Question Answering about Images using Visual Semantic Embeddings

Q2: what is grazing in a grassy area above a body of water ?
Ground truth: sheep
2-CNN-LSTM: sheep (0.6904)
CNN-BOW: sheep (0.8393)
BLIND-LSTM: zebra (0.3476)
BLIND-BOW: giraffe (0.1673)
Image features are useful

Q85: what is standing in the sand alone ?
Ground truth: bird
2-CNN-LSTM: bird (0.6608)
CNN-BOW: seagull (0.3627)
BLIND-LSTM: elephant (0.2465)
BLIND-BOW: elephant (0.1757)
Image features are useful

Q223: what is the man in a helmet eating ?
Ground truth: banana
2-CNN-LSTM: banana (0.8961)
CNN-BOW: banana (0.9729)
BLIND-LSTM: sandwich (0.2065)
BLIND-BOW: sandwich (0.1627)
Occluded object are still focusable

Q1213: what is wrapped in packaging paper for shipping ?
Ground truth: toilet
2-CNN-LSTM: toilet (0.3934)
CNN-BOW: toilet (0.3243)
BLIND-LSTM: sandwich (0.2178)
BLIND-BOW: sandwich (0.2595)
Blind model has no clue from the text alone

Q1549: what are sitting on wood UNK one is cut in half ?
Ground truth: oranges
2-CNN-LSTM: oranges (0.8196)
CNN-BOW: oranges (0.3242)
BLIND-LSTM: pizzas (0.1006)
BLIND-BOW: plate (0.0424)
Both models know plural for

Q2007: what is sitting next to a cup of coffee ?
Ground truth: banana
2-CNN-LSTM: banana (0.4267)
CNN-BOW: banana (0.6397)
BLIND-LSTM: sandwich (0.1032)
BLIND-BOW: sandwich (0.1012)
When there are two objects present the model attend to the correct one

Q4974: what is the person riding down the street ?
Ground truth: skateboard
2-CNN-LSTM: skateboard (0.8822)
CNN-BOW: skateboard (0.4611)
BLIND-LSTM: motorcycle (0.3368)
BLIND-BOW: skateboard (0.2380)
Reletively small object to focus on

Q5429: what do two women hold with a picture on it ?
Ground truth: cake
2-CNN-LSTM: cake (0.5611)
CNN-BOW: laptop (0.1443)
BLIND-LSTM: umbrellas (0.1567)
BLIND-BOW: phones (0.1447)
Does not look like a regular cake and the model has to infer from the candles

Q1965: what did people set up on the beach next to blankets ?
Ground truth: umbrellas
2-CNN-LSTM: umbrellas (0.4768)
CNN-BOW: umbrella (0.7319)
BLIND-LSTM: kites (0.1585)
BLIND-BOW: boat (0.1258)
The CNN-LSTM model knows the plural but CNN-BOW does not

Figure 1. COCO-QA Object Questions

Q1966: what are the man and woman holding up ?
Ground truth: pizza
2-CNN-LSTM: pizza (0.0775)
CNN-BOW: pizzas (0.1807)
BLIND-LSTM: phones (0.1421)
BLIND-BOW: phones (0.2272)
The CNN-BOW again messed up the counting (mostly of the questions with LSTM win over BOW is on singular/plural)

Q2950: what is the color of the man ?
Ground truth: black
2-CNN-LSTM: black (0.7063)
CNN-BOW: black (0.2809)
BLIND-LSTM: black (0.4410)
BLIND-BOW: black (0.4541)
LSTM learns to add plural form from the verb tense

Q14930: what next to some UNK UNK a lime and a mixed drink ?
Ground truth: blender
2-CNN-LSTM: blender (0.3763)
CNN-BOW: plate (0.0927)
BLIND-LSTM: standing (0.1304)
BLIND-BOW: plate (0.0410)
Many items present but the LSTM figures out the correct item

Q15178: what are two guys riding down the street ?
Ground truth: motorcycle
2-CNN-LSTM: motorcycle (0.6804)
CNN-BOW: motorcycles (0.4514)
BLIND-LSTM: motorcycles (0.2584)
BLIND-BOW: bikes (0.2329)
From the text alone it is misleading that there might exist multiple motorcycles

