Contrastive pre-training



Le modèle peut ensuite être utilisé en recherche d'information (requête iconographique exprimée en langage naturel) ou en classification, avec des requêtes textuelles ou des requêtes par l'exemple (image).



Photographies d'ateliers de restauration (BnF, 2023) : classification de photographies selon des genres (modèle CLIP, github)

- Exemple : pour chercher des photos de boîtes de conservation = *a photo of a paper preservation box for books*

WISE (VGG Oxford, 2023) : recherche en texte libre sur 1 million d'illustrations de la BL (modèle OpenCLIP). De même, on peut classer les images en utilisant des requêtes

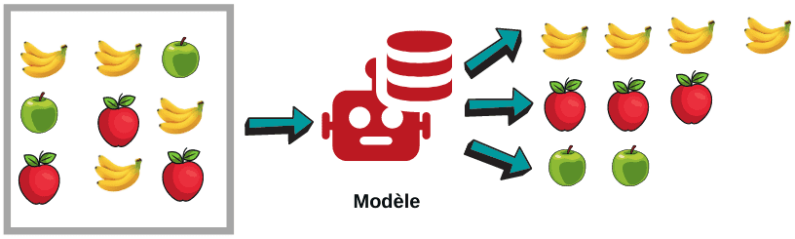
- https://meru.robots.ox.ac.uk/britishlibrary/
- Chercher des illustrations avec un critère "technique de production" (par ex. photographie)

Musée national de Norvège : recherche en texte libre sur 6 000 oeuvres (OpenAI, API GPT-4 Vision : génération de légendes et indexation sémantique des légendes)

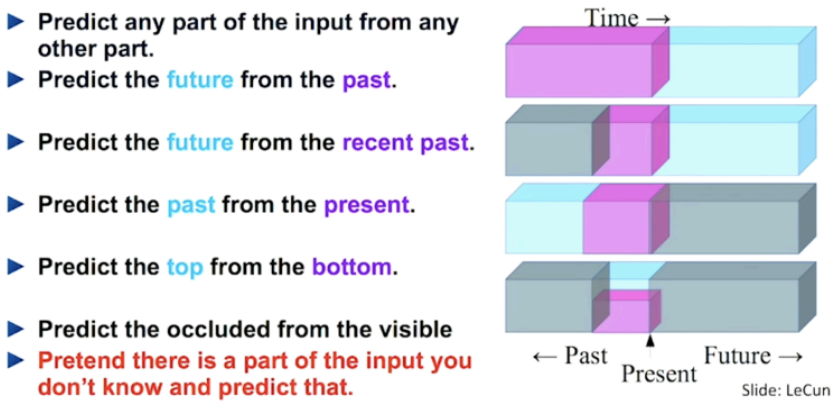
- https://beta.nasjonalmuseet.no/collection/
- Chercher des illustrations avec deux critères (technique et contenu)

Auto-supervisées

L'apprentissage supervisé implique la disponibilité de jeux d'images annotées (pour l'apprentissage). C'est fastidieux. On aimerait ne pas avoir à annoter (ou très peu) :

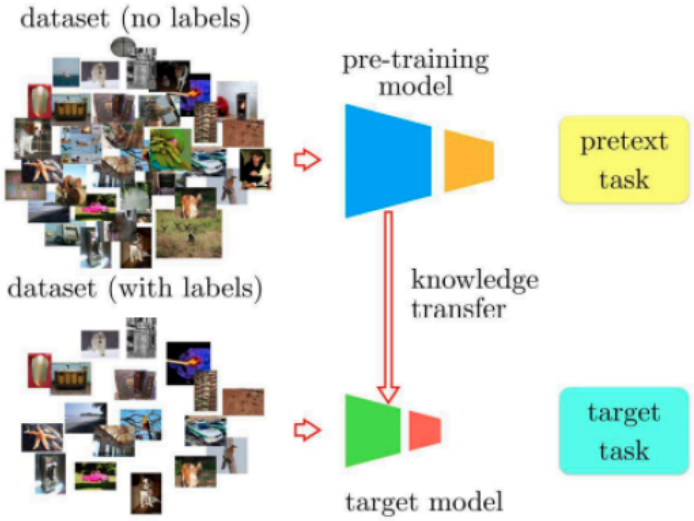


L'apprentissage auto-supervisé (*self-supervised*) est un processus d'apprentissage automatique dans lequel le modèle s'entraîne à apprendre une partie de l'entrée à partir d'une autre partie de l'entrée. Dans ce processus, le problème non supervisé est transformé en un problème supervisé en *générant automatiquement les étiquettes*. Pour exploiter l'énorme quantité de données non étiquetées, il est essentiel de définir les bons objectifs d'apprentissage afin d'obtenir une supervision à partir des données elles-mêmes. Pour ce faire, la méthode d'apprentissage auto-supervisé (“auto-attentif”) consiste à identifier toute partie cachée de l'entrée à partir de toute partie non cachée de l'entrée. Un mécanisme “d’attention” permet au modèle de prédire la partie cachée à partir de la partie connue.



