Joint Event Detection and Description in Continuous Video Streams

Huijuan Xu[1], Boyang Li[2], Vasili Ramanishka[1], Leonid Sigal[3], Kate Saenko[1]
[1]Boston University, [2]Baidu Research, [3]University of British Columbia
[1]{hxu, vram, saenko}@bu.edu, [2]boyangli@baidu.com, [3]lsigal@cs.ubc.ca

Abstract

Dense video captioning involves first localizing events in a video and then generating captions for the identified events. We present the Joint Event Detection and Description Network (JEDDi-Net) for solving this task in an end-to-end fashion, which encodes the input video stream with three-dimensional convolutional layers, proposes variable-length temporal events based on pooled features, and then uses a two-level hierarchical LSTM module with context modeling to transcribe the event proposals into captions. We show the effectiveness of our proposed JEDDi-Net on the large-scale ActivityNet Captions dataset.

1. Approach

Figure 1 provides an overview of our proposed JEDDi-Net model. The model consists of two main components: a segment proposal module and a captioning module.

Proposal Module: We employ the C3D architecture to encode the input frames in a fully-convolutional manner. To obtain feature vectors C_{tpn} for predicting proposals at each time point, we add two 3D convolutional filters with kernel size $3\times3\times3$ on top of C_{conv5b}, followed by a 3D max-pooling filter to remove the spatial dimension. Based on the 512-dimensional feature vector at each temporal location in C_{tpn}, we predict a relative offset $\{\delta c_i, \delta l_i\}$ to the center location and the length of each anchor segment $\{c_i, l_i\}_{i=1\cdots R}$, as well as a binary label indicating whether the predicted proposal contains an activity or not. This is achieved by adding two $1\times1\times1$ convolutional layers on top of C_{tpn}. A detailed diagram of the Segment Proposal Network (SPN) is shown in Figure 2.

We train the SPN network by jointly optimizing the binary proposal classification and proposal boundary regression. The cross-entropy loss, denoted as $\mathcal{L}_{cls}$, is used for binary proposal classification. The smooth L1 loss, $\mathcal{L}_{reg}$, is used for proposal boundary regression and defined as

$$\mathcal{L}_{reg}(x) = \mathbb{1}(|x|<1)\frac{1}{2}x^2 + \mathbb{1}(|x|\geq 1)(|x|-\frac{1}{2}) \quad (1)$$

where $\mathbb{1}(\cdot)$ is indicator function. The joint loss is given by

$$\mathcal{L}_{spn} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{cls}(\hat{a}_i, a_i^*) + a_i^*\left(\mathcal{L}_{reg}(\delta\hat{c}_i\text{-}\delta c_i^*) + \mathcal{L}_{reg}(\delta\hat{l}_i\text{-}\delta l_i^*)\right) \quad (2)$$

The Hierarchical Captioning Module: We encode predicted variable-length proposals into features $I_{p,t}$ through 3D SoI Pooling [2] and the fc6 layer of the C3D network. To represent visual context, we encode the entire input video segment V as a vector I_C using a max pooling layer and the shared fc6 layer. We adopt a hierarchical LSTM structure to model relationships between the generated caption sentences. The high-level Controller LSTM records the visual context and sentence decoding history. The low-level Captioning LSTM decodes every proposal into a caption word by word, while being aware of visual and language context. Figure 3 illustrates this hierarchical structure.

We optimize the normalized log likelihood over all T ground truth proposals and all K unrolled timesteps in captioning module:

$$\mathcal{L}_{caption} = -\frac{1}{KT}\sum_{t,k}\log P(w_k^t | I_{p,t}, h_t^c, w_1^t, ..., w_{k-1}^t) \quad (3)$$

End-to-End Optimization: JEDDi-Net can be trained end-to-end with the proposal and hierarchical captioning modules optimized jointly. The overall loss is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{spn} + \lambda\mathcal{L}_{caption} \quad (4)$$

2. Experiments

We evaluate JEDDi-Net on the large-scale ActivityNet Captions dataset [1], by computing the average precision (BLEU, METEOR, CIDEr and ROUGE_L) across tIoU thresholds of 0.3, 0.5, 0.7, 0.9 for the top 1000 proposals. The number of frames $\mathcal{L}$ is set to 768 sampled at 3fps. The maximum caption length is set to 30, which covers over

