언어 모델을 이용한 시각 묘사 생성과 CLIP 기반 제로샷 이미지 분류의 상관관계 연구

강준용 ○ 전찬 ○ 강유석 ○ 정영민
서강대학교 컴퓨터공학과
{jykang, jeoncharn, kingyuseok, YMCHUNG}@sogang.ac.kr

A study on relationship between CLIP-based zero-shot image classification and language-model aided visual description generation

Junyong Kang, Chan Jeon, Yuseok Kang, Yeongmin Jeong
Department of Computer Science and Engineering, Sogang University

요 약

CLIP 기반의 제로샷 이미지 분류 작업은 분류할 클래스에 대한 시각 묘사와 이미지 특징 벡터의 유사도를 계산하여, 가장 유사도가 높은 시각 묘사의 클래스로 분류한다. 본 연구에서는 언어 모델을 통해 각 클래스에 대한 시각 묘사를 추출하는 과정에 주목하여 시각 묘사 문장의 형태와 제로샷 이미지 분류의 결과의 상관관계를 관찰하고자 하였다. 이를 위해 시각 묘사를 생성하는 언어 모델에 주는 예시 prompt를 묘사 형태에 따라 여러 옵션으로 구분하였다. 실험 결과, CLIP 기반 제로샷 이미지 분류는 1) 문장보다 단어 형태의 시각 묘사가 선호되며, 2) 시각 묘사의 개수에 따라 성능과 확신도가 증가하며, 3) 다른 옵션은 시각 묘사 품질 향상에 유의미한 향상을 주지 않음을 관찰하였다.

1. 서 론

최근 딥러닝 기반 이미지 분류 모델의 비약적인 성능 향상에 비해 제로샷 이미지 분류 문제는 최근까지 컴퓨터 비전 분야에서 풀기 어려운 문제로 인식되었다. 여기서 제로샷 이미지 분류 문제란 모델이 훈련 과정에서는 주어지지 않았던 클래스를 추론 단계에서 분류하고자 하는 문제를 지칭한다. 제로샷 문제를 접근하는 연구 중엔 이미지 분류 모델이 훈련중에 보지 못한 몇 개의 클래스를 추가로 예측하는 경우도 있으나[4], 만약 추론 단계에서 예측하는 클래스에 대한 아무 제약이 없다면 제로샷 문제에서 전통적인 분류 모델을 활용하기란 매우 어렵다.

CLIP[1]은 최근 이러한 문제를 해결할 수 있는 방법으로 각광받고 있는 딥러닝 기반의 이미지-언어 모델이다. 이 모델은 대조 학습 (Contrastive Learning)을 통해 이미지와 자연어 (natural language)를 같은 공간의 특징 벡터로 부호화 (encoding) 하는 특성을 지니고 있는데, 이를 제로샷 이미지 분류에 이용할 수 있다. 구체적으로, 임의의 클래스들에 대한 자연어 문장(e.g. “A photo of {class}”)과 이미지의 특징 벡터 사이의 유사도를 계산한 뒤, 가장 유사도가 높은 클래스로 분류한다. 그러나 CLIP을 포함한 기존 연구에서는 이 방식을 이용할 뿐, 입력 데이터의 특성에 따른 성능에 대한 변화가 분석되지 않았다. 특히, 모델 사용자 입장에서 정확한 제로 샷 분류를 위해 어떤 형태의 자연어 입력을 주어야 하는지에 대한 탐구가 부족하였다.

최근 Menon and Vondrick[2]은 각 클래스에 대한 자세한 자연어 시각 묘사를 여러 개 사용하면 제로샷 분류 성능이 향상될 수 있음이 밝혀냈다. 즉, 특정 클래스의 대상에 대한 일반적인 시각적인 특징을 여러 개 기술한 다음(e.g. ‘television’ 클래스의 경우 ‘television, which has a large, rectangular screen’, ‘television, which has input ports for connecting to other devices’ 등이 시각 묘사가 될 수 있다), 이 묘사들과 이미지와의 유사도를 종합하여 분류를 수행한 것이다. 구체적으로,

$$s(c,x) = \frac{1}{|D(c)|}\sum_{d\in D(c)} \phi(d,x)$$

와 같이 이미지 x가 각 클래스 c에 속할 점수 s(c,x)를 계산한다. 이때, D(c)는 각 클래스에 대한 시각 묘사 자연어 문장을 의미하며, ϕ는 시각 묘사 d와 x에 대해 CLIP 부호기 (encoder) 를 이용해 계산한 유사도를 의미한다.