Un type d’architecture à mécanisme d’attention, le modèle *transformer*, a trouvé son essor avec les modèles de langage (GTP). L’attention au texte “passé” permet de prédire la suite probable de ce dernier. Dans leur déclinaison sur les images (*visual transformers,* VIT), le modèle est entraîné à prédire la position de portions d’images (*patchs*) :

Les tâches finales (par ex. la classification) sont entraînées à l’aide de peu de données, par transfert d’apprentissage :



Ressources

Applications :

- GallicaPix (BnF, 2017) : classification de types d'image (modèle CNN Inception)
- Newspaper Navigator (Library of Congress, 2020) : classification de types d'image (modèle CNN)
- Front page detection (National Library of Norway, 2022) : détection de pages de titre de fascicules de presse (modèle *transformer* Google ViT)
- AI Explorer (Harvard Art Museums, 2023) : classification multiclasse des contenus (API Amazon, Clarifai…)
- WISE (VGG Oxford, 2023) : recherche en texte libre sur 1M d'illustrations de la BL (modèle OpenCLIP)

Ressources pédagogiques :

- Teachable Machine
- Tutoriel Google
- Vision Transformers (Hugging Face)

Modèles

1. Modèles CNN

- VGGNet (VGG Group, Oxford, 2014)
- ResNet (Microsoft, 2015)

- U-Net (2015)
- Inception (Google, 2014)
- Yolo (Facebook, 2016)
- ConvNet (Facebook, 2022)

2. Modèles *visual transformer*

- ViT (Google Research, 2021)
- Swin-T (Microsoft, 2021)
- SAM, DINO…

3. Modèles texte image

- CLIP (OpenAI, 2021)
- OpenCLIP

Outils :

- Scripts du projet MODOAP, modèle CNN
- Google Vision AI
- Amazon AI
- LabelStudio (annotation d'images)
- …

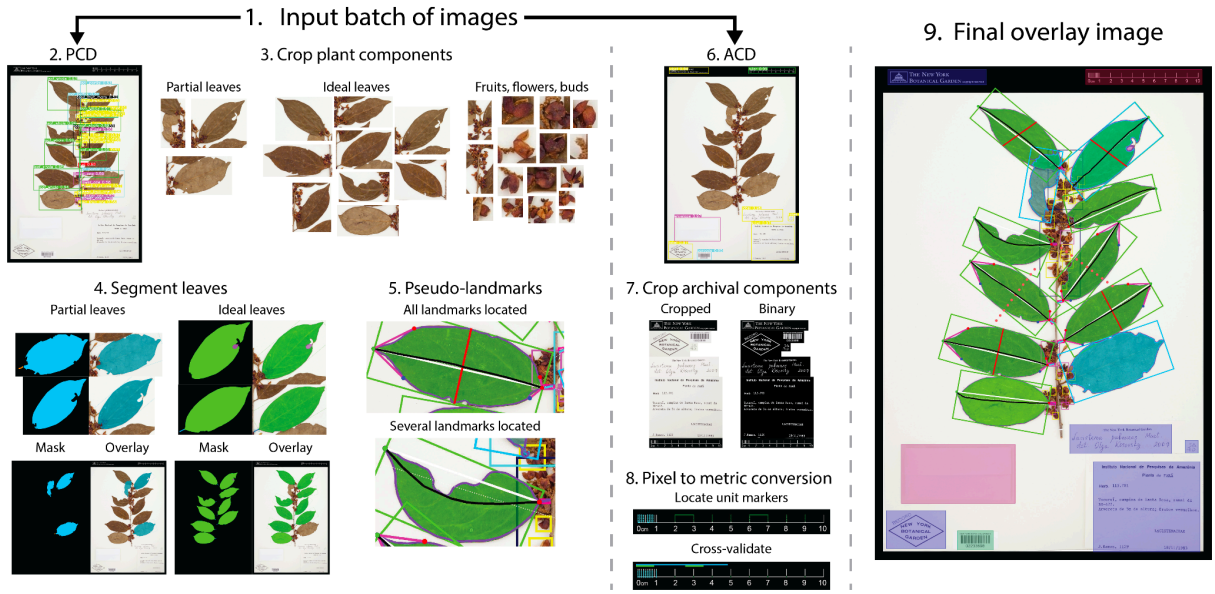
---

2. Détection d'objets (20 mn)

Objectif : identifier des objets présents dans une image

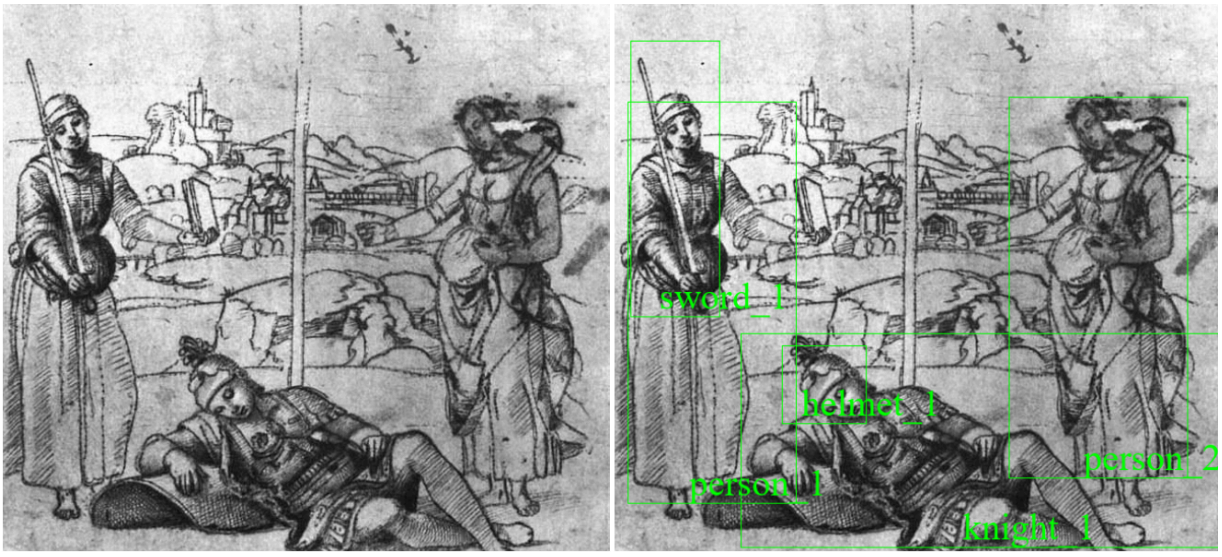
Applications :

