Learning Multimodal Representations For Unseen Actions
– Supplementary Material –

AJ Piergiovanni[1] Michael S. Ryoo[1,2]
[1]Indiana University, [2]Stony Brook University
{ajpiergi,mryoo}@indiana.edu

A. Implementation/training details

We implement our models in PyTorch. For the per-segment video CNN, we use I3D [2] to obtain a $1024 \times T$ video representation. We trained a version of I3D based on Kinetics-600, but withheld all classes that appear in ActivityNet, HMDB51, or UCF101 so that the classes are truly unseen. This resulted in a training set with 478 classes and 278k videos. Since generating videos is an extremely challenging task, the video autoencoders start with and generate the I3D feature. We use GloVe word embeddings [3] to obtain a language representation. We set $N = 4$ for the temporal attention filters and apply 4 fully connected layers. These layers are followed by L_2 normalization so that the embedding space has unit length [5]. We train the models for 200 epochs and use stochastic gradient descent with momentum to minimize the loss function with a learning rate of 0.01. After every 50 epochs, we decay the learning rate by a factor of 10. When training in the adversarial setting (e.g., Algorithm 1 in the main paper), we initialize the network training for 50 epochs on paired data followed by 200 on the paired + unpaired data.

A.1. Unseen video captioning

As our model learns a bi-directional mappings, we can apply our model to generate video captions. Existing video captioning models are unable to create realistic captions for unseen activities, as without training data they do not know the words to describe the video. Given a video, v, we can generate a caption by mapping the video to text $t = G_T(E_V(v))$. For each word, we then use nearest neighbors matching with the GloVe embeddings to obtain the words to form a sentence. In Table 1, we report the commonly used METEOR [1] and CIDEr [6] scores of our various models, measured with the unseen classes from the ActivityNet dataset. We find that learning a joint representation is beneficial and using unpaired samples further improves the task. Note that this task is extremely challenging, as it requires the model to generate captions using activity words (e.g., basketball) not seen during training.

Table 1. Comparison of several models for unseen activity captioning using the ActivityNet dataset, using METEOR and CIDEr scores. This evaluation was done on 10 unseen classes held out during training. Higher values are better.

	METEOR	CIDEr
Fixed Text Representation	3.64	8.95
Joint	4.21	9.23
All (paired)	5.31	11.21
All (paired + unpaired)	6.89	13.95

Table 2. Comparison of temporal pooling methods for 5 unseen classes in the ActivityNet dataset.

	Accuracy
Max Pooling	23.4
Sum Pooling	24.1
LSTM	42.3
Temporal Attention Filters	55.2

B. Additional Experiments

B.1. Comparison of temporal pooling methods

To confirm that temporal attention is beneficial, we compare different forms of temporal pooling (i) max-pooling, (ii) sum-pooling, (iii) LSTM, and (iv) temporal attention filters [4]. In Table 2, we compare these temporal pooling methods learning the joint embedding space. We confirm that using the temporal attention filters performs best.

B.2. Comparison of different ratios of paired and unpaired data

We compare different ratios of paired and unpaired data to see how much paired data we require and how much unpaired data is beneficial. For these experiments, we use all the loss terms (i.e., what provided us the best results). Note that in these experiments, the total number of samples was the same for each method (40k examples) so that we can directly compare the effects of unpaired data vs. paired data. Thus not all the available data was used.

In Table 3, we show the results. We find that using no paired data results in nearly random performance, but using using some paired data greatly improves the embedding

Table 3. Comparison of different ratios of paired and unpaired data methods for 5 unseen classes in the ActivityNet dataset.

Paired/Unpaired	Accuracy
100% / 0%	74.2
75% / 25%	73.2
50% / 50%	69.7
25% / 75%	62.6
0% / 100%	24.5

Table 4. Comparison using 40k paired examples and varying amounts of unpaired samples for 5 unseen classes in the ActivityNet dataset.

Unpaired Samples	Accuracy
0	77.1
10k	82.4
20k	83.9
40k	83.6
60k	83.5

space. The model using 100% paired data performs best, as all the others are using less overall paired data.