Q19533: what lays in the bed resting it 's head against a pillow ?
Ground truth: bear
2-CNN-LSTM: bear (0.3648)
CNN-BOW: bed (0.3379)
BLIND-LSTM: dog (0.5108)
BLIND-BOW: dog (0.6639)
BOW is not a good blend of visual information with the question semantics

Q24952: what is the black and white cat wearing ?
Ground truth: hat
2-CNN-LSTM: hat (0.6349)
CNN-BOW: cat (0.6880)
BLIND-LSTM: tie (0.5821)
BLIND-BOW: tie (0.3385)
BOW is not a good blend of visual information with the question semantics

Q29528: what is there filled with sauce covered food ?
Ground truth: tray
2-CNN-LSTM: tray (0.1637)
CNN-BOW: pizza (0.1872)
BLIND-LSTM: plate (0.2158)
BLIND-BOW: plate (0.2363)
Tray is hard to get

Q30568: what stuck in an apple ?
Ground truth: knife
2-CNN-LSTM: knife (0.2179)
CNN-BOW: apple (0.2666)
BLIND-LSTM: bananas (0.0894)
BLIND-BOW: apple (0.0838)
It is rare to see something stuck into something and here CNN-LSTM win over other 4 models shows a good combination of image and question semantics

Q36019: the woman wearing what UNK with a little girl as she played with her stroller UNK which also has an umbrella over it ?
Ground truth: hat
2-CNN-LSTM: hat (0.4144)
CNN-BOW: UNK (0.0371)
BLIND-LSTM: shirt (0.0738)
BLIND-BOW: plate (0.0401)
LSTM remembers the beginning part of the question although the entire question is very long

Figure 2. COCO-QA Object Questions

Q38613: what are the people watching and taking photos ?
Ground truth: elephants
2-CNN-LSTM: elephants (0.4016)
CNN-BOW: horses (0.7849)
BLIND-LSTM: picture (0.1402)
BLIND-BOW: picture (0.2065)
The object is fairly small to focus on

Q3098: what is locked up by some trees ?
Ground truth: bike
2-CNN-LSTM: bicycle (0.5127)
CNN-BOW: bike (0.7138)
BLIND-LSTM: giraffe (0.3866)
BLIND-BOW: elephant (0.1834)
In many cases where CNN-BOW wins over CNN-LSTM is just some synonyms of answers

Q14124: what does the man with no shirt hold ?
Ground truth: toothbrush
2-CNN-LSTM: phone (0.0897)
CNN-BOW: toothbrush (0.2189)
BLIND-LSTM: umbrella (0.1725)
BLIND-BOW: phone (0.1449)
CNN-LSTM fails on this one because it probably takes in too much textual information

Q14518: what will get the lot of use by its owner ?
Ground truth: mouse
2-CNN-LSTM: toothbrushes (0.1746)
CNN-BOW: mouse (0.2407)
BLIND-LSTM: dogs (0.1367)
BLIND-BOW: elephant (0.0866)
CNN-LSTM fails

Q15499: what flies across the blue sky ?
Ground truth: airplane
2-CNN-LSTM: jet (0.3949)
CNN-BOW: airplane (0.4655)
BLIND-LSTM: jet (0.3733)
BLIND-BOW: plane (0.4066)
Synonyms again

Q37066: what is walking on the sandy beach ?
Ground truth: dog
2-CNN-LSTM: bird (0.3321)
CNN-BOW: dog (0.5306)
BLIND-LSTM: bird (0.2449)
BLIND-BOW: bird (0.2159)
Both objects are present but the LSTM output similar to Blind ones which tells us that it takes more textual information

Q385: two ladies enjoying what at a roadside stand ?
Ground truth: dessert
2-CNN-LSTM: bananas (0.4363)
CNN-BOW: banana (0.9090)
BLIND-LSTM: fruit (0.0820)
BLIND-BOW: meal (0.1226)
Yellow food doesn't have to be bananas

Q1681: what is the man in a white shirt UNK striped tie holding ?
Ground truth: umbrella
2-CNN-LSTM: phone (0.2010)
CNN-BOW: phone (0.0785)
BLIND-LSTM: phone (0.1760)
BLIND-BOW: plate (0.0409)
Overfits to people holding phones

Q2608: the teddy bear on a windowsill holding what ?
Ground truth: cellphone
2-CNN-LSTM: bear (0.3890)
CNN-BOW: bear (0.8489)
BLIND-LSTM: bear (0.2484)
BLIND-BOW: bear (0.2859)
Hard for a bear to hold cellphone

Figure 3. COCO-QA Object Questions

Q828: what is towing the heavy load across town ?
Ground truth: truck
2-CNN-LSTM: train (0.8407)
CNN-BOW: train (0.5635)
BLIND-LSTM: truck (0.6577)
BLIND-BOW: truck (0.4229)
Very long truck looks like a train?

