Weakly Supervised Dense Video Captioning

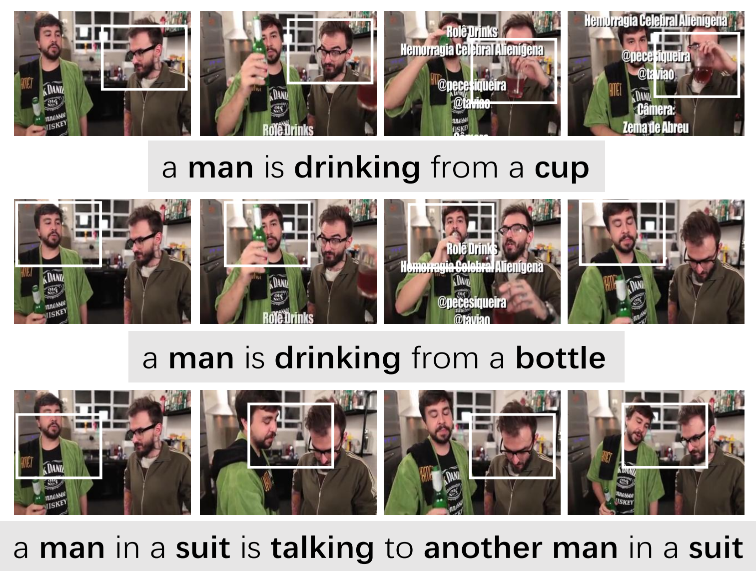
Zhiqiang Shen†, Jianguo Li‡, Zhou Su‡, Minjun Li†, Yurong Chen‡, Yu-Gang Jiang†, Xiangyang Xue†

†Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University ‡Intel Labs China

