[논문 리뷰] PaLM-E: An Embodied Multimodal Language Model

diddu · 방금 전

논문

🔎 논문 리뷰

목록 보기 4/4

💡 PaLM-E: An Embodied Multimodal Language Model_논문 리뷰

📌 Abstract

LLM들은 복잡한 작업을 수행하는 것이 가능하지만 real-world에선 예를들어 로봇과제처럼 일반적인 추론을 수행하기 위해서는 언어와 센서를 연결하는 **grounding** 문제가 발생한다.

- `Embodied Language Models`
  이 논문은 embodied language models를 제안한다. 이 모델은 실제 세계에서 언어 모델과 연속적인 센서 모달리티를 직접 통합하여 언어와 인식 사이의 연결을 구축한다. 텍스트 입력 외에도 실제 세계에서 발생하는 연속 입력을 직접 통합하여 언어 모델이 실제 세계의 순차적 의사 결정을 더 잘 이해하고 추론할 수 있게 한다.

- `Multi-modal Input`
  embodied language 모델의 입력은 Multi-modal 문장이다. real-world data인 이미지와 상태 추정 데이터는 언어 토큰과 동일한 잠재적 인코딩에 통합되며, Transformer 기반의 LLM의 self-attention layers를 통해 텍스트와 동일한 방식으로 처리된다.

- `End-to-end Training`
  이런 인코딩들은 end-to-end로 훈련되며, 미리 훈련된 LLM과 함께 다양한 embodied 과제에 대해 훈련된다.

📌 Architectural

> **Architectural Idea** : PaLM-E architecture의 핵심 개념은 이미지, 상태 추정 등의 연속적인 embodied 관찰을 사전 학습된 언어 모델의 언어 임베딩 공간으로 주입하는 것이다. 연속적인 관찰을 언어 토큰의 임베딩 공간과 동일한 차원의 벡터 시퀀스로 인코딩한다.

- `Decoder-only LLMs and Prefix-decoder-only LLMs`
  PaLM-E는 prefix 또는 prompt가 주어진 상태에서 autoregressively 텍스트 완성을 생성하는 decoder-only LLM이다. 이러한 접근 방식은 LLM이 autoregressive하므로, pre-trained model은 architecture를 변경할 필요 없이 prefix에 condition 될 수 있다.

- `Token Embedding Space and Injection of Continuous Observations`
  내부적으로, LLM은 w_i를 word token embedding space X로 임베드합니다. 이렇게 하면, 이미지 관찰과 같은 멀티모달 정보는 직접 연속적인 관찰을 언어 임베딩 공간 X로 매핑함으로써 LLM에 주입할 수 있습니다.

- `Embodying the Output: Robot Control Loop`
  PaLM-E는 입력으로 멀티모달 문장을 기반으로 텍스트를 생성하는 generative model 이다. 모델 출력을 embodiment와 연결하기 위해, embodied question answering 또는 scene description tasks와 같이 텍스트만 출력하여 해결할 수 있는 경우와, PaLM-E가 embodied planning 또는 control task를 해결하는데 사용되며 low-level commands에 condition되는 텍스트를 생성해야 하는 경우 두 가지 경우를 구분한다.

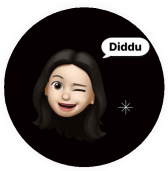
📌 Conclusion

- `Generalist vs Specialist Models - Transfer:`
  이 연구에서는 PaLM-E가 다양한 작업과 데이터셋에서 동시에 학습되면 각각의 작업에 대해 따로 학습된 모델들에 비해 성능이 크게 향상된다는 것을 보여주었다. 즉, 다양한 작업을 함께 학습하는 것이 전이 학습(transfer learning)을 촉진하며, 이를 통해 모델의 성능을 높일 수 있다.

- `Data Efficiency`
  로보틱스 데이터는 대규모 언어나 시각-언어 데이터셋에 비해 상당히 부족하다. 하지만 이 연구에서 제시된 모델은 전이학습을 통해 로보틱스 도메인에서 매우 적은 수의 학습 예제로도 로보틱스 작업을 해결할 수 있는 능력을 보여준다.

- `Retaining Language Capabilities`
  멀티모달 학습 과정에서 모델의 언어 능력 유지하는 방법으로 LLM 고정 및 입력 인코더만 학습하는 옵션과 전체 모델을 end-to-end로 학습하는 방법 등 두 가지 경로를 제시한다.

> **Conclusion** : pre-trained LLM의 임베딩 공간에 이미지와 같은 멀티모달 정보를 주입함으로써 embodied language model 구축을 제안한다. 실험 결과, 일반 VQA 및 captioning 작업에 대해 교육된 최첨단 시각-언어 모델들이 embodied reasoning tasks 에서 충분하지 않음을 보여주었다.



Diddu

이전 포스트
[논문 리뷰] iCaRL: Incremental Classifier and Representation Learning

0개의 댓글

댓글을 작성하세요

댓글 작성