[2]는 각 클래스 별 시각 묘사를 생성하기 위해, GPT-3 모델[3]을 이용하였다. GPT-3은 생성형 거대 언어 모델의 일종으로-사용자 입력에 따라 적절한 언어 출력을 생성한다. 따라서 각 클래스를 구분하기 위한 시각적 특징을 생성하도록 모델에게 입력 prompt를 주고, 이에 대한 결과

값을 사용하였다. 이때 사용한 입력 prompt는 다음과 같다.

Q: What are useful visual features for distinguishing a {category_name} in a photo?

A: There are several useful visual features to tell there is a {category_name} in a photo:

그런데 단순히 이 질문만 입력으로 주는 경우, 언어 모델이 어떠한 형태로 시각 묘사를 생성해야 하는지 혼란을 겪을 수 있다. 따라서 위의 질의응답 형식을 따르는 예시 prompt를 입력 앞에 삽입하면 언어 모델이 예시와 비슷한 형식으로 시각 묘사를 생성해주게 된다.

본 연구는 이와 같은 자연어 시각 묘사 기반의 제로샷 분류에서 시작하여, 시각 묘사 문장의 형태와 성능에 대한 상관관계를 분석하였다. 구체적으로, 언어 모델 (GPT-3)이 시각 묘사를 생성할 때 주는 입력 prompt의 구성을 체계적으로 분류 및 분석하여 다양한 형태의 시각 묘사를 생성한 다음, 이에 이미지 분류를 수행하였다. 이를 통해 본 시각 묘사의 개수, detailedness, additional instruction 과 분류 성능의 상관관계를 분석할 수 있었다.

2. 입력 prompt의 분류

[2]는 서론에서 언급한 방식으로 언어 모델에게 입력 prompt를 전달하며, 임의의 클래스에 대한 두가지 예시를 사용하였다. 본 연구에서는 예시로 들어가는 prompt의 구성 방식을 체계화하여, 표1과 같은 prompt 옵션을 만들었다.

표 1: Prompt 옵션의 분류

Option name	Value type
$\|D(c)\|$	Positive Integer
Detailedness	Boolean
Additional Instruction	(Additional) String
The number of objects	Positive Integer

먼저 $|D(c)|$ 는 예시 Prompt에 적힌 각 Object 당 시각 묘사의 개수를 의미한다. 이때 언어 모델이 생성하는 시각 묘사의 개수는 예시의 시각 묘사 개수와 같지 않을 수 있으므로 생성된 시각 묘사 개수의 전반적인 경향성을 통제하는 변수가 된다. Detailedness는 시각 묘사의 상세성에 관여하는데, True인 경우 해당 대상에 대한 자세한 문장 형태의 묘사이며, False인 경우 각 묘사가 짧은 단어 위주로 나타난다. Additional Instruction은 Prompt 예시와 별개로 입력 Prompt에 붙는 명령을 의미하며, 시각 묘사를 만들 때의 규칙 등이 될 수 있다. (예를 들어, 시각과 관련된 묘사만을 적거나, 구별적인 특징 위주로 적으라는 명령 등이 있다.) 마지막으로 The number of objects는 prompt 예시의 총 개수를 의미한다.

3. 실험 환경

앞 장에서 구성한 prompt 옵션별로 입력 prompt에 들어갈 예시를 구성하였다. 이 구성에 대한 실제 prompt 예시는 부록에 첨부하였다. 예시 prompt의 구성에 따른 공정성을 최대한 확보하기 위해, 모든 예시는 GPT-3을 이용하여 생성하였다. 이후, 클래스 별 시각 묘사를 생성할 질문 prompt에 각 옵션별로 예시를 함께 추가한 입력을 GPT-3에 넣고, 만들어진 시각 묘사를 각각 저장하였다. 마지막으로 각 옵션별로 생성된 시각 묘사 문장들을 이용해 서론에 제시된 수식의 방법으로 사전 훈련된 CLIP 모델을 이용하여 제로샷 이미지 분류 성능을 측정하였다.