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017

Illustration of *DenseVidCap*



Ground-truth

two men are drinking alcohol

two men are talking indoors

credits are shown as two men have a discussion

a man with a bottle clinks the glass of another and both take a drink

two men are talking about something and drinking something

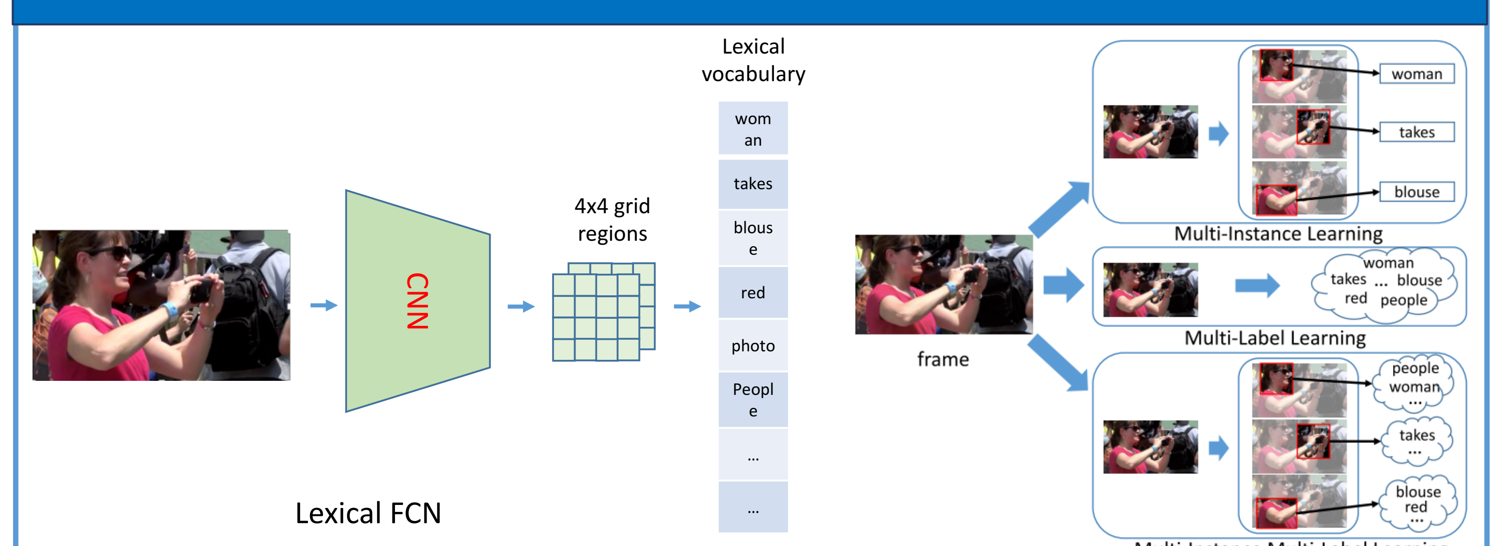
Contribution

(1) To the best of our knowledge, this is the first work for dense video captioning with only video-level sentence annotations.

(2) We propose a novel dense video captioning approach, which models visual cues with Lexical-FCN, discovers region-sequence with submodular maximization, and decodes language outputs with sequence-to-sequence learning. Although trained with weakly supervised signal, it can produce *informative* and *diverse* captions.

(3) We evaluate dense captioning results by measuring the performance gap to oracle results, and diversity of the dense captions. The best single caption outperforms the state-of-the-art results on the MSR-VTT challenge significantly.

Lexical FCN Model



We define the loss function for a bag of instances. As each bag has multiple word labels, we adopt the cross- entropy loss to measure the multi-label errors:

$$L(\mathbf{X},\mathbf{y};\theta)=-\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{y}_i\cdot\log\hat{\mathbf{p}}_i+(1-\mathbf{y}_i)\cdot\log(1-\hat{\mathbf{p}}_i)\right]$$

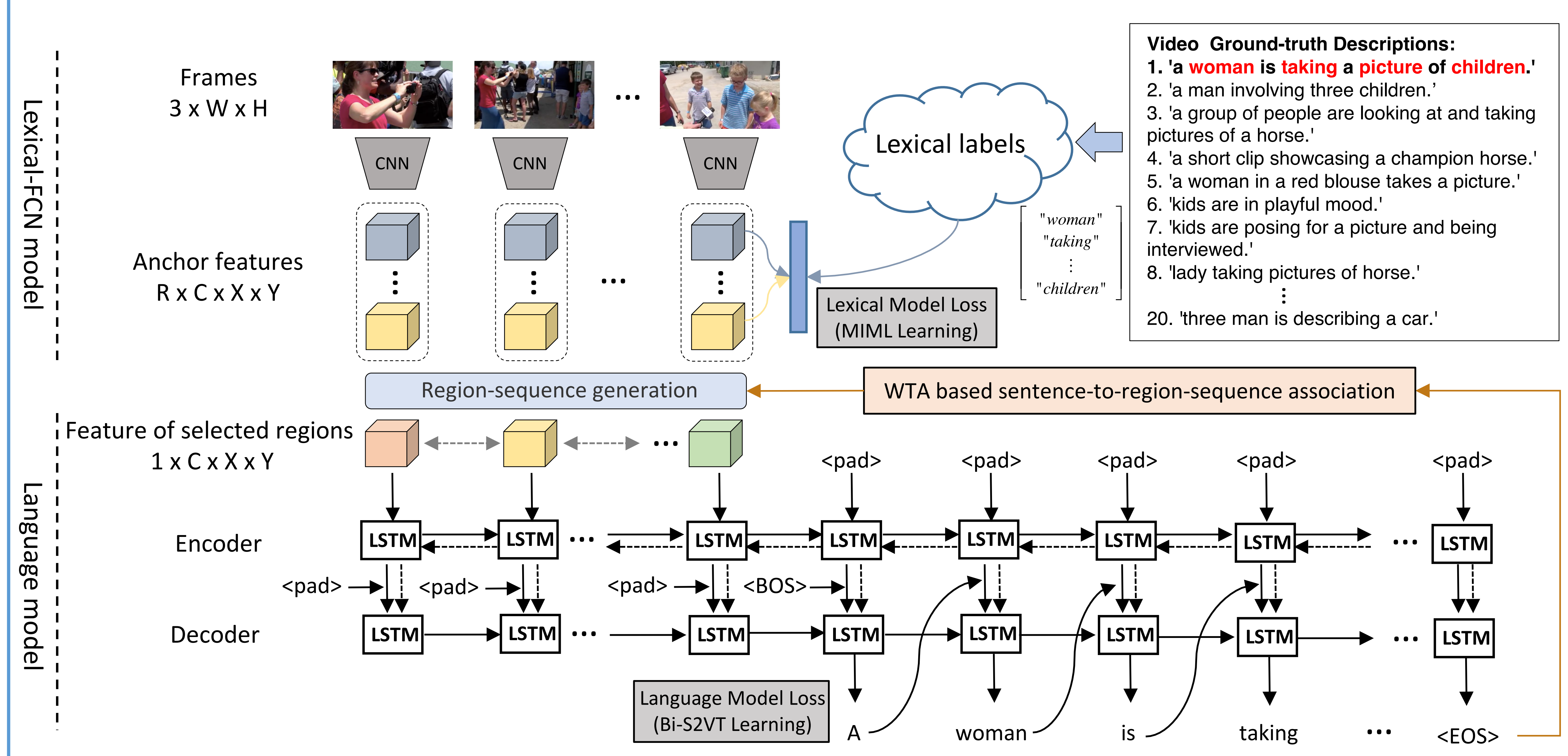
where θ is the model parameters, N is the number of bags, $\mathbf{y}_i$ is the label vector for bag $\mathbf{X}_i$, and $\hat{\mathbf{p}}_i$ is the corresponding probability vector.

