Analyse de l'encodeur pour la segmentation d'une base de données hétérogène de photographies de plaies chroniques avec peu d'annotations

Guillaume PICAUD[1,3], Marc CHAUMONT[1,2], Gérard SUBSOL[1], Luc TEOT[3]

[1] LIRMM, équipe ICAR, Univ. Montpellier, CNRS, Montpellier, France
[2] Univ. Nîmes Place Gabriel Péri, 30000 Nîmes Cedex 01, France [3] Cicat-Occitanie, Montpellier, France

{guillaume.picaud, marc.chaumont, gerard.subsol}@lirmm.fr, l-teot@chu-montpellier.fr

Résumé

La segmentation est cruciale en imagerie médicale mais l'obtention de données annotées en quantité suffisante est difficile, limitant le développement de modèles d'apprentissage profond performants. Les stratégies d'apprentissage auto-supervisé (SSL) offrent une solution prometteuse pour pallier ce manque d'annotation. L'une d'entre elles, Dinov2 pour Distillation with NO labels a permis l'élaboration de l'immense base de données LVD-142M ainsi que l'entraînement d'encodeurs, aujourd'hui en accès libre. Cependant, les images cliniques ne sont pas nécessairement bien représentées dans LVD-142M. Dans cet article, nous comparons différentes méthodes d'initialisation d'encodeurs pour la segmentation de photographies cliniques dans un contexte de manque d'annotation.

Mots-clés

Apprentissage auto-supervisé, segmentation, Dinov2, images cliniques.

Abstract

Segmentation task is crucial in medical imaging, but obtaining a sufficient quantity of annotated data is challenging, limiting the development of high-performing deep learning models. Self-supervised learning (SSL) strategies offer a promising solution to address this lack of annotation. One such strategy, Dinov2 for Distillation with NO labels, enabled the creation of the vast LVD-142M database and also the training of encoders, whose weights are now freely accessible. However, clinical images may not be well represented in LVD-142M. In this article, we evaluate the benefits of different encoder initialization methods for segmentation in a context of scarce annotated clinical data.

Keywords

Self supervised learning, segmentation, Dinov2, weight initialization, clinical images

1 Introduction

La Haute Autorité de Santé définit les plaies chroniques comme des lésions n'ayant pas atteint une cicatrisation complète après 4 à 6 semaines d'évolution. De multiples facteurs peuvent favoriser leurs apparitions au sein de populations à risque comprenant les personnes âgées, les diabétiques ainsi que les personnes à mobilité réduite. Elles posent un problème socio-économique majeur avec des conséquences sévères pour l'individu pouvant aller de l'amputation au décès du patient. L'assurance maladie a estimé à plus d'un milliard d'euros la seule gestion des escarres et ulcères à domicile pour l'année 2011. Leur prévalence est en constante augmentation, notamment en raison du vieillissement de la population.

Le "Réseau Cicat-Occitanie" fournit une assistance à la prise en charge des plaies chroniques par le biais de téléconsultation afin de mettre en contact des experts avec les équipes médicales de proximité. En plus de 20 ans d'expérience, cette initiative a généré une base de données considérable de plus de 133 000 images photographiques de plaies chroniques de tout type (escarre, ulcère, plaie du pied diabétique, etc...). Cette base de données représente une ressource précieuse mais sous-exploitée en raison du manque de standardisation du protocole d'acquisition et d'annotations disponibles. Des exemples provenant de cette base de données sont visibles avec la figure 1.

La segmentation revêt une importance cruciale dans la prise en charges des plaies chroniques car la réalisation de leurs calques tout au long du parcours de soin aide le corps médical à évaluer l'efficacité des traitements choisis et ainsi valider ou réfuter la pertinence du diagnostic établi. Cependant, le détourage manuel est une tâche complexe entraînant des écarts mesurables tant entre les annotateurs qu'entre les différentes propositions d'un même annotateur, même lorsque ces derniers sont des experts. Par ailleurs, la base de données Cicat-Occitanie est caractérisée par la diversité du matériel d'acquisition (smartphones utilisés), des scènes (domiciles différents avec des variations dans l'éclairage, la prise de vue, l'arrière-plan, la distance entre le smartphone et la plaie) et des plaies (de nature et de localisation variées), ce qui complique la tâche.

