Projets Deep Learning 2022

Yann Vernaz, Paul GAY

Cy-tech / UPPA

Vendredi 5 mars

Règles du jeu

- **de maintenant au 29/02** Proposition de sujets
- **jeudi 29 février** Choix des sujets et constitution des équipes
  - Entre 2 et 4 par équipe
  - Possibilité de proposer un sujet
- **22 mars** Soutenance des projets
  - 10 minutes Présentation et démo + 5 minutes questions
  - Choix méthodologiques et interprétation des résultats
  - (Généralement) Pas de présentation du code (ou du notebook)
  - restitution générale avec la classe

  Critères d'évaluation:
  - Réponse au cahier des charges
  - Pertinence des modèles proposés
  - Qualité de l'analyse des données
  - Maîtrise du sujet
  - Qualité du Pitch

Note: Le code pourra vous être demandé.

Comment faire votre choix

Tous les sujets ne sont pas égaux

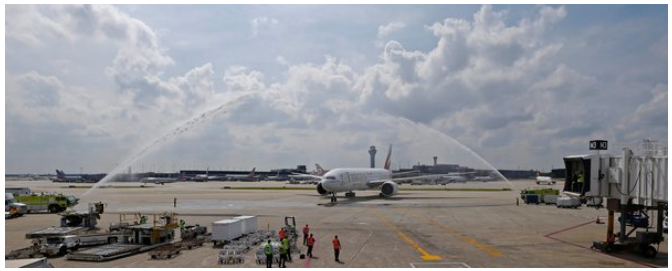
Un sujet difficile est facile, et un sujet facile est difficile

Les sujets

- 1 Apprentissage multimodal texte/image dans les médias
- 2 Apprentissage sur des graphes avec Twitter
- 3 Multi-GPUs sur un cluster de Jetson
- 4 Moteur de traduction Occitan/français.
- 5 Généralisation des modèles
- 6 Détection visuelle d'arbres malades
- 7 Moteur de recherche pour l'écologie
- 8 Classification de décharge sauvages dans les images aériennes
- 9 visualisation de l'actualité

Apprentissage multimodal texte/image dans les médias

L'objectif est de proposer un modèle multimodal permettant de faire le lien entre le texte et une image



An Emirates aircraft from Dubai arriving at Chicago's O'Hare airport. Emirates flies to nine cities in the United States from Dubai.

Typiquement votre modèles comportera :

- Un convnet qui extrairera une représentation de l'image
- Un Embedding + modèle RNN ou transformer qui extraiera une représentation du texte
- un dernier module dont le but sera de comparer ces deux représentations
- Ou partir d'un modèle pré-entrainé comme CLIP

Apprentissage multimodal texte/image dans les médias

Données:

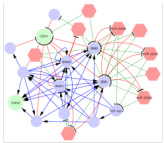
- +18000 images du new york times avec la description et l'article correspondant.

Résultat attendu:

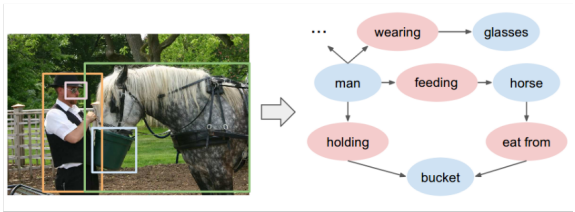
- Démonstrateur qui, étant donné une phrase, renvoie une liste d'images correspondantes.

Apprentissage sur des graphes avec Twitter

Depuis 6 ans, de nouvelles méthodes permettent d'appliquer le deep learning sur des données contenues dans des graphes.