We use a noisy-OR formulation to combine the probabilities that the individual instances in the bag are negative:

$$\hat{p}_i^w=P(y_i^w=1\mid\mathbf{X}_i;\theta)=1-\prod_{\mathbf{x}_{ij}\in\mathbf{X}_i}(1-P(y_i^w=1\mid\mathbf{x}_{ij};\theta))$$

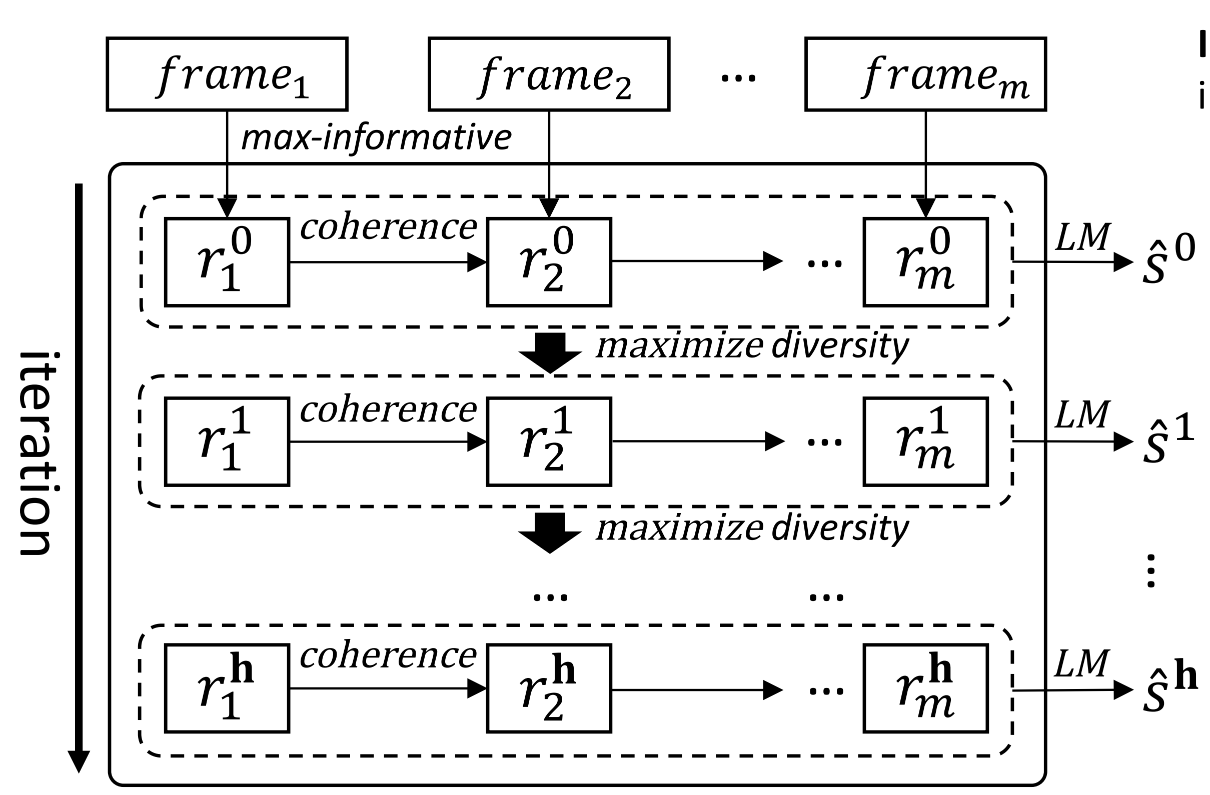
where $\hat{p}_i^w$ is the probability when word w in the i-th bag is positive.

