COSE474-2023F: Final Project Proposal
Lightweight Vision-Language Model

2020330003 최창호

1. Introduction

컴퓨터 비전의 한 분야인 Vision-Language Model은 vision과 language 분야를 아우르는 분야로 최근 빠르게 발전하고 있는 vision의 한 분야이다.vision-language model을 통해 사용자는 image들에 대해서 보다 손쉽게 컴퓨터와 상호작용을 할 수 있으며, 사용자의 요구를 보다 정확하게 표현할 수 있다. 그러나 vision과 language를 모두 다뤄야하는 만큼 학습에 필요한 데이터들이 증가하고, 새로운 방식들이 생겨나면서 학습 및 실행에 필요한 Computing Power가 굉장히 빠른 속도로 증가하고 있다. 하지만 일부 사용자는 성능이 떨어지더라도 가벼운 모델이 필요한 경우가 있으며 본 연구는 이를 위한 모델을 제시함으로써 더욱 많은 사용자들이 손쉽게 vision-language model을 사용할 수 있게 하는 것을 목표로 한다.

2. Problem definition & challenges

일반 사용자들은 자원의 한계 때문에 가벼운 모델을 사용해야 하는 경우가 있다. 그러나 최근의 vision-language model들은 굉장히 무겁기 때문에 직접 학습하거나 학습된 모델을 가져오는 것 자체가 부담이 되는 경우가 많다. 이러한 사용자들을 위하여 가벼운 model의 개발이 필요하다고 느껴 본 연구를 진행하게 되었다.

model의 성능 측정은 vision-language의 세부분야인 Visual Question Answering(VQA)에서 얼마나 좋은 성능을 보일 수 있는지로 삼는다.

3. Related Works

vision-language model 중 경량화를 시도한 SimVLM model(Wang et al., 2021)은 large-scale weak supervision을 통해 training complexity를 줄임으로써 pretrainig에 필요한 cost를 줄였다. SOTA를 달성하지는 못했지만 VQA, image captioning task 등 다양한 downstream task들에서도 훌륭한 성능을 보이고 있다.

4. Datasets

본 연구에서는 Visual Question Answering 분야에서 많이 사용하는 데이터셋인 COCO 데이터셋을 사용한다. COCO 데이터셋은 80개 이상의 다양한 객체 클래스를 포함하고 있으며 이미지에 대한 주석과 설명이 포함되어 있기 때문에 vision-language multimodal 연구를 진행하기에 적합하다.

5. State-of-the-art methods and baselines

현재 Visual Question Answering 분야의 VQA v2 test-dev에서 SOTA를 달성하고 있는 model은 PaLI(Chen et al., 2023)로 Accuracy는 84.3이다. PaLI에 이어 BEiT-3 model(Wang et al., 2022)이 84.19로 2등을 달성하였다.

본 연구에서 진행하려는 simple visual language model 중 baseline으로 삼은 SimVLM의 경우 80.03의 accuracy를 가지고 있으며, 이 SimVLM을 바탕으로 연구를 진행할 예정이다.

6. Schedule & Role

~2023-11-05: SimVLM 아이디어 및 코드 확인
~2023-11-19: Model proposal 및 학습 코드 작성
~2023-11-26: Hyperparameter 튜닝 및 모델 학습
~2023-12-03: 성능 측정 및 모델 수정
2023-12-04~: 보고서 및 발표자료 작성

References

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B. K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., and Soricut, R. Pali: A jointly-scaled multilingual language-image model, 2023.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining

with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.