**DANIEL ULISES CAMPOS DELGADO**
*ducd@fciencias.uaslp.mx*
FACULTAD DE CIENCIAS, UASLP

# Bienvenida la multimodalidad en la inteligencia artificial

Los seres humanos percibimos, comprendemos y valoramos el mundo a través de nuestros sentidos. Como ejemplo consideremos los dos escenarios siguientes. La opinión de un restaurante la construimos a través de nuestra percepción multiple entre la decoración (visual), la descripción y estructura del menú (texto), la música de fondo (audio), la preparación de la comida (gusto), y la atención del personal (visual y audio). Por otro lado, un autor al buscar transmitir una historia puede hacerlo de forma escrita, por medio de un texto; a través de un relato en audio; o de forma visual por un video. A estas diversas formas que tenemos de recibir y procesar información de nuestro entorno, se les llama modalidades. Un último ejemplo puede ser con el avance de las telecomunicaciones, donde la evolución se percibe a través del aumento progresivo de las modalidades disponibles para comunicarse. De esta manera, en los inicios del siglo XX, solo compartíamos información a través de cartas (texto). El siguiente gran avance se generó por medio del telefono (audio), donde se establecia una comunicación por voz. A finales del mismo siglo, con al auge de la telefonía celular y el internet, se tuvo la oportunidad de tener videollamadas, y así compartir voz e imagen entre los interlocutores. Ahora, asociado con las comunicaciones de sexta generación o 6G, se espera en un futuro cercano tener transmisiones holográficas y ápticas, es decir donde se compartan imágenes 3D y con retroalimentación sensorial entre los interlocutores.

Recientemente, las aplicaciones de inteligencia artificial como ChatGPT (https://chat.openai.com/) o Bard (https://bard.google.com/), se centraron en una sola modalidad para recibir y generar un diálogo con los usuarios, así inicialmente solo compredían y entablaban una conversación por medio de texto. Un primer enfoque hacia el uso de múltiples modalidades o multimodalidad en la interlocución, se tuvo con las aplicaciones que generan imágenes a través de una descripción en texto de ellas, por ejemplo DALL-E-3 (https://openai.com/dall-e-3), Craiyon (https://www.craiyon.com), Deep Dream Generator (https://deepdreamgenerator.com/) y Dreamlike (https://dreamlike.art/). Estas aplicaciones a partir de una modalidad de entrada, como es el texto, tiene una modalidad diferente de salida, que sería la imagen construida. Un avance hacia una interacción multimodal se tuvo al final del 2023, por medio de GPT-4V de la empresa OpenAI, y que se puede accesar por medio de ChatGPT en la opción "ChatGPT 4", que permite analizar las imágenes ingresadas por el usuario y establecer un diálogo a través de la información visual, así como generar imágenes por medio de texto de entrada. De esta forma, la interacción con el usuario se enriquece y se presenta ahora de forma bidireccional entre texto e imágenes, lo cual es un avance significativo hacia la multimodalidad. ¿Pero qué ocurre con otras modalidades de comunicación? Por ejemplo, si se buscara interaccionar por medio del audio y el video. Aquí es donde se centra el más reciente avance de Google DeepMind presentado a inicios de diciembre de 2023, por medio de Gemini, que promete centrarse por completo en una interacción multimodal para comprender texto, imágenes, video, audio e inclusive código de programación. Al revisar los videos demostrativos en el sito web https://deepmind.google/technologies/gemini/ la capacidad de interacción multimodal es realmenrte sorprendente. Sin embargo, para el público en general, actualmente Gemini puede accesarse parcialmente por medio de Bard.

Pero aquí mi pregunta, ¿este es el futuro de la inteligencia artificial? En mi opinión creo que sí lo es, al igual que en la evolución de las telecomunicaciones, la multimodalidad incrementó la capacidad de comunicarnos y sentirnos cerca entre nosotros; en la inteligencia artificial, la multimodalidad permitirá simplificarnos tareas rutinarias y agregar dimensiones nunca antes pensadas, por ejemplo, desarrollar herramientas de apoyo a personas con discapacidad visual, incrementar el potencial de la realidad aumentada, ayudar a la navegación de los vehículos autonomos, mejorar las capacidades del diagnóstico médico, generar sistemas de vigilancia multisensoriales, entre muchos otros. Sin embargo, un aspecto que no debemos de dejar a un lado, son los retos éticos de este tipo de tecnologías. Al respecto, ChatGPT y Bard incluyen un monitoreo de las preguntas y peticiones de los usuarios, de manera que las respuestas no violen aspectos de privacidad, generen información falsa, ni iniciten a conductas violentas. Asimismo, los creadores de estas tecnologías deben cuidar que estas aplicaciones no perpetuen esterotipos, ni provean respuestas con sesgos. Dicho lo anterior, no cabe duda que la inteligencia artificial está revolucionando nuestro mundo a un ritmo impresionante, por lo que es crucial valorar el alcance de la multimodalidad y entenderla como un completo a nuestras capacidades, pero no como un reemplazo, siendo el gran reto, la democratización de estas técnologías en países como el nuestro. UP