Salesforce의 InstructBLIP

얼마 전에 Salesforce 에서 BLIP2를 공개했었습니다. 높은 성능을 보이는 Vision-Language LLM 이었고, 테스트를 해
봤을 때 꽤 흥미로운 결과 (예를 들어 이미지 내의 사람이나 객체 갯수를 잘 카운팅하는 등) 를 보여주기도 해서 발전 속도가 놀랍다고 생각을
했었습니다. 이번에 Salesforce에서 그 후속으로 InstrcutBLIP을 공개하였습니다. \- 코드:

https://github.com/salesforce/LAVIS/tree/main/projects/instructblip 기존의 BLIP2 기반으로 여러 가지 public dataset 을 가져다가 instruct tuning이 가능한 형태로 만들어서 fine-tuning을
진행한 것이고, 매우 큰 Vision-Language 데이터셋인 Flamingo를 포함한 여러 데이터에 대해서 zero-shot 에서
SOTA를 달성하였다고 합니다. 논문을 살펴보면 구조 자체는 기존 BLIP2 와 다를 것이 별로 없습니다. BLIP의 핵심인 Q-former
부분에 기존에는 Query + Text 형태로 넣던 것에서 text를 instruction 으로 한 정도만 차이라고 볼 수 있습니다.
(Instruct-tuning 이므로 당연한 것입니다만...ㅎㅎ) 이번 연구는 모델 아키텍쳐 등이 개선되었다기 보다는, 여러 데이터셋에
대해서 fine-tuning을 진행하고 이 모델을 공개한 것에 의의가 있다고 볼 수 있습니다. 여기서는 Vicuna와 FlanT5기반으로 된
모델을 공개하였는데 MiniGPT4 (
https://github.com/Vision-CAIR/MiniGPT-4) 와 비슷하네요. 어찌되었건
최근에 Vision/Language를 같이 다루는 LLM들이 많아지고 있고 점점 더 여러 modality 로 확장되어 가는 것 같습니다.
그와는 별개로 Meta의 OPT, LLaMA나 Google의 FlanT5 등은 오픈소스 쪽에 정말 큰 기여를 하고 있는 것 같습니다.