Reproduction of Original Dissertation Results

This document is meant to serve as a sanity check for initial results produced by the primary designer of Hawk. The primary objective is to discuss any differences in such results and explain any code changes that may have affected overall performance and statistical details of the tables therein. The following sections each contain two tables of results, one for each Scout-Home bandwidth setting. All tables on the right are the original Hawk results from S̃ummer of 2022 and on the left are from August 2023. Figure 1 above depicts

Model Gen	New Model Installed (sec)	Pos Discovered	Total Positives	Transmitted (pos + neg)	Processed Tiles
1	250 (0)	11(1)	25 (0)	56 (0)	16023
2	430 (0)	20(1)	41 (0)	89 (1)	27362 (2)
3	730 (0)	30 (0)	64 (0)	168 (0)	46218 (75)
4	1111 (35)	50 (2)	91(3)	275 (5)	70183 (2197)
5	1671 (139)	76 (6)	138(10)	420 (41)	105464 (8742)
6	2171 (208)	102 (4)	177(16)	545 (60)	136950 (13128)
7	2912 (69)	147 (5)	237(8)	736 (22)	183542 (4393)
8	3612 (121)	197 (3)	306(8)	915 (37)	227592 (7651)
End	4033 (0)	220 (5)	336 (0)	1029 (2)	252192 (48)

Gen	New model installed (seconds)	Positives discovered so far	Total Positives so far	Tiles transmitted so far	Tiles processed so far
0	0	0	0	0	0
1	277 (54)	8 (1)	19 (3)	47 (16)	17759 (713)
2	530 (59)	19 (1)	39 (4)	103 (16)	32797 (556)
3	768 (55)	28 (3)	59 (4)	152 (17)	42859 (624)
4	1168 (115)	50 (2)	95 (5)	241 (34)	65116 (703)
5	1579 (63)	72 (2)	126 (8)	335 (12)	94186 (433)
6	2201 (57)	97 (5)	165 (5)	455 (15)	124783 (542)
7	3010 (44)	137 (3)	218 (5)	604 (14)	166576 (366)
8	3529 (97)	187 (5)	283 (8)	772 (31)	207074 (676)
	4149 (32)	219 (6)	336 (0)	923 (3)	252231 (0)

Figure 1: Results Comparison for 12k Bandwidth Setting

the results for the reproduced (left) and original (right) results. All data points represent "mean (standard deviation)" in that format. These tables were generated from experiments with a bandwidth setting of 12 kbps between each scout and home. This is the most conservative bandwidth setting, and thus results in the lowest number of true positive samples discovered. The following are the key observations and explanations for significant differences between these two tables:

- The most obvious difference in the 12k setting is that the new results had roughly 100 more tiles transmitted than did the original. This is because the original used the "tile retriever" option, which groups together individual tiles of the same image and assigns them to the same scout prior to the mission. This is more realistic in that if part of an object (roundabout in this case) is visible in one tile, it is likely to also be present in another adjacent tile of the same image. So while true positive tiles are more likely to be clustered together, it's also likely that many tiles with no true positive tiles are clustered together. This is why the original results also had a lower number of transmitted tiles because with tile retriever, the number of tiles across the 7 scouts are not as evenly distributed as in random retriever. With random retriever, all tiles derived from any image are put in a large pool prior to the mission and are evenly distributed (randomly) across the scouts. This approach produces a more balanced and IID data distribution where each scout processes such a similar number of tiles throughout the mission that each also transmit the exact same number of tiles.

- A small bug for which we have yet to determine the source is at the end of a mission after all tiles have been retrieved. It may be the case that a single variable is not being incremented properly due to some flag that is being set upon the retriever finishing. This prevents the number on the bottom right of the table (total number of processed objects) to (occasionally) not reach 252231.

- Larger standard deviation on new results: The second column from the left represents the time when each new model is installed and inference on streaming data resumes. The times in the new results vary from the original's somewhat because the original approach was to average the exact times of all scouts when each model was installed whereas the new results recorded the time of the first log file written at home after the first scout reported installing its new model. These differences in times caused a significantly larger standard deviation of the number of processed tiles (in the new (left) table) and other fields to a lesser degree as the numbers were thus recorded at different time snapshots in separate missions.

- An actual code change that was made had a positive result. Each scout transmits the top k results (wwhen using the top k selector) after processing a batch (1000 - but is tunable), therefore if the total

number of tiles processed by each scout is not a multiple of the batch size, there are tiles processed just before the end of the mission yet the scouts are unable to transmit another top k samples. With the change, the scouts will transmit this last group of top k tiles processed after it has processed (inferenced) all of its tiles. This resulted in an increase of 4-6 true positives received at home.

