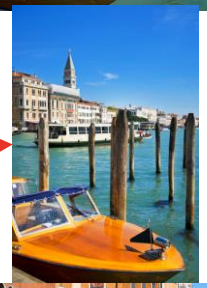
NICE
4th place result technical introduction

- CALISOLO (SOLGIL OH)

Table of Contents

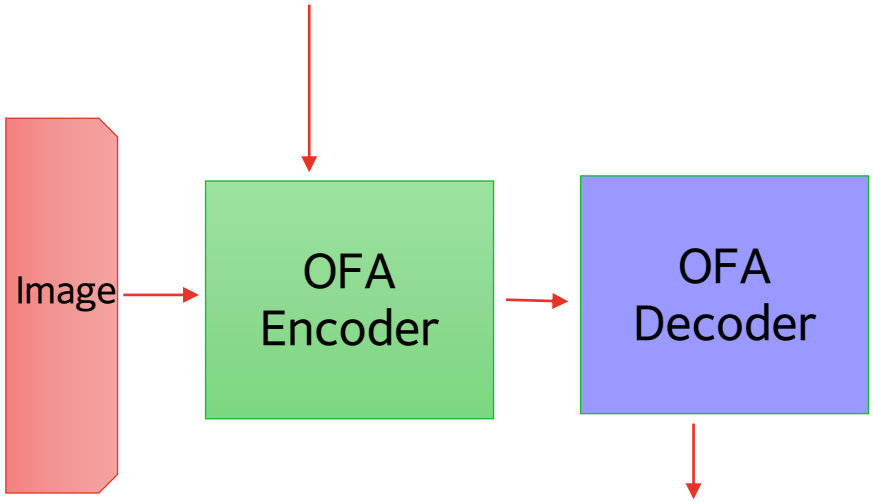
Overview of the methodology

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



Hint structure

[cosine similarity][id similarity][caption]



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

- Intuition: The characteristic of the NICE task is to capture the features of the image, but it is considered challenging to match the unique writing patterns of the NICE dataset (persona dialogue generation)
- OFA (apache 2.0 license) model-based, fine-tuned with a few shot hints to input data

Dataset characteristics

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

When looking 'Caption_gt', lots of situations where capitalization is irregular/describing shot style of photos

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

A caption that describes the shot style of the photo is most likely to have a prefix that describes format of the photo.
Judging by the context, capitalization that does not appear at the beginning is mainly for specific place names (Spain, Brazil, Amazon River), etc.

I tagged 5000 validation sets as written above.[1)]
(6-8 hours of manual reclassification after simple filters like searching for capitalization)

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) expect better performance with better tagging – possibility of automatic tagging inferred from captions generated by Shutterstock sources

Highland cattle on farm Seefeld Bavaria Germany

Location features:

Requires background knowledge, hard to infer from photos alone (habitat of the highland cattle, distribution of vegetation in photos?)

1) Which captions should have location and which shouldn't?
2) Even if I know the photo with the location, how do I know where that location is?

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 style features:

It describes the photo well, but it's hard to decide whether to describe it or not.

1) Which captions have shot_style and which don't?

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					

Discovered trends in shot_style/ location when sorting by public id

Hypothesis

1. Photos from the same provider can be inferred from the information inherent in the image and will be similar in subject matter/photo/caption.
2. The public id is the upload number on Shutterstock, and it's likely that the photo is from the same provider for consecutive uploads.

Approach

- Train from similarities between photos, utilizing the public id provided in the validation_set

Input data form

Input image

Cosine similarity Map

Ordered by cosine similarity

Caption 1

Caption 2

Caption 3

Caption 4

Cosine Hint | diff Hint | category | Caption 1

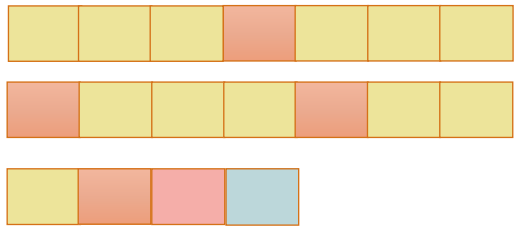
Cosine Hint | diff Hint | category | Caption 2

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] : I wanted to use word embeddings to model how much the current query image differs from the image representation in the hint set based on cosine similarity [2] between images. If I put a number like 0.355 into the input, I expect the model to be unable to determine the specific case, so I assign 4 tokens to the vocabulary to represent it, and then normalize them into groups of 4.[3]

1. [cosHint lv4] : Strong hints for nearly identical photos > 0.4
2. [cosHint lv3] : Same topic but expected to have different captions > 0.32
3. [cosHint lv2] : Similar photos but different captions > 0.29
4. [cosHint lv1] : Irrelevant photos ≤ 0.29

- [Diff Hint]: Similar to cosine hint, a configuration to pass information to the model about how different the public_id of the query image and the hint set are. Performed by assigning 3 tokens to a vocabulary and normalizing them into 3 groups

1. [diffHint lv3]: The public_id difference between the photos is very small <100
2. [diffHint lv2]: The public_id difference between the photos is small <10000
3. [diffHint lv1]: The public_id difference between the photos is large The remaining

2) The comparison was based on the output of the OFA encoder. I wanted to utilize the segment anything model (SAM), but with colab's high RAM settings (83gb), I encountered storage issues with the SAM's representation for about 30000 photos. With more resources and a better image encoder, I expect to get a finer level segmentation.

