

Challenge IA : Présentation finale

Groupe : Couture Vision

Membres : Noémie GUISNEL, Pierre JOURDIN, Clément FLORVAL,
Louis GAUTHIER, Anthony QUENTIN

February 4, 2025

Sommaire

Jeu de données

Intuitions

Principe et architecture du modèle retenu

Amélioration des résultats

User Interface

Evaluation

Expérimentations non concluantes

Conclusion

Jeu de données



Figure: Image de test



Figure: Image DAM

Objectif

- ▶ Pour une image test d'article, trouver l'article (donc la référence) correspondant dans la base DAM (2766 articles).
- ▶ L'article doit être parmi les 1, 3 ou 5 références proposées.

Groupe : Couture Vision

Jeu de données : Difficultés

- ▶ Dataset de test non labélisé (80 images).
- ▶ Grand nombre d'articles dans la base DAM (2766 articles).
- ▶ Unique photo par classe : difficile de faire de la classification.
Et en plus classification = pas scalable.
- ▶ Les images (surtout celles du test) ne sont pas standardisées
(background, orientation...)
- ▶ Images très similaires dans DAM.
- ▶ Deux images de test dont le label n'est pas présent dans DAM.

Intuitions

- ▶ Utilisation d'un modèle permettant d'extraire l'article de l'image et de mettre un fond blanc derrière.
- ▶ Utilisation d'un modèle pré-entraîné pour créer des embeddings sur les images DAM et les images de test.
- ▶ Pour chaque image de test, calcul des similarités cosinus avec tous les embeddings des images DAM et affichage des 5 meilleures similarités (Top-5).

Modèles utilisés

- ▶ **Extraction et traitement de l'article** : Utilisation de RMBG-2.0 pour rogner l'image et supprimer le fond.
- ▶ **Embeddings d'image** : Modèle ViT par Google (Base Patch-32, 86.4M paramètres).
- ▶ **Data Augmentation** : Utilisation de TRELLIS pour générer des modèles 3D.
- ▶ **Triplet Network** : Permet d'optimiser les embeddings pour réduire la distance cosinus entre des images similaires.

Traitement des images

Suppression du fond avec rembg et recentrage de l'objet

- ▶ Les images à inférer contiennent un arrière-plan indésirable.
- ▶ Utilisation de la bibliothèque rembg avec le modèle u2net pour retirer le fond. Objectif : réduire le bruit pour l'extraction des caractéristiques.
- ▶ Ensuite, un traitement d'image identifie les pixels non transparents pour **recentrer et recadrer** l'objet.



Extracted Object



Figure: Extracted object with rembg

Figure: Example of Image test

Amélioration avec RMBG-2.0

Utilisation du modèle briaai/RMBG-2.0 pour une meilleure suppression du fond

- ▶ Le modèle précédent pouvait générer des artefacts ou des découpages imprécis.
- ▶ Amélioration de la qualité du traitement en utilisant le modèle briaai/RMBG-2.0.
- ▶ Ce modèle utilise des techniques avancées pour un détourage plus précis et net.



Figure: Example of Image test

Extracted Object



Figure: Extracted object with rmbg-2.0

Comparaison des méthodes de suppression de fond

Tableau comparatif des performances techniques

Critère	U2Net (rembg)	RMBG-2.0
Architecture du modèle	U-Net modifié	développé à partir de BiRefNet
Nombre de paramètres	44M	221M
Qualité du détourage	Moyenne (bords flous)	Excellent (détails fins)
Rapidité de traitement	Rapide	Un peu plus lent
Consommation mémoire	Modérée	Plus élevée
Adaptabilité aux arrière-plans complexes	Moyenne	Très bonne

Table: Comparaison entre U2Net et RMBG-2.0

- ▶ **RMBG-2.0 et recentrage de l'objet** sont appliqués aux images du DAM et à chaque image destinée à l'inférence

Présentation du Modèle ViT Base Patch32 224

- ▶ **Nom du Modèle** : google/vit-large-patch16-224
- ▶ **Origine** : Développé par Google
- ▶ **Nombre de Paramètres** : Environ 86 millions
- ▶ **Pré-entraînement** :
 - ▶ Dataset : ImageNet
 - ▶ Contenu : Environ 1.2 million d'images couvrant 1000 classes

