Étude comparative de méthodologies issues de Mask R-CNN : Application au Corpus DeepFashion2

Warren Jouanneau, Aurélie Bugeau, Marc Palyart, Nicolas Papadakis, Laurent Vezard

Étude comparative de méthodologies issues de Mask R-CNN : Application au Corpus DeepFashion2

W. Jouanneau[1] A. Bugeau[2] M. Palyart[1] N. Papadakis[3] L. Vezard[1]

[1] Lectra, F-33610 Cestas, France
[2] Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, F-33400 Talence, France
[3] CNRS, Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France
w.jouanneau@lectra.com

1 Introduction

Pour l'industrie textile et de la mode, les images de vêtements ont une forte valeur à tous les niveaux du cycle de vie d'un produit. On les trouve par exemple, lors de la conception comme source d'inspiration, lors de séances d'essayage comme support de validation ou encore lors de la mise sur le marché comme référence visuelle. Il est donc nécessaire de faciliter leur accès et leur recherche parmi un grand nombre d'images candidates. Cela repose le plus souvent sur l'apposition manuelle de mots clefs afin de les indexer. L'automatisation de cette étape fastidieuse permettrait ainsi une économie considérable de temps et permettrait aux différents acteurs de se concentrer sur des tâches au coeur de leurs métiers.
En apprentissage supervisé, cette problématique se réduit à une classification. Des travaux se sont tournés vers l'attribution d'une classe à une image entière par le biais de réseaux de neurones à convolution [5]. Cependant, cette approche donne de meilleurs résultats si un seul vêtement est présent. Afin de retrouver les classes de plusieurs vêtements présents, la classification peut s'opérer sur des sous-images n'en contenant qu'un seul [9] (*i.e.* détection), ou sur chacun des pixels de l'image [8] (*i.e.* segmentation sémantique). En couplant ces deux approches (*i.e.* segmentation d'instances), on obtient des masques qui contiennent les pixels d'un unique vêtement. Ceci permet d'ajouter la distinction des instances d'une classe et une localisation plus fine à la détection. De plus, ces masques peuvent servir à la caractérisation des vêtements par d'autres méthodologies.
Pour l'industrie, la segmentation d'instances est une étape cruciale pour l'indexation de contenu. Nous proposons ici une évaluation des méthodes de segmentation d'instances de l'état-de-l'art pour notre cas d'usage.

2 Présentation des données

Parmi les corpus d'images de vêtements disponibles, peu incorporent les masques nécessaires à la segmentation d'instances. Le corpus DeepFashion2 [4] est actuellement celui qui propose le plus grand nombre d'images et d'annotations avec masques. Les données sont séparées en deux, 312 186 instances dans 191 961 images pour l'entraînement, 52 490 instances dans 32 153 images pour l'évaluation. Chacune des instances est labélisée suivant 13 catégories (*cf.* Table 1). Les images proviennent de particuliers ou de professionnels et comprennent des vêtements dans de nombreux contextes (*e.g.* en extérieur, studio). Il y a une grande variabilité des prises de vues et des conditions de capture (figure 1). On y retrouve trois niveaux de visibilité du vêtement, trois niveaux de zoom, si le vêtement est porté ou non et enfin la position (vue de face ou de côté). Les masques sont représentés sous la forme de contours discrétisés.

TABLE 1 – Nombre d'instances par classes *(s.s. : short sleeved, l.s. : long sleeved)*

	shirt		outwear		vest	sling	shorts	trousers	skirt	dress			
	s.s.	l.s.	s.s.	l.s.						s.s.	l.s.	vest	sling
entraînement	71645	36064	543	13457	16095	1985	36616	55387	30835	17211	7907	17949	6492
validation	12556	5966	142	2011	2113	322	4167	9586	6522	3127	1477	3352	1149

