TABLEAU C1 – PLAN DE COURS-CADRE

Ce document doit décrire brièvement le contenu du cours, les objectifs, les méthodes pédagogiques et l'évaluation.

SIGLE	IFT 6XXX
NOMBRE DE CRÉDITS	4
TITRE LONG	Vision and Language
TITRE COURT	Vision and Language

1. CONTENU DU COURS

What is Vision and Language: This will be a seminar course on recent advances in vision and language research – a sub-field of artificial intelligence that studies multimodal tasks at the intersection of computer vision and natural language processing. Some examples of these tasks include image / video captioning (automatically describing images / videos in natural language), visual question answering (automatically answering natural language questions about images / videos), visual dialog (holding a conversation with a human grounded in an image), visual commonsense reasoning (automatically answering questions involving commonsense reasoning about situations described in images) etc.

Why study Vision and Language: Vision and Language research has seen tremendous progress over the past decade, owing to the availability of large-scale datasets, development of high-capacity deep learning models and availability of computational resources. There are various motivations behind studying vision and language:

- Vision and Language tasks such as visual question answering, image captioning provide a natural testbed to evaluate how good our current visual understanding systems are and how grounded our current natural language understanding systems are.
- Vision and Language tasks have many potential applications such as serving as an aid for visually impaired users (helping them navigate the visual world by talking in natural language), serving as day-to-day assistants in our homes (imagine Siri with eyes), aiding children in learning through interactive demos.
- Vision and Language research involves multiple research challenges such as visual recognition, natural language understanding and grounding, learning joint visio-linguistic representations, reasoning about commonsense and knowledge bases, learning to overcome spurious correlations in training data.

Topics covered: Major Vision and Language tasks, datasets, modelling techniques and their shortcomings, such as:

- Tasks such as image-caption retrieval, referring expressions, image captioning, visual question answering, visual dialog, visual commonsense reasoning.
- Datasets such as Flickr30k, COCO Captions, VQA, Visual Genome, GQA, CLEVR, Visual Dialog, VCR.
- Modelling techniques such as attention, multimodal pooling, compositional networks, multimodal transformers.
- Shortcomings of current state-of-the-art models such as lack of robustness to new distributions, lack of compositional understanding and reasoning.

2. OBJECTIFS ET COMPÉTENCES VISÉS

- Gain a thorough understanding of recent advances in Vision and Language (tasks, datasets, modelling techniques, shortcomings).
- Develop the ability to read and critique research papers in Vision and Language.
- Be able to identify interesting open research questions and challenges in Vision and Language.
- Be able to execute a research project in Vision and Language.
- Enhance presentation skills.

3. PRINCIPALES MÉTHODES PÉDAGOGIQUES

- Introductory lectures in the first few classes providing an overview of major Vision and Language tasks, datasets, modelling techniques. A subset of these could also be guest lectures from researchers working on these tasks.
- Reading and reviewing research papers. After the first few classes, students will read and write technical reviews (conference style reviews) for one research paper prior to each class.
- Discussing pros and cons of research papers. In each class (after the introductory classes), two students will lead a discussion on strengths and weaknesses (one student leading each aspect) of the paper that was reviewed.
- Course projects. Each student will work on a course project (in teams of 2-3 students). These projects can range from coming up with new Vision and Language task, to developing new modelling techniques and advancing the state of the art, to applying an existing technique for a new task / dataset, to analyzing the behavior of existing models and providing new insights. For each project, there will be five deliverables spread across the term:
    - Initial presentation presenting the project idea to the class.
    - First progress update presentation to the class.
    - Second progress update presentation to the class.
    - Final project presentation to the class.
    - Spotlight project video (1 min) summarizing the project, to be submitted to the TA.

4. DÉMARCHE ÉVALUATIVE ET PONDÉRATION *(à titre indicatif seulement)*

- Paper reviews – 30%
- Paper discussion in class – 10%
- Course Project – 60%
    - Initial presentation – 10%
    - First progress update presentation – 10%
    - Second progress update presentation – 10%
    - Final presentation – 15%
    - Spotlight video – 15%