# NICE
# 4th place result technical introduction

- CALISOLO (오솔길)

# 목차

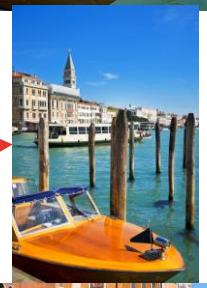
- 방법론 전반 소개
- 데이터셋 특성
- 입력 데이터 형태
- 학습경과 및 중간보고
- 회고&한계점 및 추후 연구주제

# 방법론 전반 소개

Gondoliers paddling tourists in gondola among architectural buildings in the sunny Grand Canal in Venice Italy

hint1

hint2



hint3

hint4

Hint structure

[cosine similarity][id similarity][caption]

Image

OFA Encoder

OFA Decoder

Gondoliers paddling tourists in gondola among architectural buildings in the sunny Grand Canal in Venice Italy

- Intuition: NICE task의 특징은 이미지의 특징을 잡아내는 task이기도 하지만, NICE dataset 특유의 말투에 일치하게 만들어 내는 부분이 challenging하게 여겨짐 (persona dialogue generation)
- OFA (apache 2.0 라이선스) 모델 기반, 입력 데이터에 few shot 힌트 제공하는 형태로 미세조정

# 데이터셋 특성

| | | |
|---|---|---|
| 1869046529 | Horizontal three quarter length shot of a woman having strawberries in breakfast smiles at the camera | food |
| 1590352094 | Marina and Cathedral of Palma de Mallorca at night Mallorca Spain | outdoors |
| 1586704871 | White Clouds Against Blue Summer Sky | outdoors |
| 1868716469 | A high angle vertical shot of a senior farmer watching potatoes filling into a trailer in a sunny rural field | outdoors |

Caption_gt를 살필때, 대문자가 불규칙적으로 등장하는 상황/ 사진의 스타일을 묘사하는 내용이 다수 등장

Horizontal three quarter length shot of a woman having strawberries in breakfast smiles at the camera
Marina and Cathedral of Palma de Mallorca at night Mallorca Spain
White Clouds Against Blue Summer Sky
A high angle vertical shot of a senior farmer watching potatoes filling into a trailer in a sunny rural field

shot style
location

사진의 스타일을 묘사한 캡션은 대부분 prefix에 사진의 형식을 설명
처음에 등장하지 않은 대문자는 주로 특정한 지명 (스페인, 브라질, 아마존 강) 등을 나타내는 상황으로 판단

5000건의 validation set에 위의 내용들을 판단하는 태깅 수행[1)]
(대문자 검색 등의 간단한 필터 후 수작업으로 재분류 6~8시간 소요)

| | | | |
|---|---|---|---|
| 1869046529 | Horizontal three quarter length shot of a woman having strawberries in breakfast smiles at the camera | food | Horizontal three quarter length shot of |
| 1590352094 | Marina and Cathedral of Palma de Mallorca at night Mallorca Spain | outdoors | Cathedral of Palma de Mallorca at night Mallorca Spain |
| 1586704871 | White Clouds Against Blue Summer Sky | outdoors | |
| 1868716469 | A high angle vertical shot of a senior farmer watching potatoes filling into a trailer in a sunny rural field | outdoors | A high angle vertical shot of |

1) 정확도 다소 낮으며 태깅 개선시 성능 향상 기대 - shutterstock 원천에서 생성하는 캡션으로 유추되어 기계적인 태깅 가능성 기대

Highland cattle on farm Seefeld Bavaria Germany

Location의 특징:
배경지식 등이 필요하며 사진만으로는 추론하기 어려움
(highland cattle의 서식지? 사진상 식생의 분포?)

1) 어떤 사진에 location이 들어가고 어떤 사진에는 들어가지 않을까?
2) location이 들어가는 사진을 안다고해도 그 장소를 어떻게 맞출까?

Vertical shot of a happy woman touching a flower and sitting in a meadow full of wildflowers

A mid adult woman holding a dried leaf

Shot style의 특징:
사진을 잘 설명하고 있으나, 이를 서술할지 말지 판단하기 어려움

1) 어떤 사진에 shot_style이 들어가고 어떤 사진에는 들어가지 않을까?

