Effect of Image Captioning with Description on the Working Memory

Nithiya Shree Uppara[1(✉)], Troy McDaniel[2], and Hemanth Venkateswara[1]

[1] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA
{nuppara,hemanthv}@asu.edu
[2] The Polytechnic School, Arizona State University, Mesa, AZ, USA
troy.mcdanie@asu.edu

Abstract. Working memory plays an important role in human activities across academic, professional, and social settings. Working memory is defined as the memory extensively involved in goal-directed behaviors in which information must be retained and manipulated to ensure successful task execution. The aim of this research is to understand the effect of image captioning with image description on an individual's working memory. A study was conducted with eight neutral images comprising situations relatable to daily life such that each image could have a positive or negative description associated with the outcome of the situation in the image. The study consisted of three rounds where the first and second round involved two parts and the third round consisted of one part. The image was captioned a total of five times across the entire study. The findings highlighted that only 25% of participants were able to recall the captions which they captioned for an image after a span of 9–15 days; when comparing the recall rate of the captions, 50% of participants were able to recall the image caption from the previous round in the present round; and out of the positive and negative description associated with the image, 65% of participants recalled the former description rather than the latter.

Keywords: Working memory · Image captioning · Sentiment analysis

1 Introduction

The quest to understand the human brain and the workings of human memory has intrigued philosophers and researchers for centuries. Memory is one of the most important aspects of what makes us human, and yet it is one of the most elusive and misunderstood of human faculties. Memory can be pictured as a small filing cabinet with separate memory folders where information is kept, or as a brain supercomputer with enormous capacity and speed [31]. To retrieve a memory from the past, different areas of the brain collaborate. For example, let's consider the act of driving a car which is recreated by the brain from many

C. Stephanidis et al. (Eds.): HCII 2021, LNCS 13096, pp. 107–120, 2021.
https://doi.org/10.1007/978-3-030-90328-2_7

different areas: the memory of how to get from the current location to the end of the block, the memory of how to operate the car, and the memory of driving the car while following the safety rules, which all come from different parts of the brain. Each memory element (sights, sounds, phrases, and emotions) is encoded in the same portion of the brain that created that fragment in the first place, and recalling a memory effectively reactivates the neural patterns that were established during the original encoding [31]. A lasting memory in the brain is created when all the different types of memory work together to form it. The popular Atkinson-Shiffrin model defines a 3 step model for memory including sensory memory, short-term memory or working memory, and long-term memory [2].

Working memory, in particular, has been a fascinating area of research since its introduction in the 1960s [5,15]. Various studies about memory in the fields of psychology, biology or neuroscience have not been able to completely outline a categorization of memory in terms of its functionality and mechanism [4,11,33]. Working memory has been gaining a lot of importance in mundane human activities such as in academic, professional and social settings [25]. To understand the basic definition of working memory, one must first understand the difference between long-term memory and short-term memory. Long-term memory is defined as a vast store of knowledge and a record of prior events [11]. Long-term memory capacity varies from situation to situation and from person to person. Short-term memory is the ability of the human mind to hold a finite amount of information in a very accessible state, temporarily [2]. The main difference between long-term memory and short-term memory is the duration of the situation of information stored and the capacity of the information stored [7]. The former has a huge capacity to retain information for a long duration and the latter is limited by the total number of chunks of information that can be stored at a time [11].

Working memory is not completely different from short-term memory. Working memory is defined as the memory extensively involved in goal-directed behaviors in which information must be retained and manipulated to ensure successful task execution [9]. Miller et al. [32] proposed the term working memory to refer to memory as it is used to plan and carry out behavior. An example of a common use of working memory is recalling partial calculations while solving a mathematical problem. The information stored during this process is stored only for that instance of time and is discarded from memory when the purpose is served. The factors related to the amount of time the information is stored change depending on the situation in which the information is perceived. Working memory assessments have been found to correlate with intellectual aptitudes (particularly fluid intelligence) better than short-term memory measures, and possibly better than assessments of any other psychological process [13,14,19,28,33].

One of the most important characteristics of working memory is its limited capacity [3,12]. Working memory capacity helps to predict fluid intelligence and attentional control [20,22]. For visual objects, this value has been estimated to be three or four visual objects [30,34,39,40]. Studies which examine visual

working memory by sequentially presenting items have shown that the information is either completely stored or entirely forgotten [27,37,41]. Many studies have shown that as the number of objects to be stored in working memory increases, the precision gradually decreases and it is worse for a sequential array of objects than a simultaneous array of objects [1,6,29].