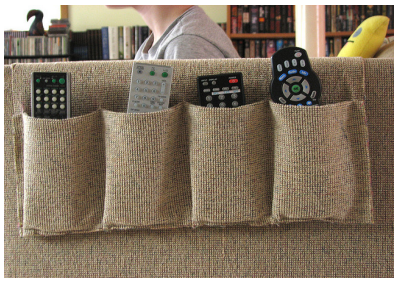
Q33: what is shining down on the passing airplane ?
Ground truth: sun
2-CNN-LSTM: airplane (0.4153)
CNN-BOW: airplane (0.2487)
BLIND-LSTM: sun (0.3013)
BLIND-BOW: car (0.1130)
Image provides no extra information and blind LSTM really focuses on the text to get the right answer

Q7231: what lined up along one wall with a handicap UNK one at the end ?
Ground truth: urinals
2-CNN-LSTM: suitcases (0.2665)
CNN-BOW: luggage (0.0518)
BLIND-LSTM: urinals (0.0969)
BLIND-BOW: plate (0.0401)
Interesting that the blind model knows the correct answer whereas the CNN model thinks they are suitcases

Q24432: what does the colorful market booth sell to a customer ?
Ground truth: bananas
2-CNN-LSTM: box (0.1577)
CNN-BOW: cake (0.1204)
BLIND-LSTM: bananas (0.0727)
BLIND-BOW: fruit (0.2503)
Very hard to answer bananas but the blind model made it

Figure 4. COCO-QA Object Questions

Q3202: how many red double decker buses parked side by side ?
Ground truth: three
2-CNN-LSTM: three (0.3344)
CNN-BOW: two (0.5916)
BLIND-LSTM: two (0.3844)
BLIND-BOW: two (0.4718)
Couting regarding to an object with a specific color



Q4305: how many remotes are inside of UNK on the arm of a couch ?
Ground truth: four
2-CNN-LSTM: four (0.5329)
CNN-BOW: five (0.2370)
BLIND-LSTM: three (0.3898)
BLIND-BOW: plate (0.0330)
Even counts to four

Q6805: how many uncooked doughnuts sit on the baking tray ?
Ground truth: six
2-CNN-LSTM: six (0.2779)
CNN-BOW: twelve (0.2342)
BLIND-LSTM: four (0.2345)
BLIND-BOW: four (0.2473)
Six is very hard unless there is a training example about six donuts

Q7270: how many guys in a room is playing with a wii ?
Ground truth: four
2-CNN-LSTM: four (0.4461)
CNN-BOW: five (0.3932)
BLIND-LSTM: two (0.4457)
BLIND-BOW: two (0.6535)
Four again

Q8112: the jet airliner wing that has how many jet engines ?
Ground truth: two
2-CNN-LSTM: two (0.2822)
CNN-BOW: one (0.2979)
BLIND-LSTM: one (0.0833)
BLIND-BOW: four (0.3918)
Tricky question - four engines in total but the picture is only showing two on one side

Q9148: how many bowls of various foods including fruits UNK vegetables UNK nuts and soup ?
Ground truth: four
2-CNN-LSTM: four (0.4394)
CNN-BOW: meal (0.1174)
BLIND-LSTM: three (0.2357)
BLIND-BOW: plate (0.0410)
Counting food is good because of the clear boundary

Q32912: how many bottles of beer is sitting on a wooden table next to bananas ?
Ground truth: three
2-CNN-LSTM: three (0.3403)
CNN-BOW: two (0.4664)
BLIND-LSTM: two (0.3781)
BLIND-BOW: four (0.3717)
Multiple objects but asking about bottles