Les méthodes par apprentissage profond représentent aujourd'hui l'approche privilégiée pour la segmentation des plaies chroniques. Des compétitions internationales comme

le DFUC pour Diabetic Foot Ulcer Challenge [1] rendent disponibles des bases de données de plusieurs milliers d'images acquises en conditions hospitalières et annotées en segmentation par des experts. Cette compétition a d'ailleurs mis en évidence les performances du modèle HardNet-MSEG [6]. Toutefois, l'absence d'initiative similaire pour des images hétérogènes dites "into the wild", limite aujourd'hui le développement d'approches supervisées suffisamment robustes face à la diversité des cas cliniques.

Pour surmonter cet obstacle, l'apprentissage auto-supervisé (self-supervised Learning, SSL) apparaît comme une piste prometteuse car il permet aux réseaux de neurones d'apprendre à représenter plus efficacement les images sans nécessiter de supervision humaine. En particulier, la méthodologie DINO pour Distillation with No labels [1, 7] a mené à l'élaboration de l'immense base de données LVD-142M à partir de laquelle l'encodeur ViT (Vision Transformer) a été entraîné [3]. Cependant, cette base de données générique ne représente pas bien les images cliniques et l'état de l'art manque de références quant aux bénéfices apportées par son utilisation dans ce domaine spécifique.

Cet article vise à explorer, dans un contexte clinique spécifique de segmentation de plaies chroniques avec peu d'annotations, l'intérêt de l'encodeur générique ViT préentraîné avec DINO sur LVD-142M face à l'encodeur HardNet-MSEG, plus léger et initialisé aléatoirement. Nous nous intéressons également à l'effet que produit un préentraînement SSL DINO sur les données cibles effectué avant la tâche finale de segmentation. Enfin, nous mesurons l'impact de la quantité de données disponible sur les performances des différents scénarios d'entraînements proposés.

FIGURE 1 – 4 exemples illustrant la diversité des photographies de plaies chroniques en terme de localisation, de nature et de conditions d'acquisition.

2 Etat de l'art

Le SSL est une approche où un encodeur est entraîné durant une tâche dite prétexte qui, au lieu d'utiliser des annotations humaines, se fonde sur des labels générés automatiquement à partir de la donnée elle-même [8, 11] disponible en grande quantité. Suite à ce préentraînement, les poids de l'encodeur servent d'initialisation pour l'apprentissage sur une tâche finale. Parmi l'ensemble des méthodes SSL présente dans l'état de l'art, nous nous intéressons ici à l'approche discriminative illustrée par la tâche prétexte DINO proposée par META.

DINO utilise 2 encodeurs d'architecture identique, dont l'un est appelé élève et l'autre enseignant. Pour la tâche prétexte, une stratégie multi-crop est appliquée à l'image d'entrée : 2 vues "globales", dont la surface représente au moins la moitié de l'image d'origine, et n vues "locales", ayant une surface inférieure à 50 %, sont générées. L'encodeur enseignant recevra les 2 vues "globales" tandis que l'élève verra toutes les vues. Chacune de ces vues est augmentée différemment à l'aide de transformations spatiales et colorimétriques. Lors de l'apprentissage, les deux encodeurs produisent une représentation de ces vues qui seront transmises à leur tête de projection respective consistant à une suite de couches linéaires (MLP). Les représentations sont alors comparées par entropie croisée et les poids de l'élève sont mis à jour par rétropropagation du gradient. Les poids de l'enseignant sont eux mis à jour via une moyenne mobile exponentielle à partir de ceux de l'élève.

A l'aide de la base de données soigneusement assemblée LVD-142M, SSL DINO a permis l'entraînement d'encodeurs ViT de différentes échelles (small 21 M, large 307 M, giant 1100 M de paramètres) voir [2]. Les Transformers sont des architectures imposantes, gourmandes en ressources informatiques et ne se démarquent des approches convolutives que lorsque les bases de données sont de très grandes tailles.

