TÍTULO: Procesamiento automático de lenguaje situado en un entorno visual			
AÑO: 2023	**CUATRIMESTRE:** 2°	**N° DE CRÉDITOS:**	**VIGENCIA:** 3 años
CARGA HORARIA: 60 horas de teoría y 10 horas de práctica			
CARRERA/S: Doctorado en Ciencias de la Computación			

FUNDAMENTOS

Con el advenimiento de la era del aprendizaje automático (DL, del inglés deep learning), tareas relacionadas a la visión por computadora (CV, del inglés computer vision) y al procesamiento automático del lenguaje natural (NLP, del inglés natural language processing) han logrado otorgar resultados prometedores para muchas aplicaciones útiles.
Históricamente, construir sistemas automáticos que sean capaces de explotar modelos multimodales y así interpretar y generar lenguaje natural, ha sido considerado un objetivo ambicioso. Sin embargo, esta última década ha visto enormes progresos en sistemas interactivos de NLP y CV; otorgando resultados muy prometedores sobre benchmarks del área de lenguaje y visión. Dada su naturaleza, esta reciente área de la inteligencia artificial, intenta ser el puente que permita convertir información y conceptos visuales, a lenguaje y viceversa; mediante la aplicación de los conocimientos y los últimos adelantos conseguidos en las disciplinas de CV y NLP.

OBJETIVOS

Este curso se centra en el campo de procesamiento automático de lenguaje situado en un entorno visual (PALSEV). Esta es una subárea de la inteligencia artificial que estudia la conexión entre el procesamiento automático de lenguaje natural situado, y el procesamiento automático de imágenes. Proporciona a los estudiantes una visión general de los avances recientes al mismo tiempo que revisa los desafíos de larga data establecidos por la comunidad de inteligencia artificial en sus inicios. Establece conexiones entre el procesamiento de lenguaje natural, la visión por computadora y la interfaces humano-robot. Cubre tanto la comprensión de lenguaje natural situado como la generación de lenguaje natural situado, así como una arquitectura unificada para estos dos componentes cruciales de los agentes de inteligencia artificial. El curso finaliza proporcionando a los estudiantes herramientas sobre la conexión entre PALSEV y la interacción humano robot, y comparando los mecanismos de negociación de significado de los seres humanos en el marco de la interpretación multimodal de lenguaje y los mecanismos de modelos multimodales de última generación.

PROGRAMA

Unidad I: Grounding en el procesamiento de lenguaje situado.
Unidad 1: El problema de grounding. Introducción al curso. ¿Por qué grounding? Desde el pasado computacional lejano hasta el cercano

Introducción al concepto del Problema del grounding y su importancia en el campo.
Exploración de la evolución del grounding en el pasado computacional.
Discusión sobre la relevancia actual del grounding en el procesamiento de lenguaje situado.

Unidad II: Modelos Computacionales para Representaciones Conceptuales Multimodales
Unidad 2: Modelos Computacionales para Representaciones Conceptuales Multimodales

Estudio de modelos y enfoques computacionales para la representación de conceptos multimodales.
Integración de información de diferentes modalidades (texto, imágenes, sonido, etc.) en la representación de conceptos.
Análisis de técnicas para desarrollar representaciones conceptuales ricas y robustas.

Unidad III: Procesamiento de Preguntas Visuales y de Expresiones Referenciales
Unidad 3: Procesamiento de Preguntas Visuales y de Expresiones Referenciales

Exploración del campo de Preguntas y Respuestas Visuales (Visual Question Answering).
Estudio de técnicas y modelos para comprender preguntas formuladas sobre imágenes y generar respuestas basadas en el contenido visual.
Ejemplos y aplicaciones prácticas de sistemas de preguntas y respuestas visuales.
Resolución de expresiones referenciales. Clasificación de expresiones referenciales. Análisis de error. Aplicaciones en sistemas de asistencia para personas no videntes.

Unidad IV: Diálogos Visuales Hombre-Computadora
Unidad 4: Diálogos Visuales Hombre-Computadora

Análisis de los Diálogos Visuales (Visual Dialogues).
Estudio de enfoques para generar y mantener una conversación visual entre humanos y sistemas basados en imágenes. Resolución colaborativa de referencias.
Exploración de técnicas para el procesamiento de lenguaje situado en el contexto de los diálogos visuales. Conceptos de región bajo discusión y pregunta bajo discusión. Teaming entre personas y agentes.