An important question to examine is the effect of positive and negative information on visual working memory. Various studies have shown that emotional content increased the chances of retaining the information for a long period of time [8,17,24]. Various experiments conducted to examine the link between emotion and working memory by inducing a change in the mood of the participants have shown a change in cognitive task performance [18,23,38]. Spies et al. [38] and Cheng et al. [10] have demonstrated that negative mood hinders the performance on tests of problem solving, working memory and attention. This may be due to intrusive thoughts and worries which distract participants from the task at hand [21,36]. Individuals may be more likely to direct attention consciously toward emotional stimuli or to elaborate on emotional information because of its personal relevance [16,26]. Depending on the task at hand, having additional emotional stimuli can ease, if task-relevant information is processed, or weaken, if task-irrelevant information is processed, working memory capacity and performance of an individual. Perlstein et al. [35] have shown that emotional content has hindered the performance on working memory tasks.

In the proposed study we examine the effect of image captioning with description on the working memory of humans. We aim to understand the impact of positive and negative outcomes on working memory associated with a neutral image, and understand the impact on long-term memory as well. We hypothesize that positive descriptions will be retained for a longer period of time compared to negative descriptions associated with the outcome of an image. We would also like to understand if additional information associated with the image helps the participants in retaining the image captions for longer time in the working memory.

2 Experimental Setup and Methodology

2.1 Participants

The total number of participants enrolled in this IRB-approved study was 65 undergraduate and graduate students from Arizona State University between ages 17–27. All participants were well acquainted with the English language and had basic computer usage skills. We used data from 50 participants for the analysis after dropping records with missing entries.

2.2 Procedure

Eight neutral images were selected where each image could have a positive and negative outcome. All images comprise situations from everyday life. Each image

is associated with two descriptions: a positive description, which is the result of a positive outcome of the situation shown in the image, and a negative description, which is the result of a negative outcome of the situation shown in the image.

The study consisted of three rounds performed with a minimum gap of three days to a maximum gap of five days between rounds. The first round consisted of two parts. In the first part, an image was randomly selected from the set of eight images, and was displayed to the subject to be captioned. In the second part of the round, the same image was displayed but along with a positive or negative description. The participant captioned the image again after reading the description associated with it. In addition, a question was asked to see if the participant understood the description correctly. The number of positive and negative descriptions were kept equal. Round 2 also consisted of two parts. The only difference between the first and second round was that the description in the first round was chosen at random for each participant whereas in the second round, the description was opposite of the description displayed in the first round. For the third or final round, the participants were given the same images without any descriptions and were asked to caption the image. The time taken to complete each round was noted (Fig. 1).

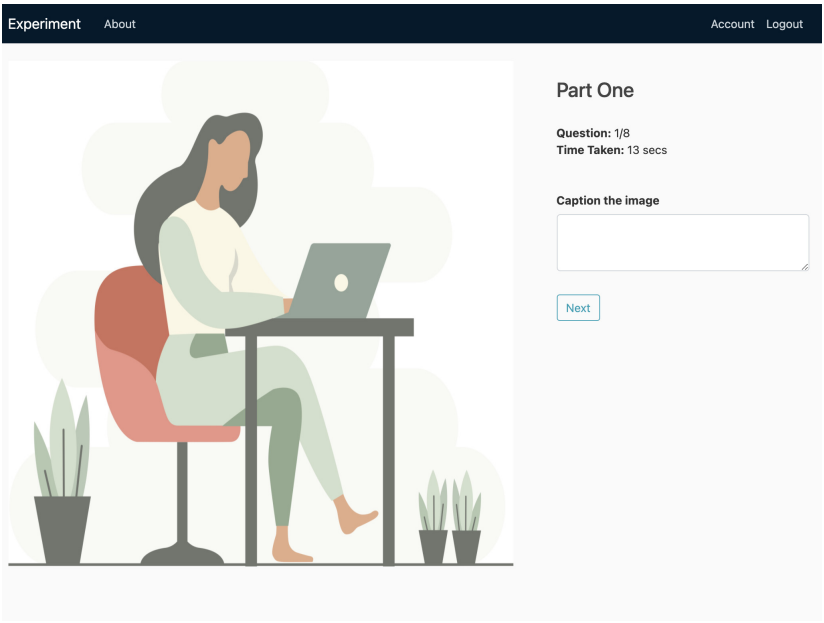

Fig. 1. Part-1 interface of Round-1 and Round-2 of the image captioning study.