Model Gen	New Model Installed (sec)	Pos Discovered	Total Positives	Transmitted (pos + neg)	Processed Tiles
1	190 (0)	7 (0)	19 (0)	70 (0)	12237 (1)
2	430 (0)	22 (2)	41 (0)	221 (2)	27363 (0)
3	590 (69)	37 (3)	52 (7)	366 (47)	37442 (4264)
4	930 (70)	56 (3)	77 (3)	563 (63)	58840 (4403)
5	1431 (69)	79 (3)	120 (6)	882 (64)	90363 (4365)
6	2051 (208)	115 (6)	168 (15)	1282 (157)	129401 (13089)
7	2752 (69)	160 (8)	221 (6)	1735 (34)	173468 (4273)
8	3472 (35)	223 (10)	296 (2)	2207 (33)	218808 (2120)
End	4073 (35)	255 (13)	336 (0)	2581 (2)	252221 (16)

Gen	New model installed (seconds)	Positives discovered so far	Total Positives so far	Tiles transmitted so far	Tiles processed so far
0	0	0	0	0	0
1	244 (49)	10 (2)	25 (2)	90 (35)	14502 (156)
2	413 (87)	20 (1)	43 (1)	160 (10)	21967 (240)
3	712 (105)	32 (2)	55 (5)	313 (65)	45424 (2161)
4	1063 (60)	48 (5)	90 (4)	540 (111)	68821 (1891)
5	1510 (104)	68 (2)	118 (7)	730 (159)	88620 (1086)
6	2264 (84)	99 (3)	156 (4)	1140 (294)	140508 (1378)
7	2884 (50)	135 (2)	207 (4)	1603 (185)	183283 (296)
8	3456 (125)	188 (3)	266 (2)	2063 (99)	221766 (1568)
	4143 (31)	235 (7)	336 (0)	1451 (44)	252231 (0)

Figure 2: Results Comparison for 30k Bandwidth Setting

- Figure 2 above represents results for 30k transmissions bandwidth in comparison to the previous 12k tables, the use of random retriever yielded a larger performance gain over tile retriever (255 to 235) of total true positives found. This is because when tiles are more evenly distributed across scouts and within each scout in time throughout the mission, TP tiles do not have to compete with each other as much for sufficient prioritization for transmission and thus more true positive tiles are found. The additional tiles transmitted by the end of the mission is also mostly accounted for by the final top k transmission at the end of mission described below the 12k tables.

Model Gen	New Model Installed (sec)	Pos Discovered	Total Positives	Transmitted (pos + neg)	Processed Tiles
1	190	9	19	210	12239(1)
2	390(69)	27(9)	37(6)	589(146)	24832(4384)
3	590(69)	41(1)	52(7)	1119(122)	37563(4573)
4	890(69)	59(4)	76(3)	1586(172)	56314(4333)
5	1331(159)	86(5)	110(15)	2483(300)	84017(9977)
6	1871(240)	125(4)	155(18)	3564(429)	118082(15121)
7	2652(317)	172(12)	214(25)	5040(556)	167200(1997)
8	3472(296)	240(18)	290(29)	6576(600)	217564(17545)
End	4093(60)	283(9)	336	7742(13)	252230(1)

Gen	New model installed (seconds)	Positives discovered so far	Total Positives so far	Tiles transmitted so far	Tiles processed so far
0	0	0	0	0	0
1	208 (11)	9 (0)	24 (2)	230 (173)	14824 (567)
2	374 (12)	24 (2)	40 (4)	520 (104)	21970 (875)
3	554 (14)	31 (4)	52 (4)	740 (121)	33633 (719)
4	1002 (61)	52 (3)	90 (7)	1570 (221)	64918 (1860)
5	1452 (68)	70 (8)	113 (5)	2453 (72)	89827 (2537)
6	2192 (42)	100 (3)	153 (2)	3674 (247)	136950 (3434)
7	2795 (57)	110 (4)	204 (2)	5207 (267)	179070 (1880)
8	3518 (60)	193 (7)	269 (6)	6235 (175)	219442 (2786)
	4147 (41)	246 (8)	336 (0)	7501 (128)	252231 (0)

Figure 3: Results Comparison for 100k Bandwidth Setting

- Similar to the comment on the 30k slide, the use of random retriever yielded an even greater performance gain over tile retriever as can be seen in Figure 3 (283 to 246) of total true positives found (over both 12k and 30k performance gains). The additional tiles transmitted by the end of the mission is also mostly accounted for by the final top k transmission at the end of mission described on the previous slide. Overall, using the random retriever results in much better performance when leveraging additional bandwidth and a larger k value for top k transmission.