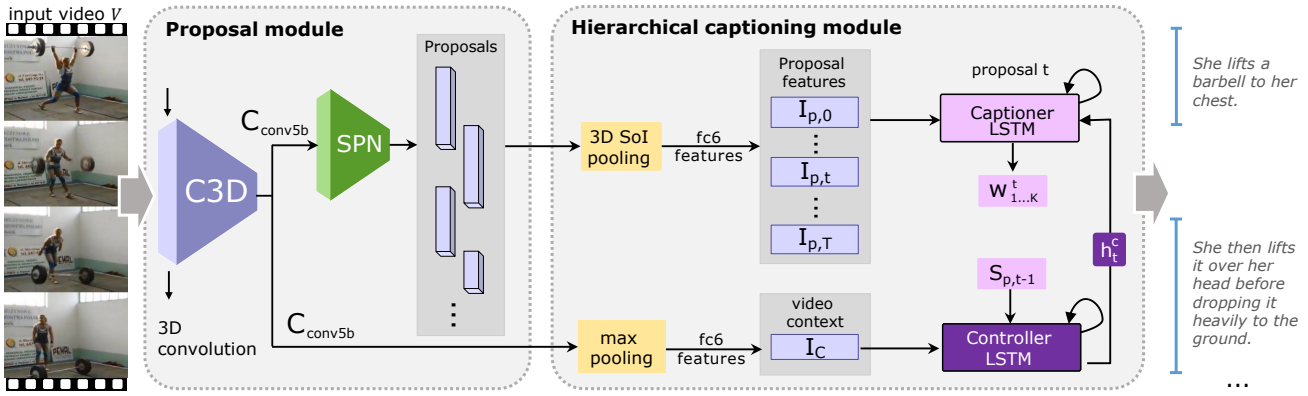

Figure 1. The overall architecture of our proposed *Joint Event Detection and Description Network (JEDDi-Net)* consists of two modules. The proposal module extracts features with 3D convolutional layers (C3D) and uses a Segment Proposal Network (SPN) to generate candidate segment proposals (see Fig. 2 for details). The hierarchical captioning module contains a controller LSTM to fuse the visual context I_c and the decoded language context $S_{p,t-1}$, and provides its hidden state h_t^c to the captioner LSTM, which decodes the next sentence. Details of these LSTMs are shown in Fig. 3.

Table 1. Dense video captioning results on ActivityNet Captions dataset (in percentage). The average Bleu_1-4 (B1-B4), METEOR (M), CIDEr (C) and ROUGE_L (R) across tIoU thresholds of 0.3, 0.5, 0.7, 0.9 are reported.

Model	B1	B2	B3	B4	M	C	R
R. Krishna et al. [1] (no context)	12.23	3.48	2.1	0.88	3.76	12.34	-
R. Krishna et al. [1] (with context)	17.95	7.69	3.86	**2.20**	4.82	17.29	-
JEDDi-Net(separate training)	16.72	6.65	2.65	1.07	7.37	14.65	16.47
JEDDi-Net(joint training)	19.27	8.69	3.78	1.54	8.30	19.81	18.86
JEDDi-Net(joint training w/ context)	**19.97**	**9.10**	**4.06**	1.63	**8.58**	**19.88**	**19.63**
JEDDi-Net(joint training w/ context) on test server	-	-	-	-	8.81	-	-

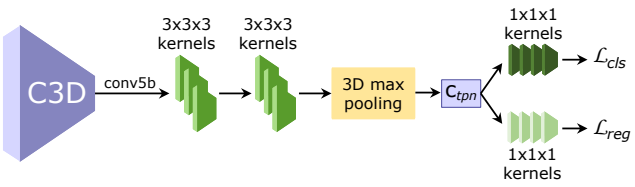

Figure 2. The structure of the Segment Proposal Network.

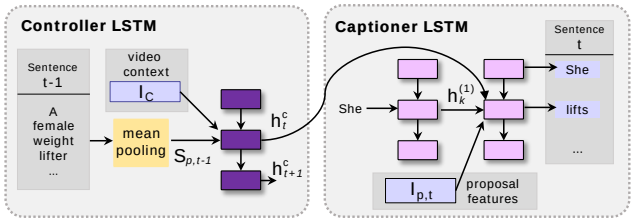

Figure 3. The LSTM structures in hierarchical captioning module.

97% of captions in the training set. The hidden state dimension is 20 in the controller LSTM and 512 in the captioner LSTMs.

In Table 1, our first 'JEDDi-Net(separate training)' model without end-to-end training already achieves reasonable results with a METEOR score 2.55% higher than the best context model in [1]. This indicates that our decoded captions are more semantically meaningful and closer to human descriptions. These results further motivate our proposal feature encoding method, which employs 3D SoI pooling directly on the conv features of the input video segment, rather than using the LSTM hidden state for a set of proposals. After end-to-end training, both 'JEDDi-Net(joint training)' and 'JEDDi-Net(joint training with context)' improve on all evaluation metrics compared to 'JEDDi-Net(separate training)'. This shows the benefits of joint parameter training for dense video captioning. Our 'JEDDi-Net(joint training with context)' model that incorporates visual and language context further improves all the language evaluation metrics compared to the no context version. Applying the same JEDDi-Net(joint training with context) on the test server yields an average METEOR score of 8.81%.

References

[1] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 2

[2] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 1