Plan prévisionnel

Dataset retenu

Le jeu de données utilisé pour ce projet est Cityscapes, une référence en segmentation sémantique urbaine. Il se compose de 5 000 images RGB haute résolution (2048×1024) prises dans 50 villes allemandes et annotées pixel par pixel. Chaque pixel est associé à une classe sémantique de 0 à 33 représentant un élément de la scène (route, trottoir, voiture, piéton, etc.).

Les 5000 images sont réparties comme suit :

- Train : 2975 images
- Validation : 500 images
- Test : 1525 images

Modèle envisagé

1. Mask2Former

Le modèle que je propose d'utiliser est Mask2Former, une architecture de segmentation unifiée introduite par Facebook AI Research en 2022. Ce modèle repose sur une approche basée sur les Transformers et permet de traiter la segmentation sémantique, panoptique et instance dans un seul framework.

Mask2Former s'est imposé comme une baseline de référence dans la littérature récente. Il a démontré d'excellentes performances sur des datasets tels que Cityscapes (*Cityscape-Adverse: Benchmarking Robustness of Semantic Segmentation with Realistic Scene Modifications via Diffusion-Based Image Editing, Naufal Suryanto et al, Nov 2024*), ADE20K ou COCO.

Le benchmark de robustesse cité ci-dessus a évalué plusieurs familles de modèles (CNN, Transformers, hybrides) face à des perturbations réalistes simulées sur Cityscapes (pluie, nuit, brouillard, flou, neige, etc.). Mask2Former s'est distingué comme l'un des modèles les plus robustes, avec une IoU macro élevée d'environ 82% sur les images originelles de Cityscapes, tout en montrant une meilleure stabilité sous perturbation que les CNNs traditionnels.

Sur un plan technique, Mask2Former repose sur un décodeur Transformer qui génère des masques prédictifs à partir d'un ensemble de requêtes dynamiques, permettant une segmentation sémantique fine, précise et respectueuse des contours d'objets et des relations contextuelles.

Ce modèle est particulièrement adapté à des contextes comme :

- La conduite autonome, où les objets sont nombreux, variés, et parfois partiellement visibles.
- La robotique mobile, avec besoin de compréhension spatiale précise.
- La vidéosurveillance intelligente ou les systèmes de cartographie urbaine.

Enfin Mask2Former est également un modèle référencé dans de nombreux articles, intégré dans des frameworks Open Source (Detectron2 de Meta AI) et cité dans des conférences majeures (CVPR, ICCV, NeurIPS). Ces éléments en font un candidat pertinent dans le cadre de ce projet.

Cependant, Mask2Former n'est pas disponible en version non pré-entrainé dans les principales bibliothèques publiques. Ce point ne permettant pas une expérimentation rigoureuse, comme souhaité dans le cadre de ce projet, un deuxième modèle sera testé.

2. SegFormer

Afin de répondre à la contrainte méthodologique d'utiliser un modèle entraînable sans pré-entraînement, je propose de compléter l'expérimentation avec un second modèle : SegFormer, une architecture légère et efficace introduite par Nvidia en 2021.

Contrairement à d'autres modèles de type Transformer qui nécessitent des tuiles d'image ou de fortes résolutions intermédiaires, SegFormer combine un encodeur hiérarchique efficace MiT (ou Mix Transformer) avec un décodeur MLP (Multi-Level Perception Head) simple et rapide, ce qui en fait une solution particulièrement adaptée à l'entraînement from scratch, même avec des ressources limitées.

Le choix de SegFormer repose sur les points clés suivants qui le rendent prometteur en performance dans le contexte de ce projet

- Architecture compacte et sans positional encoding

  Grâce à son encodeur hiérarchique basé sur le Mix Transformer, SegFormer capture les relations spatiales sans utiliser de positional encoding, ce qui le rend plus léger et plus rapide à converger.

- Pas de contraintes de résolution d'entrée

  Contrairement à DeepLab ou UPerNet, SegFormer fonctionne sans modification sur des images de résolution variable, ce qui simplifie le prétraitement des données Cityscapes.