- Détecter les composantes d'une image (pour y appliquer un traitement ultérieur) :



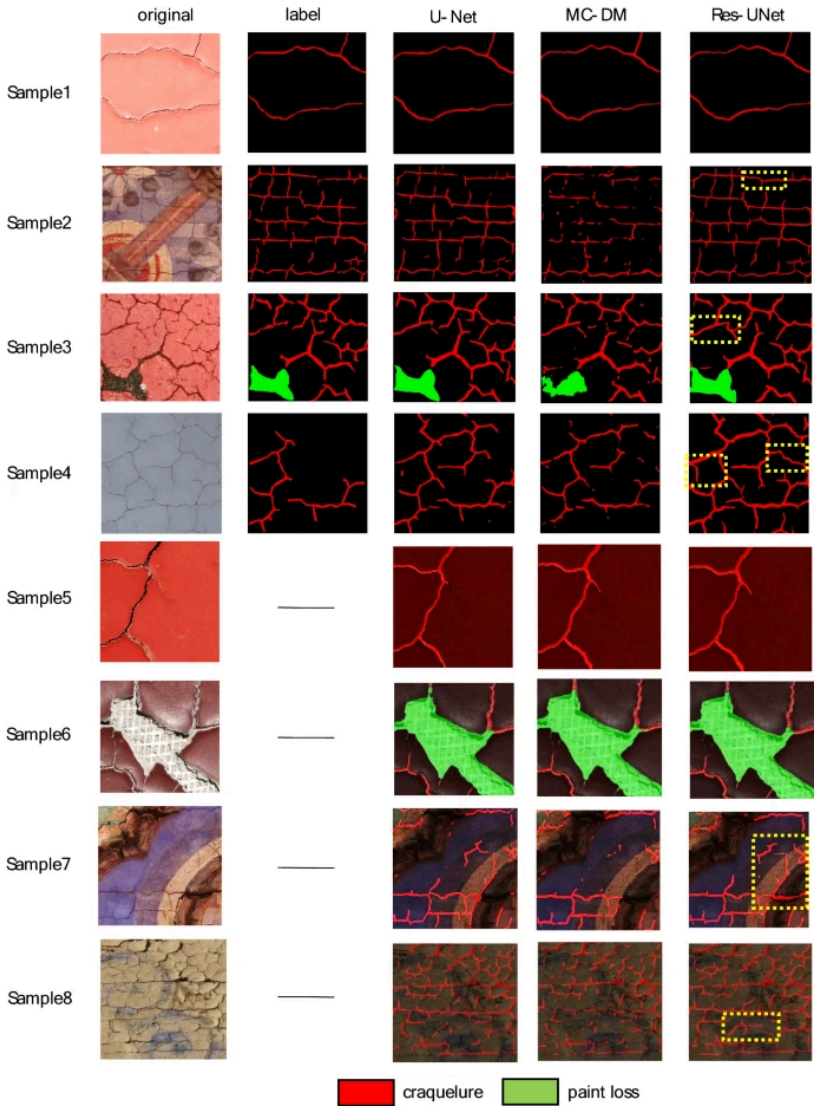
[Leaf Machine]

- Générer des métadonnées descriptives localisées dans l'image :



[Europeana/Supercomputing Center Barcelona] *: détection d'objets et génération de légendes (dans un contexte histoire de l'art)*

- *Conservation : détecter des altérations physiques*



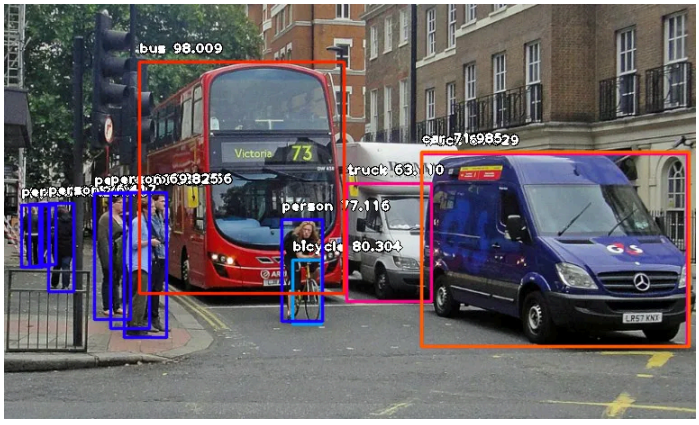
Automatic recognition of craquelure and paint loss on polychrome paintings of the Palace Museum using improved U-Net
Quan Yuan, Xiang He, Xiangna Han & Hong Guo, 2023