Amélioration des résultats

- ▶ **Data augmentation** : Certaines images de test ne sont pas prises du même angle que les images DAM.
 - ▶ Utilisation de **TRELLIS** pour générer des modèles 3D.
- ▶ Recadrage et recentrage des images.
- ▶ Tests de nombreux modèles d'embedding: Microsoft ResNet-50, DinoV2, Nomic EmbedVision... les meilleurs résultats avec ViT de Google.
- ▶ Comparaison de différentes méthodes d'agrégation des embeddings: mean, max, min, concaténation, CLS seulement... max donne les meilleurs résultats.
- ▶ Comparaison de différentes métriques de similarité: cosine et euclidean, cosine donne les meilleurs résultats.

Data augmentation avec un modèle de génération 3D

- ▶ Nous avons utilisé **TRELLIS** pour générer 2700 modèles 3D des assets du DAM.
- ▶ 8 rendus **Blender** à des angles différents pour chaque modèle.



Figure: Input for TRELLIS

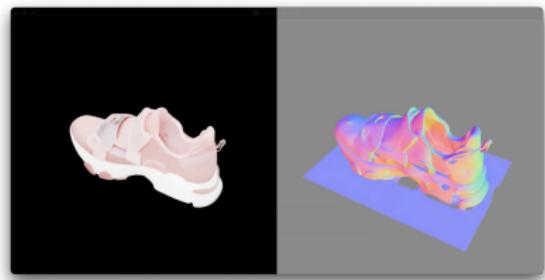


Figure: TRELLIS generated 3D model

Data augmentation avec un modèle de génération 3D



Groupe : Couture Vision

Figure: 3D renders of a bag

Navigation icons: back, forward, search, etc.

Amélioration des performances grâce aux modèles 3D

- ▶ Résultats sans l'augmentation 3D avec Google ViT: Top-1: 36%, Top-3: 47%, Top-5: 59%
- ▶ Nous avons d'abord ajouté les embeddings 3D à l'index.
- ▶ Puis nous avons retiré les embeddings 2D et avons constaté une amélioration des performances.
- ▶ Enfin, nous avons fusionné (moyenne des deux) les embeddings 2D et 3D pour obtenir les meilleurs résultats. Intuition: les embeddings des photos capturent des informations plus précises sur la texture, la couleur, et les embeddings 3D capturent des informations sur la forme et l'orientation.
- ▶ Nous avons utilisé la descente de gradients pour sélectionner les bons features ainsi que pour fusionner les embeddings 2D et 3D, mais les résultats n'étaient pas satisfaisants.
- ▶ Résultats Google ViT: Top-1: 45%, Top-3: 59%, Top-5: 64%.
- ▶ Résultats finaux avec FashionCLIP: Top-1: 61%, Top-3: 80%, Top-5: 81%.

Interface Web avec Gradio

- ▶ Interface web simple avec Gradio.
- ▶ Les utilisateurs peuvent uploader une photo et avoir le top-X des articles les plus similaires, ainsi que les références DAM.