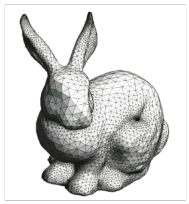


Chimie moléculaire



Vision et language

Réseaux sociaux



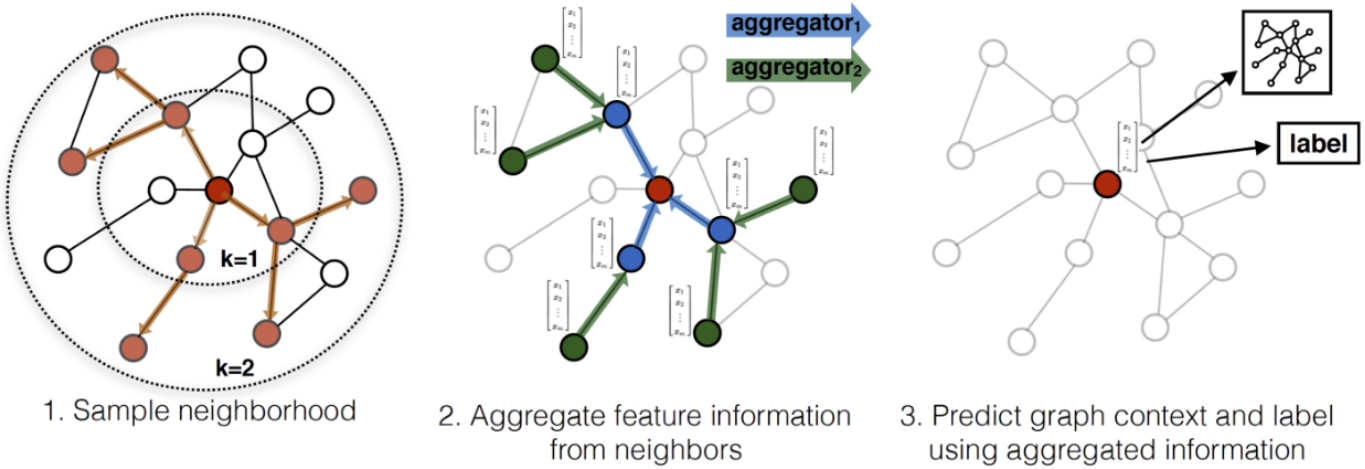
3D meshes

Applications des graphes en Deep Learning

Ce projet propose d'analyser le réseau twitter avec ce type de méthodes.

Apprentissage sur des graphes avec Twitter

Ces méthodes s'appuient sur un formalisme d'envoi de message



1. Sample neighborhood

2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

Hinton et al. Graph sage

Librairie opensource disponibles: https://www.dgl.ai/

Apprentissage sur des graphes avec Twitter

Données

Twitter est un espace public et une plateforme permettant de récupérer facilement des données et présente de nombreuses applications en sociologie.

- Plusieurs dizaines de milliers de tweets issus de la campagne présidentielle de 2017
- Chaque utilisateur est un noeud, un arc est créé pour chaque retweet ou citation.
- Chaque utilisateur est annoté avec son affiliation politique.

Résultat attendu:

- Proposer un modèle donnant l'affiliation d'un utilisateur

Multi-GPUs sur un cluster de Jetson

Ce sujet vous propose d'apprendre à entraîner un modèle sur plusieurs machines.

- Soit pour faire du Federated learning
- Soit pour paralléliser un modèle afin d'accélerer l'apprentissage.

Résultat attendus

- Une démonstration fonctionelle utilisant les différentes machines Jetson.

Ressources :
`www.tensorflow.org/guide/keras/distributed_training`

`https://pytorch.org/tutorials/beginner/dist_overview.html`

Détection de feu de forêts

Les rapports du GIEC mentionne une assèchement de certaines régions, dont la région méditerranéenne.
Renforcer la lutte contre les incendies est un enjeu majeur
Le but de ce projet est de construire un détecteur visuel d'incendie, en classant les images en deux catégories : présence de feu ou pas.

Données

- Environ 10000 images de chaque classe. Vous pouvez collectez des faux négatifs additionnels

Résultat attendu:

- Construire un classifieur
- Proposer un score de confiance et une stratégie pour diminuer le nombre de faux négatifs, et pour avoir une règle de rejet.

Alignement de texte Occitan/Français.

Il s'agit d'un sujet de traduction Occitan/Français:

```
 Cap a una pedagogia del plurilinguisme ?
Vers une pédagogie du plurilinguisme ?
```

```
Lo programa complet es a posita sul sit internet
Le programme complet est disponible sur le site internet
```

Il existe d'excellents tutoriels sur ce sujet notamment, en utilisant les transformers:

```
 https://www.tensorflow.org/tutorials/text/transformer
```

Alignement de texte Occitan/Français.

