# Less Is More: Picking Informative Frames for Video Captioning

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

[1]University of Chinese Academy of Science, [2]Key Lab of Intelligence Information Processing, Institute of Computing Technology, CAS, [3]Harbin Institute of Technology

yangyu.chen@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, wgzhang@hit.edu.cn, qmhuang@ucas.ac.cn

webpage: https://yugnaynehc.github.io/picknet

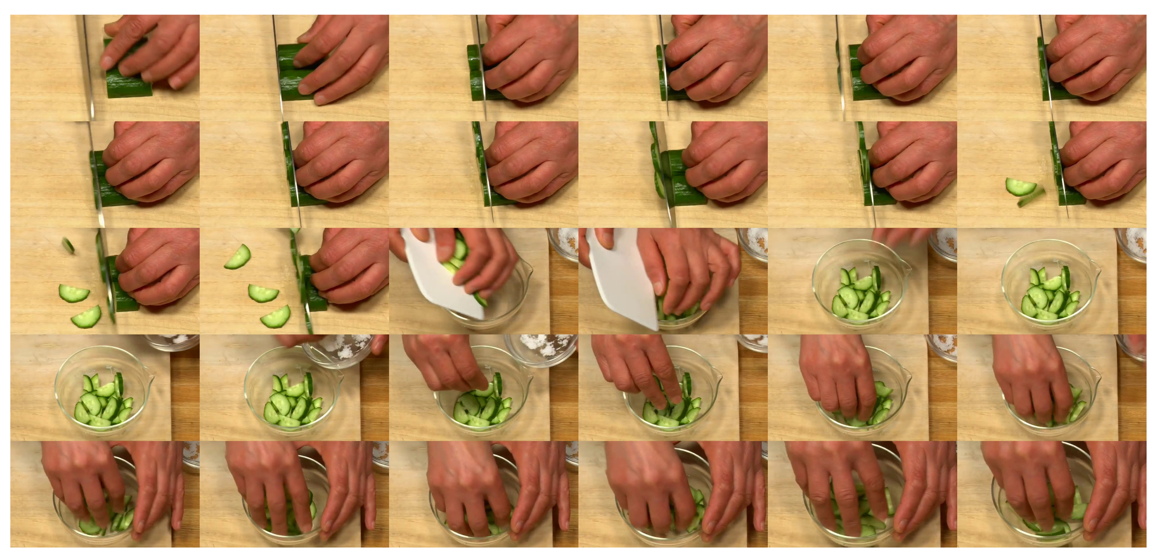
## Motivation

Existing study:

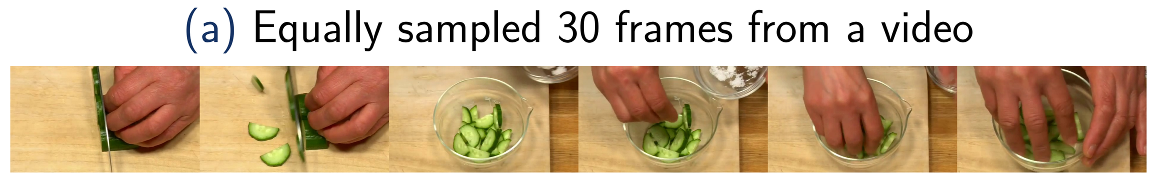
- models frame-level appearance and motion on equal interval frame sampling

But it:

- may bring about redundant visual information
- will be sensitivity to content noise
- leads to unnecessary computation cost



(a) Equally sampled 30 frames from a video



(b) Informative frames

Figure 1: Equally sampled frames contain redundancy.

We propose a plug-and-play PickNet to perform informative frame picking in video captioning.

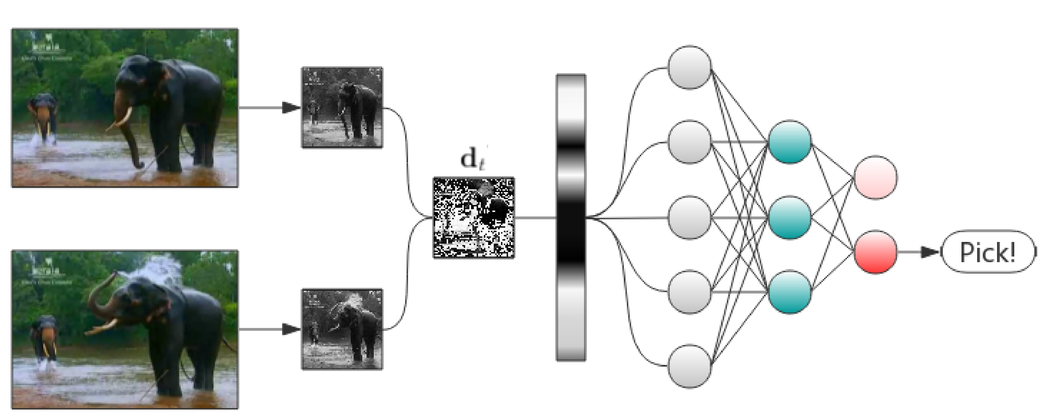



Figure 2: The architecture of PickNet.