Q28078: how many sheep is running from the dog in the grass ?
Ground truth: three
2-CNN-LSTM: three (0.3595)
CNN-BOW: five (0.4277)
BLIND-LSTM: two (0.2966)
BLIND-BOW: two (0.2700)
Multiple objects need to focus on different parts

Q3757: how many gentlemen is interacting in outdoor area with shops nearby ?
Ground truth: two
2-CNN-LSTM: three (0.4992)
CNN-BOW: two (0.6973)
BLIND-LSTM: three (0.4300)
BLIND-BOW: four (0.3170)
Mostly CNN-LSTM outputs more "three"s and CNN-BOW outputs more "two"s

Figure 5. COCO-QA Number Questions

Q9489: how many benches surrounded by leaves near the pond ?
Ground truth: two
2-CNN-LSTM: four (0.3651)
CNN-BOW: two (0.3351)
BLIND-LSTM: three (0.3202)
BLIND-BOW: four (0.3390)
Seems to have no reason to answer "four"

Q170: how many black laptop computers in UNK on a white table ?
Ground truth: three
2-CNN-LSTM: three (0.3208)
CNN-BOW: three (0.2394)
BLIND-LSTM: two (0.4606)
BLIND-BOW: plate (0.0363)
Both CNN models agree on counting

Q3449: how many men cut the cake at an event ?
Ground truth: five
2-CNN-LSTM: five (0.3079)
CNN-BOW: five (0.4365)
BLIND-LSTM: three (0.4024)
BLIND-BOW: four (0.3052)
This is not easy to get firstly because "five" is fairly large number to count and secondly because cutting cakes really doesn't imply five persons

Q24263: how many ducks are standing on the beach with water in the foreground ?
Ground truth: five
2-CNN-LSTM: five (0.3164)
CNN-BOW: five (0.4691)
BLIND-LSTM: three (0.3322)
BLIND-BOW: four (0.2999)
Again five is not easy to count for both models

Q6: how many way sign that is in front of a building ?
Ground truth: one
2-CNN-LSTM: one (0.8434)
CNN-BOW: one (0.7174)
BLIND-LSTM: one (0.8946)
BLIND-BOW: one (0.7984)
Usually signs show "one way"

Q25552: how many trains are parked next to each other on the tracks ?
Ground truth: two
2-CNN-LSTM: two (0.4889)
CNN-BOW: two (0.6430)
BLIND-LSTM: two (0.4091)
BLIND-BOW: two (0.5368)
Answer is revealed in the phrase "each other"

Q25671: how many slice is there of pizza left on the plate ?
Ground truth: one
2-CNN-LSTM: one (0.9581)
CNN-BOW: one (0.4745)
BLIND-LSTM: one (0.9584)
BLIND-BOW: one (0.8934)
Answer is revealed in the word "is"

Q24776: how many small dog is running next to four sheep ?
Ground truth: one
2-CNN-LSTM: one (0.6354)
CNN-BOW: three (0.2432)
BLIND-LSTM: one (0.8920)
BLIND-BOW: two (0.2721)
LSTM captures "is" whereas BOW doesn't

Q811: how many apples is sitting in a bowl on a table ?
Ground truth: four
2-CNN-LSTM: three (0.3703)
CNN-BOW: two (0.4607)
BLIND-LSTM: two (0.2551)
BLIND-BOW: four (0.2847)
Easy one but only blind BOW gets it

Figure 6. COCO-QA Number Questions

Q33482: how many surfers on a beach with their surfboards ?
Ground truth: four
2-CNN-LSTM: two (0.4726)
CNN-BOW: two (0.5776)
BLIND-LSTM: three (0.3930)
BLIND-BOW: four (0.3283)
Again blind BOW tend to guess "four" when it has no clue and other models just don't get thecorrect answer

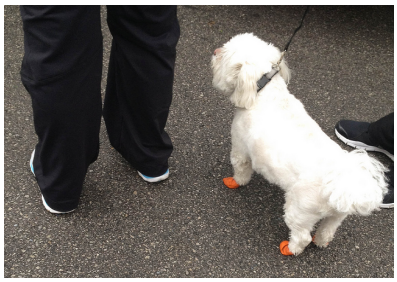
Q12255: how many engine plane is flying in lightly overcast sky ?
Ground truth: four
2-CNN-LSTM: four (0.4671)
CNN-BOW: nine (0.2623)
BLIND-LSTM: one (0.3137)
BLIND-BOW: four (0.5412)
This one is impossible to get