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 216581945 | View of snowy mountain range and blue sky | outdoors | View of | | | | | | |
| 216581957 | View of snowy mountain range and blue sky | outdoors | View of | | | | | | |
| 216582038 | View of snowy mountain range | outdoors | View of | | | | | | |
| 216582401 | View of town and bridge spanning river on sunny day Jarnac and the Charente river West Central France | outdoors | View of | Jarnac and the Charente river West Central France | | | | | |
| 216582452 | Hay being harvested into straw bales in farm field | outdoors | | | | | | | |
| 216582482 | Close up of vibrant sunflower | outdoors | Close up of | | | | | | |
| 216582731 | Countryside Bourdeilles Dordogne France | outdoors | | Bourdeilles Dordogne France | | | | | |
| 216582812 | Rooftops and river in idyllic village Bourdeilles Dordogne France | outdoors | | Bourdeilles Dordogne France | | | | | |
| 216582896 | Close up of red blooming flowers St Jean de Cole Dordogne France | outdoors | Close up | St Jean de Cole Dordogne France | | | | | |

Public id순으로 정렬시, shot_style/ location의 경향성 발견

가설

1. 같은 공급자가 제공한 사진은 이미지에 내재된 정보를 통해 추론가능하며 주제/사진/캡션방식이 유사할 것이다.
2. Public id가 shutterstock의 업로드 번호이며 연속적으로 업로드한 사진의 공급자가 같을 가능성이 높다.

접근방안

- 사진간 유사도, Validation_set 에 제공되는 public id를 활용하여 학습

# 입력 데이터 형태

Input image

Cosine similarity Map

Ordered by cosine similarity

Caption 1

Caption 2

Caption 3

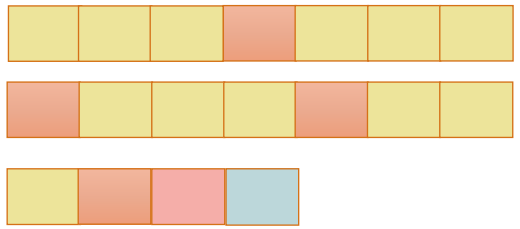
Caption 4

| Cosine Hint | diff Hint | category | Caption 1 |
|---|---|---|---|
| Cosine Hint | diff Hint | category | Caption 2 |
| Cosine Hint | diff Hint | category | Caption 3 |
| Cosine Hint | diff Hint | category | Caption 4 |

shotStyle total

Location total

Concatenate 4-shots captions with cosine similarity/ public_id difference hints



OFA Encoder

Hint Set | Cosine Hint | diff Hint | category | Caption 1

- [Cosine Hint] : 이미지 간의 코사인 유사도[2]를 기반으로 현재 쿼리 이미지와 Hint Set의 이미지 representation이 얼마나 차이나는지를 단어 임베딩으로 모델링 하고자 하였음 . 0.355 와 같은 수치를 입력에 넣을 경우 모델이 구체적인 대소관계를 판단하지 못할거라고 예상하여, 이를 표현하기 위한 4개 토큰을 단어사전에 할당한 후, 4개 그룹으로 정규화하여 수행[3]
  1. [cosHint lv4] : 거의 같은 사진을 의미하는 강력한 힌트 > 0.4
  2. [cosHint lv3] : 같은 주제지만 캡션이 다를 것이라 예상됨 > 0.32
  3. [cosHint lv2] : 유사한 사진이지만 캡션이 다름 > 0.29
  4. [cosHint lv1] : 무관한 사진 ≤ 0.29

- [Diff Hint]: cosine hint와 마찬가지로, 쿼리 이미지와 hint set의 public_id가 얼마나 차이나는지 정보를 모델에 전달하기 위한 구성. 3개 토큰을 단어사전에 할당후 3개그룹으로 정규화 하여 수행
  1. [diffHint lv3]: 사진의 public_id 차이가 아주 작음 <100
  2. [diffHint lv2]: 사진의 public_id 차이가 작음 <10000
  3. [diffHint lv1]: 사진의 public_id 차이가 큼 나머지

2) OFA 인코더의 output을 기준으로 비교하였음. SAM(segment anything model)을 활용하고자 하였으나, colab의 고용량 RAM 세팅(83gb)으로는 약 30000장 사진에 대한 SAM의 representation 저장 용량 이슈 발생. 자원이 충분하여 더욱 좋은 이미지 인코더를 사용한다면 레벨구간을 더 정밀하게 나눌 것으로 예상됨

