Visual-textual Capsule Routing for Text-based Video Segmentation: Supplementary Material

Bruce McIntosh bwmcint@gmail.com
Kevin Duarte kevin95duarte@gmail.com
Yogesh S Rawat yogesh@crcv.ucf.edu
Mubarak Shah shah@crcv.ucf.edu

Center for Research in Computer Vision
University of Central Florida
Orlando, FL, 32816

In the supplementary material we include additional qualitative and quantitative results, as well as some analysis of the sentence capsules. Also, we include figures and a more in-depth description of the network architecture.

1. Analysis of Sentence Capsules

We perform some analysis on the sentence capsules to understand the sentence capsule representations for the task of text-based video segmentation. We focus our analysis on the directional descriptors ("left" and "right") as well as color descriptors. In these sets of experiments, we observe the pose matrices' change in the 8 sentence capsules when different sentences are given to the sentence encoder.

Directional Descriptors In the test set, there exist 172 sentences which have the words "left" or "right". We can therefore make 172 coherent sentences with the word "left" and 172 sentences with the word "right" by replacing each instance of one of these words with the descriptor of interest. Once these sentences (a total of $172 \times 2 = 344$) are fed to the network, we obtain a 2 dimensional TSNE [2] visualization of the pose matrices for each sentence capsule. Figure 1 shows the 2 dimensional mappings for the pose matrices of each of the sentence capsules. It can be observed that some sentence capsule types, like 2, 6, and 7, have very distinct poses when "left" and "right" occur in the sentence. This means that these capsules are able to specialize and represent this notion of direction, even though there is no direct supervision which enforces this specialization.

Color Descriptors To examine how the sentence network interprets colors, we select 258 sentences from the test set which contain a color. Then we create two sets of sentences: 1) sentences where these colors are removed, and 2) sentences where these colors are replaced with a single color (e.g. "black"). Once again, we obtain a 2 dimensional TSNE visualization of the pose matrices for each sentence capsules, which can be found in Figure 2. There are noticeable changes in the poses, which allows the entire network correctly segment videos when color cues are present. However, these changes are not as drastic as those seen with the directional descriptors. This could help explain why the network is more responsive, in terms of its output segmentation, to directional cues than color cues.

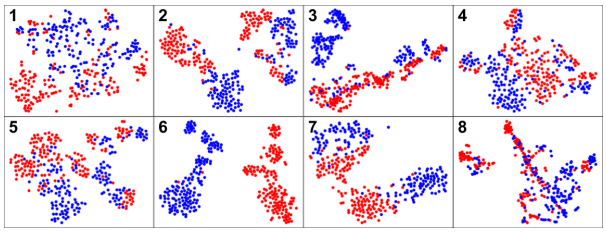

Figure 1. Visualizations of the 8 sentence capsules' pose matrices for sentences with the word "left" (red) and sentences with the word "right" (blue). Some sentence capsules specialize on these directional descriptors. This is most apparent in capsule type 6.

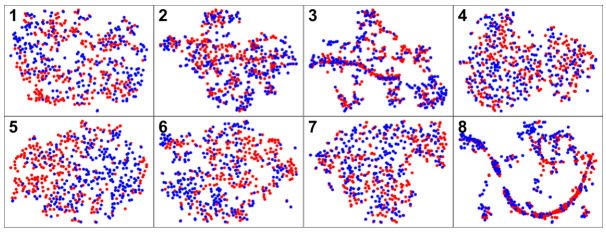

Figure 2. Visualizations of the 8 sentence capsules' pose matrices for sentences with the color "black" (red) and sentences with no color descriptor (blue).

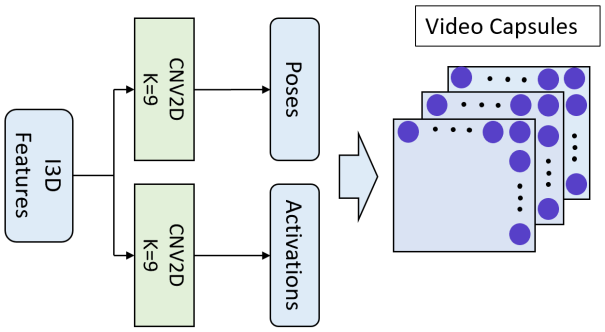

Figure 3. Video Capsule Network. Video capsules are formed from the I3D features by convolution with one layer to create the 4x4 pose matrix for each capsule, and another layer to create the activation for each capsule.

