Image Captioning - Sample Test Run

Captioning is given an empirical score, based on how appropriate and accurate it is, factoring in grammar or spelling mistakes. One observation was that the "blip_large" model often inserts one non-existent word, "arafed" or "araffle," but otherwise seems to generate good captions. Often if the one word is removed the caption would seem ideal. In this case, an extra row is added marked with an *asterisk. The score is - 0 for bad, 1 for somewhat good, 2 for seemingly ideal.

Estimated Labels

jetski.png

model	caption	score
vitgpt	a man riding a surfboard on top of a wave	1
git_base	the first race of the season	0
blip_base	a man and woman riding a jet ski	1
git_large	a man and a woman riding a jet ski.	1
blip_large	they are riding on a jet ski in the water	2

family.jpg

model	caption	score
vitgpt	a man and woman posing for a picture	1
git_base	family portrait the cast of tv sitcom	2
blip_base	a family portrait of the late - great - grandfather, father, and two - year - old - age - age	1
git_large	the family in the tv show	2
blip_large	arafed photo of a man and two women and a man sitting on a porch	1

running.jpg

model	caption	score
vitgpt	two people walking down a path in the woods	1
git_base	a group of kids running on a path in the woods.	2
blip_base	two children running down a dirt path in the woods	1
git_large	a group of people running down a path.	1
blip_large	there are three children running down a dirt path in the woods	2

house.jpg

model	caption	score
vitgpt	a living room with a couch, chairs, and a table	2
git_base	a living room with a staircase and a clock.	2
blip_base	the living room is open to the kitchen and dining area	1
git_large	a living room with a couch, coffee table, and a staircase.	2
blip_large	arafed living room with a couch, coffee table, and stairs	1
*blip_large	~~arafed~~ living room with a couch, coffee table, and stairs	2

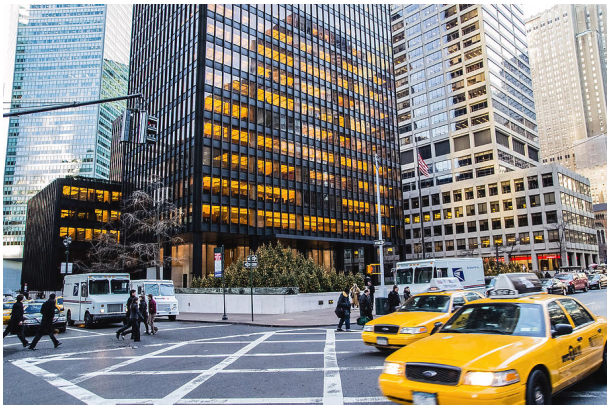
climb.jpg

model	caption	score
vitgpt	a man is sitting on a rock ledge overlooking a mountain range	0
git_base	a rock climber climbs a steep rock face.	2
blip_base	a man climbing on a rock in the mountains	1
git_large	a woman rock climbing on a cliff	2
blip_large	arafed climber on a rock face with a river in the background	1
*blip_large	~~arafed~~ climber on a rock face with a river in the background	2

pancakes.jpg

model	caption	score
vitgpt	a woman holding a plate of food with a knife	0
git_base	a girl with long hair holding a stack of pancakes.	1
blip_base	a woman is giving a thumbs up	1
git_large	a girl with a stack of pancakes on a plate.	2
blip_large	araffe woman sitting at a table with a stack of pancakes	1
*blip_large	~~araffe~~ woman sitting at a table with a stack of pancakes	2

city.jpg



model	caption	score
vitgpt	a city street with a large building and a yellow traffic light	1
git_base	a taxi cab is stopped at a crosswalk.	1
blip_base	a yellow taxi cab	1
git_large	a city street with yellow taxis and a building.	2
blip_large	araffle taxis and pedestrians cross a busy city street in the evening	1
*blip_large	~~araffle~~ taxis and pedestrians cross a busy city street in the evening	2

Overall Analysis

Each captioning model run for all images was timed, and scaled from 0 (slowest) to 2 (fastest). In the table below, the average empirical score from all images above is added, then a final average of the 2 computed (speed vs quality). Keep in mind that this test was run on a fairly slow machine with no GPUs, but model loading time was not factored-in..

model	total secs	scaled time score	average caption score	final
vitgpt	144.9	2.0	0.8	1.4
blip_base	169.6	2.0	1.0	1.5
blip_large	364.2	1.6	1.1	1.4
*blip_large	364.2	1.6	1.8	1.7
git_base	489.5	1.4	1.4	1.4
git_large	1260.9	0.0	1.7	0.9