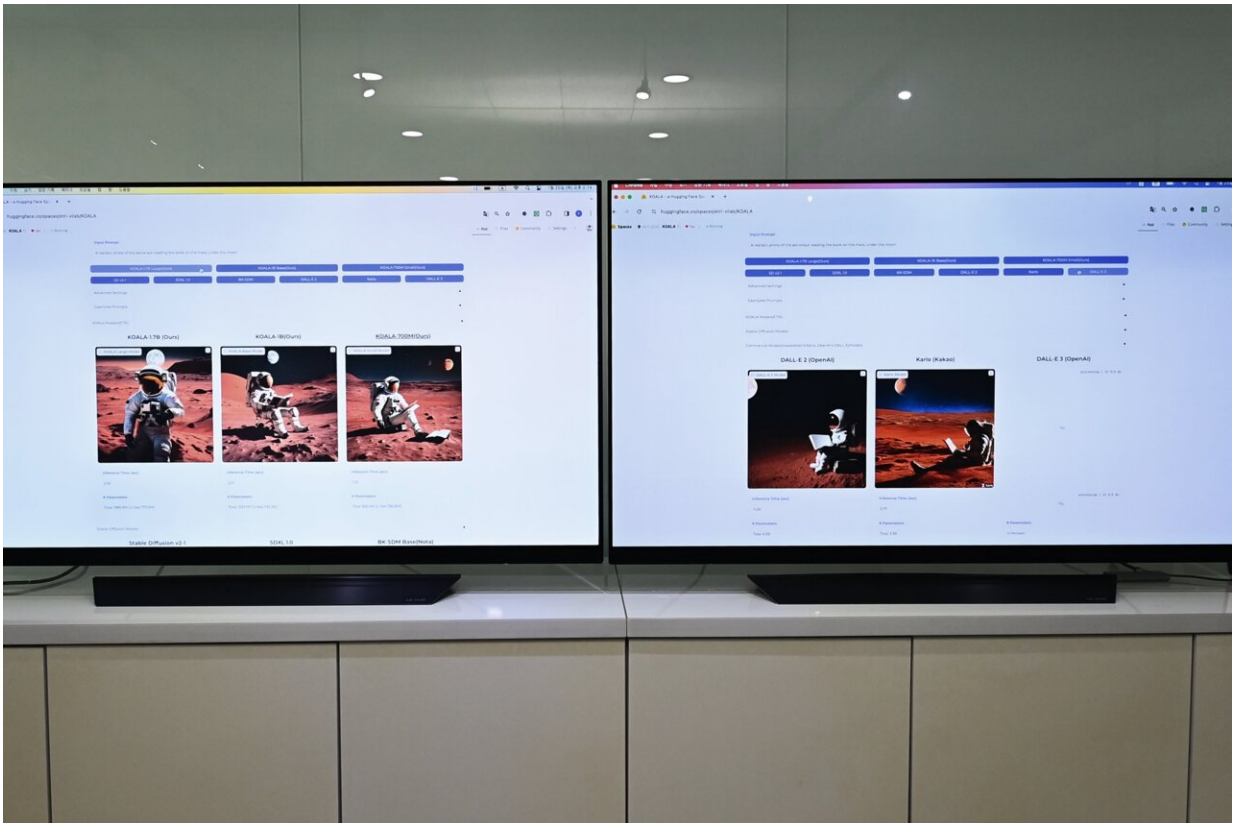
Ultra-fast generative visual intelligence model creates images in just 2 seconds

February 22 2024



ETRI unveils ultra-fast generative visual intelligence model_2. Credit: Electronics and Telecommunications Research Institute(ETRI)

ETRI's researchers have unveiled a technology that combines generative AI and visual intelligence to create images from text inputs in just 2

seconds, propelling the field of ultra-fast generative visual intelligence.

Electronics and Telecommunications Research Institute (ETRI) announced the release of five types of models to the public. These include three models of 'KOALA,' which generate images from text inputs five times faster than existing methods, and two conversational visual-language models 'Ko-LLaVA' which can perform question-answering with images or videos.