3)코사인 유사도만으로 위의 기준만큼 명확하게 유사도를 비교하는 것은 무리가 있다고 판단됨 . 어느정도 경향성으로 파악하고자 하였음

Hint Set | Cosine Hint | diff Hint | category | Caption 1

Summary hint | shotStyle total | Location total

각각의 유사한 사진에서 구성한 hint set은 코사인 유사도를 기반으로 유사 사진을 분석한 정보
마찬가지로, public_id를 기반으로 해당 사진 근방의 caption writing style을 학습하고자 Summary hint 구성

- shotStyle total:  public id 기준으로 가장 가까운 6개 사진에서 나온 shotStyle 전부 기재
- Location total: public id 기준으로 가장 가까운 6개 장소에서 나온 location과 각각의 diffHint 기재

| cos2 | cos3 | cos4 | diff1 | diff2 | diff3 | diff4 | shots | locations | cat1 | cat2 | cat3 | cat4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [cosHint lv1] | [cosHint lv1] | [cosHint lv1] | [diffHint lv1] | [diffHint lv1] | [diffHint lv1] | [diffHint lv1] | ['Close up', 'Close up of'] | [] | [outdoors] | [misc] | [outdoors] | [office] |
| [cosHint lv3] | [cosHint lv3] | [cosHint lv3] | [diffHint lv2] | [diffHint lv2] | [diffHint lv3] | [diffHint lv1] | ['Close up shot of', 'Low angle vertical shot ... | [] | [outdoors] | [outdoors] | [outdoors] | [outdoors] |
| [cosHint lv3] | [cosHint lv3] | [cosHint lv2] | [diffHint lv1] | [diffHint lv1] | [diffHint lv1] | [diffHint lv1] | ['Low angle view of'] | [] | [outdoors] | [outdoors] | [outdoors] | [outdoors] |
| [cosHint lv2] | [cosHint lv1] | [cosHint lv1] | [diffHint lv1] | [diffHint lv1] | [diffHint lv1] | [diffHint lv1] | ['Blurred view of', 'View of', 'View of', 'Vie... | [[diffHint lv2] [outdoors]Cape Town South Africa] | [outdoors] | [outdoors] | [outdoors] | [outdoors] |
| [cosHint lv2] | [cosHint lv2] | [cosHint lv2] | [diffHint lv2] | [diffHint lv2] | [diffHint lv1] | [diffHint lv2] | ['Close up of'] | [[diffHint lv2] [outdoors]Alcazar Seville Spain... | [outdoors] | [outdoors] | [outdoors] | [outdoors] |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| [cosHint lv4] | [cosHint lv3] | [cosHint lv3] | [diffHint lv2] | [diffHint lv1] | [diffHint lv1] | [diffHint lv3] | ['Aerial view of'] | [[diffHint lv3] [outdoors]Cuxiu Muni Amazon Riv... | [outdoors] | [outdoors] | [outdoors] | [outdoors] |

좌측의 표처럼 data 저장형태 구성

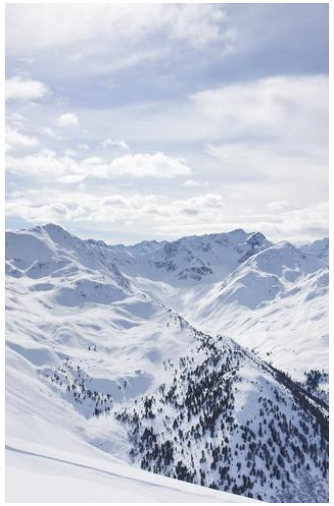
How it really looks like

Hint set

**[cosHint lv4][diffHint lv3][outdoors]**View of snowy mountain range and blue sky**[cosHint lv4][diffHint lv2][outdoors]**View of snowy mountain range
**[cosHint lv4][diffHint lv3][outdoors]**View of snowy mountain range and blue sky**[cosHint lv4][diffHint lv1][outdoors]**Snow capped mountains in Fimbatal the border between Switzerland and Austria Eschol
**[shot_style]**View of | View of | View of | **[Location][diffHint lv2][outdoors]**Jarnac and the Charente river West Central France |

