Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: ML-Engineer.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

VL Models (Vision-Language Models)

Vision language models are models that can learn simultaneously from images and texts to tackle many tasks, from visual question answering to image captioning.

What is a Vision Language Model?

Vision language models are broadly defined as multimodal models that can learn from images and text. They are a type of generative models that take image and text inputs, and generate text outputs. Large vision language models have good zero-shot capabilities, generalize well, and can work with many types of images, including documents, web pages, and more. The use cases include chatting about images, image recognition via instructions, visual question answering, document understanding, image captioning, and others. Some vision language models can also capture spatial properties in an image. These models can output bounding boxes or segmentation masks when prompted to detect or segment a particular subject, or they can localize different entities or answer questions about their relative or absolute positions. There's a lot of diversity within the existing set of large vision language models, the data they were trained on, how they encode images, and, thus, their capabilities.