Recherche de prompts intelligibles pour l'optimisation de Modèles Vision et Langue

1 Contexte

Depuis l'avènement du Deep Learning, le domaine du machine learning a connu une évolution significative. Le passage au Deep Learning a été marqué par des avancées majeures impactant durablement la vision par ordinateur ainsi que le traitement du langage. Ces dernières années, les approches d'apprentissage dites auto-supervisées ont permis l'exploitation à très grande échelle de jeux de données non annotés. En particulier, de larges modèles vision et langue (VLM) tels que CLIP [3] ont pu être appris de manière contrastive pour rapprocher des concepts et des images similaires. Ces modèles atteignent des performances remarquables dans la reconnaissance automatique de concepts décrits en langage naturel.

2 Objectif

Ce projet vise à optimiser les prompts utilisés pour la classification d'image par des VLM, en s'inspirant de méthodes récentes en *prompt learning* [4, 2, 1]. Plutôt que de dépendre d'interventions humaines pour l'optimisation des prompts décrivant les concepts recherchés, ces méthodes proposent d'apprendre directement les représentations associées à ces concepts. Ces représentations sont soigneusement optimisées par descente de gradient sur une loss de classification pour maximiser les performances de classification d'un modèle pré-entraîné.

Bien que ces approches aient montré leur efficacité en améliorant les performances de classification avec un nombre limité d'exemples, les *embeddings* résultants ne sont pas associés à un texte intelligible. Un simple mapping vers le plus proche voisin dans un dictionnaire donne au final un texte aléatoire. Afin de garantir l'explicabilité des modèles, le but de ce projet sera de tester d'une part une stratégie de régularisation de ces prompts visant à les échantillonner au sein d'une banque de templates puis grâce à un modèle de langue.

Ce projet se concentrera sur CLIP avec un encodeur visuel de type ResNet pour limiter les temps d'inférence. Nous resterons ici dans un cadre *zero-shot* voire *few-shot* et ne chercherons pas à adapter le modèle. Le dataset d'image utilisé sera CIFAR-10 voire CIFAR-100.

3 Encadrement

Le projet sera encadré par Clément Rambour (clement.rambour@sorbonne-universite.fr), lecturer à Sorbonne Université.

Références

[1] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual Prompt Tuning, July 2022. arXiv:2203.12119 [cs].

[2] S. Parisot, Y. Yang, and S. McDonagh. Learning To Name Classes for Vision and Language Models. pages 23477–23486, 2023.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to Prompt for Vision-Language Models. *Int J Comput Vis*, 130(9):2337–2348, Sept. 2022. arXiv:2109.01134 [cs].