Détection d'encre ferro gallique

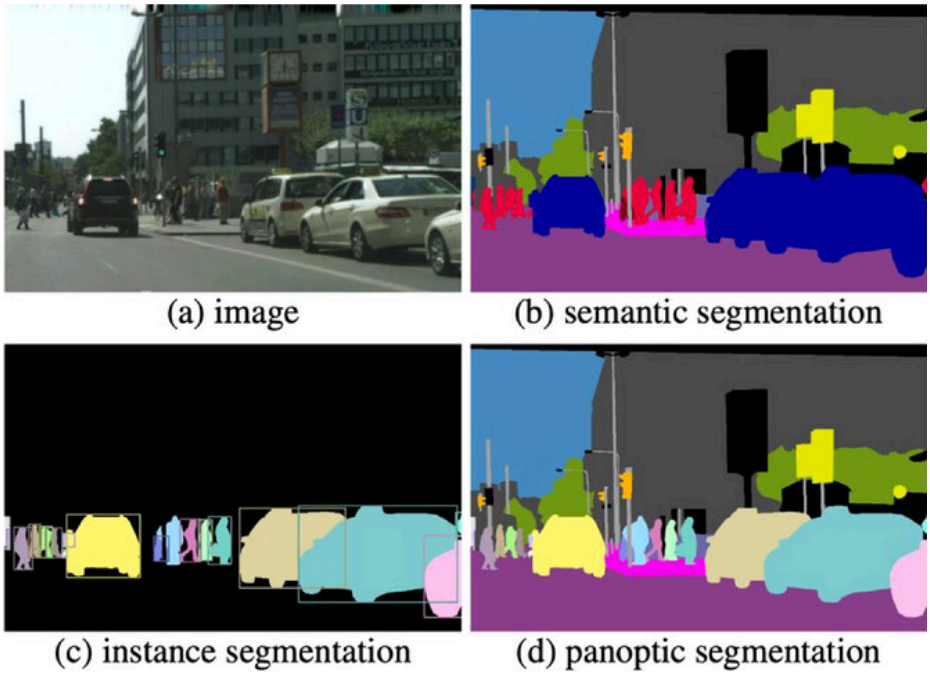
Détection d'étiquette de rondage pour océrisation et alignement avec le catalogue

Résultats **:** des zones de l'image contenant un objet :

- boîtes englobantes
- contours des objets



On parle aussi de *segmentation sémantique* quand on cherche à associer chaque pixel de l'image à une classe (y compris les "arrière-plans") ; de *segmentation d'instance* quand on a besoin d'identifier chaque objet individuellement ; et enfin de *segmentation panoptique* (combinaison des deux précédents cas) :



Approches

Modèles neuronaux

Plusieurs approches se sont succédées ou coexistent :

- Chercher des objets potentiels dans l'image puis les classer : Faster R-CNN, Mask R-CNN, Cascade R-CNN…
- Détecter et classer en même temps : Yolo, SSD, RetinaNet (plus rapide)
- Utiliser conjointement les modalités texte et image (CLIPSeg)
- Chercher des objets sans les classer (modèles visuels auto-supervisés : SAM, Dino, cf. ci-avant).

Les modèles offrant de la classification sont entraînés sur des bases d'images comprenant plusieurs dizaines de milliers de classes (en majorité des photos contemporaines). Les modèles VIT de dernière génération sont entraînés sur des bases d'images non annotées.

GallicaPix (BnF, 2017) : détection d'objet (Yolo v3, API IBM Watson et Google)

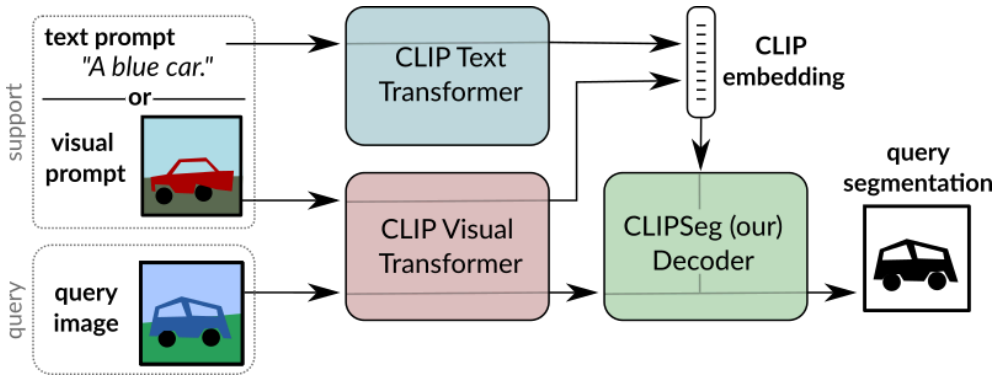
- https://gallicapix.bnf.fr/
- Choisir le corpus *Vogue*
- Choisir le mode "Yolo"
- Valuer le critère Concept : sofa, chapeau, parapluie, raquette…
- Tester l'impact de l'indice de confiance

- Visualiser les objets détectés dans chaque image (bouton *i)*

Bilan ?

Modèles multimodaux texte-image

On associe un modèle multimodal texte-image à un détecteur d'objet (cf. CLIPSeg, basé sur OpenCLIP). La requête se fait alors en langage naturel et non selon un vocabulaire limité (celui des classes des bases d'apprentissage).



Modèles auto-supervisés

Les modèles SAM, DINO et DINOv2 sont des modèles auto-supervisés qui segmentent l'image en objets. Ils peuvent être utilisés comme base d'un modèle de détection d'objet.

