Université de Rennes

Appel à projets - Année 2024

Contrats doctoraux «Ordinaires»

Direction de la Recherche et de l'Innovation

ECOLE DOCTORALE : Matisse **POLE DE RECHERCHE : Mathématique-Numérique**

Unité de recherche d'accueil

Unité de Recherche : IRISA

Equipe : LINKMEDIA

Encadrement

Directeur de thèse : Guillaume Gravier **Unité de recherche** : IRISA

Employeur/Établissement d'origine : CNRS

HDR : X oui ☐ non **Nb de thèses en cours** : 5

Mail : guig@irisa.fr **Tél.** :

Co-directeur de thèse : Pascale Sébillot **Unité de recherche** : IRISA

Employeur/Établissement d'origine : INSA Rennes

HDR : X oui ☐ non **Nb de thèses en cours** : 2

Mail : pascale.sebillot@irisa.fr **Tél.** :

Sujet

Titre : Géométrie des transformeurs en traitement automatique des langues, explicabilité et mesures de confiance

Mots clés : traitement automatique des langues ; apprentissage automatique ; transformeurs ; mécanisme d'attention ; explicabilité en IA ; mesure de confiance

Cotutelle internationale (oui/non) : non **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

Les modèles fondés sur l'auto-attention sont devenus les standards de facto en traitement automatique des langues (TAL). En particulier, les modèles pré-entrainés via une tâche de modèles de langue (LLM ou large language models), comme BERT [1] pour l'anglais ou FlauBERT [2] pour le français, sont au cœur de l'état de l'art dans de nombreuses tâches en TAL. L'étape de pré-entrainement leur permet de capturer une information linguistique riche qui est ensuite implicitement exploitée lorsque ces modèles sont adaptés pour différentes tâches applicatives comme la classification de texte, l'étiquetage de séquence (étiquetage morpho-syntaxique, détection d'entités nommées, analyse syntaxique en dépendance, etc.) ou encore l'inférence en langage naturel.

Pour éclairer ce débat, quelques travaux se sont penchés sur le fonctionnement interne de ces modèles [7— 13], notamment sur la géométrie des plongements dans les modèles d'auto-attention [11—13]. Ces dernières études ont mis en évidence que les couches successives d'auto-attention renforcent l'anisotropie des plongements (c.-à-d. leur concentration dans une direction) au sein d'une même phrase. Ce phénomène semble lié à celui du 'neural collapse' qui apparaît dans les réseaux profonds. Dans des travaux préliminaires, nous avons notamment observé sur une tâche de classification de texte que les plongements en sortie d'un modèle BERT adapté convergent dans un cône étroit, la direction du cône étant liée à la classe. Nous avons montré que cette convergence se construit progressivement au travers des couches du modèle [13].

L'objectif de la thèse est d'**approfondir la compréhension des transformations géométriques opérées au sein d'un modèle de type transformeurs et d'étudier comment ces transformations peuvent être exploitées dans deux contextes** : l'explicabilité d'une part, c'est-à-dire la capacité à mettre en avant les mots de la phrase ou les relations entre mots de la phrase qui justifient la décision prise par le modèle ; la mesure de confiance dans la décision prise par le modèle.

Concernant la meilleure compréhension des transformations géométriques au sein d'un modèle, on s'appuiera dans une première partie sur les travaux préliminaires menés sur la tâche de classification [12,13] avec plusieurs objectifs. Il s'agira tout d'abord de mettre en évidence comment le jeu entre clés et valeurs dans le mécanisme d'attention assure cette

convergence des plongements au sein d'une phrase et de relier ce phénomène à celui du *neural collapse*. Dans un second temps, il s'agira d'élargir le cadre de l'étude à des tâches autres que la classification de texte.

Dans une seconde partie, on cherchera à exploiter la compréhension fine de la géométrie des modèles pour améliorer l'explicabilité. Par exemple, dans la tâche de classification de texte, on peut se demander si certains mots ne sont pas déterminants dans la direction vers laquelle on projette la phrase, ces mots étant alors cruciaux pour l'explication. On cherchera aussi à exploiter la géométrie des modèles pour faciliter la mesure de la confiance dans la décision prise par un modèle. Cette dernière est traditionnellement liée au score d'activation de la classe en sortie d'un perceptron, ce qui s'avère en pratique une mesure peu fiable. On pourra par exemple s'appuyer sur la direction des plongements après convergence et sur la caractérisation des cônes associés aux différentes classes pour mesurer la confiance.