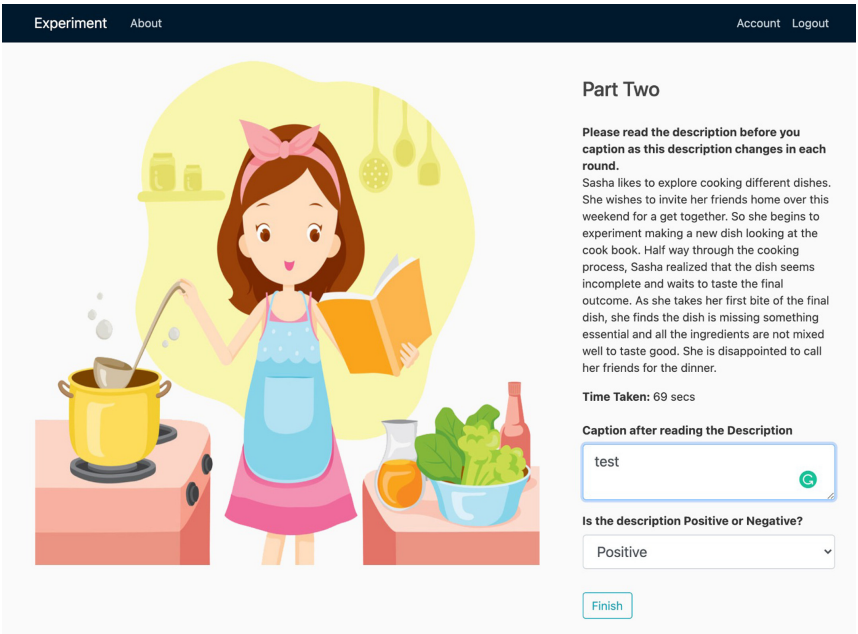

Fig. 2. Part-2 interface of Round-1 and Round-2 of the image captioning study.

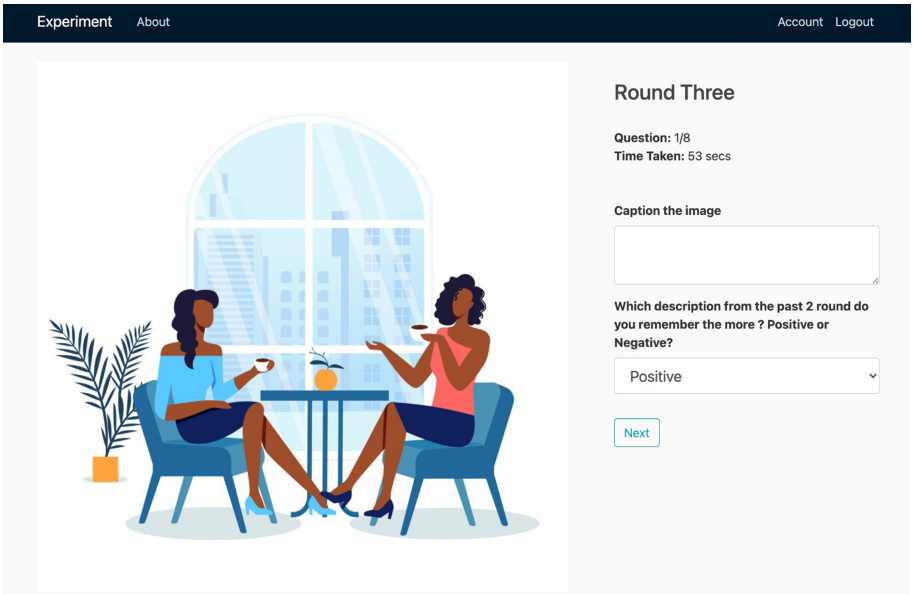

Fig. 3. Round-3 interface of the image captioning study.

2.3 Data Analysis

Each subject inputs 5 captions for each of the 8 images. The first caption is named Round-1 Part-1 (R1-P1), the second caption is named Round-1 Part-2 (R1-P2), the third caption is named Round-2 Part-1 (R2-P1), the fourth caption is named Round-2 Part-2 (R2-P2) and the fifth or final caption is named Round-3 (R3) (Figs. 2 and 3).

We propose to analyze the captions to test for similarity and to evaluate sentiment. It is possible that two captions from the same user are similar in context but differ in language. To account for such cases, we used the HuggingFace BERT Sequence Classification pre-trained model [42] to identify contextual similarity. We also used the HuggingFace BERT Sequence Classification pre-trained model to identify the sentiment of the caption.

The main purpose of conducting different rounds of the experiment with regular intervals of time is to observe the recall span and retention span of the image captions when additional information like description is provided with the image. The possible combinations for this purpose considered are the ability to remember captions from R1-P2 in R2-P1, from R2-P1 in R3 and from R1-P1 in R3.

3 Results

3.1 Round-1 Part-1 (R1-P1) vs. Round-3 (R3)

Figure 4 shows that only 25% of the captions, i.e., 101, were contextually the same, and 75% of the captions, i.e., 299, were contextually different for R1-P1 and R3. Figure 5 shows the trend in the number of the participants who captioned the image contextually the same and different for each image from the image dataset. It is interesting to note that even after looking at the same image five times in total, and captioning it four times before R3, 75% of participants could not recollect the first caption they used to caption the image.

3.2 Round-1 vs. Round-2 and Round-2 vs. Round-3

Round-1 (R1-P1 and R1-P2) vs. Round-2 Part-1 (R2-P1):
Figure 6 shows that of the participants who finished Round-1 (R1) and have seen the images twice, 36% of the participants, i.e., 143, captioned the image in R2-P1 contextually the same as in R1-P1; 13% of the participants, i.e., 54, captioned the image in R2-P1 contextually the same as in R1-P2; and the rest, 51%, of the participants, i.e., 203, captioned the image differently from the previous round. Figure 7 shows the trend in the number of the participants who captioned the image in R3 contextually the same as R1-P1, R1-P2, and the rest different for each image from the image dataset.