Overview



Overview of our *Dense Video Captioning* framework. In the language model, <BOS> denotes the begin-of-sentence tag and <EOS> denotes the end-of-sentence tag. We use zeros as <pad> when there is no input at the time step.

Region-Sequence Generation



Informativeness of a region-sequence is defined as the sum of each region's informativeness:

$$f_{\mathrm{inf}}(\mathbf{x}_v,\mathcal{A}_t)=\sum_w p^w;\quad p^w=\max_{i\in\mathcal{A}_t}p_i^w$$

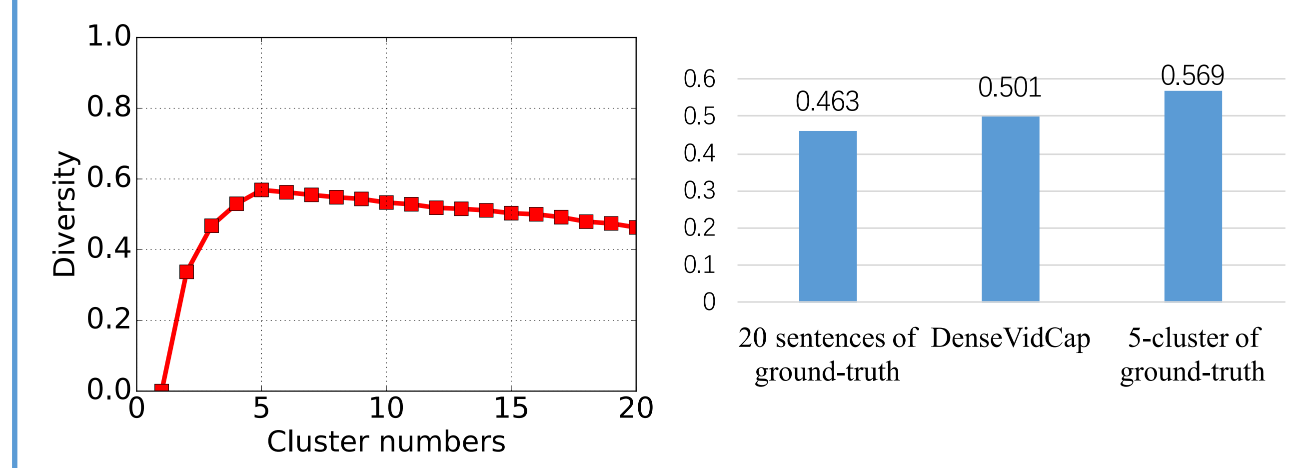
Coherence aims to ensure the temporal coherence of the region-sequence, we try to select regions with the smallest changes temporally, and define the coherence component as:

$$f_{coh}=\sum_{r_s\in\mathcal{A}_{t-1}}\langle\mathbf{x}_{r_t},\mathbf{x}_{r_s}\rangle$$