3 Méthodes

Lors de la constitution du corpus DeepFashion2, Mask R-CNN [6] a été évalué sur la segmentation d'instances [4]. Mask R-CNN s'appuie sur les performances de la méthode de classification multi-objets Faster R-CNN [11] en ajoutant la prédiction du masque de l'instance en parallèle de la prédiction de la boîte englobante et de la classe de l'instance. Nous proposons ici une évaluation de récentes évolutions issues de cette méthode.
Dans Mask R-CNN, le score du masque est identique à celui de la classification, cependant il y a rarement une corrélation entre celui-ci et la qualité du masque. Mask scoring R-CNN [7] propose une correction de ce score en ajoutant un bloc pré-

disant l'indice de Jaccard (intersection sur l'union entre les masques prédits et la vérité terrain) par régression. Cela permet au réseau d'avoir connaissance de la qualité de sa propre prédiction du masque.
Dans Mask R-CNN, l'indice de Jaccard sur les boîtes englobantes ou les masques permet de distinguer les prédictions positives des négatives. Un seuil trop élevé implique la disparition de cas positifs lors de l'entraînement et un sur-apprentissage. Pour y remédier, Cascade R-CNN [1] propose de chaîner plusieurs détecteurs associés à des niveaux croissants de seuils. Les prédictions des boîtes englobantes d'un détecteur sont alors fournies en entrée du suivant. Cascade R-CNN peut être couplé avec Mask R-CNN, en adoptant le même chaînage sur les branches de prédictions des masques. Hybrid task cascade [2] propose d'entremêler les chaînes de détection et de prédiction des masques au lieu de les chaîner séparément, et aussi d'ajouter une branche de contexte à l'architecture.

4 Résultats préliminaires

Les résultats ont été obtenus en utilisant des implémentations des architectures mises à disposition par MMDetection [3]. Concernant l'extraction des descripteurs, le même backbone (feature pyramid network reposant sur un ResNet 50 [10]) a été choisi. Dû au temps conséquent nécessaire pour l'entraînement des modèles sur Deepfashion2, les résultats présentés découlent de l'évaluation après une seule epoch.