- Tester Dinov2 (choisir la démonstration : *semantic segmentation*) sur un document non photographique dans les exemples proposés : https://dinov2.metademolab.com/
- Le modèle SAM autorise une segmentation interactive menée avec l'utilisateur. Tester sur le même document : https://segment-anything.com/demo

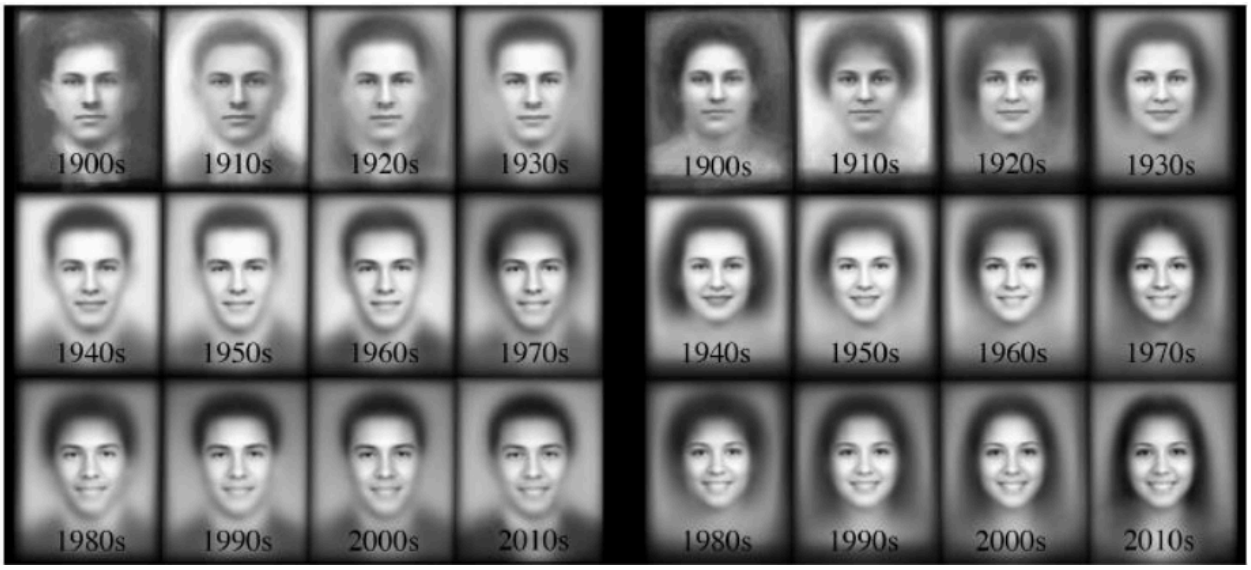
Le cas de la détection de visage

Des modèles spécialisés existent pour les personnes ou les visages humains. Ils prennent en compte des caractéristiques visuelles spécifiques.

Exemples : module dnn (opencv). Modèle SSD (script, notebook).

Applications pour le patrimoine :

- “Moyenne” de visages :
  - Médiation : Le soldat inconnu (BnF), Historial de la Grande guerre
  - SHS : Robots Reading Vogue (Yale DHLab)
  - SHS : A Century of Portraits: A Visual Historical Record of American High School Yearbooks (UC Berkeley)



- Humanités numériques (*visual studies*)

Ressources

Applications :

- GallicaPix : recherche d'information / classification et détection d'objets (API Google Vision et IBM Watson, modèle Yolo v3)
- Leaf Machine (University of Michigan Herbarium) : botanique / détection de morphologie végétale (herbiers), mesures
- Saint George on a Bike (Europeana, CSB) : recherche d'information / génération de légendes descriptives (histoire de l'art)
- VIC, ArtUK (VGG Oxford Group) : recherche d'information dans des bases iconographiques

Ressources pédagogiques :

- Détection d'objets avec Yolo (script et notebook)
- CLIPSeg
- Robotflow : entraîner un modèle avec Faster R-CNN

Modèles

- Yolo
- Faster R-CNN
- SAM
- Dinov2

---

3. Segmentation de documents (10 mn)

Objectif : identifier les zones illustrées dans un document

GallicaPix (BnF), magazine Vogue

Rara Magnetica (Institut Max Planck)

À noter : *il s'agit en fait d'un cas particulier de la détection d'objet. Et plusieurs types d'objet peuvent être détectés : illustration, texte, ornement, etc.*

Résultats **:**

- détection de boîtes englobantes,
- détection de contours (au niveau des pixels).

Détection de contours sur un manuscrit

Approches

Approches “classiques”

Par des techniques d’analyse d’images “classiques” : par exemple avec des algorithmes de détection de discontinuité dans l’image.

Utiliser un OCR…

Mais avec ABBYY FineReader sur le magazine Vogue (1920-1940), 94 % des illustrations identifiées par l’OCR n’en sont pas !

Fausse détection (Abbyy) : typographie, ornements, lignes verticales, etc.

Approche neuronale

Idem approches détection d'objet.

