**PlantTraits2024 - FGVC11**

한성환1 우석민2

1서강대학교 컴퓨터공학전공 2서강대학교 심리학전공

**Abstract**

본 대회에서는 식물의 이미지와 관련된 여러 지표가 담긴 테이블 데이터가 주어진다. 우리의 목표는 두 정보를 활용하여 생태계와 관련된 6개의 지표를 예측하는 것이다. 따라서 이미지에 대한 정보 뿐만 아니라 테이블 데이터를 동시에 활용하는 것이 요구되기에 Multi-modal model을 학습할 필요가 있었다. 우선 Table data를 분석하여 각 label과 correlation이 높은 feature들만 text prompt로 활용하였다. 처리가 끝난 데이터를 효과적으로 활용하기 위하여 image-instruction pair로 구성된 데이터셋에서 학습된 Multimodal LMM을 활용하여 주어지는 식물의 이미지를 image prompt로, 관련된 Table data를 text prompt로 사용하였다. 또한 이번 task에 대한 성능을 끌어 올리기 위하여 1) image를 llm token으로 변환해주는 projector를 훈련시킨 뒤 (Pretrain) 2) 모든 module들을 훈련시키는 Finetune 과정을 진행하였다. 이를 통해 기존보다 성능이 크게 개선되었으며 Generative model인 LMM을 Multimodal data 활용이 가능한 Regression model로 사용이 가능함을 보였다.

1. **Introduction**

PlantTraints2024 – FGVC11 [9] 에서는 주어지는 식물의 사진과 식물이 위치한 지역의 기후, 토양정보가 주어졌을 때 식물의 특성을 예측하는 것을 목표로한다. 이는 생물의 다양성을 측정하는데 중요한 자료이며 특히 복잡한 Table data를 정제, 활용하는 것 뿐만 아니라 식물의 image에서 얻을 수 있는 Visual information을 활용하는 것이 필요하다. 따라서 Multi modality로 구성된 데이터들을 활용하는 것이 필요한데 지금까지의 모델들은 image description에 좀 더 초점이 맞추어져 있고 텍스트가 가진 풍부한 정보들을 완전히 활용하기에는 어려운 점이 많았다. 이와 달리, Large Language Model의 경우 instruction tuning을 통하여 사용자가 원하는 임의의 질문에 대해서도 잘 답변할 수 있다는 것이 알려지게 되었고 [3] , 이후 image또한 LLM의 input으로 사용할 경우 Visual Questiong Answering과 같은 task 수행이 가능함을 보였다. [1] 우리는 이런 점에 주목하여 Multi modal LLM 모델을 활용하여 이번 대회에서 제공하는 table data와 plant image를 활용하기로 하였으며 이 과정에서 Pretrain / Finetuning을 통하여 성능을 끌어올릴 수 있음을 보였다. 또한 Train instability issue를 해결하려는 과정에서 현 모델의 한계 또한 발견할 수 있었다.

1. **Related Works**

**Multimodal Large Language Model**

최근LLM의 등장에 맞추어 이미지 또한 입력으로 사용 가능한 다양한 Multi-modal LLM들이 소개되었다. [1,2,6,7] 이 모델들은 공통적으로 기존에 Text만을 입력으로 받던 LLM에 이미지를 입력으로 받을 수 있도록 별도의 Projection module을 추가한 뒤, Image-text pair dataset 또는 Visual instruction dataset을 기반하여 학습되어 이미지에 대한 설명이나 좀 더 심도 있는 추론이 가능하게 되었다. 그 중 특히 LLaVA [1] 의 경우 Vicuna LLM에 image를 입력으로 사용하기 위하여 Linear layer를 활용하였고 마치 이를 이미지에 대한 토크나이저 처럼 활용하였다. 구조 자체는 단순하지만, 기존 image-text pair에서 훈련된 LMM들과 달리, GPT-4 [8]를 활용하여 대량의 Visual instruction dataset을 구축하였고 이는 마치 LLM이 사용자의 다양한 query에 답변할 수 있도록 instruction dataset에서 학습되었던 것처럼 [3] LLaVA 또한 이미지에 대한 사용자의 다양한 query에 답변하는 것을 넘어 모델이 이미지를 이해할 수 있도록 하였다.

