Weakly supervised approaches for Dense Captioning

Iulia Duță
iuliaduta94@gmail.com

November 15, 2017

Bitdefender

Introduction

Image Dense Captioning

Video Dense Captioning

Captioning: definition

- **Captioning**: the task of generating text descriptions of images/videos.

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Full-Image Captioning vs Dense-captioning I

- **Full-Image Captioning**: the task of generating a set of descriptions of the whole image/video

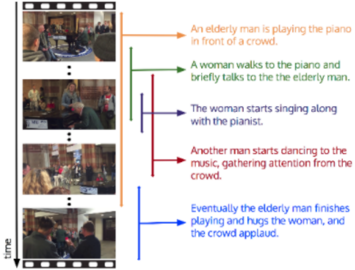
"man in black shirt is playing guitar."

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

  - easier to collect annotations
  - simpler input => simpler models
  - create general descriptions

Full-Image Captioning vs Dense-captioning II

- **Dense-captioning**: the task of generating a set of descriptions across regions of an image/ concurrent events in a video



  - hard to annotate
  - multiple instance models
  - more detailed, complementary descriptions

Bitdefender

Introduction

Image Dense Captioning

Video Dense Captioning

- Deep visual-semantic alignments for generating image description - Karpathy and Li [2014]

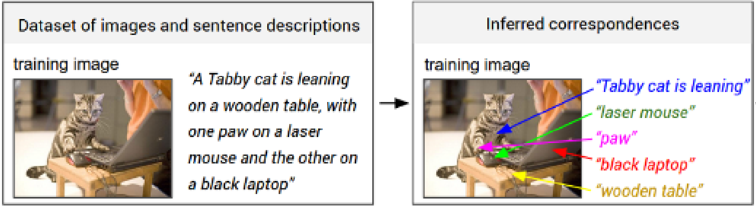
Overview

Bitdefender

- **goal**:
  - generate dense descriptions of images
- **problems**:
  - the model should build represantions for both image and language space
  - lack of datasets for dense image captioning
- **contributions**:
  - learn to infer the latent alignment between segments of sentences and the region of the image that they describe
  - create a multimodal RNN to generate dense captioning of an image

Align visual and language data I

- **given**: (full-image, sentences) pairs
- **goal**: generate relevant (visual regions, sentence snippets) pairs
- **motivation**: descriptions written by people make frequent references to certain locations in the image



Approach

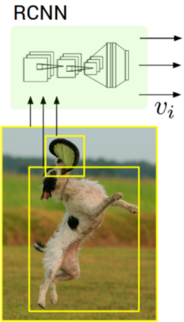
1. use RCNN + CNN for visual representation
2. use a bidirectional RNN to compute word representation
3. introduce a novel objective function

Image representation

Bitdefender

1. detect objects using Regional Convolutional Neural Network(RCNN) Girshick et al. [2013]
2. select top 19 detected locations
3. for each detected bounding box compute the representation:

$$v = W_m[CNN_{\theta_c}(I_b)] + b_m \tag{1}$$



- I_b - pixels inside each bounding box
- CNN_{θ_c} - 4096-dimensional activations of the FC immediatly before the classifier of a CNN
- W_m - embedding matrix 1600 x 4096

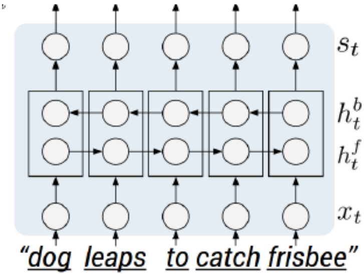
Sentence representation

- **solutions**:
  - *no context*: project every individual word into an embedding
  - *small context*: word bigram, dependency tree relations
  - *full context*: **compute representation using a Bidirectional Recurrent Neural Network (BRNN)**

Sentence representation

- **solutions**:
  - *no context*: project every individual word into an embedding
  - *small context*: word bigram, dependency tree relations
  - *full context*: **compute representation using a Bidirectional Recurrent Neural Network (BRNN)**

$$x_t = W_w 1_t \longrightarrow \text{word2vec(fix)}$$
$$e_t = f(W_e x_t + b_e) \longrightarrow \text{embedding}$$
$$h_t{}^f = f(e_t W_f h_{t-1}{}^f + b_f) \longrightarrow \text{forward pass}$$
$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b) \longrightarrow \text{backward pass}$$
$$s_t = f(W_d(f_t^f + h_t^b) + b_d) \longrightarrow \text{sentence representation}$$



Alignment model

- **remember**: no region-word annotation, the supervision is at the level of image-sentences
- **solution**: formulate an image-sentence score as a function of region-word score

Alignment model

- How similar are **i**th **region** and **t**th **word**?

$$v_i{}^T s_t$$

Alignment model

- How similar are **i**th **region** and **t**th **word**?

$$ {v_i}^T s_t $$

- How similar are **k**th **image** and **l**th **sentence**?

$$ S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} max(0, {v_i}^T s_t) $$

g_k - set of image fragments

g_l - set of sentence words

Alignment model

Bitdefender

- How similar are **i**th **region** and **t**th **word**?

$$v_i{}^T s_t$$

- How similar are **k**th **image** and **l**th **sentence**?

