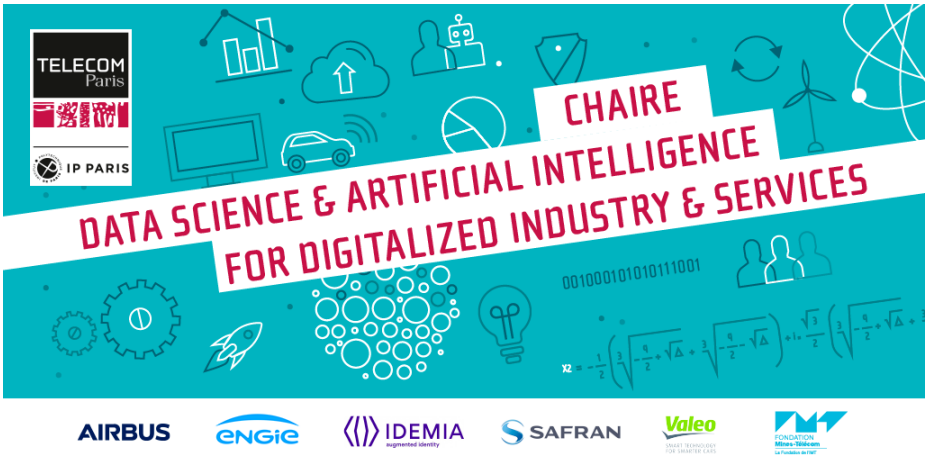

Offre de stage

Sujet : Apprentissage auto-supervisé contrastif de représentations de mots

Possibilité de poursuivre sur une thèse : Oui

La Chaire Data Science and Artificial Intelligence for Digitized Industry and Services

Portée par Florence d'Alché-Buc, enseignante-chercheur dans le département Image, Données, Signal de Télécom ParisTech, la chaire DSAI réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo Finance. Son objectif général est de développer, en liaison étroite entre les Parties, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

Description du stage

Encadrement

Matthieu Labeau

Lieu et dates du stage

Telecom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage: Début 2022

Équipe(s) d'accueil de la thèse

Département IDS, équipe S2A

Mots clés

Traitement automatique des langues, Apprentissage de représentations, Apprentissage contrastif

Sujet détaillé

Ces dernières années, on a pu constater d'énormes améliorations dans de nombreuses tâches du traitement automatique du langage. On peut attribuer une grande partie de ces progrès à deux avancées successives dans la production de représentations vectorielles denses de mots (ou plongements) : d'abord, le toolkit *word2vec* [1], implémentant de manière efficace et rapide un apprentissage non supervisé à l'aide de modèles neuronaux très simples, dont le but est de prédire si des mots sont censés apparaître ensemble. Ensuite, les représentations contextuelles de mots, correspondant à un mot considéré dans le contexte de sa phrase d'origine, obtenues à l'aide d'énormes modèles neuronaux entraînés à prédire des mots manquants dans une séquence, comme BERT [2]. L'idée centrale derrière l'entraînement de ces modèles est de prédire un mot à l'aide d'un contexte, sous la forme d'une tâche de classification (les classes étant les mots connus par le modèle, contenus dans un vocabulaire). Ainsi, la distribution de probabilité sur ces mots est habituellement obtenue à l'aide d'un *softmax* ; mais il est coûteux à calculer, et l'utilisation de *negative sampling* pour word2vec [3], et plus récemment, pour une variante de BERT [4], ont permis d'accélérer grandement l'entraînement de ces modèles. Cette manière d'apprendre peut aussi être vue comme de l'**apprentissage contrastif** : on apprend des représentations à partir de paires de points de données similaires (échantillon positif) et différents (échantillon négatif).

L'apprentissage contrastif a récemment été très utilisé pour l'apprentissage de représentations de phrases dans le traitement automatique du langage [5]; il a aussi été rapproché de l'idée de maximisation d'information mutuelle [6]. Plusieurs études théoriques et expérimentales démontrent notamment l'intérêt de bien choisir la façon de générer les échantillons négatifs [7]. Pour un modèle comme word2vec, ils sont générés à l'aide de la distribution fréquentielle des mots. Pour les modèles récents, seul l'utilisation d'un autre modèle pré-entraîné semble fonctionner. L'objectif de ce stage est d'explorer différentes façon de générer des échantillons négatifs lors de l'apprentissage contrastif de représentations de mots. On pourra commencer par une modélisation simple, en se basant sur l'idée qu'un échantillon négatif informatif doit être difficile à distinguer d'un échantillon positif ; et choisir différente manière de conditionner la génération sur le contexte d'entrée ou le mot à prédire.

Profil du candidat

Étudiant titulaire d'un master 2 recherche avec de solides connaissances en :

- Apprentissage statistique
- Traitement du langage naturel
- Programmation (Python)

et un bon niveau en anglais.

Candidatures

A envoyer à matthieu.labeau@telecom-paris.fr :

- Curriculum Vitae

- Lettre de motivation (directement dans le corps du mail)
- Relevés de notes des années précédentes
- Contact d'une personne de référence

Références

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.Distributed representations of words and phrases and their compositionality.In NIPS, 2013.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformersfor language understanding. InNAACL-HLT, June 2019

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space.CoRR, abs/1301.3781, 2013

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discrimina-tors rather than generators. InICLR, 2020

[5] Nils Rethmeier and Isabelle Augenstein. A primer on contrastive pretraining in language processing: Methods, lessonslearned and perspectives.CoRR, abs/2102.12982, 2021

[6] Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. A mutual informationmaximization perspective of language representation learning. InICLR, 2020.

[7] Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation. InProceedings of the 2021Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,June 2021.