1. **Method**

**Data Pre-Processing**

텍스트, 평행, 직사각형, 라인이(가) 표시된 사진

자동 생성된 설명

**그림 1. Label(X11)에 대한 boxplot**

Label 값은 density, mass, area 등 물리적인 값이다. Outlier을 제거하기 위해 label값이 이를 참고하여, 0 이하의 데이터와 99 percentile 이상의 데이터를 제거하였다. 각 feature 들에 대해서도 이상치가 있는 것으로 확인되었지만, 모든 feature들에서 이상치 제거를 하면 데이터의 50%가 제거되어서 label만 이상치 제거를 진행하였다. 데이터의 source가 여러 곳이고, 대회 주최자가 이런 데이터 source들을 단순히 pairing을 했기 때문에 이런 문제가 생긴 것으로 추정된다.

LLaVA 모델에 모든 데이터를 넣을 수 없었기에 미리 각 label과 feature들의 correlation을 구해 정렬하고, 상위 n%의 feature들을 구하여 데이터 셋을 구성하였다. Label마다 상위 10%, 20%, 30%, 40%, 50%, 75%를 구성하였고, 이를 이용해서 모델 훈련을 진행하였다.

Min-Max normalization, Z-score normalization 등을 진행해 보았는데, 최종적으로 가장 좋은 성능을 보이는 것이 Z-score normalization이였기 때문에 이를 선택하였다.

**Pretrain / Finetuning MLLM**

[1] 에서는 instruction dataset을 기반으로 LLM은 동결시켜 둔 채 Projector를 우선 학습하는 Pretriain 이후, LLM까지 학습하는 Finetune으로 두 단계에 걸쳐 학습을 진행하였다. 우리도 이에 맞추어 이전 섹션에서 구축한 train data를 두 단계에 걸쳐서 훈련을 진행하였다. Learning rate와 같은 상세 세팅의 경우 [1] 의 세팅을 그대로 활용하였으며 train time이 너무 길기에 1epoch만 우선 학습하였다. 학습의 경우 memory budget이 허용하는 한도내에서 배치 사이즈를 gradient accumulation까지 활용해가며 최대한으로 늘려 활용하였다. 또한 Generative model을 활용하여 Regression task를 접근하는 상황이므로 출력 text 길이가 너무 길 경우 잘못된 문장을 출력하거나 정확도가 떨어질 가능성이 있으므로 LLM 출력의 길이를 제한해주었다. Inference에 대해서도 동일하게 적용하였으며 이를 통해 LLM이 좀 더 훈련이 쉽고 테스트에서 좀 더 높은 정확도를 기대할 수 있었다.

초기에는 하나의 모델이 6개 라벨 모두를 훈련할 수 있도록 시도하였지만, 이후 실험해본 결과 하나의 모델이 하나의 라벨을 예측할 때 좀 더 높은 성능을 획득할 수 있음을 확인하였다. 따라서 최종 제출을 할 때에는 6개의 모델을 각각 훈련시킨 뒤, 테스트 데이터 또한 각각 제작하여 prediction을 하는 식으로 진행하였으며 이 방식이 가장 높은 score를 획득할 수 있었다. 이후 다양한 settings (epoch, dataset 구성)을 시도해봤으나 상위 20% 의 label과의 correlation을 가지는 feature들을 활용할 때 가장 성능이 좋게 나오는 것을 확인하였다.

1. **Result**

**Kaggle**

A screenshot of a phone

Description automatically generated

**그림 2. 최종 Leader board**

최종적으로는 **0.18457**의 score와 전체 398명의 참가자 중 **234위**의 순위를 기록하였다.

1. **Discussion**

**Train instability issue with MLLM**

Train 도중 중간에 Loss 함수를 추적하던 중 공통적으로 갑자기 훈련이 불안정해진 뒤 수렴하지 못하는 문제가 발생하는 것을 확인하였다. Pretrain을 진행하는 과정에서는 발생하지 않았으나, Finetune을 진행하는 과정에서 종종 발생한다는 것을 다수의 실험을 통해 확인하였다.

