A Short Note on the Kinetics-700-2020 Human Action Dataset

Lucas Smaira
lsmaira@google.com

João Carreira
joaoluis@google.com

Eric Noland
enoland@google.com

Ellen Clancy
clancye@google.com

Amy Wu
amybwu@google.com

Andrew Zisserman
zisserman@google.com

| Dataset       | # classes | Average | Minimum |
|---------------|-----------|---------|---------|
| Kinetics-400  | 400       | 683     | 303     |
| Kinetics-600  | 600       | 762     | 519     |
| Kinetics-700  | 700       | 906     | 532     |
| Kinetics-700-2020 | 700 | 926     | 705     |

Table 1: Statistics on the number of video clips per class for different Kinetics datasets as of 14-10-2020.

Abstract

We describe the 2020 edition of the DeepMind Kinetics human action dataset, which replenishes and extends the Kinetics-700 dataset. In this new version, there are at least 700 video clips from different YouTube videos for each of the 700 classes. This paper details the changes introduced for this new release of the dataset and includes a comprehensive set of statistics as well as baseline results using the I3D network.

1. Introduction

The Kinetics datasets are a series of large scale curated datasets of video clips, covering a diverse range of human actions. They can be used for training and exploring neural network architectures for modelling human actions in video.

Three editions have been released: Kinetics-400 [6], Kinetics-600 [1] and Kinetics-700 [2], with 400, 600 and 700 human action classes, respectively. In each case: (i) the clips are from YouTube videos, last 10s, and have a variable resolution and frame rate; and (ii) for an action class, all clips are from different YouTube videos. The statistics of the datasets are given in table 1.

Building datasets of realistic videos from YouTube presents the challenge of dealing with video disappearance – for example, due to users removing the videos or making them private. The scale of this problem is illustrated in table 2. To address this problem, we have released a new edition of the Kinetics-700 dataset, called Kinetics-700-2020, where the clips for each class have been replenished. Note, unlike in previous years we have not increased the number of classes.

The URLs of the YouTube videos and temporal intervals of all the Kinetics datasets can be obtained from https://deepmind.com/research/open-source/kinetics. The link also includes additional annotations for the A V A-Kinetics [7] and Countix [4] datasets.

2. Data Collection Process

The collection process follows that described in [2] but focuses only on the rare classes. We collect new clips for the 123 rarest classes (containing less than 700 clips), topping up those until they reach at least 700 per class. This is shown in table 1. We also show yields for these classes in Appendix A.

Since rare classes have a poor yield rate (proportion of candidate clips which are rated positive), we increased the number and quality of the text queries used to collect candidate YouTube video ids by techniques such as: using verbs in both infinitive and gerund format; removing stop words and articles; and using synonyms. The same procedure was carried out in all four query languages (English, French, Spanish and Portuguese). Augmenting the query space proved successful in helping to obtain more and better quality videos (with content more related to the class).

Removing duplicates. The same clip can occur multiple times. This happens because: (i) the same video is uploaded multiple times to YouTube; or (ii) different videos contain the same clip (e.g. compilations). This is common in instructional videos, particularly in classes such as `pour-
| Dataset & split       | # clips | # clips 14-10-2020 | % retained |
|----------------------|---------|-------------------|------------|
| Kinetics-400 train   | 246,245 | 220,033           | 89%        |
| Kinetics-400 val     | 20,000  | 18,059            | 90%        |
| Kinetics-400 test    | 40,000  | 35,400            | 89%        |
| Kinetics-600 train   | 392,622 | 371,910           | 95%        |
| Kinetics-600 val     | 30,000  | 28,366            | 95%        |
| Kinetics-600 test    | 60,000  | 56,703            | 95%        |
| Kinetics-700 train   | 545,317 | 532,370           | 98%        |
| Kinetics-700 val     | 35,000  | 34,056            | 97%        |
| Kinetics-700 test    | 70,000  | 67,302            | 96%        |
| Kinetics-700-2020 train | 545,793 | –                | –          |
| Kinetics-700-2020 val | 34,256  | –                | –          |
| Kinetics-700-2020 test | 67,858  | –                | –          |

Table 2: The number of original (left) and current (right) available video clips in the various Kinetics datasets.

ing milk’, ‘tasting wine’, ‘vacuuming car’. In order to filter those clips from the final dataset, we cluster them and look at individual clusters gifs removing duplicates. A final filtering is also done to make sure clips belong to the correct class.