We also compare augmenting our 40k paired training samples with different amounts of unpaired data. Since UCF101 and HMDB only have 13k and 7k examples, to get up to 60k samples, we also use videos from the Kinetics dataset [2]. The results, shown in Table 4, show that adding the initial 10k samples is most beneficial, while additional samples do not seem to meaningfully improve results. However, due to our training method where each batch consists of 50% paired data and 50% unpaired data, the additional unpaired data does not harm results either.

B.3. MLB-Youtube Captions

In Fig. 1, we compare t-SNE embeddings of the fixed text representation and our joint embedding space. This visually shows that learning a joint embedding space gives more distinct class distributions.

B.3.1 MLB-YouTube Captions

As a baseline for the MLB-YouTube captions dataset, we compared several different models for standard video captioning (i.e., all activity classes are seen). This task is quite challenging compared to other datasets as the announcers commentary is not always a direct description of the current events. Often the announcers tell loosely related stories and attempt to describe events differently each time to avoid repetition. Additionally, the descriptions contain on average 150 words for each 30 second interval and current captioning approaches usually only trained and tested on 10-20 word sentences. Due to these factors, this task is quite challenging the standard evaluation metrics do not account for these factors. In Table 5, we report our results on this task.

Table 5. Comparison of several models for standard, seen video captioning using the MLB-YouTube dataset, using Bleu, METEOR and CIDEr scores. Higher values are better.

	Bleu	METEOR	CIDEr
Fixed Text Representation	0.12	0.04	0.12
Joint Representation	0.14	0.08	0.15
Joint + all paired	0.15	0.10	0.18
Joint + paired + unpaired	0.10	0.02	0.08

Table 6. Comparison of various pronouns on the UCF101 dataset with 50 unseen classes.

	Accuracy
Baseline Sentences	33.4
All ‘man’	33.2
All ‘woman’	33.3
All ‘person’	33.4
Random pronoun	33.4

C. HMDB and UCF101 Sentences

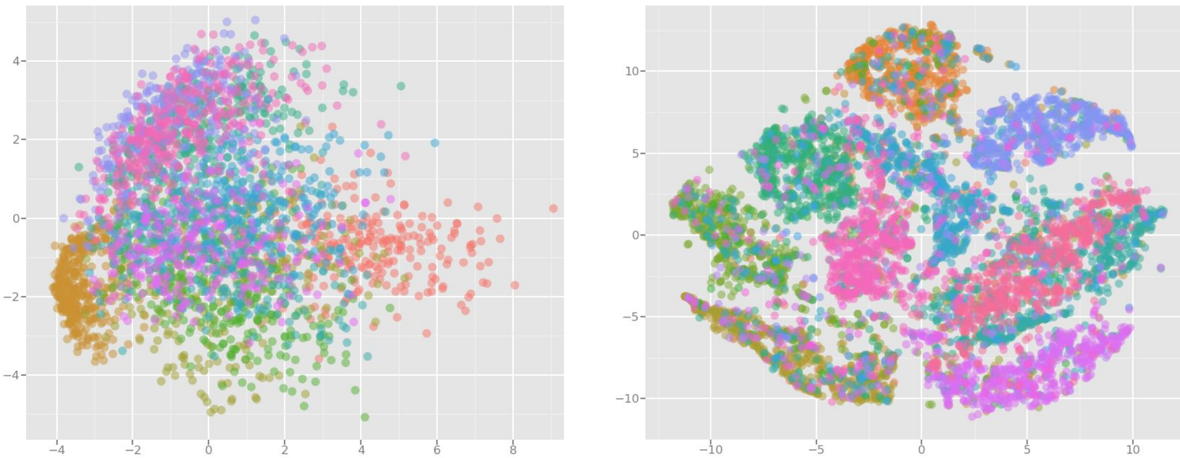
For the HMBD and UCF101 datasets, we created sentences to describe each activity class. Our sentences descriptions are included in this appendix.

These sentences are written for each activity class (by randomly selecting a single video per class) and are shared for all instances of the activity. Depending on what video was randomly chosen for the class, some sentences describe the actor as a ‘man’, ‘woman’, or ‘person’ which could confuse the model. Ideally, the CNN embedding needs to learn to ignore the impact of such pronoun changes.