HardNet, pour Harmonic Densely Connected Network [2], est une architecture convolutive améliorée de l'architecture DenseNet [4] dont le but est de réduire le temps d'inférence sans réduire les performances de l'encodeur. Pour ce faire, le nombre et la position des connections résiduelles au sein des blocs de convolutifs ont été modifiés. Dans le cadre de la compétition DFUC2022 [5], cet encodeur a été amélioré et connecté à un décodeur de segmentation appelé Lawin pour Large window attention [10]. Cette proposition nommée HarDNet-MSEG a atteint la première place de la compétition de segmentation, rendant de facto cette architecture intéressante dans l'analyse des plaies chroniques.

3 Préparation des bases de données

Un Faster RCNN [9] modifié a été entraîné comme détecteur de plaies via les données de la compétition DFUC2020 [3]. Il a été appliqué sur la base de données du réseau Cicat-Occitanie rassemblant plus de 133 000 images. Seules les images n'ayant qu'une seule plaie prédite sont conservées, soit 88 727 images venant constituer la base de données B_1. Nous sélectionnons alors aléatoirement 400 images dans B_1 afin que deux experts les annotent en segmentation manuellement à l'aide de l'outil labelme [4] constituant ainsi B_2. L'élaboration de B_1 et B_2 est illustrée dans la figure 2.

B_1 est découpée en 3 catégories dédiées aux préentraînements SSL qui utiliseront le même protocole que DINO : l'entraînement, la validation et le test avec un ratio respectif de 70%, 20%, 10%.

Les images B_2 sont issues de la catégorie test de B_1 et sont utilisées durant la tâche finale de segmentation. Elles sont réparties en 5 folds de division 70%, 10%, 20%. Afin d'évaluer l'impact de la quantité de données

1. https ://dfuc2022.grand-challenge.org/
2. https ://github.com/facebookresearch/dinov2
3. https ://dfu2020.grand-challenge.org/
4. https ://github.com/labelmeai/labelme

d'entraînement, 3 copies de ces 5 folds ont été réalisées. Chaque copie se voit ôter un certain nombre d'images d'entraînement choisies aléatoirement tandis que les parties validation et test restent inchangées. Finalement, nous obtenons 3 groupes de 5 folds où le pourcentage de données d'entraînement est respectivement de 70%, 50% et 25% de B_2.

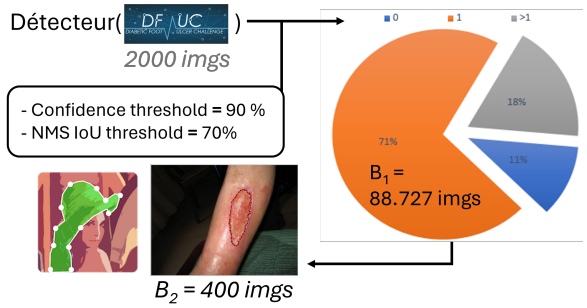

FIGURE 2 – Protocole de filtrage de la base de données Cicat-Occitanie : B_1 est réservée au SSL et B_2 à la segmentation.

4 Expériences

4.1 Scénarios d'entraînements

Concernant le choix des encodeurs, nous avons choisi les configurations ViT**s**14_reg (21 M) et ViT**l**14_reg (307 M) afin d'observer l'effet du changement d'échelle de la taille de l'encodeur. Leurs poids initiaux sont ceux issus de l'article Dinov2 [7]. Ces deux encodeurs sont comparés avec celui issu de HardNet-MSEG (3 M) dont l'initialisation est aléatoire.

La figure 3 résume les scénarios d'entraînement évalués. Les encodeurs peuvent être préentraînés ou non via la méthode SSL DINO sur B_1. Durant la tâche finale, les poids des encodeurs sont soit figés, soit optimisés au travers d'une stratégie de décongélation des poids. Par limite de mémoire GPU, les poids de l'encodeur ViT**l**14_reg n'ont pas pu être optimisés durant la tâche de segmentation.

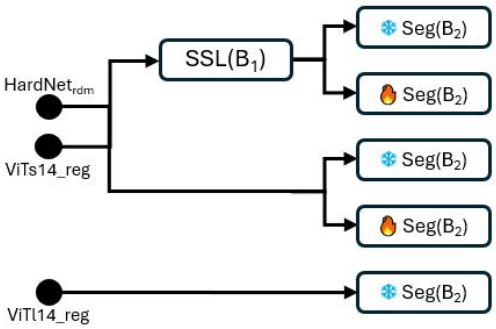

FIGURE 3 – Les scénarios d'entraînements explorés : le flocon signifie que l'encodeur reste figé tandis que la flamme désigne son optimisation via la décongélation des poids.