Q38217: how many blue jets with yellow painted on top are flying together ?
Ground truth: four
2-CNN-LSTM: four (0.3229)
CNN-BOW: six (0.1610)
BLIND-LSTM: three (0.2661)
BLIND-BOW: four (0.4019)
Objects are overlapped probably the model just guesses four planes regardless

Figure 7. COCO-QA Number Questions

Q521: what is the color of the umbrellas ?
Ground truth: yellow
2-CNN-LSTM: yellow (0.9376)
CNN-BOW: yellow (0.3751)
BLIND-LSTM: red (0.2789)
BLIND-BOW: blue (0.2287)
CNN tend to focus on bright colors even if they only occupy small portion of the image

Q1362: what is the color of the flowers ?
Ground truth: red
2-CNN-LSTM: red (0.4641)
CNN-BOW: red (0.4694)
BLIND-LSTM: purple (0.6047)
BLIND-BOW: purple (0.5514)
The blue bench is more obvious to detect but the question is asking about flowers



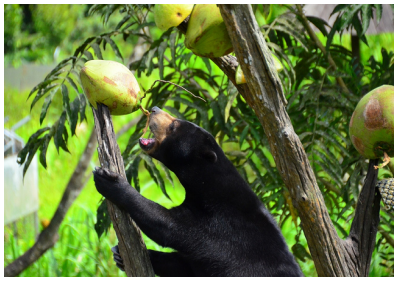
Q2023: what is the color of the dog ?
Ground truth: white
2-CNN-LSTM: white (0.6819)
CNN-BOW: white (0.9388)
BLIND-LSTM: brown (0.5489)
BLIND-BOW: brown (0.4795)
Focus on the animal not the pants

Q2683: what is the color of the table ?
Ground truth: black
2-CNN-LSTM: black (0.7793)
CNN-BOW: black (0.3785)
BLIND-LSTM: brown (0.2596)
BLIND-BOW: white (0.3189)
Attend to background of the image

Q5520: what is the color of the banana ?
Ground truth: green
2-CNN-LSTM: green (0.4476)
CNN-BOW: green (0.7215)
BLIND-LSTM: yellow (0.3524)
BLIND-BOW: yellow (0.3629)
Banana question 1

Q9822: what is the color of the bananas ?
Ground truth: yellow
2-CNN-LSTM: yellow (0.6238)
CNN-BOW: yellow (0.8101)
BLIND-LSTM: green (0.3822)
BLIND-BOW: green (0.3820)
Banana question 2

Q34206: what is the color of the bowl ?
Ground truth: red
2-CNN-LSTM: red (0.4605)
CNN-BOW: red (0.3050)
BLIND-LSTM: white (0.1952)
BLIND-BOW: white (0.4396)
Hard to look for the bowl



Q139: what is the color of the bear ?
Ground truth: black
2-CNN-LSTM: black (0.8752)
CNN-BOW: green (0.3879)
BLIND-LSTM: brown (0.5518)
BLIND-BOW: brown (0.6222)
BOW now mistakenly output the dominant color whereas the blind models give a good guess

Q1404: what is the color of the toothbrush ?
Ground truth: green
2-CNN-LSTM: green (0.4124)
CNN-BOW: blue (0.3226)
BLIND-LSTM: blue (0.1879)
BLIND-BOW: blue (0.2653)
LSTM is more attentive to details of both question and image

Figure 8. COCO-QA Color Questions

Q1586: what is the color of the wheels ?
Ground truth: yellow
2-CNN-LSTM: yellow (0.4101)
CNN-BOW: brown (0.4306)
BLIND-LSTM: orange (0.2095)
BLIND-BOW: blue (0.1855)
Attend to detail

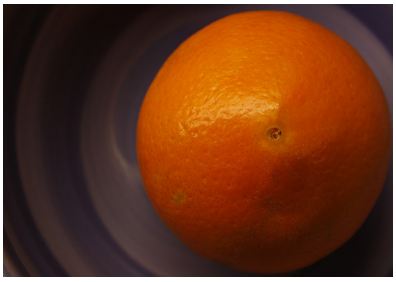
Q6358: what is the color of the boat ?
Ground truth: red
2-CNN-LSTM: red (0.5345)
CNN-BOW: blue (0.2879)
BLIND-LSTM: white (0.1832)
BLIND-BOW: white (0.2890)
Not the colorful umbrella it is asking

