**LXMERT: Learning Cross-Modality Encoder Representations from Transformers(EMNLP 2019)**

Visual Question Answering 2019 challenge 3등(공식 레퍼런스는 1등)

**Main Idea**

Multi-Modality 사이의 추론 능력 다루는 Task들

1. 비주얼적인 개념
2. 언어의 의미
3. 두 모달리티(시각, 언어) 사이의 관계를 잘 파악

이 3가지가 중요

싱글 modality에선 비전-이미지넷, 언어-BERT 등 사전 학습에서 좋은 representation 얻을 수 있다. -> 확장해서 비전-언어 사이의 연결을 사전 학습하는 방법을 제안(LXMERT)

1. Object relationship encoder
2. Language encoder
3. Cross-modality encoder

의 3가지 인코더가 있다. 각각은 트랜스포머로 구성, 그리고 비전과 언어의 의미를 연결할 수 있는 능력 학습을 위해 5가지 tasks로 pre-training

각각의 task로 intra-modality 뿐만 아니라 inter-modality(cross modality) 또한 학습 가능  
이 모델을 여러 VQA 데이터셋에 fine-tuning 했을 때 SOTA 가능