AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mr **SEBBAN Othmane**
Soutiendra : **le Samedi 03/10/2026** à **11H00**
*Lieu : **FSDM – Centre Visioconférence***

Une thèse intitulée :

Efficient Transformer-Based Vision-Language Models for Image and Video Captioning in Assistive Technologies

*En vue d'obtenir le **Doctorat***

*FD : **Sciences et Technologies de l'Information et de la Communication***
*Spécialité : **Informatique***

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
Pr. ZINEDINE Ahmed	Faculté des Sciences Dhar El Mahraz, Fès	PES	Président
Pr. EL FAR Mohamed	Faculté des Sciences Dhar El Mahraz, Fès	PES	Rapporteur
Pr. BERRADA Ismail	Université Mohammed VI Polytechnique, Benguerir	MCH	Rapporteur
Pr. SALIH ALJ Yassin	AL Akhawayn University, Ifrane	PES	Rapporteur
Pr. FARDOUSSE Khalid	Faculté Charria, Fès	PES	Examinateur
Pr. BENNANI Taj Mohamed	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Examinateur
Pr. AZOUGH Ahmed	Ecole Supérieure d'ingénieurs Léonard De Vinvci (ESILV) Paris, France	MCH	Expert
Pr. LAMRINI Mohamed	Faculté des Sciences Dhar El Mahraz, Fès	PES	Directeur de thèse

Résumé :

Cette thèse aborde les limites des technologies d'assistance existantes destinées aux personnes malvoyantes en proposant un cadre multimodal intégré et efficace pour la compréhension visuelle. Nous concevons et mettons en œuvre SeeAround, un système mobile capable de fournir des descriptions audio contextuelles en temps réel en combinant des modules de légendage d'images, de reconnaissance optique de caractères (OCR), de traduction automatique et de synthèse vocale. Nous démontrons que les architectures basées sur les transformateurs, en particulier Vision Transformer (ViT) combiné à GPT-2, surpassent les modèles CNN-RNN traditionnels dans la génération de légendes d'images riches en sens et sensibles au contexte, tout en conservant un bon compromis entre précision et efficacité computationnelle.

De plus, nous présentons V2CT-SF (Video-to-Commonsense Transformer with Scene Filtering), un nouveau cadre de légendage vidéo qui intègre le filtrage de scène et une architecture à double décodeur. Les résultats expérimentaux montrent que le modèle proposé capture efficacement à la fois les aspects descriptifs et ceux relevant du bon sens du contenu vidéo, ce qui conduit à une génération de légendes plus cohérentes et plus significatives sur le plan contextuel. Dans l'ensemble, ce travail démontre que la combinaison de la perception visuelle, de la modélisation spatio-temporelle et du raisonnement logique au sein d'un cadre unifié améliore considérablement la qualité et la facilité d'utilisation des technologies d'assistance, tout en garantissant un déploiement efficace dans des environnements réels et aux ressources limitées.

Mots clés :

Déficience visuelle ; Technologies d'assistance ; Légendes d'images ; Légendes de vidéos ; Raisonnement logique ; V2CT-SF; Assistance visuelle en temps réel; Vision Transformer (ViT); Modèles basés sur Transformer; GPT-2.

EFFICIENT TRANSFORMER-BASED VISION-LANGUAGE MODELS FOR IMAGE AND VIDEO CAPTIONING IN ASSISTIVE TECHNOLOGIES

Abstract :

This thesis addresses the limitations of existing assistive technologies for visually impaired individuals by proposing an integrated and efficient multimodal framework for visual understanding. We design and implement SeeAround, a mobile-based system capable of delivering real-time contextual audio descriptions by combining image captioning, optical character recognition (OCR), machine translation, and text-to-speech modules. We demonstrate that transformer-based architectures, particularly Vision Transformer (ViT) combined with GPT-2, outperform traditional CNN-RNN models in generating semantically rich and context-aware image captions while maintaining a favorable trade-off between accuracy and computational efficiency.

Furthermore, we introduce V2CT-SF (Video-to-Commonsense Transformer with Scene Filtering), a novel video captioning framework that integrates scene filtering and a dual-decoder architecture. Experimental results demonstrate that the proposed model effectively captures both descriptive and commonsense aspects of video content, resulting in more coherent and contextually meaningful caption generation. Overall, this work demonstrates that combining visual perception, spatio-temporal modeling, and commonsense reasoning within a unified framework significantly enhances the quality and usability of assistive technologies, while ensuring efficient deployment in real-world and resource-constrained environments.

Key Words :

Visual Impairment; Assistive Technology; Image Captioning; Video Captioning; Commonsense Reasoning; V2CT-SF; Real-Time Visual Assistance; Vision Transformer (ViT); Transformer-based Models; GPT-2.