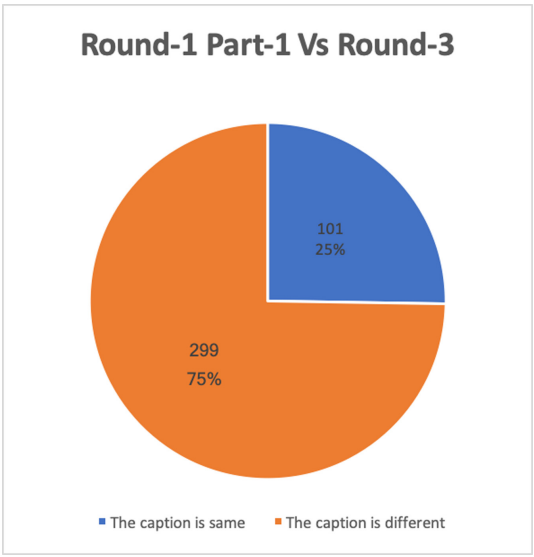

Fig. 4. Distribution of image captions from R1-P1 which are contextually the same as R3.

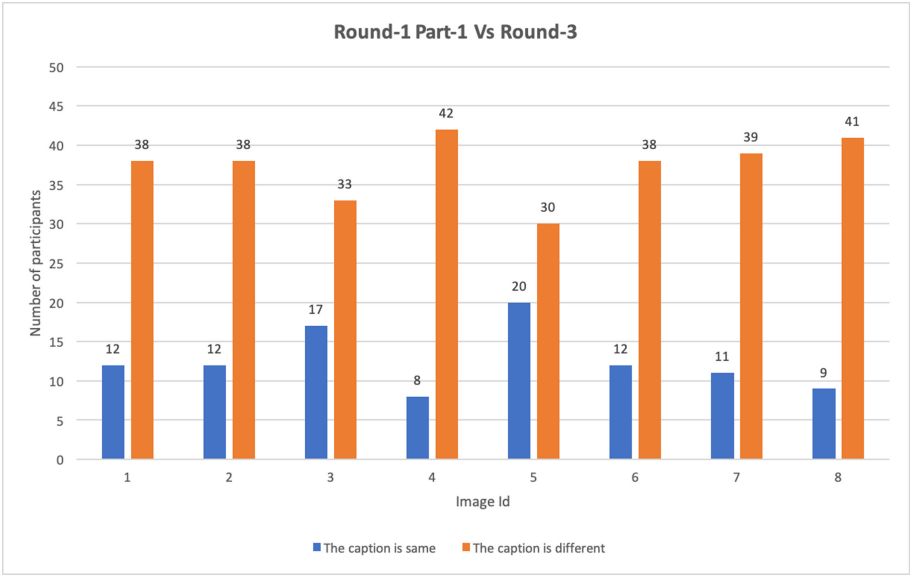

Fig. 5. Trend in the number of participants who captioned the image in R3 contextually the same as R1-P1 and different for each image from the image dataset.

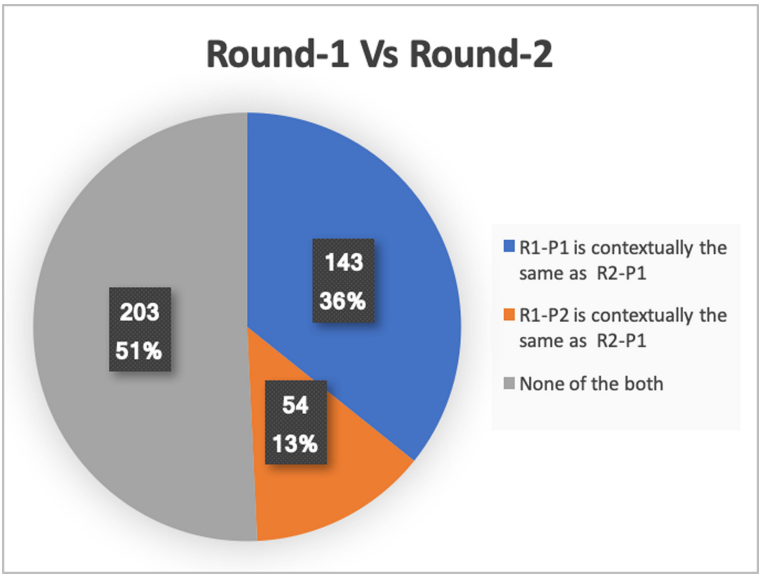

Fig. 6. Distribution of image captions from R1-P1 and R1-P2 which are contextually the same as R2-P1.

