Less is More: Picking Informative Frames for Video Captioning

ECCV 2018

Yangyu Chen[1], Shuhui Wang[2*], Weigang Zhang[3] and Qingming Huang[1,2]

[1]University of Chinese Academy of Science, Beijing, 100049, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
[3]Harbin Inst. of Tech, Weihai, 264200, China
yangyu.chen@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, wgzhang@hit.edu.cn, qmhuang@ucas.ac.cn

2018-07-30

Video Captioning

- Seq2Seq translation:
  - encoding: use CNN and RNN to encode video content
  - decoding: use RNN to generate sentence conditioning on encoded feature

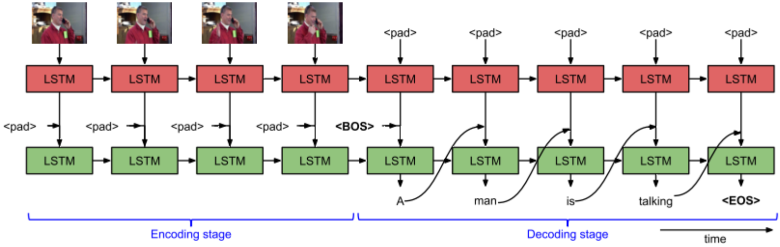
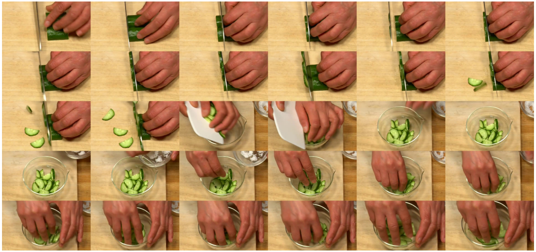


Figure 1: Standard encoder-decoder framework for video captioning[1]

[1] S. Venugopalan et al. “Sequence to sequence - video to text”. In: *Proceedings of IEEE International Conference on Computer Vision*. Santiago: IEEE Computer Society Press, 2015, pp. 4534–4542.

Motivation

- Frame selection perspective: there are many frames with duplicated and redundant visual appearance information selected with equal interval frame sampling, and this will also involve remarkable computation expenditures.



(a) Equally sampled 30 frames from a video

(b) Informative frames

Figure 2: Video may contains many redundant information. The whole video can be represented by a small portion of frames (b), while equally sampled frames still contain redundant information (a).

Motivation

- Downstream task perspective: temporal redundancy may lead to an unexpected information overload on the visual-linguistic correlation analysis model, hence using more frames may not always lead to better performance.

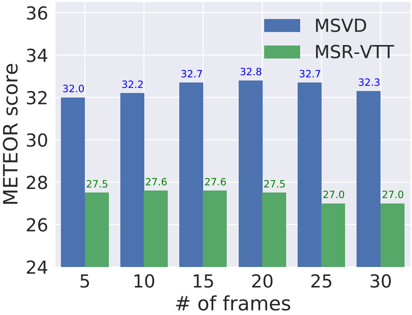


Figure 3: The best METEOR score on the validation set of MSVD and MSR-VTT when using different number of equally sampled frames. The standard Encoder-Decoder model is used to generate captions.

Picking Informative Frames for Captioning

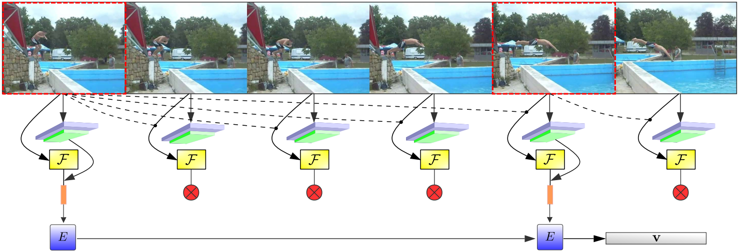
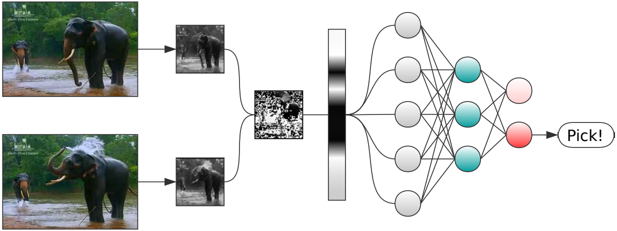


Figure 4: Insert PickNet into the encode-decode procedure for captioning.

- Insert PickNet before encoder-decoder.
  - Perform frame selection before processing downstream task.
  - Without annotations, we can try reinforcement training to optimize picking policy.