Données

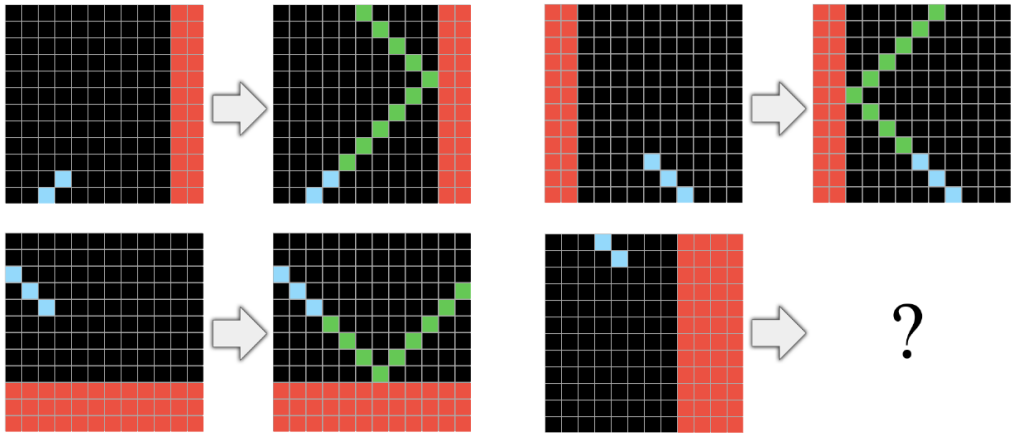
- 46 548 phrases dans trois variétés de l'occitan :
  - 28 127 en gascon
  - 11 863 en limousin
  - 6 558 en provençal

Résultats attendus

- Application qui propose une phrase en occitan étant donnée une phrase en français et vice versa
- Observation des poids d'attention pour l'intérprétation du modèle.
- Optionnel 1 : comparer les résultats en fonction des variétés.
- Optionnel 2 (ou second sujet à part entière) proposer une méthode permettant d'apparier deux phrases mot à mot.

Généralisation des modèles

L'une des grands problèmes du machine learning est le manque de généralisation des modèles. L'année dernière a été lancé un challenge dans le but de se concentrer sur cet aspect: **The Abstraction and Reasoning Corpus (ARC)**.



Comprendre les règles d'un jeu en quelques exemples

`https://github.com/fchollet/ARC`

Généralisation des modèles

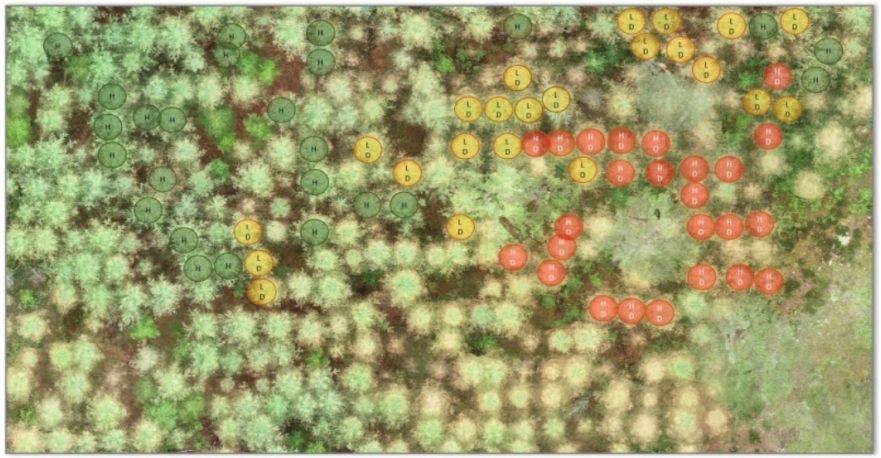
Le corpus est composée de données d'apprentissage et de test. La particularité est que les tâches d'apprentissage et de test ne sont pas les mêmes.

Résultats attendus:

- Résoudre le problème serait très difficile. Dans le cadre de ce projet, il s'agit de proposer une solution imparfaite qui marchera dans certains cas.
- Les étudiants seront d'avantage évalués sur la compréhension du problème et l'originalité de leur solution que sur les performances obtenues.