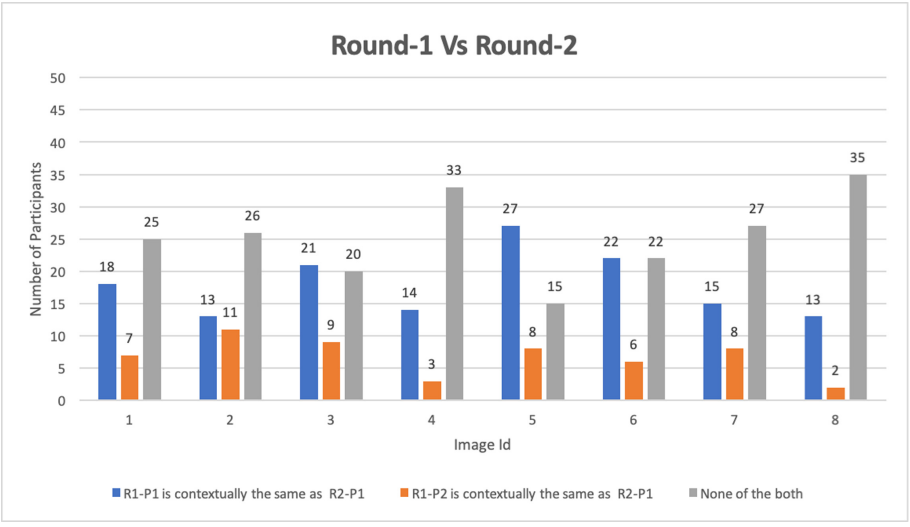

Fig. 7. Trend in the number of participants who captioned the image in R2 contextually the same as R1-P1, R1-P2 and not the same.

Round-2 (R2-P1 and R2-P2) vs. Round-3 (R3):
Figure 8 shows that of the participants who completed Round-2 (R2) and have seen the images four times, 37% of the participants, i.e., 146, captioned the image in R3 contextually the same as in R2-P1; 15% of participants, i.e., 61, captioned the image in R3 contextually the same as R2-P2; and the rest, 49%, of participants, i.e., 193, captioned the image differently from the previous round. Figure 9 shows the trend in the number of participants who captioned the image in R3 contextually the same as in R2-P1, R2-P2, and none of both, i.e., the rest are different for each image from the image dataset.

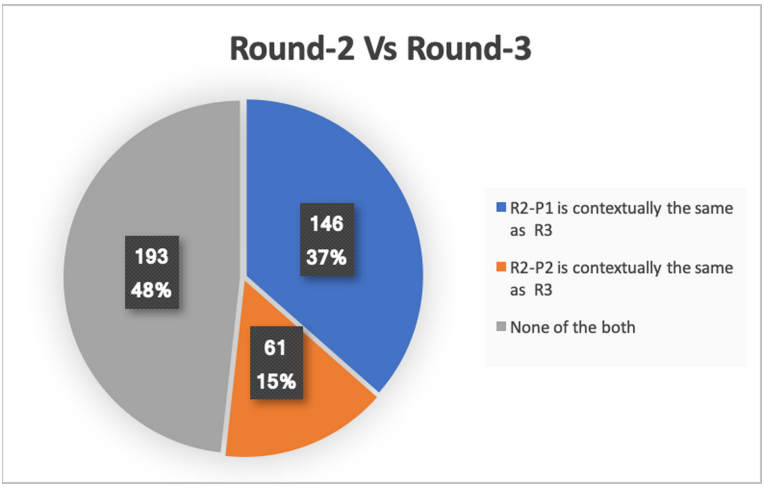

Fig. 8. Distribution of image captions from R2-P1 and R2-P2 which are contextually the same as R3.

An interesting observation is that even after seeing the image with extra information like the description associated with the image in R1 and R2, the majority of participants tend to remember the image caption they captioned in R1-P1 and R2-P2 respectively.

3.3 Trends in Round-1 Part-2 and Round-2 Part-2

Figure 10 shows that among the total number of participants whose image caption from R1-P2 is contextually the same as R1-P1, 67% of participants, i.e., 36, captioned the image with respect to the positive description associated to it, and 33% of participants, i.e., 18, captioned the image with respect to the negative description associated to it.

Figure 11 shows that among the total number of participants whose caption from R2-P2 is contextually the same as in R3, 56% of the participants, i.e., 34, captioned the image with respect to the positive description associated to it,