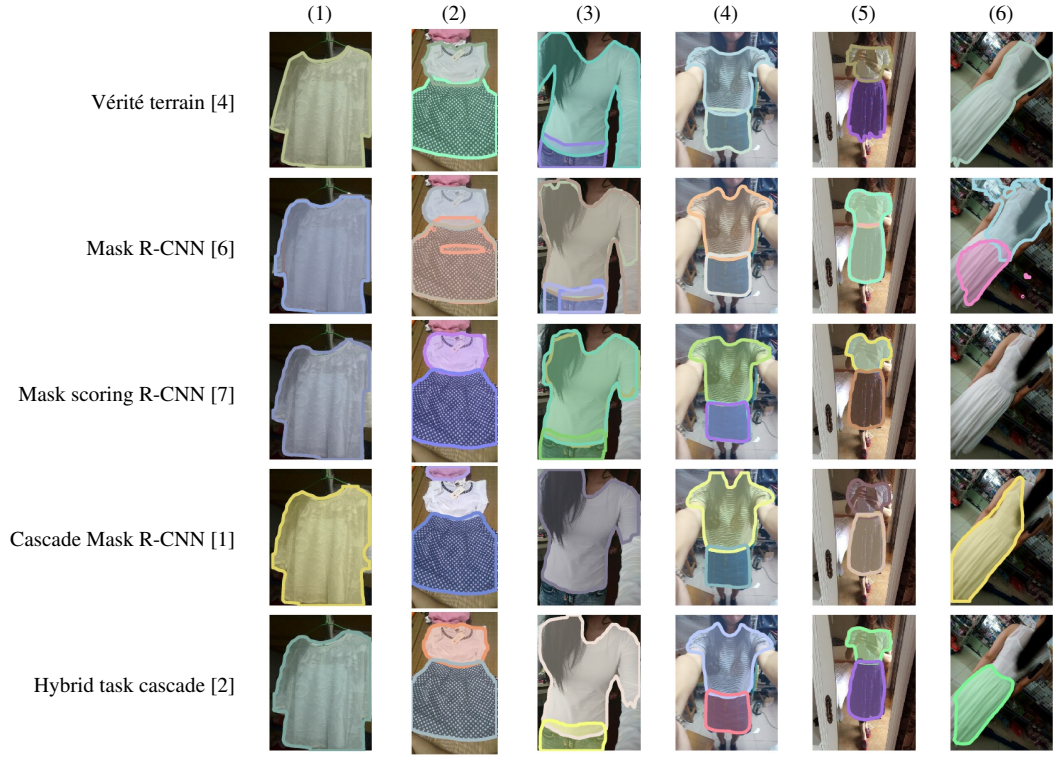

FIGURE 1 – Images où figurent les masques réels et prédits avec un score supérieur à 0.4. Vêtement(s) (1) sur cintre, (2) à plat, (3) avec occlusion dûe aux cheveux, (4) de face, (5) avec occlusion dûe aux bras, (6) de dos

La figure 1 montre la qualité des masques inférés à partir des différents modèles. Les masques prédits par les modèles de Mask R-CNN et de Cascade Mask R-CNN semblent subjectivement moins précis. On peut ainsi noter des imperfections ou des absences de détection sur la jupe et le haut en colonne (2), le pantalon en colonne (3), le col en colonne (4), la jupe et le haut en colonne (5). La robe en colonne (6) semble poser problème pour l'ensemble des modèles, cela pouvant être dû à l'orientation du sujet peu commun dans le corpus.
La table 2 présente différentes métriques observées lors de l'entraînement et de l'évaluation. La mAP correspond au moyennage des précisions moyennes pour des indices de Jaccard allant de 0.5 à 0.95 par incrément de 0.05. La mAP50 et la mAP75 correspondent seulement à celles aux indices 0.5 et 0.75 respectivement. On peut noter ainsi que Mask scoring R-CNN et Hybrid task cascade ont des mAP et mAP75 proches, avec des scores supérieurs en mAP50 pour la première méthode. Cela traduit donc une meilleure mAP aux indices supérieurs à 0.75 pour Hybrid task Cascade. La table 3 présente le détail des mAP par classes. Les meilleurs mAP se répartissent de nouveau entre les méthodologies Mask scoring R-CNN et Hybrid Task, sauf deux exceptions. Cependant, pour certaines classes ces résultats sont à prendre avec précaution. Les mAP peuvent en effet être proches entre deux méthodologies (*e.g.* label shorts), ou présenter de faibles valeurs, dues à une possible ambiguïté entre labels (*e.g.* label sling dress) ou une disparité et un faible nombre d'instances pour certaines classes (*cf.* Table 1 labels

short sleeved outwear et sling). La machine d'expérimentation disposait d'une NVIDIA Tesla P100. Le temps nécessaire pour l'inférence varie du simple au double entre les modèles reposant sur Mask R-CNN et Hybrid Task Cascade. La variation relative en temps n'est pas aussi élevée pour l'entraînement.

TABLE 2 – Résultats des évaluations

architectures	mAP	mAP50	mAP75	temps d'inférence	temps d'entraînement
Mask R-CNN	0.244	0.361	0.277	0.11 s/image	10.94 h/epoch, 0.41 s/batch
Mask scoring R-CNN	0.256	0.382	0.294	0.11 s/image	11.20 h/epoch, 0.42 s/batch
Cascade Mask R-CNN	0.243	0.362	0.276	0.13 s/image	14.68 h/epoch, 0.55 s/batch
Hybrid Task Cascade	0.260	0.375	0.295	0.22 s/image	16.87 h/epoch, 0.63 s/batch

TABLE 3 – mAP par classes *(s.s. : short sleeved, l.s. : long sleeved)*

	shirt		outwear		vest	sling	shorts	trousers	skirt	dress			
	s.s.	l.s.	s.s.	l.s.						s.s.	l.s.	vest	sling
Mask R-CNN [6]	0.558	0.339	0.0	0.174	0.175	**0.019**	**0.378**	0.395	0.330	0.281	0.158	0.311	0.053
Mask scoring R-CNN [7]	0.516	0.288	0.003	**0.188**	**0.282**	0.012	0.355	0.402	**0.398**	0.269	0.153	0.346	**0.110**
Cascade Mask R-CNN [1]	0.521	0.288	0.0	0.183	0.248	0.011	0.344	0.404	0.346	0.257	0.148	0.333	0.071
Hybrid task cascade [2]	**0.584**	**0.341**	**0.007**	0.157	0.230	0.017	0.373	**0.418**	0.318	**0.317**	**0.171**	**0.348**	0.094

