AVIS DE PRESENTATION DE THESE EN SOUTENANCE POUR L'OBTENTION DU DIPLOME NATIONAL DE DOCTEUR

Madame Lady Viviana BELTRAN BELTRAN

Présentera ses travaux intitulés :

« Espaces sémantiques visuels et textuels communs pour l'analyse du contenu Multimodal »

Spécialité : Informatique et applications

Le 17 juin 2021 à 14h00

Lieu :

La Rochelle Université
Pôle Communication, Multimédia et Réseaux
Amphithéâtre Michel Crépeau
44 Av. Albert Einstein
17000 LA ROCHELLE

Composition du jury :

Mme BUGEAU Aurélie	**Professeure, Université de Bordeaux**
M. COUSTATY Mickaël	**Maître de conférences, La Rochelle Université**
M. DOUCET Antoine	**Professeur, La Rochelle Université**
Mme EGLIN Véronique	**Professeure, Université Claude Bernard Lyon**
M. JOURNET Nicholas	**Maître de conférences, HDR, Université de Bordeaux**
M. MARINAI Simone	**Professeure, Université de Florence**
M.RAMEL Jean-Yves	**Professeur, Université de Tours**

Résumé :

L'apprentissage multimodal implique l'utilisation de multiples sens (tactile, visuel, auditif, etc.) au cours du processus d'apprentissage pour mieux comprendre un phénomène. Dans le domaine du calcul, nous avons besoin de systèmes pour comprendre, interpréter et raisonner avec des données multimodales, et bien qu'il y ait eu d'énormes progrès dans le domaine, de nombreuses capacités souhaitées restent hors de notre portée. L'objectif de ces systèmes est d'exploiter différents types de données sémantiquement liés pour produire de meilleures prédictions pour un phénomène d'intérêt. Par exemple, pour les utilisateurs ayant des handicaps sensoriels tels que visuels, pour effectuer des tâches quotidiennes telles que faire un achat ou trouver une place dans une ville, l'information visuelle de leur environnement doit être transformée en une modalité différente avec une signification plus sémantique. Un système à cette fin pourrait utiliser des informations auditives fournies par l'utilisateur qui spécifient quelles informations sont requises et qui peuvent être facilement transformées en données textuelles, et des informations visuelles telles que des images obtenues de son environnement pour aider l'utilisateur à prendre une décision. Il s'agirait donc d'un système multimodal exploitant les informations de trois modalités différentes: auditif + texte + images.

En ce qui concerne le calcul, travailler avec des données multimodales présente plusieurs défis. Cette thèse se concentre sur l'avancement de la recherche sur l'apprentissage multimodal à travers diverses contributions scientifiques: nous simplifions la création de modèles d'apprentissage profond en proposant des cadres qui trouvent un espace sémantique commun pour les modalités visuelles et textuelles en utilisant l'apprentissage profond comme outil de base; Nous proposons des stratégies compétitives pour aborder les tâches de recherche intermodale, de réponse visuelle aux questions de texte de scène et d'apprentissage d'attributs; Nous abordons divers problèmes liés aux données tels que le déséquilibre et l'apprentissage lorsque les données annotées sont insuffisantes. Ces contributions visent à combler le fossé entre les humains (comme les utilisateurs non experts) et l'intelligence artificielle pour s'attaquer aux tâches quotidiennes.

Notre première contribution vise à évaluer l'efficacité d'un système multimodal qui reçoit des images et du texte et récupère les informations multimodales pertinentes. Cette approche nous permet de réaliser une étude complète pour évaluer l'efficacité d'un système de récupération cross-modal avec le deep learning comme outil de base. La fonction cross-modal permet de formuler les requêtes sous forme d'images ou de texte et de récupérer les données multimodales pertinentes. Avec cette approche, nous pouvons évaluer la capacité du modèle à produire des représentations multimodales efficaces et à gérer toute requête multimodale avec un seul modèle. Par la suite, dans notre deuxième contribution, nous adaptons le système pour effectuer une tâche récente appelée réponse visuelle aux questions de texte de scène (ST-VQA). L'objectif est d'apprendre aux modèles VQA traditionnels à lire le texte contenu dans des images naturelles. Cette tâche nous oblige à effectuer une analyse sémantique entre le contenu visuel et les informations textuelles contenues dans les questions associées pour donner la bonne réponse. Nous trouvons cette tâche très pertinente dans le contexte multimodal car elle nous oblige vraiment à développer conjointement des mécanismes qui raisonnent sur le contenu visuel et textuel.

Nos dernières contributions mettent en évidence des problèmes liés aux données. Les données sont l'un des facteurs les plus importants pour viser de bonnes performances. Par conséquent, nous avons déterminé qu'une compétence pertinente consiste à comprendre comment nettoyer et analyser correctement les données et créer des stratégies qui peuvent en tirer parti. Nous abordons des problèmes très courants et fréquents tels que le bruit, le déséquilibre et l'insuffisance des données annotées. Pour évaluer nos stratégies, nous considérons le problème de l'apprentissage des attributs. L'apprentissage des attributs peut compléter la reconnaissance au niveau des catégories et donc améliorer le degré de perception des objets visuels par les machines. Dans la première étude, nous couvrons deux aspects clés. Déséquilibre et données étiquetées insuffisantes. Nous proposons des adaptations aux stratégies d'apprentissage déséquilibrées classiques qui ne peuvent pas être directement appliquées lors de l'utilisation de modèles d'apprentissage profond multi-attributs. Dans la deuxième étude, nous proposons une nouvelle stratégie pour exploiter les relations classe-attribut pour apprendre les prédicteurs d'attributs de manière semi-supervisée. L'apprentissage semi-supervisé permet d'exploiter les grandes quantités de données non étiquetées disponibles dans de nombreux cas d'utilisation en combinaison avec des ensembles généralement plus petits de données étiquetées.