실험 모델로는 사전 훈련된 CLIP의 이미지/텍스트 부호기를 사용했으며, 평가 데이터는 ImageNet[5]를 이용하였다. ImageNet은 서로 다른 1000개의 클래스가 있는 데이터 셋으로, 한 실험 당 각 클래스 별 시각 묘사를 생성하도록 언어 모델(GPT-3)을 1000번 사용하였다.

4. 실험 결과

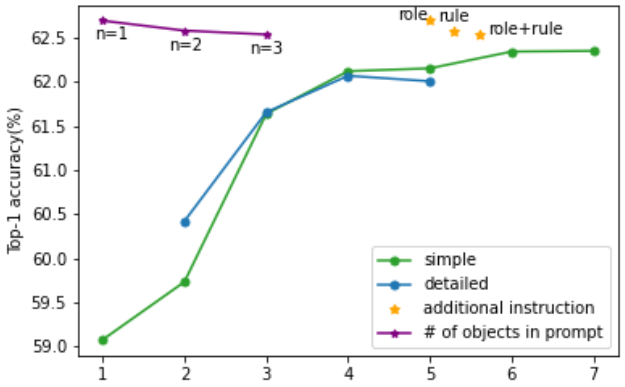

그림 1: Prompt 옵션에 따른 제로샷 이미지 분류 성능

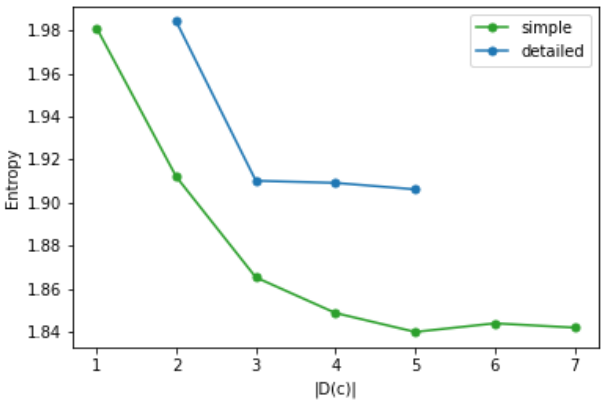

그림 2: 시각 묘사 개수에 따른 예측 분포의 엔트로피의 변화

그림 1은 각 prompt 옵션별 성능을 보여준다. 먼저 초록색과 파란색 곡선은 각각 prompt가 단어 위주인 경우와 상세한 경우를 나타내며, 초록색 곡선의 7은 [2]의 설정을 재현한 결과이다. 평균적으로, 단어 위주의 prompt가 상세한 경우에 비해 약 0.43% 더 높은 성능을 보여준다. 이는 CLIP 언어 부호기가 단어 위주의 데이터셋에서 대조 학습으로 훈련되었기 때문에 여러 정보가 혼합된 문장보다 이미지와 일치하는 속성에 해당하는 단어를 더 선호한다는 것을 의미한다. (상세한 경우 실험에서 생성된 시각 묘사의 개수가 대체로 입력

예시만큼 나오지 않은 경우는 생략하였다.)

두번째로, 두 곡선에서 전반적으로 시각 묘사의 개수가 많을수록 성능이 향상됨을 관찰할 수 있었다. 예시로, $|D(c)| = 1$ 인 경우 성능은 Deatiledness = False에서 59.07%, $|D(c)| = 7$ 인 경우 성능이 62.35%를 기록함으로 3.28%p 향상되었다.

또한, 그림 2는 각 클래스별 점수 $s(c,x)$를 scaling (−10 ~ 10) 하고 softmax 함수를 거쳐 확률분포로 정규화 한 분포의 entropy $\sum_c -p(C=c|x)\cdot \ln p(C=c|x)$ 를 나타낸다. Entropy는 낮을수록 모델의 확신도(confidence)가 강하다는 것을 의미하는데, 그림 2를 통해 시각 묘사의 개수가 많을수록 model의 결정이 더욱 높은 확신도를 가지게 됨을 확인할 수 있었다. 다만, 성능과 확신도의 향상은 시각 묘사의 개수가 4개 이상일 때부터는 크게 증가하지 않는 것을 관찰하였다.

한편 Additional Instruction과 the number of objects 옵션의 경우 $|D(c)| = 7$을 기반으로 각각 Rule과 Role을 더하거나 예시의 개수를 조절하여 실험하였는데, 유의미한 성능 향상의 경향이 나타나지는 않았다.