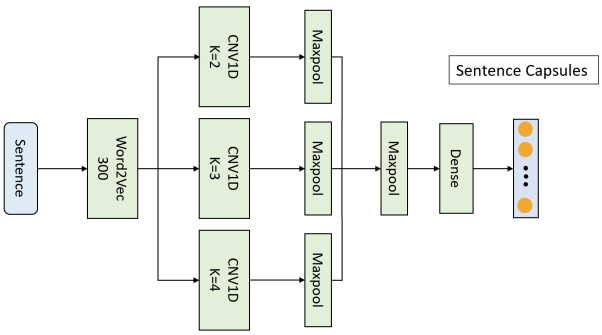

Figure 4. Sentence Network. Each word from a natural language sentence is converted into a size 300 word2vec vector. The vectors go through a convolutional network and are then reshaped into the poses and activations of capsules representing the sentence.

2. Network Architecture

When designing our network, we found several components are key to obtaining our state-of-the-art results. We explain some of these components, and how they impact our results. Furthermore, we include several figures (3, 4) which illustrate the construction of the video capsules and the sentence network respectively.

2.1. Video Encoder

Our network uses a pretrained I3D network to encode the video into a set of feature maps of size 28×28. Originally we used the simpler C3D network [4], to encode the feature maps, but it achieved substantially worse results - a mean IoU of 34%. This shows that the features extracted from the I3D network, are more useful for our video capsules than were the features from the C3D network.

2.2. Sentence Network

We also tested several different configurations for our sentence network. We began with using the capsule networks (both Capsule-A and Capsule-B) described in Figure 2 of [5] which showed strong results on text classification. These network have several capsule layers, but their use led to poor results: a mean IoU of 35.7% for the Capsule-A network, and a mean IoU of 36.4% for the Capsule-B network. Although these network performed well on the text classification task, a different set of text features must be learned to merge with visual features. Therefore, our sentence network with conventional convolutional layers, max-pooling, and a fully connected layer, is better able to extract textual features for the task of actor and action video segmentation from a sentence.

3. Evaluations

We also include several tables here. Table 1 shows the results of our network trained on the bounding box annotations from A2D, and tested on JHMDB. The network's outputs are more box-like because it was trained on bounding boxes as opposed to pixel-wise segmentations; therefore, the network has strongest results when tested using bounding box ground-truths. Table 2 contains all the metrics for the ablation experiments. Table 3 contains all the metrics for the ReferItGame experiments, which were described in the main text.

Ablation for Skip Connections To understand the effectiveness of the parameterized skip connections from the I3D encoder, an experiment was run with these skip connections removed. This resulted in a 3% reduction in mean IoU score and mean average precision. The decrease in performance shows that the skip connections are necessary for the network to preserve fine-grained details from the input video.

4. Qualitative Results

We have generated several videos with the segmentations produced by our networks for both the A2D and JHMDB datasets. Each video contains the input sentences, color coded to match the ground-truth colors. The first row of each video has the ground-truth bounding boxes (for A2D) or ground-truth segmentation masks (for JHMDB). The second row is the output of the network which was trained on the key frames of the A2D dataset, which had pixel-wise segmentation ground-truths. The third row is the output of the network which was trained on all the frames of the A2D dataset, using bounding-box ground-truths. In our analysis of the qualitative results, we will refer to the prior network as the "Key Frame network" and the latter network as the "Bounding Box network".The video clips can be found in

	Video Overlap					v-mAP	Video IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
All frames	16.3	2.0	0.1	0.0	0.0	2.3	37.9	35.6
All frames (bbox)	46.7	32.1	15.8	3.2	0.0	16.3	44.5	44.4

Table 1. Results for network trained on the bounding box annotations for A2D with sentences, and evaluated on JHMDB. The first row is the network tested against the ground-truth pixel-wise segmentations from the JHMDB dataset. The second row is the network tested against bounding boxes around the ground-truth segmentations from the JHMDB dataset. Since our network was trained with bounding boxes, it performs better when it is evaluated against bounding box ground-truths.