5 Conclusion et perspectives

Le corpus Deepfashion2 a quelques limites : les sujets sont principalement féminins, les masques contiennent des pixels qui n'appartiennent pas aux vêtements (*e.g.* main, cheveux), la taxonomie choisie pour les labels pose question, et certaines classes sont faiblement représentées. Il pourra ainsi être utile de se tourner vers d'autres corpus d'images de vêtements ou d'en constituer un nouveau présentant des segmentations plus précises et une granularité plus fine et hiérarchique (*e.g.* segmentation aux niveaux des manches, des cols, des jambes appartenant aux vêtements).

Les premiers résultats sont néanmoins encourageants dans une optique d'exploitation industrielle. Les méthodes de segmentation d'instances étudiées semblent en mesure de fournir des extractions de vêtements suffisamment précises pour permettre leur caractérisation ultérieure. Les architectures suivant la méthodologie de Mask scoring R-CNN et Hybrid task cascade se distinguent. En effet, ces méthodologies ont les meilleures mAP en évaluation. Toutefois, Hybrid task cascade prédit des masques de meilleures qualité (*i.e.* mAP aux indices de Jaccard supérieur à 0.75). Cependant, si les résultats entre ces approches restent proches, Mask scoring R-CNN aura un avantage en terme de complexité.

Se concentrer sur ces méthodes et potentiellement expérimenter d'autres backbones est donc une piste à suivre. L'évaluation de la segmentation d'instances tente de reproduire celle utilisée pour l'évaluation de classification en déterminant les vrais et faux positifs par seuillage d'une mesure de la qualité ou de la similarité d'un masque. Il peut alors être intéressant de tenter de s'abstraire de la binarisation d'une variable continue en proposant d'autres métriques ou une autre approche d'évaluation afin de quantifier au mieux la qualité des prédictions. De plus, ces métriques pourront être enrichies au cas spécifique de la segmentation d'instances de vêtements.

Références

[1] Z. CAI et N. VASCONCELOS. Cascade R-CNN : High quality object detection and instance segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2019).

[2] K. CHEN et al. Hybrid task cascade for instance segmentation, *IEEE Conf. on Comp. Vis. and Pat. Recogn.* 2019.

[3] K. CHEN et al. MMDetection : Open MMLab Detection Toolbox and Benchmark, 2019. arXiv : 1906.07155.

[4] Y. GE et al. Deepfashion2 : A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, *IEEE Conf. on Comp. Vis. and Pat. Recogn.* 2019.

[5] P. GUTIERREZ et al. Deep learning for automated tagging of fashion images, *Europ. Conf. on Comp. Vis.* 2018.

[6] K. HE et al. Mask R-CNN, *IEEE Conf. on Comp. Vis. and Pat. Recogn.* 2017.

[7] Z. HUANG et al. Mask scoring R-CNN, *IEEE Conf. on Comp. Vis. and Pat. Recogn.* 2019.

[8] W. JI et al. Semantic Locality-Aware Deformable Network for Clothing Segmentation, *Int. J. Conf. on Artif. Int.* 2018.

[9] B. LAO et K. JAGADEESH. Convolutional neural networks for fashion classification and object detection, *Chinese Conf. on Comp. Vis.* 2015.

[10] T.-Y. LIN et al. Feature pyramid networks for object detection, *IEEE Conf. on Comp. Vis. and Pat. Recogn.* 2017.

[11] S. REN et al. Faster R-CNN : Towards real-time object detection with region proposal networks, *Adv. in Neur. Inf. Proc. Systems*. 2015.