

# Supplemental materials for midterm presentation

## Group 6 - Hieu Nguyen

- Our python code is working after I change the library from gym (which has been deprecated) into Gymnasium (v1.1.1) and change some function variable name to match the newer version of Tensorflow too because the format of the model from the repository I believe was developed using Tensorflow v2.13 or older that still use the .h for saved trained model. Instead we use Tensorflow v2.16 which support .keras model. There are several old and new version compatible problem but by changing the code slightly we are able to get it working.
- For example in the architecture of the network we use different declaration

```
def _build_model(self):
    model = Sequential()
    model.add(Reshape((1, 80, 80), input_shape=(self.state_size,)))
    model.add(Convolution2D(32, 6, 6, subsample=(3, 3), border_mode='same',
                            activation='relu', init='he_uniform'))
    model.add(Flatten())
    model.add(Dense(64, activation='relu', init='he_uniform'))
    model.add(Dense(32, activation='relu', init='he_uniform'))
    model.add(Dense(self.action_size, activation='softmax'))
    opt = Adam(lr=self.learning_rate)
    # See note regarding crossentropy in cartpole_reinforce.py
    model.compile(loss='categorical_crossentropy', optimizer=opt)
    return model
```

Changed to

```
def _build_model(self):
    model = Sequential()
    model.add(Reshape((1, 80, 80), input_shape=(self.state_size,)))
    model.add(Conv2D(32, kernel_size=(6, 6), strides=(3, 3), padding='same',
                    activation='relu', kernel_initializer='he_uniform'))
    model.add(Flatten())
    model.add(Dense(64, activation='relu', kernel_initializer='he_uniform'))
    model.add(Dense(32, activation='relu', kernel_initializer='he_uniform'))
    model.add(Dense(self.action_size, activation='softmax'))
    opt = Adam(learning_rate=self.learning_rate)
    model.compile(loss='categorical_crossentropy', optimizer=opt)
    return model
```

- Working code evidence:

This is the terminal when running the code to train without enabling the interface for faster execution.

```
To enable the following instructions: ARMV8 ARMV812F ARMV812_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
Model: "sequential"
```

Layer (type)	Output Shape	Param #
reshape (Reshape)	(None, 1, 80, 80)	0
conv2d (Conv2D)	(None, 1, 27, 32)	92,192
flatten (Flatten)	(None, 864)	0
dense (Dense)	(None, 64)	55,360
dense_1 (Dense)	(None, 32)	2,080
dense_2 (Dense)	(None, 6)	108

```

Total params: 149,830 (585.27 KB)
Trainable params: 149,830 (585.27 KB)
Non-trainable params: 0 (0.00 B)
Epoch: 1 - Score: -13.00
Epoch: 2 - Score: 1.00
Epoch: 3 - Score: -6.00
Epoch: 4 - Score: 3.00
Epoch: 5 - Score: -6.00
Epoch: 6 - Score: -16.00
Epoch: 7 - Score: -17.00
Epoch: 8 - Score: -9.00
Epoch: 9 - Score: -5.00
Epoch: 10 - Score: -13.00
WARNING:absl:You are saving your model as an HDF5 file via `model.save()` or `keras.saving.save_model(model)`. This file format is considered legacy. We recommend using instead the native Keras format, e.g. `model.save('my_model.keras')` or `keras.saving.save_model(model, 'my_model.keras')`.
Epoch: 11 - Score: -3.00
Epoch: 12 - Score: -8.00
Epoch: 13 - Score: -6.00
Epoch: 14 - Score: -12.00
Epoch: 15 - Score: -17.00
Epoch: 16 - Score: -3.00

```

By setting render\_mode to “human” the environment will render the visualization of the state.

```
env = gym.make("ALE/Pong-v5", render_mode="human")
```