3) Cosine similarity alone is not as clear a comparison of similarity as the above criteria. I wanted to get some sense of trend.

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

Summary hint | shotStyle total | Location total

The hint set constructed from each similar photo provides information about the analysis of similar photos based on cosine similarity. Similarly, to learn the caption writing style of a photo's neighbors based on its public_id, I configured a Summary hint

- shotStyle total: List all shotStyles from the six closest photos by public id
- Location total: The six closest locations based on public id and their respective diffHints

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]

Configure the data storage as shown in the table on the left

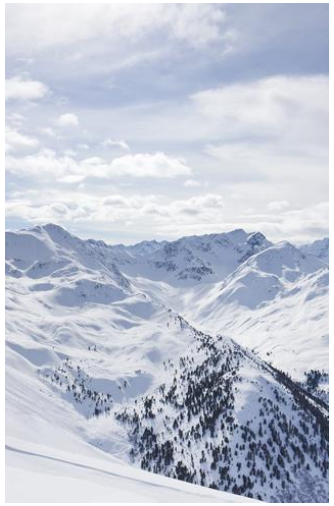
How it really looks like

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



Based on the image on the left and the text above, the OFA model inferences the correct answer

With cosine similarity, it is likely that this case will also generate "View of snowy mountain range" because similar photos are often captioned "View of snowy mountain range"

However, this may not be the case for very similar photos with identical captions. I utilize tokens that indicate the level of the hint to help the model learn which hint to refer to when there are multiple additional hints behind it (cosHint lv4, diffHint lv3, etc.)

The NICE dataset provides general categories, which I also utilize by adding special tokens such as [outdoors] after the hint.[4)]

4) Early versions didn't use categories, then added them because I thought they might contain some information, but I didn't observe any performance changes.

Progress and mid-report

#	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)

Achieved a CIDEr score of 277.87 on an intermediate case submitted with the above methodology

Analyzed outputs and compared published scores with other applicants to explore ways to improve the model

Intuition again

BLEU scores were higher than other applicants[5], compare answers from several checkpoints created as I worked to formulate the following hypotheses.

1. Text hints are too strong, resulting in a difference between the convergence rate of image features in the encoder and the convergence rate of text embedding.

For example, the image on the left is a florist, and at a certain checkpoint, it correctly inferences this, but as training progresses, it tends to predict doctor by referring to doctor as a hint.

```
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) Since BLEU compares the number of words in the entire question, it was determined that the model followed the writing style well, but occasionally described other than the given image.

Attempt to adjust for text embedding while preserving image features through hyperparameter settings

Mainly adjusted learning rate, total epochs, weight decay, freeze image encoder, etc. but failed[6)7)8)]

Constructed the final answer by selecting a few checkpoints with good progress and voting the answers based on cosine similarity.

Results: CIDEr (287.69) finishing in 4th place.

#	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) Due to the short time remaining, the version with high lr and short epochs was difficult to utilize due to the significant decrease in performance
7) Without a separate eval_set, it was difficult to provide proper feedback to the model and relied on intuition by looking at the output.
8) Ideas such as freezing Word_embedding and encoder, freezing only one and then unfreezing it during training need further testing.

Limitations and future research topics

1) expect better performance with better tagging – possibility of automatic tagging inferred from captions generated by Shutterstock sources
2) The comparison was based on the output of the OFA encoder. I wanted to utilize the segment anything model (SAM), but with colab's high RAM settings (83gb), I encountered storage issues with the SAM's representation for about 30000 photos. With more resources and a better image encoder, I expect to get a finer level segmentation.
3) Cosine similarity alone is not as clear a comparison of similarity as the above criteria. Need to explore other criteria.
4) Early versions didn't use categories, then added them because I thought they might contain some information, but I didn't observe any performance changes.
5) It was determined that the model followed the writing style well, but occasionally described other than the given image. (BLEU)
6) Due to the short time remaining, the version with high lr and short epochs was difficult to utilize due to the significant decrease in performance
7) Without a separate eval_set, it was difficult to provide proper feedback to the model and relied on intuition by looking at the output.
8) Ideas such as freezing Word_embedding and encoder, freezing only one and then unfreezing it during training need further testing.

9) I tried SCST tuning, which is popularly known as CIDEr optimization methodology, but it did not converge properly due to implementation issues and wasted some time.

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)