Shot style described in 6 adjacent captions

locations described in 6 adjacent captions



좌측의 이미지와, 위의 텍스트를 기준으로 OFA 모델은 정답을 추론

코사인 유사도를 기준으로 유사 사진들은 'View of snowy mountain range'라는 캡션을 주로 달았기 때문에, 해당 케이스는 마찬가지로 'View of snowy mountain range'를 생성할 가능성이 높음

하지만, 아주 유사한 사진과 동일한 캡션을 다는 케이스가 아닐 수 있기 때문에
여러 부가적인 힌트가 뒤에 다수 존재하며 이 중 어떤 힌트를 참조할 것인지 모델에 도움을 주기 위해서 힌트의 레벨을 표기한 토큰을 활용하여 학습 (cosHint lv4, diffHint lv3 etc)

NICE dataset에서는 general categories를 제공하는데, 이 역시 활용하고자 힌트 뒤에 [outdoors]와 같이 스페셜 토큰으로 구성하였음 4)

4) 초기 버전에서는 categories를 사용하지 않다가 어떠한 정보가 내재될 수 있을거라 예상하여 이를 추가하였으나 성능상의 변화를 관측하지는 못함

# 학습경과및 중간보고

| # | User | Entries | Date of Last Entry | Team Name | Bleu_1 ▲ | Bleu_2 ▲ | Bleu_3 ▲ | Bleu_4 ▲ | ROUGE_L ▲ | CIDEr ▲ | METEOR ▲ | spice ▲ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PEJI | 2 | 04/21/23 | MMU | 52.7543 (5) | 42.8612 (2) | 36.2810 (2) | 31.5089 (2) | 51.5034 (3) | 285.6904 (1) | 26.9244 (2) | 40.7550 (1) |
| 2 | calisolo | 3 | 04/23/23 | Otsuka AI | 54.2277 (1) | 44.1486 (1) | 37.3150 (1) | 32.2381 (1) | 51.6011 (2) | 277.8786 (2) | 27.1880 (1) | 40.1471 (3) |

상기 방법론으로 제출한 중간 케이스에서 CIDEr 277.87점 달성 ( 최종 5위정도의 score)

출력물 분석과 공개된 점수를 다른 지원자와 비교하여 모델 개선 방향 탐구

# Intuition again

BLEU 스코어가 다른 지원자에 비해 높았고[5], 작업하면서 만들어진 몇가지 체크포인트의 답안을 비교하여 다음과 같은 가설 수립.

1. 텍스트 힌트가 너무 강력하기 때문에, 인코더에서의 이미지 피쳐 수렴 속도와 텍스트 임베딩 수렴 속도에 차이가 발생

예를들어, 좌측의 이미지는 florist이고, 특정 체크포인트에서는 이를 정확하게 추론하나 학습이 진행될수록  힌트로 제공하는 doctor를 참고하여 doctor로 예측하는 경향이 있다고 판단

```
1297797287 1297787657 1297787624 1586681501 1586681510
[cosHint lv2][diffHint lv2][medical]Studio cut out of female doctor with stethoscope smiling at camera
[cosHint lv1][diffHint lv2][misc]Studio cutout of car mechanic with air hammer on white background smiling at camera
[cosHint lv1][diffHint lv1][office]cutout Of Male Tailor
[cosHint lv1][diffHint lv1][office]cutout Of Male Teenage Student Studying Fashion
[shot_style]cutout | Cut out of | Cut out of | cutout of | [Location][NULL]

caption at 21: A female doctor with a bouquet of flowers smiling at the camera
caption atnew: Studio cutout of female florist holding bouquet of flowers looking at camera
decoder final: Studio cutout of female doctor with flowers smiling at camera
caption at s3  : Cut out of female doctor holding bouquet of flowers smiling at camera
caption at s4  : cutout of smiling florist holding bouquet of flowers
```

Submission 3 answer

5) BLEU는 전체 문제에서의 등장 단어수를 카운팅하기 때문에 말투는 잘 따라하였지만 주어진 이미지가 아닌 다른 설명을 하는 빈도가 높다고 판단

Hyperparameter 세팅을 통해 이미지 피쳐는 최대한 보존하면서 텍스트에 대한 조정 시도

Learning rate, total epochs, weight decay, freeze image encoder  등을 주로 조정하였으나 실패 6)7)8)