- Performances très compétitives

  Dans l'article original, SegFormer atteint un mIoU de 82.37 % sur Cityscapes tout en restant très rapide à l'inférence. Cela le positionne au niveau des meilleures solutions actuelles, avec un rapport précision/vitesse très favorable.

- Modèle facilement entraînable from scratch

  L'architecture modulaire permet d'entraîner SegFormer depuis zéro (sans poids pré-entraînés), tout en offrant une convergence stable et des performances correctes dès les premières époques.

Références bibliographiques

Pour construire ce projet de segmentation sémantique, j'ai mené une veille sur les évolutions récentes du domaine, à partir de sources fiables et reconnues. Voici les références clés sur lesquelles je m'appuierai pour orienter mes choix :

1. **Mask2Former: Masked-attention Mask Transformer for Universal Image Segmentation** (*Bowen Cheng et al., arXiv, 2022* – arXiv:2112.01527)

Cet article introduit Mask2Former, un modèle innovant qui unifie les approches de segmentation sémantique, panoptique et par instance. Grâce à l'utilisation de Transformers et de masked attentions qui restreignent l'attention à l'intérieur des régions d'intérêt, il surpasse les méthodes classiques de CNNs comme DeepLabV3+ sur plusieurs benchmarks dont Cityscapes, ADE20K et COCO.

2. **Cityscapes-Adverse: Benchmarking Robustness of Semantic Segmentation with Realistic Scene Modifications via Diffusion-Based Image Editing** (*Suryanto et al., arXiv, novembre 2024* – arXiv:2411.00425)

Ce travail propose un benchmark de robustesse sur le dataset Cityscapes, en y ajoutant des perturbations réalistes générées par modèles de diffusion (pluie, neige, nuit, flou). Il met en évidence la résilience de certains modèles récents, dont Mask2Former, face à des conditions d'environnement dégradées, ce qui est crucial pour des cas d'usage en conduite autonome.

3. **SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers** (Enze Xie et al., NeurIPS 2021 – arXiv:2105.15203)

Cet article introduit SegFormer qui propose un backbone MiT qui permet de combiner la précision des Transformers avec la compacité et la hiérarchie des CNNs + decodeur MLP. Performant et rapide, sans besoin de positional encoding. SegFormer est très compétitif même sans pré-entraînement.

DataSpace

Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)

Dans le cadre de ce projet, ma démarche s'inscrit dans une logique d'amélioration du travail réalisé lors du projet 8 de la formation IA Engineer, dans lequel j'ai implémenté la segmentation sémantique sur le dataset Cityscapes à l'aide d'un modèle U-Net. J'y avais testé diverses variantes (encodeur pré-entraîné VGG16, data augmentation, tuning d'hyperparamètres) et établi une baseline robuste sur Cityscapes. Les performances du modèle (mIoU, précision par classe, etc.) sur l'ensemble de validation Cityscapes me fourniront un point de comparaison fiable pour évaluer l'intérêt d'un modèle plus avancé.

Dans ce nouveau projet, je propose de mettre en œuvre :

- **Mask2Former**, utilisé tel que disponible, avec pré-entraînement.
- **SegFormer,** sans pré-entraînement, pour une comparaison non biaisée.

Le projet prendra la forme d'une preuve de concept (PoC) :

- Je réaliserai une expérimentation avec entrainement + validation de SegFormer non pré-entrainé,
- Je réaliserai une seconde expérimentation avec Mask2Former pré-entrainé sur un sous-ensemble de Cityscapes,
- Je comparerai les résultats obtenus (visuellement et quantitativement) avec ceux de la baseline U-Net du projet 8, afin de valider l'intérêt du modèle avancé,
- Enfin, je présenterai la comparaison des résultats dans un dashboard Streamlit que je déploierai.

Cette démarche permettra de démontrer la faisabilité technique et la valeur ajoutée de Mask2Former et de SegFormer, un modèle plus léger, dans le contexte de la segmentation sémantique urbaine.