- PickNet produces a Bernoulli distribution for selecting decision:

$$\mathbf{s}_t = W_2(\max(W_1 \mathrm{vec}(\mathbf{d}_t) + \mathbf{b}_1, \mathbf{0})) + \mathbf{b}_2 \quad (1)$$

$$a_t \sim \mathrm{softmax}(\mathbf{s}_t) \quad (2)$$

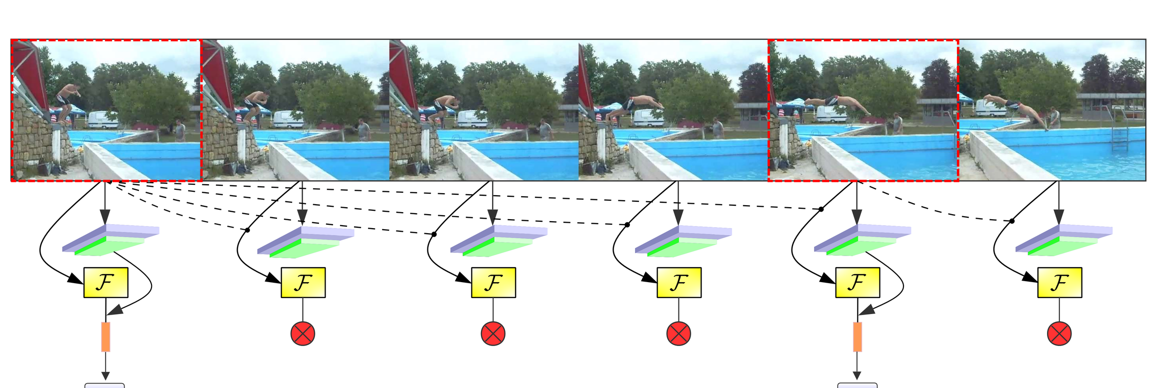


Figure 3: The framework. $\mathcal{F}$ denotes PickNet, $E$ is the encoder unit and $\mathbf{v}$ is the encoded video representation.

## Method

### Rewards

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

$$r_v(\mathcal{V}_i) = \frac{1}{\binom{N_p}{2}} \sum_{k=1}^{N_p-1} \sum_{m>k}^{N_p} \left(1 - \frac{\mathbf{x}_k^{\mathbf{T}} \mathbf{x}_m}{\|\mathbf{x}_k\|_2 \|\mathbf{x}_m\|_2}\right) \quad (3)$$

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

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

$$r_l(\mathcal{V}_i, S_i) = \mathrm{CIDEr}(c_i, S_i) \quad (4)$$

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

- Picking limitation: the final reward $r(\mathbf{a}^s) =$

$$\begin{cases} \lambda_l r_l(\mathcal{V}_i, S_i) + \lambda_v r_v(\mathcal{V}_i) & N_p \in [N_{\min}, N_{\max}] \\ R^- & \text{otherwise,} \end{cases} \quad (5)$$

  - $N_p$ is the number of picked frames, $R^-$ is the punishment, and $\mathcal{V}_i = \{\mathbf{x}_t | a_t^s = 1, \forall a_t^s \in \mathbf{a}^s\}$.

- Supervision stage: training encoder-decoder.

$$L_X(\mathbf{y}; \omega) = -\sum_{t=1}^{m} \log(p_\omega(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots \mathbf{y}_1, \mathbf{v})) \quad (6)$$

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

- Reinforcement stage: training PickNet.

$$L_R(\mathbf{a}^s; \theta) = -\mathbb{E}_{\mathbf{a}^s \sim p_\theta}[r(\mathcal{V}_i)] = -\mathbb{E}_{\mathbf{a}^s \sim p_\theta}[r(\mathbf{a}^s)] \quad (7)$$

  - $\mathbf{a}^s$ is the action sequence and $\theta$ is the parameter of PickNet.
  - Using REINFORCE algorithm to estimate gradient:

$$\nabla_\theta L_R(\mathbf{a}^s; \theta) = -\mathbb{E}_{\mathbf{a}^s \sim p_\theta}[r(\mathbf{a}^s) \nabla_\theta \log p_\theta(\mathbf{a}^s)] \quad (8)$$