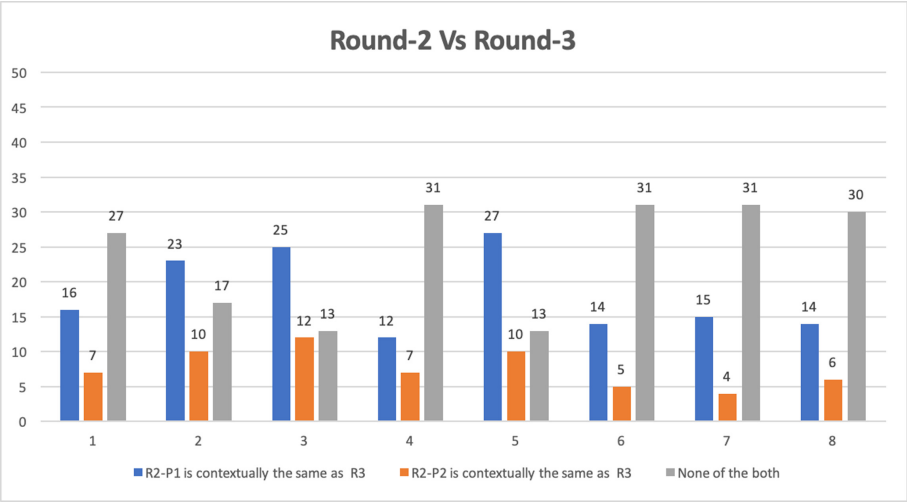

Fig. 9. Trend in the number of the participants who captioned the image in R3 contextually the same as R2-P1, R2-P2 and not the same.

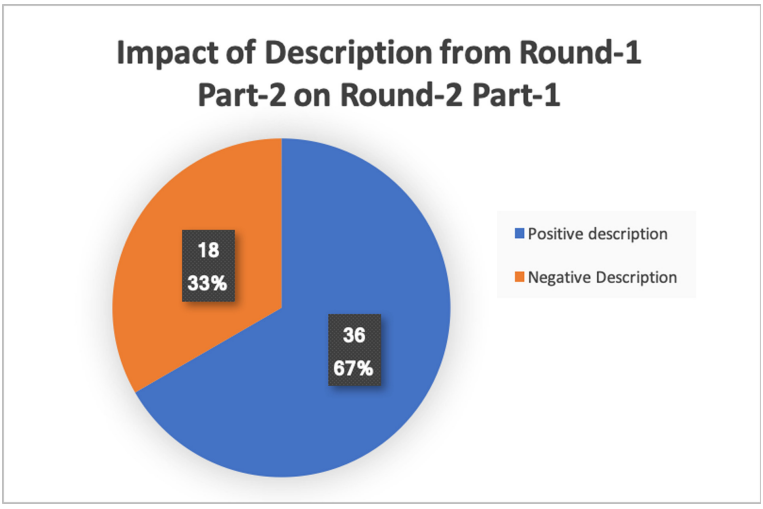

Fig. 10. Impact of positive and negative description from R1-P2 on R2-P1.

and 44% of the participants, i.e., 27, captioned the image with respect to the negative description associated to it.

It is interesting to note that out of the captions remembered by participants from R1-P2 and R2-P2, positive descriptions tend to have more impact on participants, causing them to remember the image caption for a longer duration compared to negative descriptions.

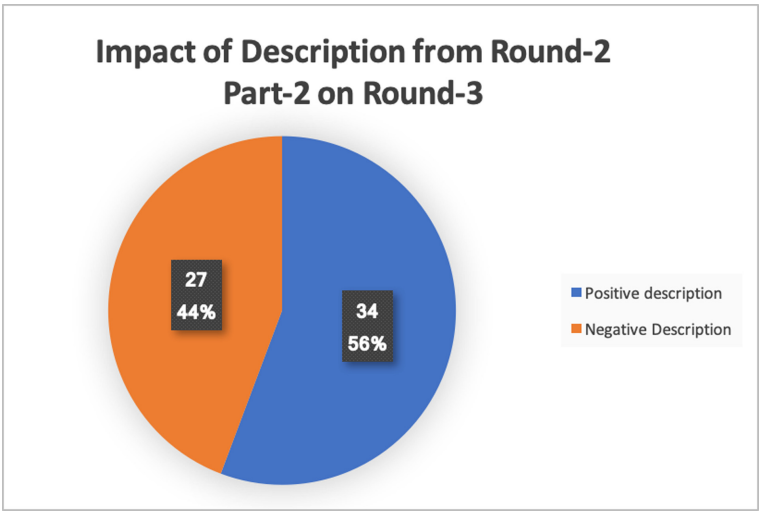

Fig. 11. Impact of positive and negative description from R2-P2 on R3.

4 Discussion

Only 25% of participants were able to recall the captions which they captioned for an image after a span of 9–15 days. It is interesting that even though the visual working memory capacity of a human is considered to be three to four objects, participants tended to retain some of the information for 9–15 days. This may be due to the fact that some participants were able to relate the situations from the images leading them to correlate the image with one or more experiences from their past, consistent with [16,26]. Due to this, even though the images were of no purpose to them, they tended to remember the captions for a long period of time, posing an interesting question to examine whether the image captions were saved to working memory or long-term memory. One other possible reason for retaining the image caption would be due to the additional information, i.e., the description, provided with the images. If providing description is a potential reason for participants to retain the information, it is fascinating to note that if the hypothesized reason behind remembering the image caption is that of the description, participants tended to recall the image caption which was captioned without the description.