Détection visuelle d'arbres malades

Le Porte-case du mélèze (*Lepidoptera Coleophoridae* est un papillon responsable d'importants dommage dans les forêts de mélèze en Suède. Afin d'étudier cette épidémie, les chercheurs utilisent des images issues de drones afin de déterminer l'étendue des zones touchées.



Il s'agit donc d'un problème de détection d'objets ou il faut classer les arbres en 4 catégories : Healthy (H), Light Damage (LD), High Damage (HD), and Other.

Détection visuelle d'arbres malades

Données 101,878 arbres annotés sur 1,543 images.

		Annotations				
Dataset name	**Images**	Larch Trees	Other Trees	Healthy (H)	Light damage (LD)	High damage (HD)
Bebehojd_20190527	172	8,749	2,570	142	6,279	2,328
Ekbacka_20190527	215	6,423	2,702	327	5,126	970
Jallasvag_20190527	50	2,580	475	625	1,743	212
Kampe_20190527	179	15,508	3,219	1,905	11,318	2,285
Nordkap_20190527	224	11,262	2,987	488	8,668	2,106
Bebehöjd_20190819	149	7,014	2,949			
Ekbacka_20190819	181	6,768	3,868			
Jallasvag_20190819	39	1,533	1,486			
Kampe_20190819	113	8,134	1,537			
Nordkap_20190819	221	8,656	3,458			
	1,543	**76,627**	**25,251**	**3,487**	**33,134**	**7,901**

Résultats attendus:

- Distinguer les mélèzes des autres arbres
- Catégoriser les mélèzes suivant leur santé.

Moteur de recherche pour l'écologie

La dernière bibliothèque est un site de citations en rapport avec le changement climatique.

https://la-derniere-bibliotheque.org/partage/

Le but de ce projet est d'améliorer le moteur de recherche existants grâce à un nouveau corpus sur l'écologie.

Un jeu de données d'évaluation et un outil a été constitué pour le data challenge IAPau 2022.

Moteur de recherche pour l'écologie

Données Environ 1 million de tweets extraits de plusieurs communautés parlant d'écologie. (Vous pouvez trouver d'autres données)
Résultats attendus:

- Ré-apprendre un nouveau modèle de langage avec Fastext
  - Étudier une stratégie de sélection de données
- Comparer avec un modèle de type Camembert ou Flaubert.

Classification de décharge sauvages dans les images aériennes

Données Environ 10000 images aériennes

Examples of the presence of waste in potentially illegal sites. Red circles indicate suspicious objects. In all the images accumulations of various materials and scattered waste are present.

Résultats attendus:

- Construire un classifieur d'images

Visualisation de l'actualité de la transition énergétique

Données Une sociologue collecte les titres de journaux sur cette thématique avec les liens.

Elle nous a fourni les titres depuis janvier à octobre 2023

February 1, 2023 at 10:48AM	La voiture électrique poursuit sa conquête du marché européen
February 1, 2023 at 10:48AM	Les ventes de voitures électriques ont encore augmenté en Europe en 2022
February 1, 2023 at 10:48AM	Mix énergétique : le détail des 3 scénarios possibles d'ici 2050, selon le rapport BP Energy ...
February 1, 2023 at 10:56AM	Can new cheap, frequent "laser" monitoring of critical components extend Nuclear plant lifetimes by decades?
February 1, 2023 at 10:56AM	Electricité : pour les consommateurs, la hausse des factures devient concrète

Résultats attendus:

- Construire un système de clustering et de visualisation de cette actualité

Visualisation de l'actualité de la transition énergétique

Une sociologue collecte les titres de journaux ...