4.2 Implémentation

Les expériences décrites ont été réalisées à l'aide d'une carte graphique NVIDIA RTX A6000 de 48 Go de mémoire. Les entraînements SSL à la DINO sont réalisés à l'aide de la librairie lightly [5]. Une augmentation à la volée est effectuée durant les 300 époques des entraînements SSL avec un mini-batch de 128 images. Chacune aboutit à la création de 8 vues. La résolution des 2 vues "globales" est fixée à 224x224 et à 98x98 pour les 6 vues "locales". La tête de projection est composée de 3 couches linéaires. Sa dimension d'entrée dépend de la dimension du tenseur de sortie de chaque encodeur tandis que les dimensions des autres couches restent inchangées entre les expériences et valent respectivement 512, 128 et 2048.

Pour les entraînements supervisés sur B_2 suivant un préentraînement SSL sur B_1, les poids de départ des encodeurs correspondent à ceux ayant minimisé la fonction de coût sur les données de validation SSL. La segmentation est réalisée sur 150 époques en 5 folds cross validation avec le décodeur Lawin : 4 tenseurs caractéristiques sont extraits de l'encodeur et sont adaptés aux 4 entrées attendues du décodeur. La taille du mini-batch dépend de la taille mémoire GPU occupée par l'encodeur : 12 pour HardNet-MSEG et pour les encodeurs ViT lorsque leurs poids sont congelés mais 2 pour ViT**s**14_reg lors de sa décongélation progressive. Les performances sur l'ensemble test seront évaluées avec la métrique Dice, une mesure couramment utilisée en segmentation pour évaluer la similarité entre la prédiction du modèle et la vérité terrain. Elle se calcule par la formule suivante où *pred* désigne les pixels prédits par le modèle comme appartenant à la plaie et *vt* les pixels désignés par l'annotateur comme appartenant à une plaie :

$$DICE = \frac{|pred \cap vt|}{|pred| + |vt|}$$

4.3 Résultats

Durant les préentraînements SSL, un phénomène de surapprentissage apparaît au bout d'une centaine d'époques, quel que soit l'encodeur. La stratégie d'arrêt précoce est réglée à 30 époques pour limiter le temps de calcul. Les durées des préentraînements ont été d'environ 30 h pour une centaine d'époque. Quel que soit le scénario, l'optimisation des algorithmes utilisant HardNet varie entre 1 h et 2 h en fonction de la quantité de données. Le temps d'optimisation des scénarios avec ViT**s**14_reg varie entre 1 h et 3 h en fonction de la quantité de données mais aussi de l'état figé ou décongelé de l'encodeur. Ces temps sont similaires pour l'optimisation du scénario avec ViT**l**14_reg. Le tableau 1 présente les performances obtenues par les différents scénarios d'entraînements proposés en fonction de la métrique Dice sur la tâche de segmentation sur B_2.

4.4 Discussion

Dans le tableau 1, les lignes 6 et 8 montrent que l'augmentation de l'échelle du modèle ViT préentraînée "à la DINO" permet une amélioration des performances. Le passage de ViT**s**14_reg à ViT**l**14_reg est donc lié à une amélioration des capacités d'extraction des caractéristiques des images de plaies chroniques. Cependant, d'après les lignes 3 et 8, un modèle convolutif léger tel que HardNet-MSEG, optimisé sur B_1 et sans préentraînement SSL sur B_2 au préalable,

5. https ://github.com/lightly-ai/lightly

Encodeur	SSL(B_1)	Optimisation (B_2)	Train=25%	Train=50%	Train=70%	ligne
HardNet-MSEG$_{rdm}$	✔	❄	0.724±0.032	0.737±0.011	0.755±0.017	1
		🔥	0.756±0.034	0.784±0.009	0.798±0.013	2
	✘	🔥	0.694±0.038	0.736±0.033	0.771±0.023	3
ViTs14_reg	✔	❄	0.594±0.055	0.644±0.035	0.665±0.034	4
		🔥	0.674±0.017	0.709±0.023	0.720±0.025	5
	✘	❄	0.571±0.043	0.646±0.028	0.653±0.020	6
		🔥	0.685±0.025	0.721±0.021	0.729±0.006	7
ViTl14_reg	✘	❄	0.637±0.042	0.639±0.024	0.701±0.032	8

