Sujet de stage M2 recherche : Modèles multimodaux et distillation

Distillation de modèles multimodaux pour la recherche texte-image d'œuvres d'arts. Comparaison de modèles multimodaux lourds et distillés.

La recherche d'images à partir de requêtes formulées en langage naturel est devenue un enjeu central pour l'accès aux collections numériques d'œuvres d'art. Des modèles d'intelligence artificielle dits vision–langage, tels que CLIP (Radford et al., 2021) ou BLIP-2 (Li et al., 2023), permettent aujourd'hui d'associer efficacement du texte à des images, ouvrant la voie à des interfaces de recherche plus intuitives pour les musées, les institutions patrimoniales et les amateurs d'art. Toutefois, ces modèles récents sont souvent volumineux, coûteux en ressources de calcul et difficiles à déployer dans des contextes opérationnels contraints.

Ce stage s'inscrit dans ce contexte et vise à évaluer l'intérêt de modèles multimodaux plus légers, obtenus par des techniques de distillation, pour des tâches de recherche texte–image appliquées aux œuvres d'art. À partir d'un corpus d'environ 60 000 peintures, accompagnées de descriptions textuelles et de métadonnées structurées, le ou la stagiaire mettra en place un pipeline de recherche de référence basé sur un modèle vision–langage de grande taille, puis comparera ses performances à celles d'un ou plusieurs modèles distillés, plus compacts.

La distillation de connaissances (Hinton et al., 2015) permet de transférer, dans un contexte restreint et des applications spécifiques, les capacités d'un modèle large vers un modèle compact, tout en préservant une partie significative des performances pour le contexte et les applications retenus.

Le travail consistera à concevoir des protocoles expérimentaux permettant de mesurer à la fois la qualité de la recherche (pertinence des images retournées, diversité des résultats) et l'efficacité computationnelle des modèles (temps de réponse, consommation mémoire). Une analyse critique des résultats permettra d'identifier les compromis entre performances et coûts, ainsi que les scénarios d'usage dans lesquels des modèles distillés constituent une alternative crédible aux modèles lourds. Le stage donnera également lieu à la réalisation d'un prototype simple de moteur de recherche texte–image, illustrant concrètement les différentes approches étudiées.

Objectifs

Évaluer quantitativement et qualitativement la pertinence d'un modèle distillé par rapport à son modèle enseignant pour une tâche de recherche multimodale d'images de peintures via interface conversationnelle.

1. Concevoir et implémenter un pipeline de distillation adapté aux modèles vision-langage pour le domaine artistique
2. Développer une architecture de recherche intégrant le graphe de connaissances du corpus
3. Comparer les performances de plusieurs méthodes selon des métriques adaptées
4. Analyser le compromis performance/efficacité computationnelle des modèles distillé

Livrables

- Prototype fonctionnel de moteur de recherche
- Code documenté et reproductible
- Jeu de requêtes d'évaluation pour la recherche texte–image
- Mémoire de niveau recherche (état de l'art, analyse des résultats), présentations orales

Informations pratiques

- Lieu : CNAM, Paris – Laboratoire CEDRIC (équipe ISID), en collaboration avec Télécom Paris (équipe IDS)
- Durée : 4 à 6 mois
- Période : démarrage entre février et avril 2026
- Encadrement : Nada Mimouni (CNAM, laboratoire CEDRIC, équipe ISID), Jean-Claude Moissinac (Telecom Paris, équipe Images, Données, Signal (IDS))

Profil recherché

- Étudiant(e) en Master 2 en intelligence artificielle, data science, vision par ordinateur ou domaine proche
- Bonnes bases en Python
- Connaissances en apprentissage automatique et deep learning
- Intérêt pour les modèles multimodaux texte–image et les applications patrimoniales

Candidature

Les candidatures (CV, relevés de notes et court message de motivation) sont à adresser à Jean-Claude Moissinac (jean-claude.moissinac@telecom-paris.fr) avec en copie Nada Mimouni (nada.mimouni@cnam.fr)

Références

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning* (pp. 8748-8763). PMLR. https://arxiv.org/pdf/2103.00020

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning* (pp. 19730-19742). PMLR. https://arxiv.org/pdf/2301.12597

Garcia, N., Renoust, B., & Nakashima, Y. (2020). Context-aware embeddings for automatic art analysis. *Proceedings of the 2020 International Conference on Multimedia Retrieval* (pp. 25-33). https://arxiv.org/abs/1904.04985

Mots-clés : Distillation de modèles, Vision-Langage, Recherche multimodale, Graphe de connaissances, Patrimoine culturel