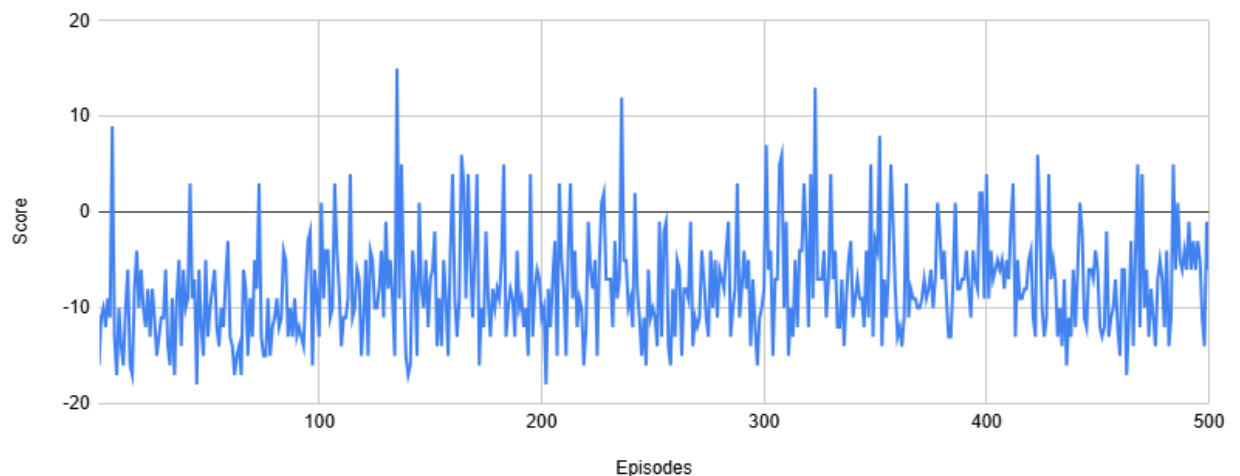
In the original version each episode will be play until someone reach 21

points and then subtract the score to get the final score of that episode like in the first picture.

- Since we can not re use the trained model on the repository the author upload, we have to train it from the beginning and after 500-1000 episodes the model slightly improve but still no where near beating the computer. We extracted the result into sheet and create a plot with the result like we show in our presentation.

Episodes	Score
1	-16
2	-11
3	-10
4	-12
5	-9
6	-11
7	9
8	-13
9	-17
10	-10
11	-14
12	-16
13	-10
14	-6
15	-16
16	-17
17	-8
18	-4
19	-10
20	-6
21	-10
22	-12
23	-8
24	-13
25	-8
26	-11
27	-15
28	-13
29	-11
30	-11
31	-6
32	-14

Score vs. Episodes



- The analysis after watch the model play the game is that the model did improve on getting score by 1 bounce only but if the computer can

catch the ball and bounce it back again the model don't know how to react yet. Because the model can only score after 1 bounce is not enough to win the whole set of the game which will most of the time contain multiple bounces back and forth therefore the result is still low. On the brightside the model did learned how to score with 1 bounce resulted in some lucky match win against the computer hence the high spike in the chart. We believe that with more training episodes the model will started to learn and beat the computer at some point.

- So after some few research studies we found another solution for training Pong using reinforcement learning by professor Karpathy using slightly different method but same preprocess and game loop. We also have to update the code to match our current version of Tensorflow to get it to work.
- Here are the architecture comparison table of 2 methods

Feature	PGAgent - Keras	Karpathy's
Neural Network Type	Convolutional Neural Network (CNN)	Fully Connected (2-layer)
Input Preprocessing	Converts to grayscale, downsamples, and flattens	Similar preprocessing
Hidden Layers	Conv2D → Dense (64) → Dense (32)	200 ReLU neurons
Output Activation	Softmax (multi-action policy)	Sigmoid (binary action)
Action Selection	argmax on probabilities	Probability-based (binary decision)
Gradient Update	Uses categorical_crossentropy loss	Custom backprop with manual weight updates

