# Hugging Face Mutimodal

멀티모달

Hugging Face

# 멀티모달(Multimodal)

# 멀티모달(Multimodal)

학습은 텍스트, 이미지, 음성 등 여러 종류의 데이터를 동시에 처리하고 이해하는 인공지능 기술을 의미.

인간의 복합적인 감각 처리 방식을 모방하여 다양 모달리티(형태)의 정보를 통합적으로 이해하고 활용하는 것을 목표로 한다.

**1. 이미지 캡셔닝(BLIP-2: Bootstrapping Language-Image Pre-training)**

: 이미지의 내용을 자연어로 설명

**2. 이미지 생성(Stable Diffusion)**

텍스트 설명을 바탕으로 이미지를 생성하는 방버

# 이미지 캡셔닝 : BLIP-2

이미지 캡셔닝(Image captioning, Image-to-text)은 컴퓨터 비전과 자연어 처리 기술을 결합해 주어진 이미지의 내용을 설명하는 문장이나 캡션(이미지나 그림 등의 내용을 간단하게 설명하는 텍스트)을 생성하는 작업

BLIP(Bootstrapping Lanauage-Image Pre-training)

이미지와 텍스트 간의 관계를 효과적으로 학습해 우수한 이미지 캡셔닝 성능을 보이는 최신 모델.

약 3억개의 이미지-텍스트 쌍으로 사전 학습됐으며 이를 통한 시각 및 언어 표현을 통합적으로 학습.

대규모 데이터세트로 사전 학습되어 강력한 표현력을 갖추고 있으며 이미지와 텍스트의 상호 작업을 직접적으로 모델링해 두 모달리티 간의 관계를 잘 파악 할 수 있다.

이미지 캡셔닝 외에도 시각 질의 응답, 이미지-텍스트 검색등 다양한 과제에 활용할 수 있다.

# 이미지 생성 : Stable-Diffusion

이미지 생성(Image generation)은 자연어 처리와 컴퓨터 비젼 기술을 융합해 텍스트 프롬프트 기반으로 새로운 이미지를 만들어 내는 과정.

텍스트를 시각적 정보로 변환하는 것을 목표로 하며 창작, 디자인, 시각화 등 다양한 분야에서 활용.