Q10492: what is the color of the plane ?
Ground truth: red
2-CNN-LSTM: red (0.5145)
CNN-BOW: blue (0.6872)
BLIND-LSTM: white (0.2756)
BLIND-BOW: white (0.2865)
BOW goes for the dominant color of the image again

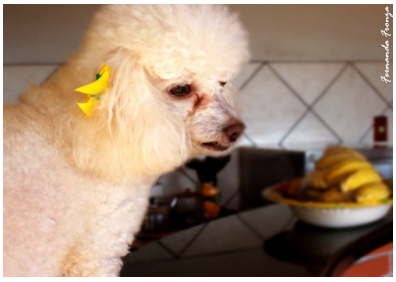
Q1037: what is the color of the skies ?
Ground truth: blue
2-CNN-LSTM: blue (0.4425)
CNN-BOW: blue (0.6370)
BLIND-LSTM: blue (0.3826)
BLIND-BOW: blue (0.4494)
Sometimes colors are easy to guess

Q591: what is the color of the shorts ?
Ground truth: black
2-CNN-LSTM: blue (0.4518)
CNN-BOW: blue (0.3261)
BLIND-LSTM: black (0.1963)
BLIND-BOW: black (0.2141)
CNN can't tell that shorts are worn on lower part of the body

Q2211: what is the color of the salad ?
Ground truth: green
2-CNN-LSTM: white (0.3428)
CNN-BOW: white (0.3321)
BLIND-LSTM: green (0.3224)
BLIND-BOW: green (0.3310)
CNN is really looking at the background instead of the foreground



Q17918: what is the color of the plate ?
Ground truth: white
2-CNN-LSTM: orange (0.4249)
CNN-BOW: orange (0.4252)
BLIND-LSTM: white (0.6865)
BLIND-BOW: white (0.7853)
Not asking about the forgrond dominant object but CNN just can't help



Q2562: what is the color of the ribbon ?
Ground truth: yellow
2-CNN-LSTM: white (0.8117)
CNN-BOW: white (0.8068)
BLIND-LSTM: red (0.2475)
BLIND-BOW: red (0.1830)
Very hard question and even if it gets right it might just focusing on the bananas

Q3705: what is the color of the flowers ?
Ground truth: blue
2-CNN-LSTM: orange (0.6924)
CNN-BOW: orange (0.3380)
BLIND-LSTM: purple (0.6047)
BLIND-BOW: purple (0.5514)
Very unusual scene setting but again CNN doesn't focus on the right object

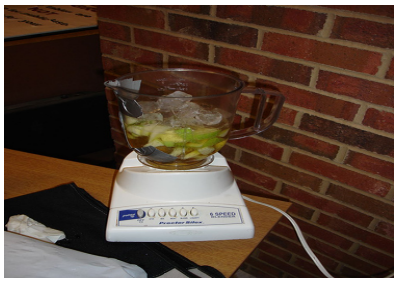
Figure 9. COCO-QA Color Questions

Q5826: what is the color of the sign ?
Ground truth: blue
2-CNN-LSTM: red (0.3877)
CNN-BOW: red (0.3503)
BLIND-LSTM: red (0.4757)
BLIND-BOW: red (0.3888)
Seems like all models are using textual information about a sign instead of the image

Figure 10. COCO-QA Color Questions

Q42: where is a white tiger ?
Ground truth: pool
2-CNN-LSTM: pool (0.1984)
CNN-BOW: pool (0.2602)
BLIND-LSTM: bathroom (0.0935)
BLIND-BOW: room (0.2206)
Image features are useful

Q857: where is the woman ?
Ground truth: station
2-CNN-LSTM: station (0.0833)
CNN-BOW: station (0.2196)
BLIND-LSTM: bed (0.0644)
BLIND-BOW: room (0.1590)
Image features are useful



Q805: where is fruit and ice ?
Ground truth: blender
2-CNN-LSTM: blender (0.1350)
CNN-BOW: blender (0.3828)
BLIND-LSTM: bowl (0.3528)
BLIND-BOW: bowl (0.5479)
Inclusion of object (not just room and scene)

