AAAI-18
Association for the Advancement of Artificial Intelligence 2018
February 2 - 7, 2018, New Orleans, USA

Reviews For Paper

Paper ID	1646
Title	T-C3D: Temporal Convolutional 3D Network for Real-time Action Recognition

Masked Reviewer ID: Assigned_Reviewer_1
Review:

Question	
[Summary] Please summarize the main claims/contributions of the paper in your own words.	The paper defines a method for highly efficient action recognition. Building on the popular 3D-CNN approach (Tran et al. 2015; 2017; Varol, Laptev, and Schmid 2017), each video is divided into a fixed number of segments, then each segment is sampled to form a clip. Since most video frames are not sampled, computational efficiency is greatly improved. In addition, the method does not compute optical flow, which is popular in two-stream action recognition methods and is very slow. These two tricks lead to fast computation. The approach aggregates information across the clips as well to form a classification decision based on the entire video, using a choice of standard aggregation schemes. Overall there is moderate algorithmic novelty here, with detailed attention to comparing different methods within various important stages such as clip sampling and cross-clip aggregation. The experimental evaluation is very thorough, with results compared against very recent approaches on large, current action recognition datasets. A brand-new dataset, Kinetics, is used for pre-training and the resulting performance gain is quantified vs. random initialization. The paper is clear and thoroughly convincing.
[Relevance] Is this paper relevant to an AI audience?	Relevant to researchers in subareas only
[Significance] Are the results significant?	Significant
[Novelty] Are the problems or approaches novel?	Somewhat novel or somewhat incremental
[Soundness] Is the paper technically sound?	Technically sound
[Evaluation] Are claims well-supported by theoretical analysis or experimental results?	Very convincing
[Clarity] Is the paper well-	

organized and clearly written?	Excellent
[Detailed Comments] Please elaborate on your assessments and provide constructive feedback.	The work will be of interest to researchers working in video, although there is little to generalize beyond video or vision work here. Within action recognition the results will be significant, as the accuracy is close to the state-of-the-art but much more efficient. It is moderately surprising that such high levels of sampling could maintain reasonable performance. The paper does not have compelling novelty, but it pushes recent ideas considerably further and will generate interest. The text is very clearly written. The relationship to prior work is outstanding, with a detailed section plus repeated grounding of various components against the literature. The first segment sampling method chooses non-consecutive frames to form into clips. These clips are then fed into the 3D-CNN, which performs temporal convolution on them. Why would this work? I would expect the temporal discontinuities to produce significantly different convolutional responses on different samplings of the same video. It's surprising that it works as well as it does, only suffering 0.3% in accuracy vs. consecutive frame sampling. The data augmentation techniques look to be very useful. Good idea to apply image-based ones to video here. The various experiments exploring different aspects of the method are outstanding. Showing the benefit of pre-training on Sports-1M and Kinetics vs. random is compelling. Showing the relative accuracy vs. FPS of different samplig schemes, including none, clearly shows the tradeoffs being made. The paper claims "real-time action recognition", which implies two properties: 1) keeping up with video frame rates e.g. 30Hz; 2) performing recognition as video frames are received, as opposed to only after the entire video has been presented. The paper achieves 1, in that the number of video frames divided by processing time is far greater than 30. It does not mention 2, however. The method may be suitable to online processing where a recognition estimate becomes available as soon as possible and is refined over time, but this should be discussed or the title should be changed.
[QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.	See discussion above.
[OVERALL SCORE]	Accept (Top 15%)
[CONFIDENCE]	Reviewer is an expert in the area

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
	This paper sets forth a temporal deep CNNs that is able to recognize the

[Summary] Please summarize the main claims/contributions of the paper in your own words.	human actions in real time. The major contribution lies in accelerating the C3D action recognition system via a temporal encoding technique. The proposed system is sensibly tested and compared with the state-of-the-art systems based on two public action recognition datasets, and the results verified the contribution. Overall, this is an interesting work and the real-time action recognition from video is indeed in demand. However, the contribution seems minor and i was not convinced that the system has made a significant improvement against the existing work.
[Relevance] Is this paper relevant to an AI audience?	Relevant to researchers in subareas only
[Significance] Are the results significant?	Moderately significant
[Novelty] Are the problems or approaches novel?	Somewhat novel or somewhat incremental
[Soundness] Is the paper technically sound?	Technically sound
[Evaluation] Are claims well-supported by theoretical analysis or experimental results?	Somewhat weak
[Clarity] Is the paper well-organized and clearly written?	Poor
[Detailed Comments] Please elaborate on your assessments and provide constructive feedback.	My major concerm came from the ambiguous statement of the contributions. After reading the paper, i still did not get the point what makes the proposed system faster than the existing 3D CNNs for human action recognition? Authors claim in the first contribution "We propose a real-time 3D-CNN based action recognition architecture to learn video representation at a multitude of granularity. The learned descriptor is able to model not only the temporal evolution of appearance between short clips but also the overall temporal dynamics of the entire video.", which indicates that a new descriptor combining appearance and temproal information may help improve the efficiency. However, when i read the Section 3. 1, i found that maybe partitioning the whole video into a few shot clips and processing them in parallel is the key to speeding up the system. The last contribution mentioned in the paper "We only employ RGB frames as the input of CNN to process action recognition at real-time while achieving comparable performance to the state-of-the-art methods." where is the contribution here? RGB input?
[QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.	I hope that authors could answer the following questions in the rebuttal: 1. Clearly explain the real contribution, which should be convincing. 2. Fig. 2 is not meaningful. I could not figure out what features are extracted here? 3. From the limited equations provided by the paper, i could not figure out how Q in Eq. (1) looks like? 4. Why authors need to elaborate four poolings, where none of them is the contribution of your paper. 5. The typos and grammatical errors are here and there. Quite some polish works are needed.
[OVERALL SCORE]	Reject

[CONFIDENCE]	Reviewer is knowledgeable in the area