경과가 좋았던 몇가지 체크포인트를 선택하여 코사인 유사도 기준으로 답변을 voting하여 최종 답변 구성

결과:  CIDEr(287.69)  최종 4등 마무리

| # | User | Entries | Date of Last Entry | Team Name | Bleu_1 ▲ | Bleu_2 ▲ | Bleu_3 ▲ | Bleu_4 ▲ | ROUGE_L ▲ | CIDEr ▲ | METEOR ▲ | spice ▲ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | jinx | 5 | 05/01/23 | no | 56.0839 (3) | 46.5881 (2) | 40.0586 (2) | 35.1730 (2) | 55.5685 (2) | 325.7216 (1) | 29.1455 (2) | 44.4351 (2) |
| 2 | stack-top | 5 | 05/01/23 | Retriever | 58.0129 (1) | 47.8769 (1) | 40.9018 (1) | 35.7796 (1) | 56.3780 (1) | 324.9277 (2) | 30.0329 (1) | 45.5456 (1) |
| 3 | PEJI | 4 | 05/01/23 | MMU | 56.4908 (2) | 46.5371 (3) | 39.6859 (3) | 34.5996 (3) | 54.9832 (3) | 316.2290 (3) | 28.9407 (3) | 43.8281 (3) |
| 4 | calisolo | 5 | 05/01/23 | Otsuka AI | 55.7849 (4) | 45.4753 (4) | 38.4468 (4) | 33.1970 (4) | 52.9229 (4) | 287.6926 (4) | 27.9402 (4) | 41.3584 (4) |

6) 남은 시간이 촉박하여 lr을 높게세팅하고 짧은 epochs로 수행한 버전은 성능이 많이 감소하여 활용 어려움
7) eval_set을 따로 구성하지 않아, 모델에 적절한 피드백을 제공하기 어려웠고 결과물을 눈으로 살피는 직관에 의존하게 됨
8) Word_embedding과 encoder를 freeze 하는 경우, 한쪽만 freeze하다가 학습도중 unfreeze하는 경우 등의 아이디어는 추가 테스트 필요

# 회고&한계 및 추후 연구주제

1) 태깅 정확도 다소 낮으며 태깅 개선시 성능 향상 기대  - shutterstock 원천에서 생성하는 캡션으로 유추되어 기계적인 태깅 가능성 기대
2) OFA 인코더의 output을 기준으로 비교하였음. SAM(segment anything model)을 활용하고자 하였으나,  colab의 고용량 RAM 세팅(83gb)으로는 약 30000장 사진에 대한 SAM의 representation 저장 용량 이슈 발생. 자원이 충분하여 더욱 좋은 이미지 인코더를 사용한다면 레벨구간을 더 정밀하게 나눌 것으로 예상됨
3) 코사인 유사도만으로 명확하게 유사도를 비교하는 것은 무리가 있다고 판단됨 . 다른 방법론 탐구 필요
4) 초기 버전에서는 categories를 사용하지 않다가 어떠한 정보가 내재될 수 있을거라 예상하여 이를 추가하였으나 성능상의 변화를 관측하지는 못함
5) 전체 문제에서의 등장 단어수를 카운팅하기 때문에 말투는 잘 따라하였지만 주어진 이미지가 아닌 다른 설명을 하는 빈도가 높다고 판단(BLEU)
6) 남은 시간이 촉박하여 lr을 높게세팅하고 짧은 epochs로 수행한 버전은 성능이 많이 감소하여 활용 어려움
7) eval_set을 따로 구성하지 않아, 모델에 적절한 피드백을 제공하기 어려웠고 결과물을 눈으로 살피는 직관에 의존하게 됨
8) Word_embedding과 encoder를 freeze 하는 경우, 한쪽만 freeze하다가 학습도중 unfreeze하는 경우 등의 아이디어는 추가 테스트 필요

9) CIDEr optimization 방법론으로 널리 알려진 SCST 튜닝을 시도하였으나 구현상의 문제로 제대로 수렴이 되지않아 시간을 다소 낭비

Github(project code & reproduction) : https://github.com/calisolo/Levels_image_captioning_NICE

Environment:  Colab pro plus + google drive
( A100 40gb x 1ea,  83gb RAM, 100GB drive)

Backbone model: https://github.com/OFA-Sys/OFA (OFA)
https://arxiv.org/pdf/2202.03052.pdf (OFA -paper)