	Overlap					mAP	IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
No skip connections	49.5	41.2	29.1	14.6	1.5	26.9	56.7	43.1
No L_c nor Masking	49.4	42.5	32.7	19.6	3.3	28.8	57.6	43.6
No Masking (with L_c)	48.3	41.4	31.2	18.4	3.1	27.8	56.6	42.5
Concatenation	22.9	15.4	7.6	2.0	0.1	9.9	35.1	25.0
Multiplication	38.4	30.1	20.9	9.7	0.8	19.4	48.2	35.0
Filter Poses	49.1	42.3	32.6	19.1	3.2	29.1	57.2	42.7
Filter Activations	48.8	42.7	33.4	20.1	3.8	29.2	56.8	43.0
Our Network	**52.6**	**45.0**	**34.5**	**20.7**	**3.6**	**30.3**	**56.8**	**46.0**

Table 2. All metrics for the ablations trained and tested on the A2D dataset with sentences. We test the effect of parameterized skip connections, capsule masking, and the classification loss. We also test conventional conditioning methods on our capsule network to evaluate the effectiveness of the proposed multi-modal capsule routing procedure. The final row contains the results of our network without any changes.

their corresponding subdirectory in the Videos folder.

4.1. Single Actor

The networks seem to perform best when there is a single actor in the scene. If this is the case, we find that the Key Frame network produces very fine-grained segmentations which maintain the boundaries of the actors; Meanwhile, the Bounding Box network successfully segment a box around the actor. This behaviour can be observed in the videos in the A2DSingleActor folder, which contains examples from the A2D dataset, and the JHMDB folder, which contains examples from the JHMDB dataset.

4.2. Multiple Actors

The network can also perform well with multiple actors in the scene as seen in the A2DMultiActor videos. In these cases, the segmentations are not as precise, but the general location of the actors is being segmented by both networks. We note that the Bounding Box network, which was trained with all the frames of dataset, tends to produce more consistant multi-actor segmentations: as seen in videos "video3_multi_a2d" and "video5_multi_a2d". In both of these cases, each instance is of the same actor class, and the Key Frame network seems to incorrectly segment one of the instances.

4.3. Failure Cases

As mentioned in the main text, our network tends to fail on the A2D dataset when there are multiple instances of the same actor class. We present 5 videos in which our network fails in the A2DFailure folder. The probability of failure is increased when the sentence queries are vague, or could describe multiple actors within the scene, like in the video "video1_failure_a2d". Several birds can fit the description of "sparrow sitting on the grass" or "sparrow is walking on the brown grass". In these cases it would be very difficult for even a human to correctly segment the video. In many cases, the failure occurs when there are many similar actors near each other, like in the videos "video2_failure_a2d.mp4" and "video5_failure_a2d". In the first, there are multiple people running next to each other, while the second contains several cars moving near each other.

Since the JHMDB dataset only has a single actor in each video, the failures encountered are not from "difficult" queries or videos, but rather a result of the mismatch between the training and testing data. A2D videos tend to have humans perform an action requiring large amounts of motion - like walking, running, jumping or rolling - while the A2D videos have many videos in which the action has little motion - like brushing hair or archery. Thus, both the

	Overall IoU	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9
Hu *et al*. * [1]	56.83	43.86	35.75	26.65	16.75	6.47
Shi *et al*. [3]	59.09	45.87	39.80	32.82	23.81	11.79
Our Network	55.7	43.4	36.2	28.3	19.6	9.7

Table 3. Results on ReferItGame dataset. *This result is using Deeplab101 as a backbone, as described in [3]. We achieve comparable results, even with a network designed for video inputs.

videos, and the input textual queries, are quite different between the datasets, which can cause a large performance discrepancy during evaluation. Examples of failure cases on the JHMDB dataset can be found in the JHMDBFailure folder.

References

[1] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 4

[2] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 1

[3] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 4

[4] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[5] Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*, 2018. 2