Geographical diversity. We provide an analysis of the geographical distribution of the videos in the final dataset at the granularity of continents. The location is assigned based on where the video was uploaded from. The results are shown in table 3 based on the fraction of videos containing that information (around 90%).

Geographical diversity increased slightly over the years, especially the percentage of videos from Latin America, probably because we started querying for videos in Portuguese in the Kinetics-600 edition and also Spanish in the Kinetics-700 edition. The multiple language queries were introduced to increase diversity and yield. Overall, still more than half of the videos were uploaded from North America, possibly because of querying in English from the start (with Kinetics-400) but maybe also due to the greater popularity of YouTube in North America.

### 3. Benchmark Performance

As a baseline model we used I3D [3], with standard RGB videos as input (no optical flow). We trained the model from scratch on the Kinetics-700-2020 training set using different numbers of training examples: 100, 200, 300, 400, 500, 600 and all (some classes have up to 1000 training examples). We report performance on the validation and test sets. Results are shown in table 4.

Top-1 and top-5 accuracy improve steadily with more examples per class, even given that I3D is a model with few parameters: around 12M. In contrast, for example, a ResNet-50 model [5] has nearly double the parameters at

![Figure 1: Performance of an I3D model with RGB inputs on the Kinetics-700-2020 dataset using different number of training examples and and evaluating using 8 linearly spaced segments per clip.](image)

### 4. Conclusion

We have described the new Kinetics-700-2020 dataset, which in terms of clip counts is considerably more balanced than the current Kinetics-700 with all classes now having a minimum of 700 examples. We have also demonstrated the benefits of having more training clips in improving I3D classification performance. The Kinetics datasets were originally introduced to aid architectural development
| Continent     | Kinetics-400 | Kinetics-600 | Kinetics-700 | Kinetics-700-2020 |
|---------------|--------------|--------------|--------------|-------------------|
| Africa        | 0.8%         | 0.9%         | 1.0%         | 1.0%              |
| Asia          | 11.8%        | 11.3%        | 11.5%        | 11.7%             |
| Europe        | 21.4%        | 19.3%        | 19.6%        | 19.5%             |
| Latin America | 3.4%         | 5.7%         | 7.6%         | 7.7%              |
| North America | 59.0%        | 59.1%        | 56.8%        | 56.6%             |
| Oceania       | 0.8%         | 3.7%         | 3.5%         | 3.5%              |

Table 3: Geographical data distribution, per continent.

| # train examples | Valid     | Test      |
|------------------|-----------|-----------|
| 100              | 38.8/63.0 | 36.9/61.1 |
| 200              | 48.6/72.4 | 46.8/70.9 |
| 300              | 52.4/76.0 | 50.8/74.6 |
| 400              | 54.1/77.6 | 52.6/76.0 |
| 500              | 55.7/79.0 | 54.0/77.7 |
| 600              | 58.1/81.1 | 56.8/79.9 |
| Kinetics-700-2020 | 59.3/82.0 | 58.2/80.9 |
| Kinetics-700     | 58.0/81.7 | 57.6/80.7 |

Table 4: Performance of an I3D model with RGB inputs on the Kinetics-700-2020 dataset valid and test set using different number of training examples and evaluating in 8 regularly spaced clips. Each row shows top-1 / top-5 accuracy in percentage.

A. Yield success rate per class

This is the ranked list of classes to which new clips have been added, where the first number is the probability that a candidate clip was voted positive for that class by three or more human annotators and the second number indicates a probability that an example is published in the dataset, after deduplication and the final filtering.