Diversity measures the degree of difference between a candidate region-sequence and all the existing region-sequences, which is defined with the Kullback-Leibler divergence as:

$$f_{div}=\sum_{i=1}^{N}\int_w p_i^w\log\frac{p_i^w}{q^w}dw$$

Diversity of Dense Captions



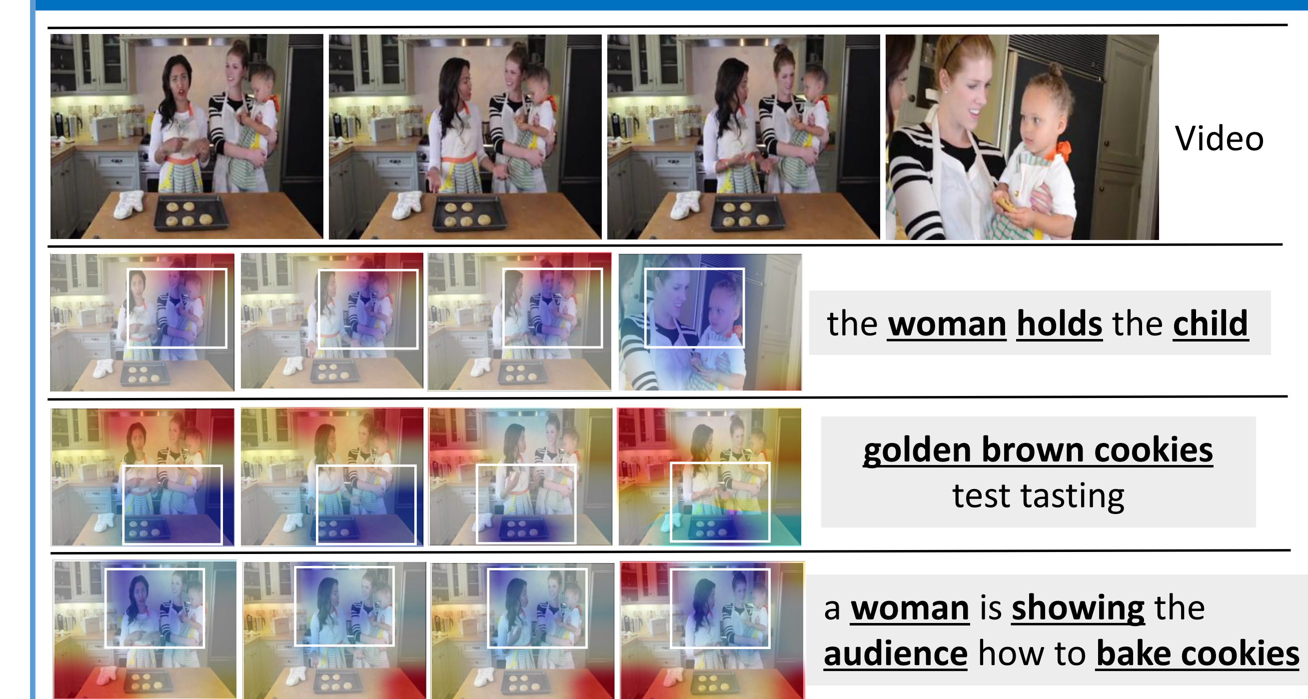
Left: Diversity score of clustered ground-truth captions under different cluster numbers. Right: Diversity score comparison of our automatic method (middle) and the ground-truth.

The diversity is calculated as:

$$D_{div}=\frac{1}{n}\sum_{\mathbf{s}^i,\mathbf{s}^j\in\mathbf{S};\,i\neq j}(1-\langle\mathbf{s}^i,\mathbf{s}^j\rangle)$$

where $\mathbf{S}$ is the sentence set with cardinality n, and $\langle\mathbf{s}^i,\mathbf{s}^j\rangle$ denotes the cosine similarity between $\mathbf{s}^i$ and $\mathbf{s}^j$.

Visualization



Visualization of learned response maps from the last CNN layer (left), and the corresponding natural sentences (right). The blue areas in the response maps are of high attention, and the region-sequences are highlighted in white bounding-boxes.

Results on MSR-VTT 2016 dataset

Model	METEOR	BLEU@4	ROUGE-L	CIDEr
ruc-uva	26.9	38.7	58.7	45.9
VideoLAB	27.7	39.1	60.6	44.1
Aalto	26.9	39.8	59.8	45.7
V2t_navigator	28.2	40.8	60.9	44.8
Ours	28.3	41.4	61.1	48.9