A screenshot of a graph

Description automatically generated

**그림3. Step-Loss 함수 그래프. 중간에 훈련이 불안정해지는 것을 확인할 수 있다.**

훈련이 불안정해짐에 따라 model의 variation이 커지는 것을 확인하였고 이는 충분히 문제라고 판단, 해결을 위한 방안들을 추가적으로 검토해보았다.

MLLM이 제시된 지 얼마 되지 않아 ICLR 2024에서도 관련된 논문들을 찾아보기 어려웠지만 최근 CVPR 2024 accept paper 위주로 검토해본 결과 몇가지 시도들이 있었다. [4] 에서는 Visual Projector가 이미지를 LLM이 사용가능한 token형태로 변환하여 줄 때 Question을 inject하는 방식을 제안하였으며, 이 과정에서 Gated mechanism [10]을 활용하는 것이 훈련 안정화에 효과적이었다고 언급하고 있다. 또한 [5] 에서는 기존의 LLaVA의 성능을 개선하는데 있어 CLIP의 hierarchy 구조를 활용하여 성공을 거뒀지만 훈련이 불안정하여 Layer normalization을 적용하였고, 이를 통해 안정화 할 수 있었다고 언급하고 있다. 우리는 이를 바탕으로 성능 개선과 동시에 훈련 안정화를 위하여 아래와 같은 구조로 LLaVA의 프로젝터를 변경하였다.

A diagram of a computer program

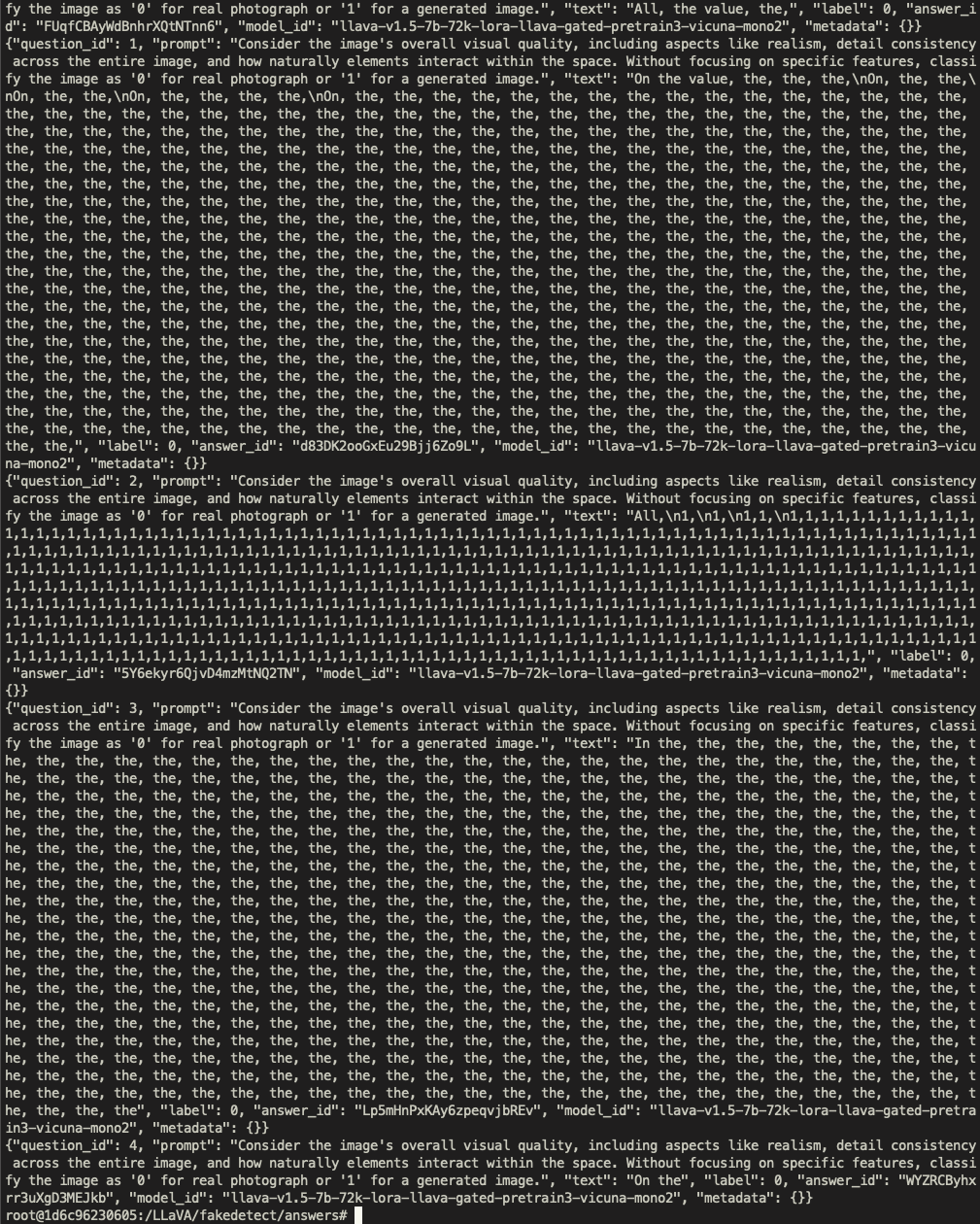
Description automatically generated with medium confidence