Q9648: where does the man with glasses and a tie clip sit ?
Ground truth: car
2-CNN-LSTM: car (0.7772)
CNN-BOW: car (0.4117)
BLIND-LSTM: bowl (0.1078)
BLIND-BOW: room (0.3101)
This one is hard to see it is in a car

Q15891: where did the cat curl up UNK laying ?
Ground truth: sink
2-CNN-LSTM: sink (0.8603)
CNN-BOW: sink (0.4068)
BLIND-LSTM: bed (0.3623)
BLIND-BOW: plate (0.0401)
Easy one for CNN

Q25218: where are the ripe bananas sitting ?
Ground truth: basket
2-CNN-LSTM: basket (0.4965)
CNN-BOW: basket (0.4929)
BLIND-LSTM: bowl (0.6415)
BLIND-BOW: bowl (0.5170)
Inclusion of object

Q25397: where are people looking are different products ?
Ground truth: store
2-CNN-LSTM: store (0.2376)
CNN-BOW: store (0.2212)
BLIND-LSTM: kitchen (0.1704)
BLIND-BOW: window (0.1182)
Hard to infer

Q35349: where is an elephant approaching ?
Ground truth: mirror
2-CNN-LSTM: mirror (0.1540)
CNN-BOW: mirror (0.8029)
BLIND-LSTM: zoo (0.1094)
BLIND-BOW: zoo (0.1345)
The model knows that the elephant is inside the mirror

Q137: where is the unique motorcycle parked ?
Ground truth: garage
2-CNN-LSTM: garage (0.3119)
CNN-BOW: motorcycle (0.7361)
BLIND-LSTM: street (0.1918)
BLIND-BOW: street (0.1905)
BOW relies too much on image features and LSTM combines image and text well

Figure 11.

Q23418: where do the girl and her dog sit ?
Ground truth: chair
2-CNN-LSTM: chair (0.4588)
CNN-BOW: bed (0.2899)
BLIND-LSTM: bed (0.2878)
BLIND-BOW: bed (0.2042)
Funny picture

Q38461: where is the lady making a funny face ?
Ground truth: restaurant
2-CNN-LSTM: restaurant (0.6185)
CNN-BOW: kitchen (0.3851)
BLIND-LSTM: mirror (0.1753)
BLIND-BOW: kitchen (0.2680)
It makes sense for the blind model to guess "mirror"

Q6428: where is the woman taking her picture ?
Ground truth: mirror
2-CNN-LSTM: bathroom (0.7403)
CNN-BOW: bathroom (0.4690)
BLIND-LSTM: mirror (0.8559)
BLIND-BOW: mirror (0.7299)
Answer could be bathroom as well but interesting to see the blind models choosing "mirror"

Q26714: where do orange flowers fill vases ?
Ground truth: shop
2-CNN-LSTM: vase (0.6365)
CNN-BOW: vase (0.9500)
BLIND-LSTM: vase (0.4533)
BLIND-BOW: vase (0.4836)
Many location based questions are noisy in the sense that there exist many valid answer

Q23637: where does the full view of some upright toilets ?
Ground truth: restroom
2-CNN-LSTM: bathroom (0.6114)
CNN-BOW: bathroom (0.5171)
BLIND-LSTM: bathroom (0.1748)
BLIND-BOW: room (0.1441)
Synonyms are also the noise of the dataset

Q6144: where does the yellow cat sit ?
Ground truth: sink
2-CNN-LSTM: sink (0.1054)
CNN-BOW: kitten (0.2359)
BLIND-LSTM: sink (0.3562)
BLIND-BOW: sink (0.1931)
As in the color case again the BOW fails to understand the question and directly uses image feature

Q172: where are the sink and toilet ?
Ground truth: bathroom
2-CNN-LSTM: bathroom (0.8315)
CNN-BOW: bathroom (0.6268)
BLIND-LSTM: bathroom (0.7750)
BLIND-BOW: bathroom (0.8103)
Some questions are fairly easy and can be answered without looking at the image

Q8887: where do delicate blue flowers sit ?
Ground truth: vase
2-CNN-LSTM: vase (0.9371)
CNN-BOW: vase (0.9150)
BLIND-LSTM: vase (0.9253)
BLIND-BOW: vase (0.8517)
Easy one

Figure 12.