PickNet



Given an input image $\mathbf{z}_t$, and the last picking memory $\tilde{\mathbf{g}}$, PickNet produce a Bernoulli distribution for selecting decision:

$$\mathbf{d}_t = \mathbf{g}_t - \tilde{\mathbf{g}} \tag{1}$$

$$\mathbf{s}_t = W_2(\max(W_1\mathrm{vec}(\mathbf{d}_t) + \mathbf{b}_1, \mathbf{0})) + \mathbf{b}_2 \tag{2}$$

$$a_t \sim \mathrm{softmax}(\mathbf{s}_t) \tag{3}$$

$$\tilde{\mathbf{g}} \leftarrow \mathbf{g}_t \tag{4}$$

where W_* and $\mathbf{b}_*$ are parameters of our model, $\mathbf{g}_t$ is the flattened gray-scale image, $\mathbf{d}_t$ is the difference between gray-scale images.
Other network structures (*e.g.*, LSTM/GRU) can also be applied.

Rewards

- Visual diversity reward: the average cosine distance of each frame pairs

$$r_v(\mathcal{V}_i) = \frac{2}{N_p(N_p-1)} \sum_{k=1}^{N_p-1} \sum_{m>k}^{N_p} \left(1 - \frac{\mathbf{x}_k^{\mathbf{T}} \mathbf{x}_m}{\|\mathbf{x}_k\|_2 \|\mathbf{x}_m\|_2}\right) \tag{5}$$

  - where $\mathcal{V}_i$ is a set of picked frames, N_p the number of picked frames, $\mathbf{x}_k$ the feature of k-th picked frame.

- Language reward: the semantic similarity between generated sentence and ground-truth

$$r_l(\mathcal{V}_i, S_i) = \mathsf{CIDEr}(c_i, S_i) \tag{6}$$

  - S_i is a set of annotated sentences, c_i is the generated sentence

- Picking limitation

$$r(\mathcal{V}_i) = \begin{cases} \lambda_l r_l(\mathcal{V}_i, S_i) + \lambda_v r_v(\mathcal{V}_i) & \text{if} \quad N_{\mathsf{min}} \le N_p \le N_{\mathsf{max}} \\ R^- & \text{otherwise}, \end{cases} \tag{7}$$

  - N_p is the number of picked frames, R^- is the punishment

Training

- Supervision stage: training the encoder-decoder.

$$L_{\mathsf{X}}(\mathbf{y};\omega) = -\sum_{t=1}^{m} \log(p_\omega(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{y}_{t-2},\ldots\mathbf{y}_1,\mathbf{v})) \quad (8)$$

  - ω is the parameter of encoder-decoder, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m)$ is an annotated sentence, $\mathbf{v}$ is the encoded result

- Reinforcement stage: training PickNet.
  - the relation between reward and action$\mathcal{V}_i = \{\mathbf{x}_t | a_t^s = 1 \land \mathbf{x}_t \in v_i\}$

$$L_R(\mathbf{a}^s;\theta) = -\mathbb{E}_{\mathbf{a}^s \sim p_\theta}[r(\mathcal{V}_i)] = -\mathbb{E}_{\mathbf{a}^s \sim p_\theta}[r(\mathbf{a}^s)] \quad (9)$$

  - θ is the parameter of PickNet$\mathbf{a}^s$ is the action sequence

- Adaptation stage: training both encoder-decoder and PickNet.

$$L = L_{\mathsf{X}}(\mathbf{y};\omega) + L_R(\mathbf{a}^s;\theta) \quad (10)$$

The combinatorial explosion of direct frame selection is avoided.

REINFORCE

- Use REINFORCE[2] algorithm to estimate gradients.
- Gradient expression:

$$\nabla_\theta L_R(\mathbf{a}^s;\theta) = -\mathbb{E}_{\mathbf{a}^s \sim p_\theta}\left[r(\mathbf{a}^s)\nabla_\theta \log p_\theta(\mathbf{a}^s)\right] \tag{11}$$

- Based on chain-ruler:

$$\nabla_\theta L_R(\mathbf{a}^s;\theta) = \sum_{t=1}^{T} \frac{\partial L_R(\theta)}{\partial \mathbf{s}_t}\frac{\partial \mathbf{s}_t}{\partial \theta} = \sum_{t=1}^{T} -\mathbb{E}_{\mathbf{a}^s \sim p_\theta} r(\mathbf{a}^s)(p_\theta(a_t^s) - \mathbf{1}_{a_t^s})\frac{\partial \mathbf{s}_t}{\partial \theta} \tag{12}$$

