Stage ingénieur en Computer Vision / NLP (Safran E&D)

En tant qu'êtres humains, nous avons la capacité de produire une représentation visuelle d'un objet en nous appuyant uniquement sur des informations textuelles. Cette aptitude nous a conduit à nous poser quelques questions :

1. **Les modèles de type LLM (Large Language Models) peuvent-ils générer des représentations relatives à des informations visuelles à partir de descriptions textuelles uniquement ?** Certains travaux intéressants, explorent cette question [1].
2. **Les modèles multimodaux comme CLIP créent-ils un espace de représentation véritablement partagé pour le texte et les images, ou s'agit-il simplement d'un espace dans lequel cohabitent ces deux modalités ?** Autrement dit, les représentations textuelles et visuelles se rapprochent-elles selon une mesure de similarité, mais sans réelle fusion ? Si l'on pouvait prouver que ces modèles produisent effectivement un espace de représentation véritablement partagé, cela ouvrirait la porte à une exploration encore plus intéressante :
3. **Un modèle multimodal entraîné à reconnaître les mêmes classes dans des ensembles d'images et de texte pourrait-il améliorer ses performances visuelles en bénéficiant d'une augmentation de données purement textuelle ?** Cette possibilité, si elle s'avère viable, permettrait de repousser les limites de la data augmentation classique qu'on utilise en renforçant les capacités du modèle sur la partie image grâce à une information textuelle supplémentaire.

Approche multimodale pour l'amélioration des performances en vision via de la data augmentation textuelle

L'objectif est de proposer une approche pour la classification d'images en exploitant simultanément des données textuelles via un réseau de neurones multimodal. La méthode repose sur l'apprentissage d'un espace de représentation **partagé** entre les modalités texte et image, permettant d'**améliorer les performances du réseau sur la tâche de vision** par une augmentation indirecte des données via une augmentation des données textuelles.

1. Création de la base de données image/texte

La première étape de ce projet consiste à créer une base de données de classification image/texte. Pour ce faire, une base de données existante de classification d'images sera utilisée. Pour chaque image, une description textuelle détaillée sera générée à l'aide d'outils automatiques de description d'images, tels que ChatGPT. Cette démarche permet de produire des descriptions riches et informatives qui capturent les détails visuels, les contextes et les significations des images.

2. Architecture multimodale

Dans un deuxième temps, un réseau de neurones multimodal est construit, comprenant deux branches principales :

- **La branche image** : celle-ci se compose d'un modèle de réseau convolutif traditionnel (CNN) conçu pour extraire les caractéristiques visuelles de l'image.
- **La branche texte** : un modèle de type Transformer sera utilisé pour extraire les représentations sémantiques du texte.

Ces deux branches sont fusionnées dans des **couches de projection communes** qui permettent d'apprendre un espace de représentation partagé entre les images et les textes. Le réseau est ensuite entraîné de manière conjointe sur les deux ensembles de données distincts : un ensemble d'images et

un ensemble de textes, tous deux étiquetés avec les mêmes classes. L'entraînement initial garantit que le réseau est capable de classifier les images et les textes avec une compréhension alignée des deux modalités.

3. Data augmentation via réentraînement multimodal

Une fois l'entraînement initial validé sur des jeux de test composés d'images et de textes, une nouvelle phase de réentraînement est proposée pour exploiter pleinement le potentiel de l'architecture multimodale. Dans cette phase, le jeu de données textuelles est considérablement enrichi avec de nouvelles informations, sans pour autant augmenter la taille du jeu d'images. Cela est rendu possible grâce à l'espace de représentation partagé : en apprenant de nouvelles informations textuelles, le réseau peut adapter ses représentations communes, ce qui a pour effet d'améliorer également les capacités de classification visuelle.

Références

1. Sharma, P., Shaham, T. R., Baradad, M., Fu, S., Rodriguez-Munoz, A., Duggal, S., ... & Torralba, A. (2024). A vision check-up for language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14410-14419).

2. Xu, N., Mao, W., Wei, P., & Zeng, D. (2020). MDA: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks. *IEEE Intelligent Systems*, *36*(6), 3-12.

3. Xiao, C., Xu, S. X., & Zhang, K. (2023, November). Multimodal data augmentation for image captioning using diffusion models. In *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications* (pp. 23-33).

4. Hao, X., Zhu, Y., Appalaraju, S., Zhang, A., Zhang, W., Li, B., & Li, M. (2023). Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 379-389).

5. Wang, W., Yang, Z., Xu, B., Li, J., & Sun, Y. (2023). ViLTA: Enhancing vision-language pre-training through textual augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3158-3169).

Profil du candidat

Le candidat idéal est un étudiant en dernière année de Master ou d'école d'ingénieur, spécialisé en intelligence artificielle (vision par ordinateur, ou traitement du langage naturel). Il doit maîtriser les réseaux de neurones profonds (DNN), avoir de l'expérience avec des frameworks de deep learning comme PyTorch ou TensorFlow.

Compétences requises

Curiosité intellectuelle et goût pour la recherche
Excellentes bases en Mathématiques appliquées, Machine Learning, IA
Forte motivation pour le ML appliquée à la vision et au NLP
Maîtrise du langage Python et du framework PyTorch ou TensorFlow

Informations

Durée du stage	6 mois
Pays	France
Régions	Ile-de-France

Départements	Essonne (91)
Ville	Massy
Contact	Samir BOUINDOUR samir.bouindour@safrangroup.com