TABLE 1 – Performances en DICE des modèles sur la tâche de segmentation de B_2 : le flocon signifie que l'encodeur reste figé tandis que la flamme désigne son optimisation via la décongélation des poids.

prévaut sur ViTl14_reg initialisé via SSL DINO sur LVD-142M, quel que soit la quantité de données d'entraînement durant la tâche finale supervisée. Cela signifie que LVD-142M n'est pas adaptée pour des applications cliniques aussi spécifiques que l'analyse des plaies chroniques. Par ailleurs, quel que soit le choix de l'encodeur, de l'optimisation sur B_2 et de la quantité de données d'entraînement de B_2, un préentraînement via la méthodologie SSL DINO sur les images de plaies chroniques B_1 s'accompagne d'une amélioration de la métrique Dice en segmentation. Cette amélioration peut être source d'économies d'annotations. En effet, comme le montre les lignes 2 et 3, le modèle HardNet-MSEG préentraîné en SSL DINO sur B_1 puis optimisé en segmentation sur B_2 avec uniquement 25% de données annotées obtient des performances similaires avec le même encodeur mais sans préentraînement et ayant 70% de données annotées pour son entraînement en segmentation.

5 Conclusion

Dans cet article, nous avons évalué l'intérêt d'utiliser un encodeur entraîné "à la DINO" sur une base de données clinique spécifique puis sur une tâche de segmentation. Nous avons comparé dans différents scénarios d'entraînements les performances de l'encodeur ViT, dont les poids sont issus de l'article Dinov2, à l'encodeur léger HardNet-MSEG, dont les poids sont initialisés aléatoirement. Les résultats montrent qu'il n'est pas utile d'employer des architectures DINO préentraînées sur LVD-142M car des modèles légers peuvent être supérieurs sur des tâches spécifiques. De plus, le préentraînement d'un encodeur par du SSL "à la DINO" sur une base spécifique avec peu d'annotations présente un intérêt. Il serait intéressant de poursuivre ce travail en étudiant l'impact de l'augmentation de la base de données pour l'entraînement SSL grâce à l'ajout de toutes les bases de données publiques liées aux lésions dermatologiques.

Nous remercions l'ANRT ainsi que le réseau Cicat-Occitanie pour financer et soutenir la thèse CIFRE.

Références

[1] Mathilde CARON et al. "Emerging properties in self-supervised vision transformers". In : *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, p. 9650-9660.

[2] Ping CHAO et al. "Hardnet : A low memory traffic network". In : *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, p. 3552-3561.

[3] Alexey DOSOVITSKIY et al. "An image is worth 16x16 words : Transformers for image recognition at scale". In : *arXiv preprint arXiv :2010.11929* (2020).

[4] Gao HUANG et al. "Densely connected convolutional networks". In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 4700-4708.

[5] Connah KENDRICK et al. "Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation". In : *arXiv preprint arXiv :2204.11618* (2022).

[6] Ting-Yu LIAO et al. "HarDNet-DFUS : Enhancing Backbone and Decoder of HarDNet-MSEG for Diabetic Foot Ulcer Image Segmentation". In : *Diabetic Foot Ulcers Grand Challenge*. Springer, 2022, p. 21-30.

[7] Maxime OQUAB et al. "Dinov2 : Learning robust visual features without supervision". In : *arXiv preprint arXiv :2304.07193* (2023).

[8] Utku OZBULAK et al. "Know Your Self-supervised Learning : A Survey on Image-based Generative and Discriminative Training". In : *arXiv preprint arXiv :2305.13689* (2023).

[9] Shaoqing REN et al. "Faster r-cnn : Towards real-time object detection with region proposal networks". In : *Advances in neural information processing systems* 28 (2015).

[10] Haotian YAN et al. "Lawin transformer : Improving semantic segmentation transformer with multi-scale representations via large window attention". In : *arXiv preprint arXiv :2201.01615* (2022).

[11] Chuyan ZHANG et al. "Dive into the details of self-supervised learning for medical image analysis". In : *Medical Image Analysis* 89 (2023), p. 102879.