We conducted experiments comparing randomly replacing the pronouns to determine if there was any bias introduced by the pronouns. We show the results in Table 6. We find that the choice of pronouns does not impact performance, as our model automatically learns to focus more on verbs rather than pronouns. When examining the temporal attention filters on the sentences, we found that they placed very little ‘attention’ on the start of the sentence, where the pronoun usually is, suggesting that the pronoun has very little effect on the embedding space we learned.

HMDB:

1. chew: a woman is chewing on bread
2. golf: a man swings a golf club
3. sword exercise: a person is playing with a sword
4. walk: a person is walking
5. jump: a person jumps into the water
6. pour: a man pours from a bottle
7. laugh: a man is laughing
8. shoot gun: a person rapidly fires a gun



(a) (b)

Figure 1. t-SNE mapping of **(a)** fixed text representation and **(b)** joint embedding with all paired losses for the MLB-YouTube dataset. The joint embedding space provides most distinct representations for the activities. Each color represents the activity class of the video (e.g., swing, hit, foul ball, etc.).

9. run: a person is running
10. turn: a person turns around
11. ride bike: a man is riding a bike on the street
12. swing baseball: a boy hits a baseball
13. draw sword: a person draws a sword
14. sit: a person sits in a char
15. fencing: two men are fencing
16. dribble: a boy dribbles a basketball
17. stand: a person stands up
18. pushup: a man does pushups
19. sword: two people are fighting with swords
20. pullup: a boy does pullups in a doorway
21. smile: a man smiles
22. shake hands: two people shake hands
23. shoot ball: a person shoots a basketball
24. kick: a person kicks another person
25. somersault: a person does a somersault
26. flic flac: a boy does a backflip
27. hug: two people hug
28. hit: a boy swings a baseball bat
29. dive: a person jumps into a lake
30. drink: a man drinks from a bottle
31. punch: a woman punches a man
32. wave: a person waves their hand
33. talk: a person is talking
34. kiss: a man and woman kiss
35. catch: a boy catches a ball
36. smoking: a woman smokes a cigarette
37. eat: a man eats pizza
38. throw: a person throws a ball
39. climb stairs: a man is running down the stairs
40. kick ball: a person kicks a soccer ball
41. ride horse: a girl is riding a horse
42. fall floor: a man is pushed onto the ground
43. brush hair: a girl is brushing her hair
44. situp: a man does situps
45. cartwheel: a guy runs and jumps and flips
46. pick: a man picks a book
47. push: a boy pushes a table
48. climb: a man is climbing up a wall
49. handstand: three girls do handstands
50. clap: a woman claps her hands

51. shoot bow: a person shows a bow and arrow

UCF101:

1. MilitaryParade: people are marching and waving a flag
2. TrampolineJumping: kids are jumping on a trampoline
3. PlayingDaf: a person moves a circle and hits it
4. SalsaSpin: poeple are dancing and spinning
5. CuttingInKitchen: a person is in the kitchen using a knife
6. ApplyEyeMakeup: a woman is putting on makeup
7. PlayingViolin: a person plays the violin
8. YoYo: a person plays with a yoyo
9. PlayingCello: a person is playing the cello
10. Bowling: a person is bowling
11. UnevenBars: a woman is spinning and flying on bars
12. BalanceBeam: a woman is on the balance beam
13. SkyDiving: people are falling out of the sky
14. SumoWrestling: two fat people are wrestling
15. PushUps: a man does pushups
16. FloorGymnastics: a girl does gymnastics
17. ApplyLipstick: a woman is putting on lipstick
18. BreastStroke: a woman is swimming
19. GolfSwing: a man swings a golf club
20. PlayingDohl: a person hits on a drum
21. HorseRiding: a woman rides a horse
22. PlayingFlute: a person blow into a flute
23. PizzaTossing: a man is making a pizza
24. CleanAndJerk: a person is lifting weights
25. WritingOnBoard: a person is writing on the wall
26. CricketShot: a person hits a ball with a bat
27. FieldHockeyPenalty: a girl in the field shoots a ball
28. HammerThrow: a person spins and throws an object
29. BodyWeightSquats: a man is squatting
30. CliffDiving: a person jumps off a cliff
31. Typing: a person is typing at a computer
32. MoppingFloor: a man mops the floor
33. TaiChi: people are doing tai chi
34. PlayingPiano: a person plays piano
35. Punch: someone punches another person
36. Nunchucks: a person swings nun chucks
37. RopeClimbing: a person climbs a rope
38. Swing: a baby is swinging
39. Knitting: a woman is knitting
40. Rafting: people are rafting on a river
41. PlayingGuitar: a person strums a guitar
42. ShavingBeard: a man shaves his beard
43. JugglingBalls: a person is juggling balls
44. Diving: a boy dives into a pool
45. JumpingJack: a person jumps and swings his arms
46. VolleyBallSpiking: people hit a volleyball
47. PoleValut: a person runs with a pole and launches into the air
48. SkateBoarding: a man is skateboarding
49. BoxingPunchingBag: a man is punching a bag
50. IceDancing: people are ice skating
51. WallPushups: a person does pushups against a wall
52. FrisbeeCatch: a person jumps and catches a frisbee
53. Drumming: people are drumming
54. JumpRope: a girl is jumping rope
55. HeadMassage: a person gets their head massaged
56. PlayingTabla: a person plays two drums
57. TableTennisShot: people are playing table tennis
58. PommelHorse: a person spins around on their hands
59. HighJump: a man jumps over a bar and lands on his back
60. BasketballDunk: a man jumps and dunks the basketball
61. BoxingSpeedBag: a man punches a bad in the air quickly
62. PullUps: a person does hangs on a bar and pulls up

63. RockClimbingIndoor: a person is climbing up rocks
64. BlowingCandles: a boy blows out candles on a cake
65. Skiing: people are skiing on a mountain
66. WalkingWithDog: a person walks a dog
67. Basketball: men are playing basketball
68. SoccerJuggling: a person is playing with a soccer ball
69. Fencing: people are fencing
70. Billiards: a man is playing billiards
71. BaseballPitch: a man throws a baseball
72. BlowDryHair: a woman is drying her hair
73. CricketBowling: a person throws a cricket ball
74. BandMarching: people are walking down the street playing music
75. PlayingSitar: a person plays a funny guitar
76. ThrowDiscus: a person spins and throws a disk
77. StillRings: a man holds in the air on rings
78. Lunges: a person bends to the ground with one knee
79. Skijet: a person rides a jetski in the ocean
80. BabyCrawling: a baby is crawling on the floor
81. Mixing: a woman is mixing in a bowl
82. Hammering: a person is hitting nails with a hammer
83. Shotput: a person spins and launches a ball
84. Archery: a man shoots a bow and arrow
85. Surfing: a man is surfing in the ocean
86. FrontCrawl: a person is swimming freestyle
87. HulaHoop: a person spins a hoop around their waist
88. JavelinThrow: a person throws a spear
89. Rowing: people are in a canoe and rowing
90. Kayaking: a person is kayaking on a lake
91. ParallelBars: a man does gymnastics on the parallel bars
92. HorseRace: horses are racing around a track
93. HandstandWalking: a person stands on their hands and walk
94. BrushingTeeth: a boy brushes his teeth
95. LongJump: a person runs and jumps into a sand pit
96. Biking: people are riding bikes
97. HandstandPushups: a person does pushups upside down
98. BenchPress: a man is lifting weights
99. Haircut: a person is getting a hair cut
100. TennisSwing: a woman hits a tennis ball

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[4] A. Piergiovanni, C. Fan, and M. S. Ryoo. Learning latent sub-events in activity videos using temporal attention filters. In *Proceedings of the American Association for Artificial Intelligence (AAAI)*, 2017.

[5] M. Tygert, A. Szlam, S. Chintala, M. Ranzato, Y. Tian, and W. Zaremba. Convolutional networks and learning invariant to homogeneous multiplicative scalings. *arXiv preprint arXiv:1506.08230*, 2015.

[6] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.