When comparing the recall rate of the captions between the first and second rounds, and the second and third rounds, more than 50% of participants were able to recall the image caption from previous rounds. Out of the 50%, an average of 36% of the captions recalled were the image captions which were captioned without seeing the descriptions. This helps to understand that even after seeing extra description related to a given image, the first impression of the image made on participants has more impact and a higher chance to be retained in the working memory than the caption which had been captioned after seeing

the description. This also leads to the question that given an image without any description, why is it easy for a human to perceive the image than the description associated with it and relate it.

The primary purpose of using two different descriptions for an image was to understand the impact of the sentiment of the description on image captioning. As hypothesized, out of the participants who recalled the image caption with description in R2 and R3, an average of 60% of participants remembered the caption associated with the positive description rather than the negative description. We may conclude that given two outcomes, one positive and the other negative, the human brain on average tends to remember and retain the positive information corresponding to the situation rather than the negative information. This also leads to an interesting question that the working memory capacity of a human tends to change with the sentiment of the objects associated with it, which is consistent with [8,17,24,35].

5 Conclusions

The findings from the results highlight that participants tend to retain information for longer periods than the expected duration for working memory, which may be because participants were able to relate the images with their everyday life scenarios. Figure 10 and Fig. 11 give insight that the positive description enabled participants to retain and recall more information than the negative description associated with the image. The inferences from this study are limited due to there being no evidence of the mood of each participant while participating in the study. Even though there are some limitations to this study, the results contribute to the growing research on working memory.

References

1. Allen, R., Baddeley, A., Hitch, G.: Is the binding of visual features in working memory resource-demanding? J. Exp. Psychol. Gen. **135**, 298–313 (2006). https://doi.org/10.1037/0096-3445.135.2.298
2. Atkinson, R.C., Shiffrin, R.M.: Human memory: A proposed system and its control processes. In: Psychology of Learning and Motivation, vol. 2, pp. 89–195. Elsevier (1968)
3. Baddeley, A.: Working memory: looking back and looking forward. Nat. Rev. Neurosci. **4**, 829–839 (2003). https://doi.org/10.1038/nrn1201
4. Baddeley, A.: Working memory. Curr. Biol. **20**(4), R136–R140 (2010)
5. Baddeley, A.: Working memory: theories, models, and controversies. Ann. Rev. Psychol. **63**(1), 1–29 (2012). https://doi.org/10.1146/annurev-psych-120710-100422. pMID: 21961947
6. Blalock, L., Clegg, B.: Encoding and representation of simultaneous and sequential arrays in visuospatial working memory. Q. J. Exp. Psychol. **2006**(63), 856–62 (2010). https://doi.org/10.1080/17470211003690680
7. Broadbent, D.E.: Percept. Commun. Pergamon Press, New York (1958)

8. Buchanan, T., Adolphs, R.: The role of the human amygdala in emotional modulation of long-term declarative memory. Adv. Cons. Res. **44**, 9–34 (2002). https://doi.org/10.1075/aicr.44.02buc
9. Chai, W.J., Abd Hamid, A.I., Abdullah, J.M.: Working memory from the psychological and neurosciences perspectives: a review. Front. Psychol. **9**, 401 (2018)
10. Cheng, P., Holyoak, K.: Pragmatic reasoning schemas. Cogn. Psychol. **17**, 391–416 (1985). https://doi.org/10.1016/0010-0285(85)90014-3
11. Cowan, N.: What are the differences between long-term, short-term, and working memory? Prog. Brain Res. **169**, 323–38 (2008)
12. Cowan, N.: The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav. Brain Sci. **24**(1), 87–114 (2001). https://doi.org/10.1017/S0140525X01003922
13. Daneman, M., Carpenter, P.A.: Individual differences in working memory and reading. J. Verbal Learn. Verbal Behav. **19**(4), 450–466 (1980)
14. Daneman, M., Merikle, P.: Working memory and language comprehension: a meta-analysis. Psychon. Bull. Rev. **3**, 422–433 (1996). https://doi.org/10.3758/BF03214546
15. D'Esposito, M., Postle, B.R.: The cognitive neuroscience of working memory. Ann. Rev. Psychol. **66**(1), 115–142 (2015). https://doi.org/10.1146/annurev-psych-010814-015031. pMID: 25251486
16. Doerksen, S., Shimamura, A.: Source memory enhancement for emotional words. Emotion **1**, 5–11 (2001). https://doi.org/10.1037//1528-3542.1.1.5. (Washington, D.C.)
17. Dolan, R.: Emotion, cognition, and behavior. Science **298**, 1191–1194 (2002). https://doi.org/10.1126/science.1076358. (New York, NY)
18. Elliman, N., Green, M., Rogers, P., Finch, G.: Processing-efficiency theory and the working-memory system: Impairments associated with sub-clinical anxiety. Pers. Individ. Differ. **23**, 31–35 (1997). https://doi.org/10.1016/S0191-8869(97)00016-0
19. Engle, R., Tuholski, S.W., Laughlin, J., Conway, A.: Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. J. Exp. Psychol. Gen. **128**(3), 309–331 (1999)
20. Engle, R., Tuholski, S., Laughlin, J., Conway, A.: Working memory, short-term memory and general fluid intelligence: a latent variable approach. J. Exp. Psychol. Gen. **130**, 169–183 (1999). https://doi.org/10.1037/0096-3445.128.3.309
21. Eysenck, M., Calvo, M.: Anxiety and performance: the processing efficiency theory. Cogn. Emot. **6**, 409–434 (1992). https://doi.org/10.1080/02699939208409696
22. Fukuda, K., Vogel, E., Mayr, U., Awh, E.: Quantity, not quality: the relationship between fluid intelligence and working memory capacity. Psychon. Bull. Rev. **17**, 673–679 (2010). https://doi.org/10.3758/17.5.673
23. Gray, J.: Emotional modulation of cognitive control: approach-withdrawal states double dissociate spatial from verbal 2-back task performance. J. Exp. Psychol. Gen. **130**, 436–52 (2001). https://doi.org/10.1037/0096-3445.130.3.436
24. Hamann, S.: Cognitive and neural mechanisms of emotional memory. Trends Cogn. Sci. **5**, 394–400 (2001). https://doi.org/10.1016/S1364-6613(00)01707-1
25. Harden, L.: A review of research on working memory and its importance in education of the deaf. Ph.D. thesis, Program in Audiology and Communication Sciences, Washington University (2011). http://digitalcommons.wustl.edu/pacs_capstones/627
26. Heuer, F., Reisberg, D.: Vivid memories of emotional events: the accuracy of remembered minutiae. Mem. Cogn. **18**, 496–506 (1990). https://doi.org/10.3758/BF03198482

