
Programme de Formation

Deep learning pour le traitement automatique des langues



Organisation

Durée : 18 heures

Mode d'organisation : Présentiel

Contenu pédagogique



Public visé

Docteurs, ingénieurs ou titulaires d'un master travaillant dans la recherche et/ou le développement.



Objectifs pédagogiques

- Connaître la théorie et la pratique des réseaux profonds (deep learning) appliqués au traitement automatique des langues
- Maîtriser les enjeux à la fois théoriques et opérationnels afin de comprendre le potentiel des réseaux de neurones profonds ainsi que leurs limites dans le cadre du traitement automatique des langues
- Mettre en oeuvre des modèles récents avec les bibliothèques Pytorch et Transformers de HuggingFace



Description

Ces dernières décennies, les réseaux de neurones artificiels et plus généralement le Deep Learning ont renouvelé les perspectives de recherche en traitement automatique des langues (TAL). La plupart des applications en TAL nécessitent de modéliser des données structurées qui se caractérisent par des distributions particulières, parcimonieuses et avec des espaces de réalisation de grande dimension.

Dans ce contexte, les réseaux de neurones et récemment les grands modèles de langue (LLM) ont permis des avancées importantes en introduisant la notion de représentations continues pour le TAL et ce pour de nombreuses tâches comme par exemple l'analyse syntaxique, la reconnaissance d'entités nommées, la classification de documents, la reconnaissance automatique de la parole et la traduction automatique. Ces progrès se sont amplifiés dès lors que les modèles neuronaux ont dépassé le cadre de l'apprentissage de représentation pour évoluer vers des architectures de plus en plus profondes, permettant de modéliser de bout en bout des tâches d'inférence complexes.



- Jour 1 (6 h) – Traitement automatique du langage naturel (TALN) : historique et introduction aux approches neuronales
- Jour 2 (6 h) – TALN : représentations vectorielles et modèles transformers (LLM)
- Jour 3 (6 h) – Adaptations efficaces (PEFT) de LLM et assistants « intelligents » (IA agentique)

Jour 1 (6 h) – Traitement automatique du langage naturel: historique et introduction aux approches neuronales

Résumé

Dans ce premier volet, nous présentons les problématiques du traitement automatique du langage naturel (TALN). Nous décrivons les principales tâches que les chercheurs ont tenté de résoudre au cours des dernières décennies. Nous nous attarderons tout d'abord aux méthodes de prétraitement du texte et nous introduirons le principe des modèles de langue. Une seconde partie sera consacrée à l'apprentissage profond, afin de fournir le bagage nécessaire à la compréhension des séances suivantes. Enfin, nous examinerons les approches neuronales appliquées au TALN.

CM – 3 h : remise à niveau sur les réseaux de neurones et le deep learning pour le TALN

- Introduction au TALN (définition, enjeux) ;
- Prétraitement du texte (normalisation, tokenisation, vocabulaire) ;
- Introduction aux modèles de langue (modèles N-gram et réseaux récurrents) ;
- Notions de base du deep learning (couches, poids, fonction de perte).

TP – 3 h : apprentissage profond pour le TALN

- Tokenisation – vocabulaire, sous-mots (BPE, SentencePiece) ;
- Premier modèle de langue – création d'un modèle N-gram ;
- Premiers pas avec PyTorch – installation, tenseurs, autograd et mise en place d'un classifieur.

Jour 2 (6 h) TALN : représentations vectorielles et modèles transformers

Résumé

Dans ce volet, nous présentons les approches récentes appliquées au traitement du TALN. En particulier, nous verrons comment représenter les unités linguistiques (mots, jetons / tokens) afin d'intégrer une information sémantique. Dans un second temps, nous décrirons précisément les architectures des modèles de langue modernes, à savoir le modèle transformer. Enfin, nous expliciterons le processus d'apprentissage des modèles de fondation et la manière dont ceux-ci peuvent être adaptés à une tâche cible.

CM – 3 h : représentation du texte et architecture des modèles de langue (Transformer)

- Word-embeddings (plongements non-contextuels vs contextuels) ;
- Architecture Transformer (blocs, hyper-paramètres) ;
- Modèles de langue: objectifs de pré-entraînement (Next Token Prediction et Masked Language Modeling) ;
- Entraînement d'un modèle Transformer (pré-entraînement, fine-tuning).

TP – 3 h : mise en pratique

- Représentation vectorielle d'unités linguistiques: Word2Vec ;
- Classification avec un modèle Transformer (fine-tuning d'un modèle BERT) ;
- Entraînement d'un modèle auto-régressif pour la génération de texte.

Jour 3 (6 h) Adaptations efficaces (PEFT) de LLM et assistants « intelligents » (IA agentique)

Résumé

Dans cette dernière séance, nous nous concentrerons sur les très grands modèles de langue (LLM). Nous aborderons d'abord les méthodes d'alignement visant suivre les instructions de l'utilisateur (instruction-tuning). De façon plus pratique, nous présenterons des approches efficaces pour adapter ces modèles à un problème précis. Enfin, nous discuterons des approches d'IA agentique, qui permettent de doter les modèles d'outils pour interagir avec plusieurs sous-systèmes.

CM – 3 h : les modèles génératifs et l'IA agentique

- Une histoire d'alignement – du pré-entraînement à l'adaptation aux instructions (instruct tuning);
 - Adapter un modèle à une tâche de manière efficace – stratégies d'adaptation efficace (LoRA, prefix-tuning) et méthodes d'optimisation temps/mémoire ;
 - IA agentique – architectures multi-modèles, assistant intelligent et outils ;
 - Quel modèle pour quelle utilisation – critères de sélection (taille, latence, licence, domaine) et discussion des problèmes liés à l'usage de ces modèles.
- TD – 3 h : mise en pratique
- Adapter un modèle de langue à une tâche précise en utilisant les approches PEFT (parameter-efficient fine-tuning) avec LoRA ;
 - Création d'outils pour exploiter les modèles génératifs dans le cadre d'un objectif donné (assistant de code et assistant pour la recherche documentaire).



Prérequis

Connaissances de base en apprentissage automatique, probabilités, statistiques
Connaissance de Python



Modalités pédagogiques

Alternance de cours (9 h) et de travaux pratiques (9 h).



Moyens et supports pédagogiques

Des fichiers au format PDF ainsi que des Notebooks Jupyter seront mis à disposition des participants.
Il est demandé aux apprenants de venir avec leur propre ordinateur sur lequel les logiciels et outils (libres) nécessaires seront préalablement installés.



Modalités d'évaluation et de suivi

Un suivi individualisé par des évaluations formatives est assuré. Une attestation de fin de formation est délivrée à la fin du parcours.



Informations sur l'admission

L'admission à cette formation ne fait l'objet d'aucun examen, test ou sélection préalable ; l'inscription est validée après réception du dossier complet et confirmation par l'organisme de formation.



Informations sur l'accessibilité

Notre organisme s'engage à garantir l'accessibilité de ses formations à distance et en présentiel aux personnes en situation de handicap. Un référent handicap est mobilisable afin d'analyser les besoins spécifiques et de mettre en place, lorsque cela est possible, les adaptations pédagogiques, techniques ou organisationnelles nécessaires.