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} max(0, v_i{}^T s_t)$$

g_k - set of image fragments

g_l - set of sentence words

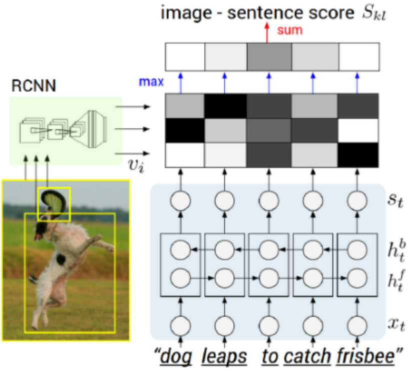
$$\Downarrow$$

$$S_{kl} = \sum_{t \in g_l} \max_{i \in g_k}(0, v_i{}^T s_t)$$

Alignment loss

- assume that (image k, sentence k) is a good match
- **Loss** to optimize:

$$C(\theta) = \sum_k [\sum_l max(0, S_{kl} - S_{kk} + 1) + \sum_l max(0, S_{lk} - S_{kk} + 1)] \tag{2}$$



From words to text-segment

- **problems**:
  - each word are assigned independently to a region
  - there are words that has no correspondence in image (stopwords)
  - naturally, continuous sequences of words are more likely allign to a single bounding box. Not in our case.

From words to text-segment

- **problems**:
  - each word are assigned independently to a region
  - there are words that has no correspondence in image (stopwords)
  - naturally, continuous sequences of words are more likely allign to a single bounding box. Not in our case.
- **solution**: formulate an energy function that encourage neighbouring words to be aligned to the same region

$$E(a) = \sum_{j=1..N} {v_{a_j}}^T s_j + \sum_{j=1..N-1} \beta 1[a_j = a_{j+1}]$$

$$a^* = \operatorname{argmax} E(a)$$

a_j - bounding box aligned to j^{th} word

β - controls the affinity towards longer phrases

From words to text-segment

- **problems**:
  - each word are assigned independently to a region
  - there are words that has no correspondence in image (stopwords)
  - naturally, continuous sequences of words are more likely allign to a single bounding box. Not in our case.
- **solution**: formulate an energy function that encourage neighbouring words to be aligned to the same region

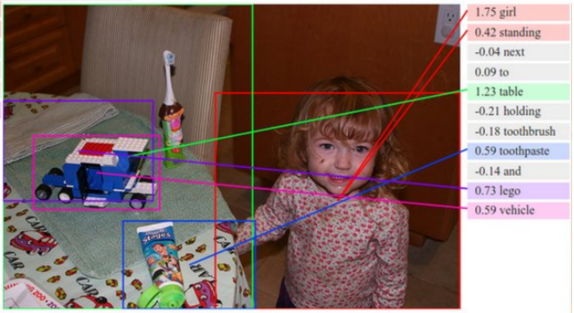
$$E(a) = \sum_{j=1..N} {v_{a_j}}^T s_j + \sum_{j=1..N-1} \beta 1[a_j = a_{j+1}]$$

$$a^* = \operatorname{argmax} E(a)$$

a_j - bounding box aligned to j^{th} word
β - controls the affinity towards longer phrases

- **goal**: given v_i and s_t (previous optimization), find best alignments **a** that maximize the energy - dynamic programming

Results



- the similarity measure $S_{kl} = \sum_{t \in g_l} \max_{i \in g_k}(0, {v_i}^T s_t)$ encourage **discriminative** entitites and discriminative words to have **higher magnitudes**

Results

Bitdefender

- **task**: Image-Sentence ranking experiments
- **experiment**: given a query, sort based on S_{kl}
- **metrics**:
  - R@K - fraction of times a correct item was found in top K
  - Med r - median rank of the closest ground truth in the list

	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med** *r*	**R@1**	**R@5**	**R@10**	**Med** *r*
Flickr30K								
SDT-RNN (Socher et al. [49])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [25]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [38]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Donahue et al. [8]	17.5	40.3	50.8	9	-	-	-	-
DeFrag (Karpathy et al. [24])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [24]	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
Our model: DepTree edges	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
Our model: BRNN	**22.2**	**48.2**	**61.4**	**4.8**	**15.2**	**37.7**	**50.5**	**9.2**
Vinyals et al. [54] (more powerful CNN)	23	-	63	5	17	-	57	8
MSCOCO								
Our model: 1K test images	38.4	69.9	80.5	1.0	27.4	60.2	74.8	3.0
Our model: 5K test images	16.5	39.2	52.0	9.0	10.7	29.6	42.2	14.0

Table 1. Image-Sentence ranking experiment results. **R@K** is Recall@K (high is good). **Med** *r* is the median rank (low is good). In the results for our models, we take the top 5 validation set models, evaluate each independently on the test set and then report the average performance. The standard deviations on the recall values range from approximately 0.5 to 1.0.