1. stacking dice 38.68% 37.93%
2. steering car 65.90% 35.44%
3. putting on sari 48.46% 32.54%
4. punching person (boxing) 30.81% 30.81%
5. steer roping 38.29% 30.79%
6. making slime 37.83% 27.71%
7. filling eyebrows 36.47% 27.26%
8. washing hair 34.15% 26.83%
9. square dancing 46.13% 27.11%
10. scrapbooking 35.94% 25.46%
11. jumping sofa 24.85% 24.49%
12. threading needle 30.63% 24.32%
13. brushing floor 31.14% 23.06%
14. eating nachos 33.97% 22.73%
15. playing with trains 46.50% 22.72%
16. metal detecting 25.96% 22.12%
17. using atm 25.19% 21.91%
18. grinding meat 28.40% 20.99%

Acknowledgements:

The collection of this dataset was funded by DeepMind.

References

[1] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about Kinetics-600. arXiv preprint arXiv:1808.01340, 2018.
[2] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the Kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987, 2019.
[3] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? New Models and the Kinetics Dataset. In IEEE International Conference on Computer Vision and Pattern Recognition CVPR, 2017.
[4] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, 2016.
[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Sulleyman, and A. Zisserman. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.
[7] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv preprint arXiv:2005.00214, 2020.
|   | Activity                                      | Rating1 | Rating2 |
|---|----------------------------------------------|---------|---------|
| 19. | base jumping                                | 30.65%  | 20.69%  |
| 20. | springboard diving                          | 32.88%  | 20.55%  |
| 21. | tie dying                                    | 23.53%  | 19.61%  |
| 22. | luge                                        | 23.42%  | 18.92%  |
| 23. | playing piccolo                             | 25.99%  | 18.88%  |
| 24. | sucking lolly                               | 30.30%  | 18.83%  |
| 25. | polishing furniture                         | 24.31%  | 18.78%  |
| 26. | calculating                                 | 25.21%  | 18.70%  |
| 27. | looking at phone                            | 26.76%  | 18.31%  |
| 28. | chiseling wood                              | 23.26%  | 17.44%  |
| 29. | picking apples                              | 19.75%  | 17.28%  |
| 30. | swimming with sharks                        | 25.10%  | 17.19%  |
| 31. | decoupage                                   | 21.13%  | 17.01%  |
| 32. | coloring in                                | 39.26%  | 16.92%  |
| 33. | poking bellybutton                          | 17.88%  | 16.56%  |
| 34. | chiseling stone                             | 21.19%  | 16.56%  |
| 35. | doing laundry                               | 22.31%  | 16.53%  |
| 36. | tiptoeing                                   | 21.99%  | 16.31%  |
| 37. | waxing armpits                              | 22.29%  | 16.28%  |
| 38. | curling eyelashes                           | 23.80%  | 16.26%  |
| 39. | pulling rope (game)                         | 17.02%  | 16.13%  |
| 40. | filling cake                                | 21.09%  | 15.99%  |
| 41. | opening coconuts                            | 22.31%  | 16.53%  |
| 42. | bending back                                | 16.46%  | 15.92%  |
| 43. | sausage making                              | 23.31%  | 15.73%  |
| 44. | passing American football (in game)         | 18.94%  | 15.61%  |
| 45. | laying stone                                | 21.26%  | 15.28%  |
| 46. | playing blackjack                           | 22.20%  | 15.07%  |
| 47. | changing gear in car                       | 18.82%  | 14.76%  |
| 48. | home roasting coffee                        | 17.34%  | 14.49%  |
| 49. | cutting cake                                | 17.83%  | 14.44%  |
| 50. | playing rounders                            | 16.39%  | 14.23%  |
| 51. | treating wood                               | 17.59%  | 13.67%  |
| 52. | vacuuming car                               | 18.14%  | 13.41%  |
| 53. | picking blueberries                         | 17.31%  | 13.03%  |