Rara Magnetica (Max Planck Institute for the History of Science, 2023) : création d'une base d'illustrations dédiée au magnétisme (16e-17e siècles)

- Ouvrir l'application de visualisation
- Consulter les pages sources
- Consulter par mot-clé

Tester l'outil CorDeep :

- https://cordeep.mpiwg-berlin.mpg.de/
- Ouvrir un document Gallica, par exemple https://gallica.bnf.fr/ark:/12148/bpt6k570128q/f1.highres
- Enregistrer l'image, l'ouvrir dans CorDeep

Ressources

Modèles

1. Modèles préentrainés sur des corpus patrimoniaux

- dhSegment (EPFL, 2027)
- docExtractor : notebook (ENPC, 2020) : documents historiques synthétiques
- CorDeep (Max Planck Institute, 2022, modèle YOLO v5) : documents 15e-17e siècles
- …

2. Modèles génériques à réentrainer (*fine tuning*)

- Labex "Les Passés dans le Présent", projet MODOAP (méthode Mask R-CNN)
- Detectron2 (Meta AI, 2019) : bibliothèque de modèles
- LayoutParser : bibliothèque de modèles (CNN) et de jeux de données
- Yolo v8

3. Modèles entraînés en autosupervision

- LayoutLMv3 (Microsoft, 2022, transformer multimodal texte/image)
- SEER et VISSL (Meta, 2021)

- CLIPSeg (2021, basé sur CLIP)
- DINOv2 (Meta AI, 2023)
- SAM (Meta AI, 2023, segmentation avec prompting)
- …

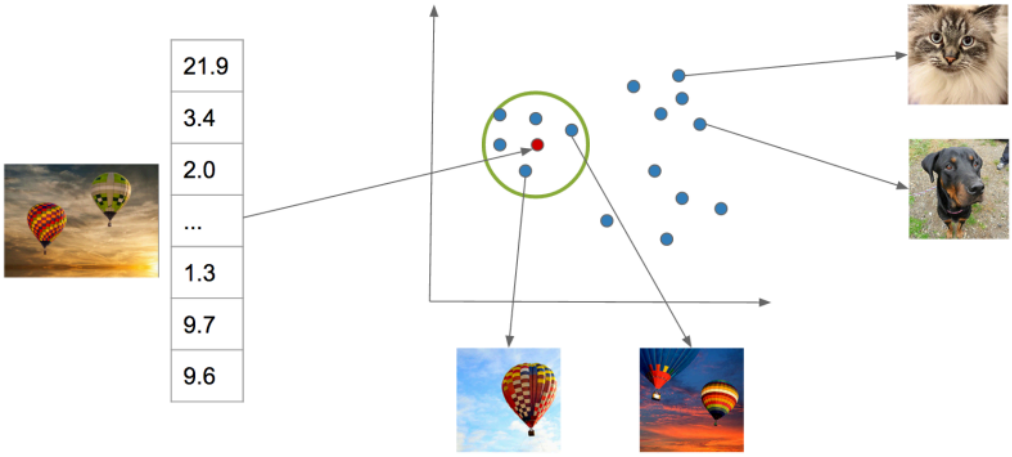
---

4. Similarité visuelle

Objectif : *recherche d'images se ressemblant*
Applications : *recherche d'information*
Résultats : *des images*

Approches

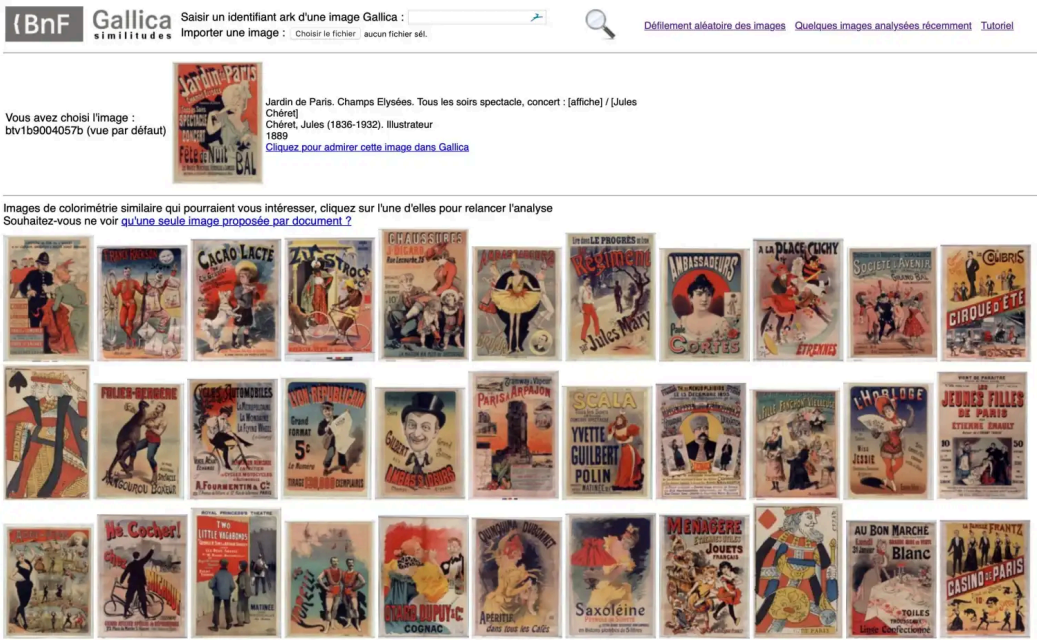
Les images sont caractérisées par une carte d'identité visuelle (un vecteur de nombres). On ne cherche plus à les regrouper en classes étiquetées mais plutôt à créer un espace "virtuel" où les images visuellement proches sont proches. Cette ressemblance est estimée par un calcul de distance entre les vecteurs numériques.



Pour visualiser (à l'œil humain) la proximité entre images, on projette les vecteurs à *n* dimensions en dimensions 2 ou 3.