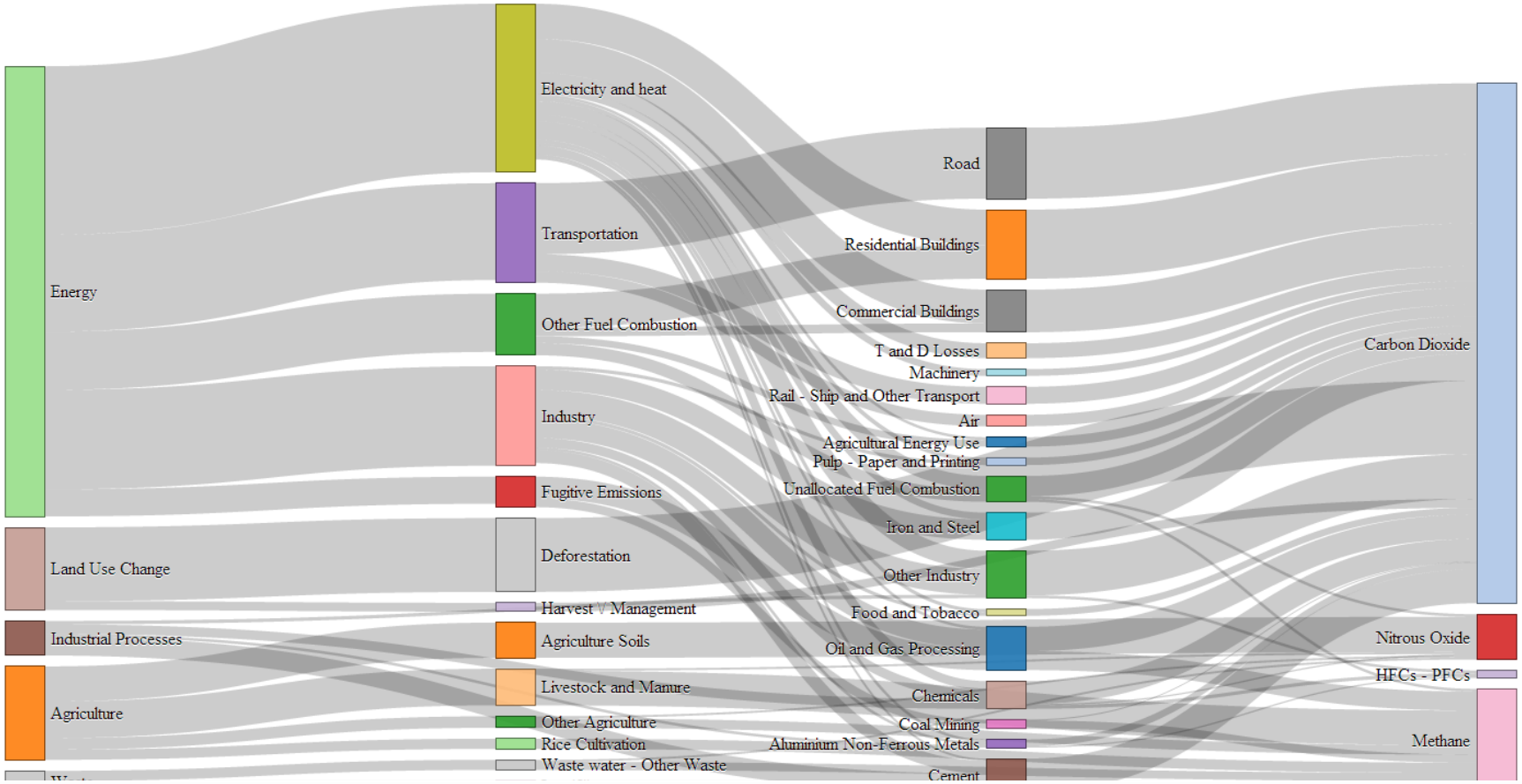
Clothilde Chagny, CEA

IThesee – Prométhée : Projet d'Observatoire Multidisciplinaire de l'Énergie pour une Transition Humaine, Économique et Écologique

Visualisation de l'actualité de la transition énergétique

Résultats attendus:

- Construire un système de clustering et de visualisation de cette actualité



Et aussi....

Proposés par les étudiants de cytech les années précédentes :

- Agent autonome appris par Active learning pour un jeu vidéo
- Génération de texte de raps

Ces sujets sont aussi intéressants :

- Apprentissage actif
- Apprentissage fédéré
- Redéfinition de coeur cuda
- Intérprétation des résultats des modèles deep learning
- Votre propre sujet...

Avant le jeudi 29 février

Envoyer le nom du sujet que vous avez choisi ainsi que les noms des membres de votre équipe à l'adresse paul.gay@univ-pau.fr

Pour aujourd'hui, n'oubliez d'éteindre vos GPUs !!