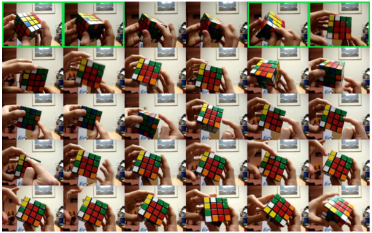
- Apply Monte-Carlo sampling:

$$\nabla_\theta L_R(\mathbf{a}^s;\theta) \approx -\sum_{t=1}^{T} r(\mathbf{a}^s)(p_\theta(a_t^s) - \mathbf{1}_{a_t^s})\frac{\partial \mathbf{s}_t}{\partial \theta} \tag{13}$$

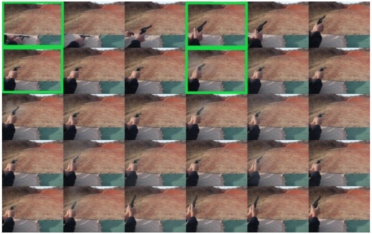
[2] R. J. Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3-4 (1992), pp. 229–256.

Picking Results

Ours: a woman is seasoning meat
GT: someone is seasoning meat



Ours: a person is solving a rubik's cube
GT: person playing with toy



Ours: a man is shooting a gun
GT: a man is shooting

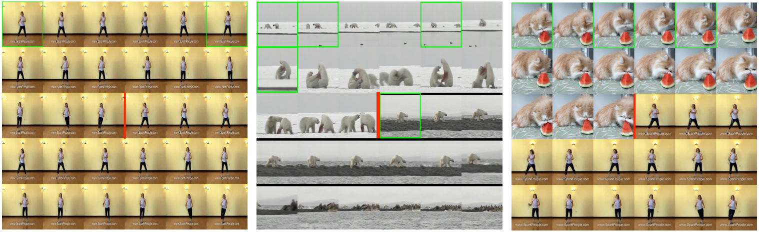
Ours: there is a woman is talking with a woman
GT: it is a movie

Figure 5: Example results on MSVD and MSR-VTT. The green boxes indicate picked frames.

Picking Results

We investigate our method on three types of artificially combined videos:

- a) two identical videos;
- b) two semantically similar videos;
- c) two semantically dissimilar videos.



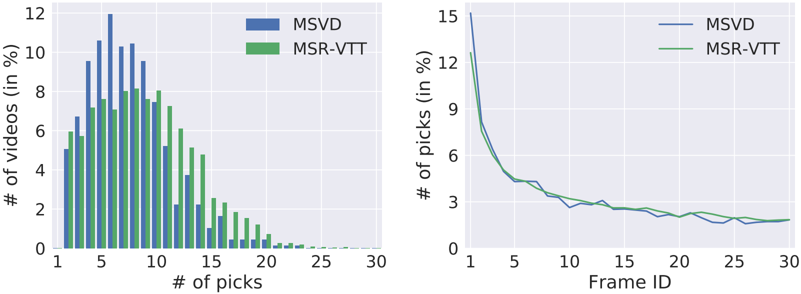
(a) **Ours: a woman is doing exercise**
Baseline: a man is dancing

(b) **Ours: two polar bears are playing**
Baseline: a bear is running

(c) **Ours: a cat is eating**
Baseline: a girl is doing a

Figure 6: Example results on joint videos. Green boxes indicate picked frames. The baseline method is Enc-Dec on equally sampled frames.

Analysis



(a) Distribution of the number of picks.

(b) Distribution of the position of picks.

Figure 7: Statistics on the behavior of our PickNet.

- In the vast majority of the videos, less than 10 frames are picked.
- The probability of picking a frame is reduced as time goes by.

Performance

model	BLEU4	ROUGE-L	METEOR	CIDEr	time
Previous Works					
LSTM-E	45.3	-	31.0	-	5x
p-RNN	49.9	-	32.6	65.8	5x
HRNE	43.8	-	33.1	-	33x
BA	42.5	-	32.4	63.5	12x
Baselines					
Full	44.8	68.5	31.6	69.4	5x
Random	35.6	64.5	28.4	49.2	2.5x
k-means (k=6)	45.2	68.5	32.4	70.9	1x
Hecate	43.2	67.4	31.7	68.8	1x
Our Models					
PickNet (V)	46.3	69.3	32.3	75.1	1x
PickNet (L)	49.9	69.3	32.9	74.7	1x
PickNet (V+L)	52.3	69.6	33.3	76.5	1x

Table 1: Experiment results on MSVD. All values are reported as percentage(%). L denotes using language reward and V denotes using visual diversity reward. k is set to the average number of picks $\bar{N}_p$ on MSVD. ($\bar{N}_p \approx 6$)