Approches "classiques"

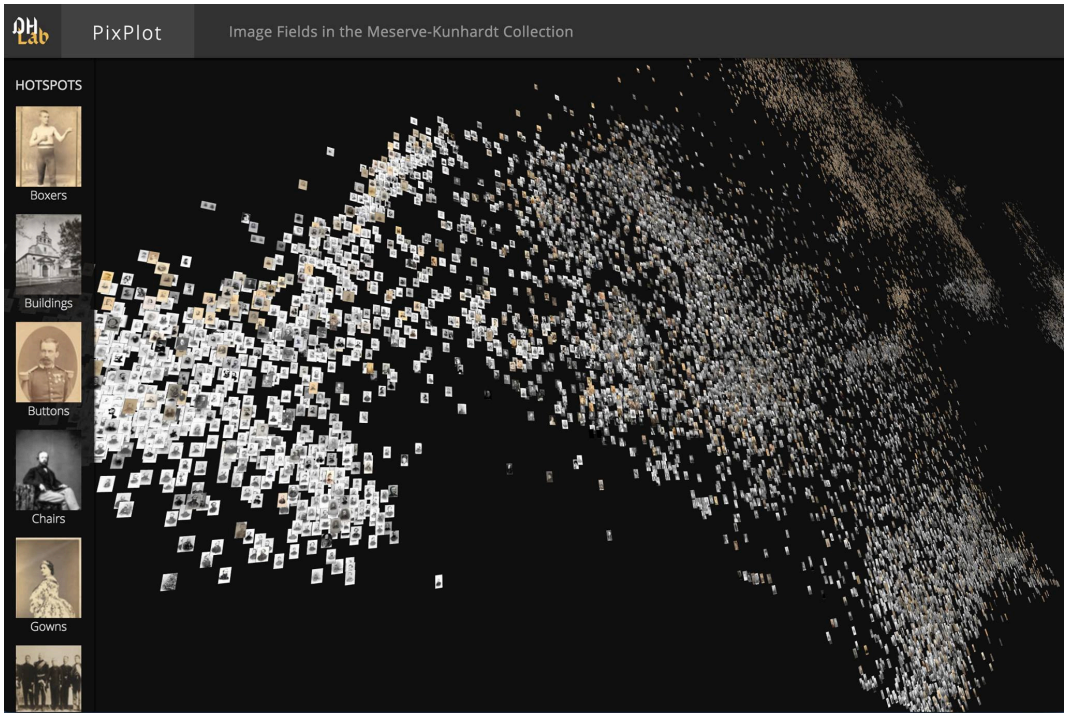
Des caractéristiques de forme, de texture, de colorimétrie, etc. sont extraites des images.



GallicaSimilitudes (2018) : calcul de contraste, palette de couleurs, etc.

Approches neuronales

Un modèle neuronal (CNN, ViT, CLIP…) est utilisé pour calculer les caractéristiques visuelles. Puisque ces modèles ont été entraînés sur des millions d'images, ils ont "appris" à lire les contenus visuels et leur sortie sont des cartes d'identité visuelles.



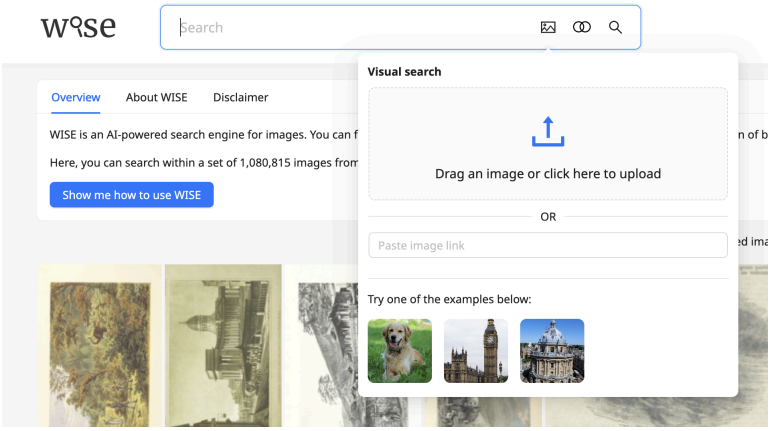
PixPlot (Yale)

PixPlot (DHLab Yale, 2017) : visualisation par nuage d'images 2D (modèle CNN, projection UMAP) : application sur un fonds photographique (Elie Kagan, bib. La Contemporaine, Nanterre)

- https://modoap.huma-num.fr/pixplot_kagan/
- Naviguer dans le nuage
- Utiliser les thèmes éditorialisés (onglet de gauche)

WISE : 1 million d'illustrations de la BL indexées avec le modèle OpenCLIP. Ici, le modèle est utilisé pour l'extraction des caractéristiques visuelles :

- Utiliser le mode "image"
- Choisir l'image "chien" en entrée
- Les résultats montrent le bénéfice d'un modèle multimodale texte-image : des photos de chien et des dessins de chien sont "rapprochées" par leur légende



Rara Magnetica (Max Planck Institute for the History of Science, 2023) : création d'une base d'illustrations dédiée au magnétisme (16e-17e siècles) :

- Ouvrir l'application de visualisation (VIKUS, modèle CLIP et format IIIF)
- Basculer en mode nuage d'images (onglet "Similarity")

Similarité "fine"

Le concept de similarité visuel est pluriel (et subjectif). Dans certains cas d'usage, on souhaite retrouver des images identiques (ou quasi). Ainsi du cas de la recherche de reproductions de photos d'agence (éventuellement recadrées) dans des imprimés.