Unidad V: Anclaje Colaborativo
Unidad 5: Anclaje Colaborativo, Evidencia de Anclaje Negativa y Evidencia de Anclaje Positiva

Discusión sobre el anclaje colaborativo en el procesamiento de lenguaje situado.
Análisis de la evidencia de anclaje negativa y positiva en el contexto del anclaje.
Exploración de cómo se utiliza la evidencia de anclaje para mejorar la comprensión y generación de lenguaje situado. Anotación de evidencia negativa y positiva en diálogos multimodales hablados y escritos.

Unidad VI: Consideraciones Éticas y Conclusiones
Unidad 6: Consideraciones Éticas

Exploración de los aspectos éticos y responsabilidad en el Procesamiento de Lenguaje Situado.
Discusión sobre el sesgo y la equidad en los sistemas de lenguaje situado.
Reflexión sobre la privacidad, seguridad y responsabilidad social en el uso de sistemas de lenguaje situado.
Conceptos e implicancias éticas de la teoría de la cortesía, entrainment y otros aspectos sociales del lenguaje situado.
Desafíos éticos y técnicos del área.

PRÁCTICAS

El curso consistirá de 2 trabajos prácticos entregables escritos. Uno será un estado del arte de una tarea de lenguaje situado en visión elegida por el estudiante y una implementación de un algoritmo de aprendizaje automatizado del estado del arte para la tarea elegida. Otro será una evaluación de un artículo científico elegido por la profesora.

BIBLIOGRAFÍA

Bibliografía básica. Toda la bibliografía consiste de artículos de ciencia abierta y los pdfs están disponibles para los estudiantes.

Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346.
F. Pulvermüller (2005) Brain mechanisms linking language and action
Kafle, K., Shrestha, R., & Kanan, C. (2019). Challenges and prospects in vision and language research. Frontiers in Artificial Intelligence, 2, 28.
Marco Baroni (2016) Grounding Distributional Semantics in the Visual World
Lisa Beinborn, Teresa Botschen and Iryna Gurevych (2018) Multimodal Grounding for Language Processing
Hossain, Sohele, Shiratuddin, Laga (2019) A Comprehensive Survey of Deep Learning for Image Captioning
Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research (JAIR) 55 (2016), 409–442.
Kushal Kafle and Christopher Kanan (2017) Visual Question Answering: Datasets, Algorithms, and Future Challenges
Raffaella Bernardi and Sandro Pezzelle (2021) Linguistic issues behind visual question answering
Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2021). Visual question answering using deep learning: A survey and performance analysis. In S. Singh, P. Roy, B. Raman, & P. Nagabhushan (Eds.), Computer vision and image processing. CVIP 2020, volume 1377 of communications in computer and information science. Springer.
Chen, Lao, Duan (2020) Multimodal Fusion of Visual Dialog: A Survey

Bibliografía complementaria

Michael J. Mayo (2003) Symbol Grounding and its Implications for Artificial Intelligence
Lazaridou, Bruni and Baroni Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world ACL 2014
Douwe Kiela, Alexis Conneau, Allan Jabri, Maximilian Nickel (2018) Learning Visually Grounded Sentence Representations
Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, Jingjing Liu Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan Show and Tell: A Neural Image Caption Generator
Will Monroe, Robert X.D. Hawkins, Noah D. Goodman and Christopher Potts Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding
Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077–6086).
Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel

Yedidsion, Justin Hart, Peter Stone, Raymond J. Mooney (2019) Improving Grounded Natural Language Understanding through Human-Robot Dialog

MODALIDAD DE EVALUACIÓN

El curso consistirá de 2 trabajos entregables escritos. Uno será un estado del arte de una tarea de lenguaje situado en visión elegida por el estudiante y una implementación de un algoritmo de aprendizaje automatizado del estado del arte para la tarea elegida. Otro será una evaluación de un artículo científico elegido por la profesora. Luego tendrá un examen oral obligatorio, individual e integrador de los contenidos del curso.

REQUERIMIENTOS PARA EL CURSADO

conocimientos avanzados de aprendizaje automatizado
habilidades maduras de programación
conocimientos avanzados de bases de datos en distintas modalidades