Étude de la Fusion des Représentations Latentes Texte-Image en IA Générative pour l'Imagerie Médicale

Proposition de Thèse - École Doctorale EEA

2025

1 Contexte

L'intégration des modalités texte et image dans les systèmes d'intelligence artificielle a révolutionné de nombreux domaines. En imagerie médicale, cette approche multimodale promet d'améliorer significativement le diagnostic et le suivi des pathologies en associant les informations visuelles des images médicales (IRM, TDM) à des descriptions textuelles cliniques détaillées. Les modèles comme CLIP [Radford et al., 2021] ont montré leur efficacité dans des contextes généraux, tandis que MedCLIP [Wang et al., 2022] a été spécialement adapté pour les images médicales. Cependant, leur application nécessite encore des adaptations spécifiques en raison de la faible représentativité des données, de leur complexité et de leur hétérogénéité.

Dans des pathologies complexes comme la sclérose en plaques (SEP), les biomarqueurs tels que l'évolution des lésions dans la matière blanche, l'atrophie des structures profondes (thalamus, noyaux caudés), et les lésions corticales jouent un rôle clé [Kalinin et al., 2020, Treaba et al., 2019]. Une meilleure exploitation des représentations texte-image pourrait faciliter l'analyse longitudinale de ces biomarqueurs et renforcer l'explicabilité des modèles génératifs.

2 État de l'art

Les récents travaux en vision-langage appliqués à l'imagerie médicale se concentrent sur plusieurs axes :

Génération d'images par prompts textuels

Des approches comme les modèles de diffusion latente (LDM) [Rombach et al., 2022] utilisent des descriptions textuelles pour guider la génération d'images. Ces modèles montrent un potentiel pour produire des images réalistes adaptées aux besoins cliniques, ainsi que pour révéler des schémas pathologiques cachés [Xing et al., 2024] .

Fusion multimodale et modèles de fondation

Les modèles de fondation multimodaux (MFM) intègrent des données visuelles et textuelles dans un espace latent partagé [Li et al., 2024]. Ces modèles, tels que Med-PaLM et Visual ChatGPT, peuvent améliorer l'interprétation et l'analyse des images médicales grâce à une explicabilité renforcée par le texte et l'image. La correspondance entre le texte et l'image générée est un défi majeur. Les métriques actuellement utilisées comme FID et BLEU-1 sont à adapter aux spécificités des données médicales.

Segmentation

SAM [Kirillov et al., 2023] et sa variante adaptée aux images médicales, MedSAM [Ma et al., 2024], offrent des solutions prometteuses pour des tâches de segmentation complexes. Néanmoins, leur capacité à s'adapter aux diverses applications médicales reste un domaine de recherche actif, en particulier pour des applications multimodales.

Génération de descriptions textuelles et explicabilité

Un modèle génératif combinant traitement du langage naturel et apprentissage profond peut produire des descriptions automatiques d'images médicales [Barreto et al., 2023]. Cependant, la qualité des descriptions générées reste encore inférieure à celle des experts humains. Des architectures avancées doivent être explorées pour renforcer l'explicabilité des modèles et inspirer plus de confiance dans leur utilisation clinique.

Malgré ces avancées, les interactions complexes entre texte et image dans l'espace latent restent insuffisamment explorées, notamment en termes de contrôle des contributions respectives des deux modalités.

3 Description de la recherche envisagée

Cette thèse vise à étudier la combinaison des représentations texte-image dans l'espace latent pour des applications en IA générative médicale. Les axes principaux sont :

- **Génération d'images médicales tridimensionnelles** : Créer des modèles capables de produire des images réalistes à partir de prompts textuels, incluant des descriptions de pathologies spécifiques (lésions, atrophie).
- **Analyse et segmentation avec génération de rapports** : Automatiser la création de rapports radiologiques explicatifs basés sur l'analyse d'images médicales.
- **Explicabilité** : Développer des mécanismes permettant de visualiser et d'expliquer l'impact des modalités texte et image sur les décisions des modèles.
- **Modélisation temporelle** : Simuler l'évolution longitudinale des biomarqueurs, comme l'apparition de nouvelles lésions et la réduction progressive de la matière grise profonde dans la SEP.

4 Résultats attendus

- **Représentations latentes robustes** : Développer un espace latent partagé texte-image facilitant la génération et l'analyse d'images médicales.
- **Amélioration des jeux de données** : Générer des données synthétiques pour renforcer l'apprentissage supervisé dans des contextes médicaux spécifiques.
- **Impact clinique** : Fournir des outils pour analyser et suivre l'évolution des biomarqueurs dans des pathologies comme la SEP, contribuant ainsi à un diagnostic et un suivi plus précis.
- **Explicabilité accrue** : Proposer des visualisations et des rapports textuels facilitant l'interprétation des décisions des modèles par les cliniciens.

5 Références Bibliographiques

References

Artur Gomes Barreto, Juliana Martins de Oliveira, Francisco Nauber Bernardo Gois, Paulo Cesar Cortez, and Victor Hugo Costa de Albuquerque. A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks. *Bioengineering*, 10(9):1098, 2023.

I Kalinin, G Makshakov, and E Evdoshenko. The impact of intracortical lesions on volumes of subcortical structures in multiple sclerosis. *American Journal of Neuroradiology*, 41(5): 804–808, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Constantina A Treaba, Tobias E Granberg, Maria Pia Sormani, Elena Herranz, Russell A Ouellette, Céline Louapre, Jacob A Sloane, Revere P Kinkel, and Caterina Mainero. Longitudinal characterization of cortical lesion development and evolution in multiple sclerosis with 7.0-t mri. *Radiology*, 291(3):740–749, 2019.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Xiaodan Xing, Junzhi Ning, Yang Nan, and Guang Yang. Deep generative models unveil patterns in medical images through vision-language conditioning. *arXiv preprint arXiv:2410.13823*, 2024.