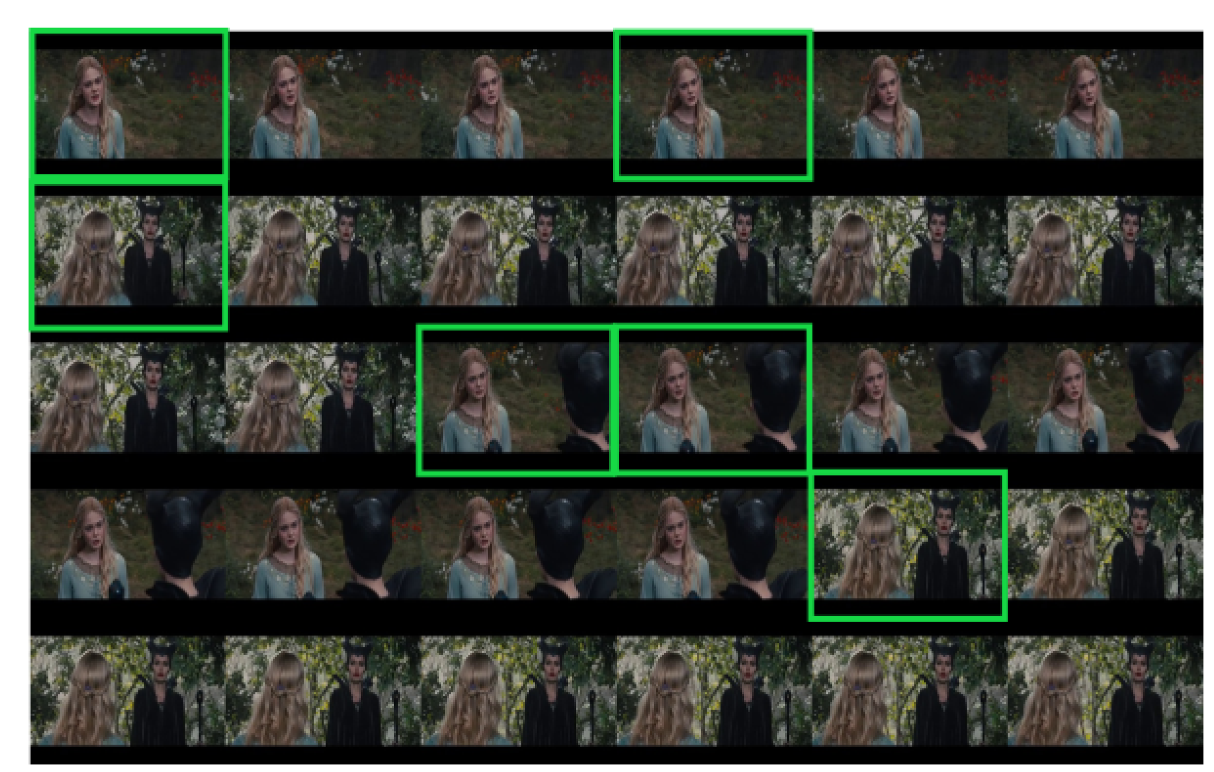
$$\approx -\sum_{t=1}^{T} r(\mathbf{a}^s)(p_\theta(a_t^s) - \mathbf{1}_{a_t^s}) \frac{\partial \mathbf{s}_t}{\partial \theta} \quad (9)$$

- Adaptation stage: training both modules.

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

## Visualization

**Ours: a person is playing a video game**
**GT: a game is being played**



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

Figure 4: Example results on the test set of MSR-VTT. The green boxes indicate picked frames.

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

Figure 5: Example results of online video captioning.

## Performance

| Model | BLEU4 | ROUGE-L | METEOR | CIDEr | Time |
|---|---|---|---|---|---|
| Previous Work | | | | | |
| 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 |
| Baseline Models | | | | | |
| 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. 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. ($\bar{N}_p \approx 6$)

| Model | BLEU4 | ROUGE-L | METEOR | CIDEr | Time |
|---|---|---|---|---|---|
| Previous Work | | | | | |
| 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 |
| Baseline Models | | | | | |
| 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. C denotes using the provided category information. $k$ is set to the average number of picks $\bar{N}_p$ on MSR-VTT. ($\bar{N}_p \approx 8$)

## Conclusion

- Flexibility. A plug-and-play RL-based PickNet is designed to select informative frames for video understanding tasks.
- Efficiency. The PickNet can largely cut down the convolution operations and makes this method more applicable for real-world video processing.
- Effectiveness. Experiment shows that our model can achieve comparable performance compared to state-of-the-art with less frames.