27. Johnson, A., Miles, C.: Serial position effects in 2-alternative forced choice recognition: functional equivalence across visual and auditory modalities. Memory **17**, 84–91 (2008). https://doi.org/10.1080/09658210802557711. (Hove, England)
28. Kyllonen, P.C., Christal, R.E.: Reasoning ability is (little more than) working-memory capacity?! Intelligence **14**(4), 389–433 (1990)
29. Lecerf, T., de Ribaupierre, A.: Recognition in a visuospatial memory task: the effect of presentation. Eur. J. Cogn. Psychol. **17**, 47–75 (2005). https://doi.org/10.1080/09541440340000420
30. Luck, S., Vogel, E.: The capacity of visual working memory for features and conjunctions. Nature **390**, 279–81 (1997). https://doi.org/10.1038/36846
31. MacDonald, M.: Your Brain: The Missing Manual: The Missing Manual. O'Reilly Media, Sebastopol (2008)
32. Miller, G., Galanter, E., Pribram, K.: Plans and the structure of behavior. Am. J. Psychol. **75** (1960). https://doi.org/10.2307/1419559
33. Oberauer, K., Cowan, N.: Working memory capacity: (2005). Exp. Psychol. **54**, 245–246 (2007). https://doi.org/10.1027/1618-3169.54.3.245
34. Pashler, H.: Familiarity and visual change detection. Percept. Psychophys. **44**, 369–78 (1988)
35. Perlstein, W., Elbert, T., Stenger, V.: Dissociation in human prefrontal cortex of affective influences on working memory-related activity. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 99, pp. 1736–41, March 2002. https://doi.org/10.1073/pnas.241650598
36. Seibert, P., Ellis, H.: Irrelevant thoughts, emotional mood states, and cognitive task performance. Mem. Cogn. **19**, 507–13 (1991). https://doi.org/10.3758/BF03199574
37. Smyth, M., Hay, D., Hitch, G., Horton, N.: Serial position memory in the visual-spatial domain: reconstructing sequences of unfamiliar faces. Q. J. Exp. Psychol. Hum. Exp. Psychol. **58**, 909–30 (2005). https://doi.org/10.1080/02724980443000412
38. Spies, K., Hesse, F., Hummitzsch, C.: Mood and capacity in baddeley's model of human memory. Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie **204**, 367–381 (1996)
39. Vogel, E., Woodman, G., Luck, S.: Storage of features, conjunctions, and objects in visual working memory. J. Exp. Psychol. Hum. Percept. Perform. **27**, 92–114 (2001). https://doi.org/10.1037//0096-1523.27.1.92
40. William, P., Phillips, W.A.: on the distinction between sensory storage and short-term visual memory. Percept. Psychophys. **16**, 283–290 (1974). https://doi.org/10.3758/BF03203943
41. William, P., Christie, D.: Components of visual memory. Q. J. Exp. Psychol. **29**, 117–133 (1977). https://doi.org/10.1080/00335557743000080
42. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online, October 2020. https://www.aclweb.org/anthology/2020.emnlp-demos.6