Wasserstein embeddings for language model visualization and document clustering

Lieu : Laboratoire Hubert Curien UMR 5516, Saint-Etienne

Mot-clés : Machine Learning (ML), Natural Langage Processing (NLP), Clustering, Optimal Transport

Résumé : En intelligence artificielle, une part importante de la recherche en traitement automatique de la langue (par la suite NLP) consiste à trouver des modèles permettant de traiter de très grands volumes de données textuelles (ex : classification de textes, chatbot, questions/réponses, etc.). Pour cela, l'état de l'art consiste en deux réseaux de neurones successifs. Un premier, peu profond et entraîné de manière non supervisée, apprend à associer les mots d'une langue à un point dans un espace euclidien de 100 à 300 dimensions (word embeddings) - deux mots sémantiquement proches se voient associer deux points proches. Ces représentations modélisent la sémantique latente d'une langue - elles sont ensuite utilisées dans d'autres réseaux (profonds et entraînés de manière supervisées), ad hoc à chaque tâche de NLP visée. Dans cette thèse, nous nous intéressons à améliorer les représentations de mots dans ces différents réseaux de neurones. Les questions posées par les limitations des représentations actuelles sont les suivantes : Comment observer en deux dimensions, pour mieux les comprendre, les vecteurs de mots d'une langue avec le minimum de pertes de régression (alternative aux approches PCA et t-SNE) ? Comment mieux modéliser les mots rares (rareté dans la langue ou par appartenance à un domaine de spécialistes) ? Comment gérer la polysémie suivant le type de document considéré, quand un simple mot peut revêtir différent sens suivant le contexte d'emploi ? Pour cela, nous nous intéresserons aux espaces de Wasserstein comme un espace de représentation intermédiaire avant de replonger les mots dans le plan. Nous nous intéresserons également à l'apprentissage conjoint de plongements de mots et de catégorisation de documents de domaine (Web, santé, et sécurité) dans un espace de Wasserstein - dans ce contexte, les mots pourront revêtir des représentations différentes suivant la catégorie dans laquelle ils sont employés.

Profil recherché :
Le(a) candidat(e) devra posséder des connaissances solides en apprentissage automatique avec notamment de bonnes bases en apprentissage statistique et en mathématiques. Il devra également avoir un bon niveau en programmation python et être capable de développer des outils efficaces potentiellement complexes. Le candidat devra aussi posséder un bon niveau d'anglais et avoir à la fois un intérêt pour des aspects théoriques et pratiques.

Encadrants :
Charlotte Laclau, charlotte.laclau@univ-st-etienne.fr
Christophe Gravier, christophe.gravier@univ-st-etienne.fr

Candidature : Les candidats intéressés sont invités à envoyer un CV, un relevé de notes (Licence + Master) avec classements (Master 2 également, éventuellement partiel).
Attention urgent - Deadline pour prendre contact : vendredi 1 mars

See English version bellow

Wasserstein embeddings for language model visualization and document clustering

Site: Laboratoire Hubert Curien UMR 5516, Saint-Etienne

Keywords: Machine Learning (ML), Natural Language Processing (NLP), Clustering, Optimal Transport

Summary: In artificial intelligence, an important part of the research into natural language processing (NLP) consists of finding models able to handle very large volumes of textual data (eg : text classification, chatbot, questions / answers, etc.). To this end, the state of the art consists of an architecture composed of two successive neural networks. A first, shallow and trained in an unsupervised way, learns to associate the words of a language to a point in a Euclidean space of 100 to 300 dimensions (word embeddings) - two semantically close words are associated with two close points. These representations model the latent semantics of one language - they are then used in the second networks (deep and supervised), ad hoc to each target NLP task. In this thesis, we are interested in improving the representations of words in these different neural networks. The questions posed by the limitations of the current representations are as follows: How to observe in two dimensions, to better understand them, the word vectors of a language with the minimum loss of regression (alternative to the PCA and t-SNE approaches)? How to better model rare words (rarity in the language or because it belongs to a field of specialists)? How to manage the polysemy according to the type of document considered, when a simple word can take different senses according to the context of occurrence? For this, we will focus on the spaces of Wasserstein as an intermediate representation space before embedding the words into the plane. We will also be interested in the joint learning of word embedding and categorization of domain documents (Web, health, and security) in a Wasserstein space - in this context, the words may take different representations according to the category in which they appear.

Required profile: The candidate must have a solid knowledge of machine learning techniques, including a good background in statistical learning and mathematics. He will also have a good level of python programming and be able to develop potentially complex tools. The candidate must also have a good level of English and have both an interest in theoretical and practical aspects.

Supervisors :
Charlotte Laclau, charlotte.laclau@univ-st-etienne.fr
Christophe Gravier, christophe.gravier@univ-st-etienne.fr

Application: Interested candidates are invited to send a CV, a transcript of their grades (License + Master) with their rank (Master 2 also, possibly partial) to the email addresses of both supervisors.
Attention - Deadline to make contact – Friday, March 1st