Droit et Liberté, 01/05/1961

Les modèles d'indexation doivent prendre en compte les détails des contenus visuels. Cela est souvent réalisé avec des descripteurs visuels locaux (par ex. SIFT), qui seront tolérants au recadrage, à la recherche d'extraits, etc.

DEMO

Fonds Elie Kagan (60k photos, La Contemporaine, Nanterre; dispositif interactif développé pour une exposition, 2022) : recherche de reproductions de photo dans la presse contemporaine avec le moteur SNOOP (et un modèle SIFT)

- https://modoap.huma-num.fr/Kagan_Contemporaine/reprises.php
- Identifier des cas de photos recadrées

DEMO

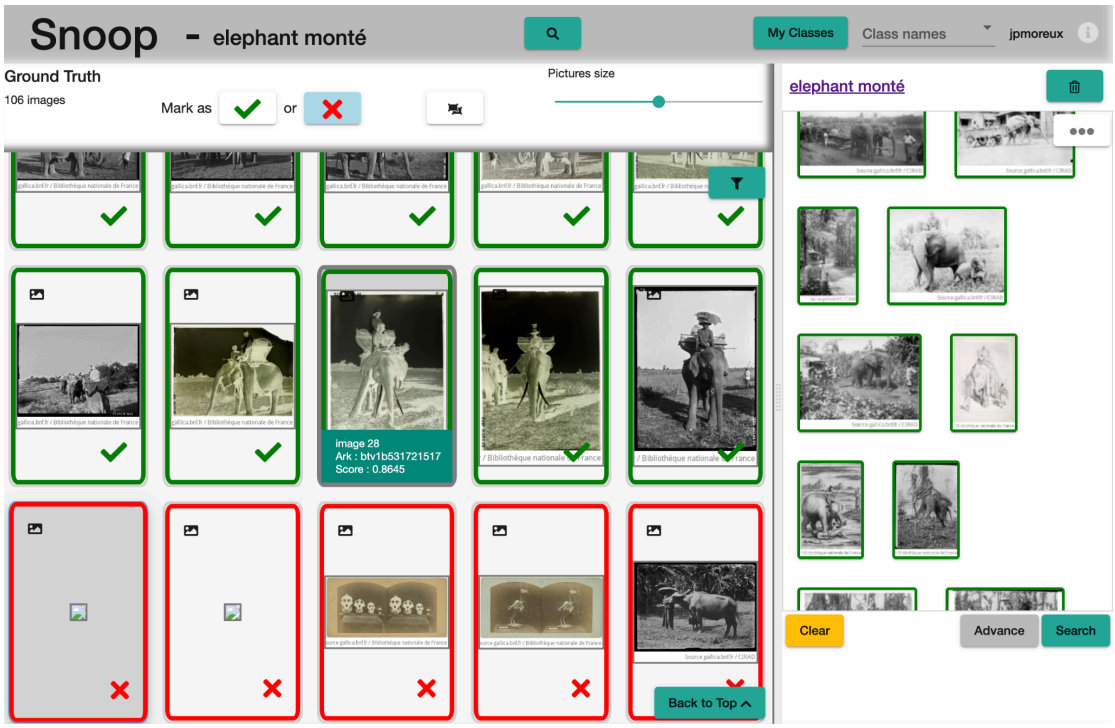
VISE (VGG Image Search Engine, Oxford, 2017) : fournit des correspondances d'images identiques (modèle SIFT). Exemple avec le moteur Bodleian Ballads, 900 pages :

- http://imagematch.bodleian.ox.ac.uk:8000
- Choisir une page
- Tracer une boîte autour d'un bois gravé
- Chercher d'autres instances du bois gravé

Interaction avec les utilisateurs

Faire interagir l'utilisateur en situation de recherche iconographique avec les moteurs de recherche permet d'obtenir une meilleure réponse des dispositifs techniques.

GallicaSnoop (INA-Inria-BnF, 2018) : moteur SNOOP associé à une boucle d'annotation d'images positives/négatives



imgs.ai (université de Marburg, 2022) : moteur de recherche visuelle sur des collections muséales (plusieurs modèles CNN et CLIP), avec interaction de l'utilisateur :

- https://imgs.ai/interface
- Sélectionner une image cible
- Cliquer sur Positive
- Continuer à sélectionner des images positives

Ressources

Applications

- GallicaSimilitudes (BnF, 2018) (n'est plus disponible en ligne)
- GallicaSnoop (INA, Inria, BnF, 2018) : moteur SNOOP (PlantNet) avec interaction de l'utilisateur

- Siamese (KBLab, Melvin Wevers, 2017) : publicités illustrées de journaux (modèle CNN)
- WISE : (VGG Oxford, 2022) : 1 million d'illustrations de la BL indexées avec le modèle OpenCLIP
- Maken (Bibliothèque nationale de Norvège, 2021) : images indexées avec le modèle OpenCLIP
- imgs.ai (université de Marburg, 2022) : moteur de recherche visuelle sur des collections muséales (plusieurs modèles CNN et CLIP)

Outils

- PixPlot (DHLab Yale, 2017) : modèle CNN, projection UMAP
- VISE (VGG Image Search Engine, Oxford, 2017)
- Dinov2, instance retrieval : https://dinov2.metademolab.com/demos?category=retrieval