**그림 4. 제시된 Visual Projector**

Finetuning 도중 발생하는 급격한 Projector weight의 변화가 train instability에 영향을 준다고 판단, 2개의 MLP를 구성한 뒤 하나의 MLP는 Pretrain 이후 동결하였고 나머지 하나의 MLP는 Finetune 단계에서 훈련을 진행하였다. 이 때 별도의 훈련 가능한 파라미터를 추가하여 학습되는 MLP가 얼마나 반영되는 정도를 조절할 수 있도록 구성하였다. 또한 Layer normalization이 추가되어 추가적인 안정화가 가능하도록 진행하였다.

**Limitation of MLLM as generative model**

하지만 여전히 몇가지 문제가 여전히 발생하였는데 우선 첫째로, inference bug가 발생하였다. Visual projector를 변경함에 따라 Pretrain weight를 활용하지 못하고 처음부터 학습을 진행해야했는데, 훈련은 안정화 됬으나 inference 단계에서 bug가 발생하는 것을 확인하였다.



**그림 5. 발생한 inference bug**

출력 한도를 잠시 풀어서 확인해본 결과 출력 한도까지 계속하여 반복해서 출력하는 것을 확인하였고 Generative model을 훈련하는 과정에서 Instruction tuning이 필요하고 데이터가 많이 필요하지만 그러지 못하였다는 점이 오히려 Generative model의 불안정성으로 인하여 발생한 것으로 보인다. 이는 추후 데이터셋을 좀 더 구성한 뒤, 시도해 봄으로써 검증할 필요가 있어 보인다.

1. **Conclusion**

이번 Kaggle competition에 참가함으로써 Multi modal data를 처리를 하는 과정에서 Multimodal LLM을 활용하였으며, 이를 Regression model로써 활용할 수 있는 가능성을 확인하였다. 또한 발생한 Train instability issue를 해결하는 과정에서 Gated mechanism, Layer Normalization과 같은 방법론들을 적용하였고 이 과정에서 아직은 Generative model로써 가지는 한계가 있어 완전한 활용이 어려움을 확인하였다. 추후 연구로써 제시된 모델에 충분한 분량의 instruction dataset을 기반으로 성능 검증을 진행할 예정이다.

1. **Reference**

[1] Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2024).

[2] Dai, Wenliang, et al. "Instructblip: Towards general-purpose vision-language models with instruction tuning." *Advances in Neural Information Processing Systems* 36 (2024).

[3] Wei, Jason, et al. "Finetuned Language Models are Zero-Shot Learners." International Conference on Learning Representations. 2021.

[4] Ganz, Roy, et al. "Question Aware Vision Transformer for Multimodal Reasoning." arXiv preprint arXiv:2402.05472 (2024).

[5] Cai, Mu, et al. "Making large multimodal models understand arbitrary visual prompts." arXiv preprint arXiv:2312.00784 (2023).

[6] Zhu, Deyao, et al. "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models." The Twelfth International Conference on Learning Representations. 2023.

[7] Awadalla, Anas, et al. "Openflamingo: An open-source framework for training large autoregressive vision-language models." arXiv preprint arXiv:2308.01390 (2023).

[8] Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).

[9] https://www.kaggle.com/competitions/planttraits2024

[10] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.