5. 결론 및 제언

본 연구에서는 언어 모델(GPT-3)을 통해 클래스별 시각 묘사를 생성한 뒤 CLIP으로 이미지와 시각 묘사들의 유사도를 종합하여 제로샷 이미지 분류를 수행하는 작업에서, 언어 모델이 시각 묘사를 생성할 때 들어가는 입력 prompt 옵션을 구성하여 시각 묘사 문장의 형태와 이미지 분류의 상관관계를 고찰하였다. 그 결과, 1) CLIP은 단어 형태의 시각 묘사를 문장 형태보다 더욱 선호하며, 2) 시각 묘사의 개수가 많다면 분류 성능과 확신도가 향상될 수 있으며, 3) 부가적인 명령이나 prompt의 예시 개수는 시각 묘사의 품질에는 큰 영향을 주지 않는다는 상관관계를 관찰했다. 이와 같은 관찰은 사용자 입장에서 어떠한 자연어 prompt를 입력해야 할지 가이드라인을 제시하는 의미가 있다. 마지막으로 향후 연구에서는 CLIP 모델의 구조와 평가 데이터셋을 달리하여 본 연구의 결론이 더욱 검증될 필요가 있다고 하겠다.

참고 문헌

[1] Alec Radford 외: Learning Transferable Visual Models From Natural Language Supervision, Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8748-8763, 2021.

[2] Sachit Menon, Carl Vondrick: Visual Classification via Description from Large Language Models. The Eleventh International Conference on Learning Representations, 2023.

[3] Tom Brown 외: Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

[4] Guanyu Yang, Zihan Ye, Rui Zhang, Kaizhu Huang: A comprehensive survey of zero-shot image classification: methods, implementation, and fair evaluation, Applied Computing and Intelligence 2022, Volume 2, Issue 1: 1-31.

[5] Olga Russakovsky, Jia Deng 외: ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

부록.

아래는 실험에 사용한 옵션별 prompt의 일부이다.

Option Name	Example prompt	Generated visual description (class : *Tench*)
|*D*(*c*)|=4 Detailedness= False	*Q: What are useful visual features for distinguishing a lemur in a photo?* *A: There are several useful visual features to tell there is a lemur in a photo:* *- four-limbed primate* *- wet and hairless nose with curved nostrils ₩n - long tail ₩n - large eyes* *Q: What are useful visual features for distinguishing a television in a photo?* *A: There are several useful visual features to tell there is a television in a photo:* *- electronic device* *- a large, rectangular screen* *- a stand or mount to support the screen* *- input ports for connecting to other devices*	*- large, rounded scales* *- long, slender body* *- two barbels on the chin* *- two dorsal fins*
|*D*(*c*)|=3 Detailedness= True	*Q: What are useful visual features for distinguishing a lemur in a photo?* *A: There are several useful visual features to tell there is a lemur in a photo:* *- Lemurs often have distinct facial markings. Many species exhibit a contrasting white face mask against a darker surrounding fur, making their facial features prominent.* *- Most lemurs sport long, bushy tails. Some species, like the ring-tailed lemur, have alternating black and white bands, which is a signature feature.* *- Unique to lemurs, they often cling to trees in a vertical posture, using their long limbs and digits to grip onto the bark.* *…(similar example for "television")*	*- Tench have a large, rounded head with a protruding lower jaw and a small, upturned mouth.* *- Tench have a large, flat, and slightly rounded tail fin. They also have two dorsal fins, the first being short and the second being longer and more pointed.*
Additional Instruction: Role	*You are a visual description generator. For a given category or object, create visual descriptions that distinctly explain that category or object.* *Q: What are useful visual features for distinguishing a lemur in a photo?* *… (same as previous examples)*	*- large scales* *- two dorsal fins* *- a long, slender body* *- a large, flat head* *- a large, upturned mouth* *- a barbel on the chin* *- a forked tail*
Additional Instruction: Rule	*Create only static visual features.* *Avoid creating dynamic visual features, such as movements.* *Create features that are not shared in other object categories.* *Create features without considering their background features.* *If more than one object exists, create each unique visual features.* *Create a particularly unique feature of the design or shape of this object.* *… (same as previous examples)*	*- large scales* *- two dorsal fins* *- a long, slender body* *- a large, flat head* *- a large, upturned mouth* *- barbels on the chin and lower jaw* *- a short, forked tail*