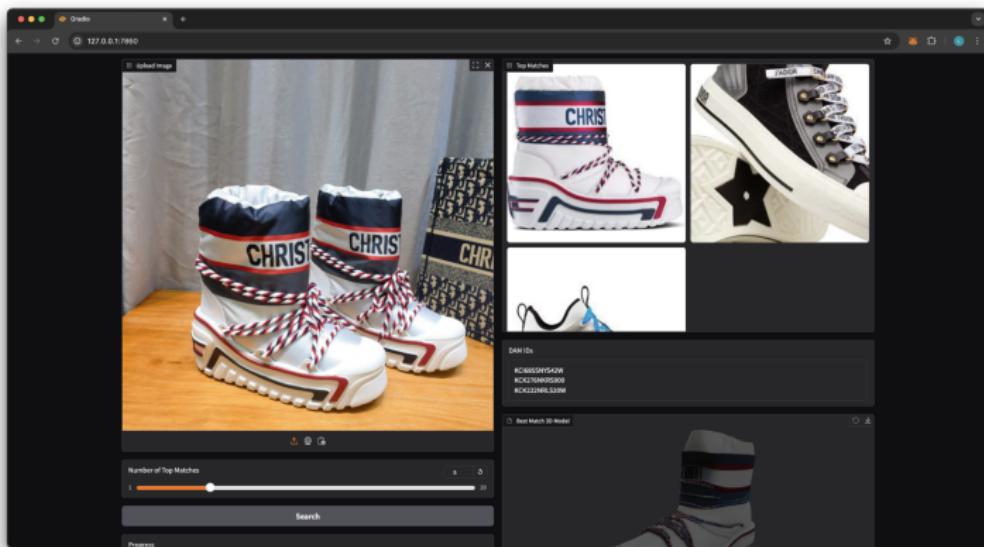


Figure: Interface Gradio

Comparaison des Modèles

Modèle	Top-1 (%)	Top-3 (%)	Top-5 (%)
RMBG 2.0 - Microsoft ResNet50	21.50	32.75	40.25
ViT Large - RMBG 2.0 - 2D	38.75	53.75	61.25
ViT Large - RMBG 2.0 - 3D - Mean Cos	41.25	62.5	68.75
ViT Large - RMBG 2.0 - 3D - Max Cos	45	58	63
FashionCLIP - RMBG 2.0 - 3D - Max Cos	61	80	81

Table: Comparaison des performances des différents modèles

Interprétation des Résultats

- ▶ **Data Augmentation** : L'augmentation des données, via la 3D, contribue à une meilleure généralisation du modèle.
- ▶ **ViT** Les modèles ViT sont les modèles les plus performant pour réaliser les embeddings
- ▶ **MaxCos** L'utilisation de MaxCos est meilleure que MeanCos pour la comparaison des embeddings.

Matrice de Confusion

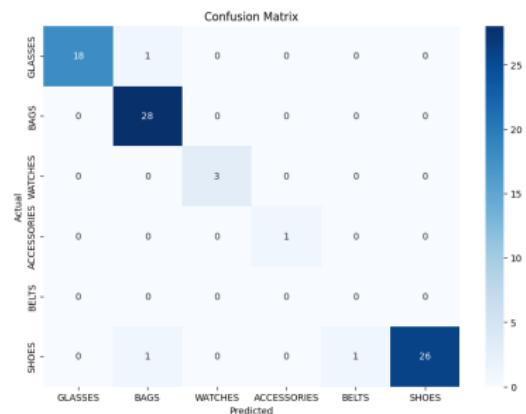


Figure: Matrice de confusion illustrant les prédictions du modèle

Analyse des Résultats

- ▶ **Bonne capacité de distinction globale :** la majorité des images sont bien classées.
- ▶ **Confusion entre classes similaires :**
 - ▶ Articles de couleur proche ou motifs similaires.
 - ▶ Sacs / Ceintures en cuir.
- ▶ L'écart entre le Top-5 et le Top-1 montre que le modèle capture les bonnes caractéristiques mais a du mal à affiner le classement final.

Expérimentations non concluantes

- ▶ Tentative d'amélioration du Top-1 par NLP :
 - ▶ Utilisation du modèle BLIP Base pour générer des descriptions d'articles.
 - ▶ Problème : Les descriptions générées sont trop pauvres pour améliorer significativement les résultats.

Image de test



The black sunglasses with gold temples

Top-1



A pair of sunglasses with a black frame and gold arms

Top-2 (Article à trouver)



A pair of sunglasses with a black frame and gold details

Expérimentations non concluantes

Fine-tuning de ViT :

- ▶ Faible convergence malgré un taux d'apprentissage réduit.
- ▶ Très long à entraîner et beaucoup de classes à prédire, pas assez de données.



Figure: Exemple de prédition après fine-tuning

Expérimentations non concluantes

- ▶ Siamese Network : entraîné sur la data 3D, résultats OK mais pas concluants.
- ▶ Réduction de dimension : PCA, AE, SAE, VAE... Meilleurs résultats avec les AE, mais pas concluants.
- ▶ Autres modèles moins performants testés : DinoV2, ResNet, Clip, VitMsn.

Conclusion

- ▶ Modèle performant avec rendus 3D et embeddings 2D/3D.
- ▶ Précision Top-1 de 61% avec FashionCLIP.
- ▶ Pistes futures :
 - ▶ Tester un modèle multimodal plus puissant pour générer des descriptions (Janus, BLIP 3B).
 - ▶ Data augmentation (flip, rotations, zooms) pour chaque image DAM. Déjà testé mais très long.
 - ▶ Amélioration du dataset DAM : webscraping ?