The 'KOALA' model significantly reduced the parameters from 2.56B (2.56 billion) of the public SW model to 700M (700 million) using the knowledge distillation technique. A high number of parameters typically means more computations, leading to longer processing times and increased operational costs. The researchers reduced the model size by a third and improved the generation of high-resolution images to be twice as fast as before and five times faster compared to DALL-E 3.

ETRI has managed to reduce the model's size(1.7B (Large), 1B (Base), 700M (Small)) considerably and increase the generation speed to around 2 seconds, enabling its operation on low-cost GPUs with only 8GB of memory amidst the competitive landscape of text-to-image generation both domestically and internationally.

ETRI's three 'KOALA' models, developed in-house, have been released in the HuggingFace environment.

In practice, when the research team input the sentence "a picture of an astronaut reading a book under the moon on Mars," ETRI-developed KOALA 700M model created the image in just 1.6 seconds, significantly faster than Kakao Brain's Kallo (3.8 seconds), OpenAI's DALL-E 2 (12.3 seconds), and DALL-E 3 (13.7 seconds).

ETRI also launched a website where users can directly compare and

experience a total of 9 models, including the two publicly available stable diffusion models, BK-SDM, Karlo, DALL-E 2, DALL-E 3, and the three KOALA models.

Furthermore, the research team unveiled the conversational visual-language model 'Ko-LLaVA,' which adds visual intelligence to conversational AI like ChatGPT. This model can retrieve images or videos and perform question-answering in Korean about them.

The 'LLaVA' model was developed in an international joint research project with the University of Wisconsin-Madison and ETRI, presented at the prestigious AI conference NeurIPS'23, and utilizes the open-source LLaVA(Large Language and Vision Assistant) with image interpretation capabilities at the level of GPT-4.

The researchers are conducting extension research to improve Korean language understanding and introduce unprecedented video interpretation capabilities based on the LLaVA model, which is emerging as an alternative to multimodal models including images.

Additionally, ETRI pre-released its own Korean-based compact language understanding-generation model (KEByT5). The released models (330M (Small), 580M (Base), 1.23B (Large)) apply token-free technology capable of handling neologisms and untrained words. Training speed was enhanced by more than 2.7 times, and inference speed by more than 1.4 times.

The research team anticipates a gradual shift in the generative AI market from text-centric generative models to multimodal generative models, with an emerging trend towards smaller, more efficient models in the competitive landscape of model sizes.

The reason why ETRI is making this model public is to foster an

ecosystem in the related market by reducing the model size, which traditionally would require thousands of servers, thereby facilitating usage among small and medium-sized enterprises.

In the future, the research team expects high demand for Korean cross-modal models that integrate visual intelligence technology into prominent open-language models of generative AI.

The team highlighted that the core patent of this technology is based on knowledge distillation, a technology that enables small models to perform the role of large models by accumulating knowledge using AI.

After making this technology public, ETRI plans to transfer it to image generation services, creative education services, content production, and businesses.

Lee Yong-Ju, director of ETRI's Visual Intelligence Research Section, stated, "Through various endeavors in generative AI technology, we plan to release a range of models that are small in size but excel in performance. Our global research aims to break the dependency on existing large models and provide domestic small and medium-sized enterprises with the opportunity to effectively utilize AI technology."

Professor Lee Yong-Jae from the University of Wisconsin-Madison, who oversees the LLaVA project, mentioned, "In leading the LLaVA project, we conducted research on open-source-based visual-language models to make it accessible to more people, competing against GPT-4. We plan to continue our research on multimodal generative models through international joint research with ETRI."

The research team aims to showcase world-class research capabilities, moving beyond the conventional types of generative AI that convert text inputs into textual responses. They plan to extend their research to types

that respond with sentences to images or videos, and types that respond with images or videos to sentences.

Provided by National Research Council of Science and Technology