| 54. | dealing cards                               | 15.42%  | 12.98%  |
| 55. | laying decking                              | 13.60%  | 12.13%  |
| 56. | poaching eggs                               | 15.37%  | 12.04%  |
| 57. | swimming with dolphins                      | 14.13%  | 11.96%  |
| 58. | petting horse                               | 16.60%  | 11.95%  |
| 59. | lighting candle                             | 12.43%  | 11.86%  |
| 60. | taking photo                                | 15.29%  | 11.59%  |
| 61. | dyeing eyebrows                            | 14.31%  | 11.48%  |
| 62. | gospel singing in church                    | 14.83%  | 11.30%  |
| 63. | sieving                                    | 13.38%  | 11.04%  |
| 64. | cutting orange                              | 16.80%  | 11.02%  |
| 65. | carving marble                              | 15.26%  | 10.96%  |
| 66. | shoot dance                                 | 12.12%  | 10.92%  |
| 67. | grooming cat                                | 17.52%  | 10.86%  |
| 68. | tasting wine                                | 11.90%  | 10.71%  |
| 69. | combing hair                                | 20.36%  | 10.69%  |
| 70. | uncorking champagne                         | 16.31%  | 10.61%  |
| 71. | skiing mono                                 | 13.30%  | 10.47%  |
| 72. | putting wallpaper on wall                   | 14.29%  | 10.27%  |
| 73. | scrubbing face                              | 12.57%  | 10.20%  |
| 74. | surveying                                   | 12.42%  | 9.99%   |
| 75. | looking in mirror                           | 12.60%  | 9.72%   |
| 76. | mushroom foraging                           | 11.29%  | 9.68%   |
| 77. | ski ballet                                  | 9.76%   | 8.62%   |
| 78. | playing road hockey                         | 11.25%  | 8.50%   |
| 79. | applying cream                              | 8.91%   | 7.70%   |
| 80. | carving wood with a knife                    | 8.31%   | 7.42%   |
| 81. | using inhaler                               | 7.32%   | 7.32%   |
| 82. | milking goat                                | 10.78%  | 7.19%   |
| 83. | assembling bicycle                          | 7.76%   | 7.13%   |
| 84. | squeezing orange                            | 9.36%   | 7.08%   |
| 85. | pulling espresso shot                       | 7.19%   | 6.90%   |
| 86. | baby waking up                              | 8.03%   | 6.80%   |
| 87. | pouring wine                                | 9.42%   | 6.75%   |
| 88. | shopping                                    | 9.53%   | 6.72%   |
| 89. | seasoning food                              | 7.58%   | 6.72%   |
| 90. | adjusting glasses                           | 8.11%   | 6.68%   |
| 91. | being in zero gravity                       | 8.58%   | 6.66%   |
| 92. | blending fruit                              | 7.05%   | 6.54%   |
| 93. | mixing colours                              | 7.48%   | 6.51%   |
| 94. | spinning plates                             | 8.03%   | 6.45%   |
| 95. | ice swimming                                | 7.25%   | 6.11%   |
| 96. | doing sudoku                                | 7.40%   | 5.75%   |
| 97. | letting go of balloon                       | 6.09%   | 5.71%   |
| 98. | fixing bicycle                              | 5.62%   | 5.62%   |
99. entering church 6.28% 5.55%
100. chasing 5.98% 5.32%
101. playing shuffleboard 6.35% 5.31%
102. playing mahjong 11.31% 5.20%
103. peeling banana 6.06% 5.14%
104. closing door 6.25% 4.99%
105. shredding paper 6.55% 4.91%
106. card stacking 6.13% 4.90%
107. saluting 8.59% 4.89%
108. capsizing 6.26% 4.82%
109. delivering mail 5.12% 4.57%
110. listening with headphones 8.69% 4.56%
111. tossing salad 4.85% 4.49%
112. pouring milk 8.12% 4.28%
113. playing nose flute 5.72% 4.25%
114. carrying weight 6.73% 4.13%
115. shooting off fireworks 4.67% 4.08%
116. answering questions 5.87% 4.07%
117. testifying 12.50% 4.04%
118. herding cattle 4.48% 4.04%
119. putting on shoes 5.40% 3.84%
120. photobombing 4.43% 2.96%
121. bouncing ball (not juggling) 2.87% 2.59%
122. coughing 2.82% 2.11%
123. twiddling fingers 3.74% 2.02%