Performance

model	BLEU4	ROUGE-L	METEOR	CIDEr	time
Previous Works					
ruc-uva	38.7	58.7	26.9	45.9	4.5x
Aalto	39.8	59.8	26.9	45.7	4.5x
DenseVidCap	41.4	61.1	28.3	48.9	10.5x
MS-RNN	39.8	59.3	26.1	40.9	10x
Baselines					
Full	36.8	59.0	26.7	41.2	3.8x
Random	31.3	55.7	25.2	32.6	1.9x
k-means (k=8)	37.8	59.1	26.9	41.4	1x
Hecate	37.3	59.1	26.6	40.8	1x
Our Models					
PickNet (V)	36.9	58.9	26.8	40.4	1x
PickNet (L)	37.3	58.9	27.0	41.9	1x
PickNet (V+L)	39.4	59.7	27.3	42.3	1x
PickNet (V+L+C)	41.3	59.8	27.7	44.1	1x

Table 2: Experiment results on MSR-VTT. All values are reported as percentage(%). C denotes using the provided category information. k is set to the average number of picks $\bar{N}_p$ on MSR-VTT. ($\bar{N}_p \approx 8$)

Time Estimation

Model	Appearance	Motion	Sampling method	Frame num.	Time
			Previous Work		
LSTM-	VGG (0.5x)	C3D (2x)	uniform sampling 30 frames	30 (5x)	5x
p-RNN	VGG (0.5x)	C3D (2x)	uniform sampling 30 frames	30 (5x)	5x
HRNE	GoogleNet (0.5x)	C3D (2x)	first 200 frames	200 (33x)	33x
BA	ResNet (0.5x)	C3D (2x)	every 5 frames	72 (12x)	12x
			Our Models		
Baseline	ResNet (1x)	$\times$	uniform sampling 30 frames	30 (5x)	5x
Random	ResNet (1x)	$\times$	randomly sampling	15 (2.5x)	2.5x
k-means (k=6)	ResNet (1x)	$\times$	k-means clustering	6 (1x)	1x
Hecate	ResNet (1x)	$\times$	video summarization	6 (1x)	1x
PickNet (V)	ResNet (1x)	$\times$	picking	6 (1x)	1x
PickNet (L)	ResNet (1x)	$\times$	picking	6 (1x)	1x
PickNet (V+L)	ResNet (1x)	$\times$	picking	6 (1x)	1x

Table 3: Running time estimation on MSVD. OF means optical flow. BA uses ResNet50 while our models use ResNet152. k is set to the average number of picks $\bar{N}_p$ on MSVD. ($\bar{N}_p \approx 6$)

Time Estimation

Model	Appearance	Motion	Sampling method	Frame num.	Time
Previous Work					
ruc-uva	GoogleNet (0.5x)	C3D (2x)	every 10 frames	36 (4.5x)	4.5x
Aalto	GoogleNet (0.5x)	C3D+IDT (2x)	one frame every second	36 (4.5x)	4.5x
DenseCap	ResNet (0.5x)	C3D (2x)	sampling 90 frames	90 (10.5x)	10.5x
MS-RNN	ResNet (1x)	C3D (2x)	uniform sampling 40 frames	40 (5x)	10x
Our Models					
Baseline	ResNet (1x)	$\times$	uniform sampling 30 frames	30 (3.8x)	3.8x
Random	ResNet (1x)	$\times$	randomly sampling	15 (1.9x)	1.9x
k-means (k=8)	ResNet (1x)	$\times$	k-means clustering	8 (1x)	1x
Hecate	ResNet (1x)	$\times$	video summarization	8 (1x)	1x
PickNet (V)	ResNet (1x)	$\times$	picking	8 (1x)	1x
PickNet (L)	ResNet (1x)	$\times$	picking	8 (1x)	1x
PickNet (V+L)	ResNet (1x)	$\times$	picking	8 (1x)	1x

Table 4: Running time estimation on MSR-VTT. IDT means improved dense trajectory. DenseCap uses ResNet50 while our models use ResNet152. k is set to the average number of picks $\bar{N}_p$ on MSR-VTT. ($\bar{N}_p \approx 8$)

Online Captioning

- When PickNet select one frame, it means that new information appears.
- Then the encode-decoder is triggered by PickNet and a more detailed description is generated.

(a) a cat is licking its lips → a woman is a baby → a woman is a baby → a woman is feeding a baby → a woman is playing with a kitten

(b) a boy is running → a boy is running → a boy is running → the boys are dancing → three persons are dancing

(c) a man is a sword → a boy is doing a → a man with a sword stabs a target → a man is stabbing a silhouette with a sword ×2

Conclusion

- Flexibility. a plug-and-play reinforcement-learning-based PickNet to pick informative frames for video understanding tasks.
- Efficiency. The architecture can largely cut down the usage of convolution operations. It makes our method more applicable for real-world video processing.
- Effectiveness. Experiment shows that our model can achieve comparable or even better performance compared to state-of-the-art while only a small number of frames are used.

Thanks!