- And the performance comparison



Mean Score vs. Eps

Eps	Mean Score
0	-20.5
2500	-19.5
5000	-18.5
7500	-16.5
10000	-12.5
11000	-8.5

The screenshot displays the PyCharm IDE interface with a project named "Midterm Presentation". The left sidebar shows the Explorer view with files like "save\_model.py", "ale\_py-0.10.2-cp311-win\_amd...", "karpathy.py", "pong.py", "pong2.py", "save.p", and "test.py". The main editor window shows the code for "karpathy.py", which implements a reinforcement learning agent for the Pong game. The code includes functions for discounting rewards, calculating gradients, and updating model parameters.

```
# compute the discounted reward backwards through time
discounted_epr = discount_rewards(epr)
# standardize the rewards to be unit normal (helps control the gradient estimator variance)
discounted_epr -= np.mean(discounted_epr)
discounted_epr /= np.std(discounted_epr)

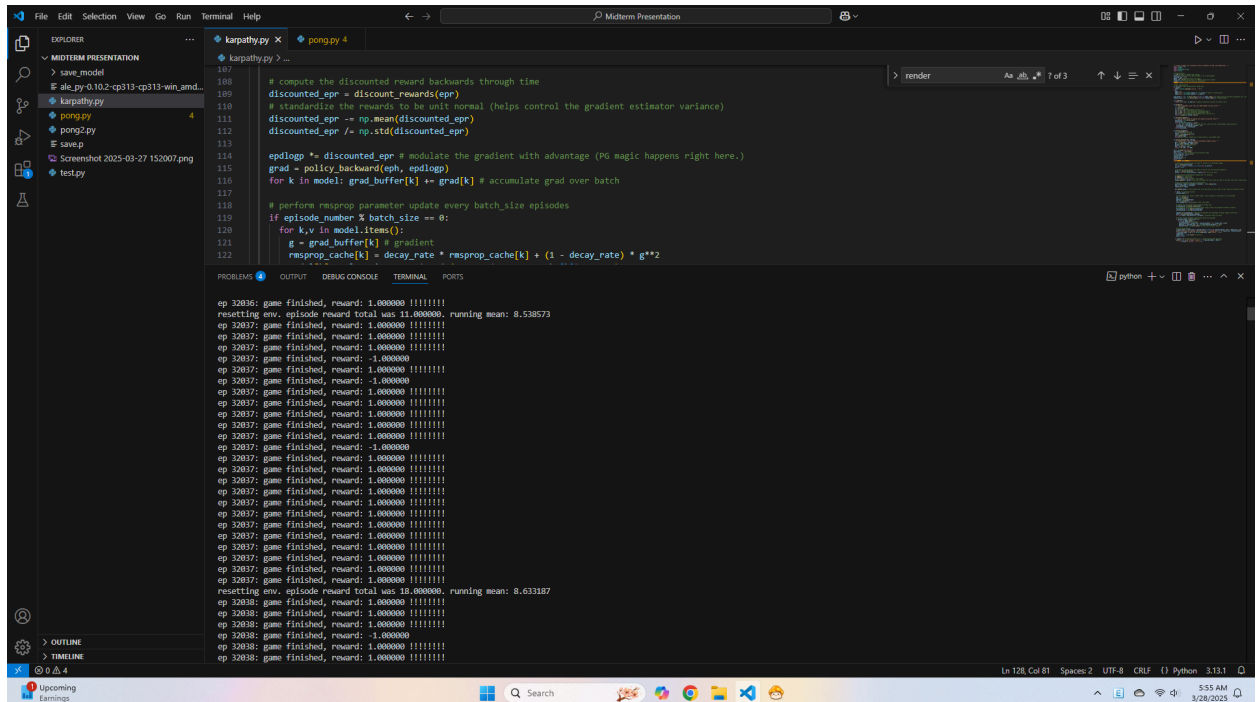
edploop = discounted_epr # modulate the grad with advantage (PG magic happens right here.)
grad = policy_backward(eph, edploop)
for k in model: grad_buffer[k] += grad[k] # accumulate grad over batch

# perform rmsprop parameter update every batch_size episodes
if episode_number % batch_size == 0:
    for k,v in model.items():
        g = grad_buffer[k] # gradient
        rmsprop_cache[k] = decay_rate * rmsprop_cache[k] + (1 - decay_rate) * g**2
```

The bottom panel shows the terminal output, indicating successful training results:

```
ep 16540: game finished, reward: 1.000000 !!!!!!!!
resetting env. episode reward total was 5.000000, running mean: 3.671897
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: -1.000000
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: 1.000000 !!!!!!!!
ep 16541: game finished, reward: -1.000000
```

And 8.5 mean score by 42000 episodes but unfortunately I did not implement a result saving function for it so I can only add 2 snapshots of the running code with align time stamp for evidence.



```
107
108
109 # compute the discounted reward backwards through time
110 discounted_epr = discount_rewards(epr)
111 # standardize the rewards to be unit normal (helps control the gradient estimator variance)
112 discounted_epr -= np.mean(discounted_epr)
113 discounted_epr /= np.std(discounted_epr)
114
115 epdlogp *= discounted_epr # modulate the gradient with advantage (PG magic happens right here.)
116 grad = policy_backward(eph, epdlogp)
117 for k in model: grad_buffer[k] += grad[k] # accumulate grad over batch
118
119 # perform rmsprop parameter update every batch_size episodes
120 if episode_number % batch_size == 0:
121     for k,v in model.items():
122         g = grad_buffer[k] # gradient
123         rmsprop_cache[k] = decay_rate * rmsprop_cache[k] + (1 - decay_rate) * g**2
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

The result get for the chart above is when I retrain the model from the beginning and try to map it as a chart before the deadline so it is not complete but from both the chart and the snapshot we can see that after 42000+ episode the model we trained started to win a lot more than the computer. I also record a short video of a pretrain and trained model playing the game it self and will attach it in the zip file.

- Finally thank you professor for introduce us to an interesting subject and please let us know if you need any more proofs.

Result links:

[https://docs.google.com/spreadsheets/d/1BS7hY0kv2BqskTP3oVNnSH1oamYaD7HVzzVMHeOd\\_zw/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1BS7hY0kv2BqskTP3oVNnSH1oamYaD7HVzzVMHeOd_zw/edit?usp=sharing)