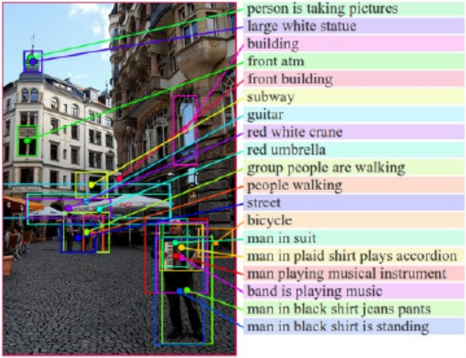
Generate description I

- Two kind of description:
  - **full-image** captioning: input = full image

boy is doing backflip on wakeboard.

Generate description II

- **dense** captioning: input = regions from previous model



Model

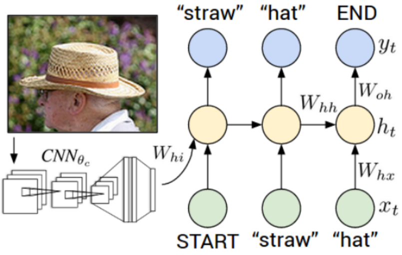
- standard architecture: CNN + RNN

$$
\begin{aligned}
b_v &= W_{hi}[CNN_{\theta_c}(I)] \\
h_t &= f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t = 1) \circ b_v) \\
y_t &= softmax(W_{oh}h_t + b_o)
\end{aligned}
$$

Model

- standard architecture: CNN + RNN

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + 1(t=1) \circ b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o)$$



Results

- **task**: Generate sentence from full image
- **experiment**: Given one image, generate sentence
- **metrics**: BLEU, METEOR, CIDEr

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	15.7	38.3
Mao et al. [38]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [54]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [8]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [12]	—	—	—	—	—	—	—	—	—	—	—	21.1	20.7	—
Chen and Zitnick [5]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	20.4	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

Table 2. Evaluation of full image predictions on 1,000 test images. **B-n** is BLEU score that uses up to n-grams. High is good in all columns. For future comparisons, our METEOR/CIDEr Flickr8K scores are 16.7/31.8 and the Flickr30K scores are 15.3/24.7.

Results

woman plays volleyball

women compete in volleyball match in london 2012 olympics

woman in bikini is jumping over hurdle

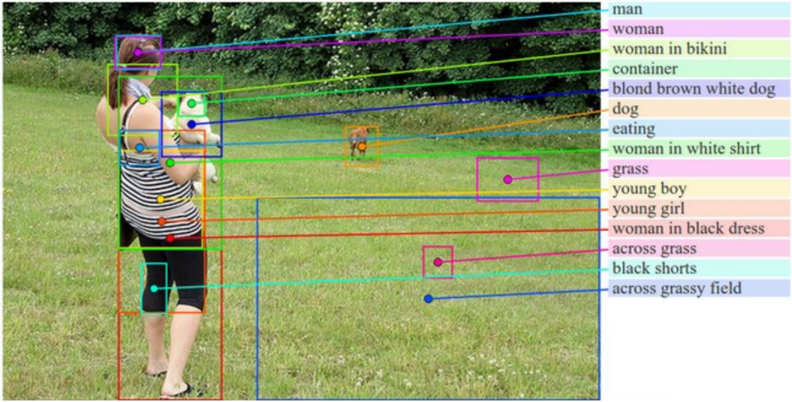
Results

- **task**: Generate snippets of text from regions of image
- **experiment**: Given one image, generate (regions, snippets) basen on alignments model, than generate captioning for each one
- Create a new dataset from AMT only for test time

Model	B-1	B-2	B-3	B-4
Human agreement	61.5	45.2	30.1	22.0
Nearest Neighbor	22.9	10.5	0.0	0.0
RNN: Fullframe model	14.2	6.0	2.2	0.0
RNN: Region level model	**35.2**	**23.0**	**16.1**	**14.8**

Table 3. BLEU score evaluation of image region annotations.

Results


man
woman
woman in bikini
container
blond brown white dog
dog
eating
woman in white shirt
grass
young boy
young girl
woman in black dress
across grass
black shorts
across grassy field

Bitdefender

Introduction

Image Dense Captioning

Video Dense Captioning

- Weakly Supervised Dense Video Captioning - Shen et al. [2017]

Overview

- **goal**: generate dense captioning for video
- **problems**:
  - no dense annotation for video-sequence corespondence
  - no explicit segmentation of video into sequences
- **contributions**:
  - novel dense video captioning approach
  - firs dense video captioning model with only video-level sentence annotation
  - create diverse captioning

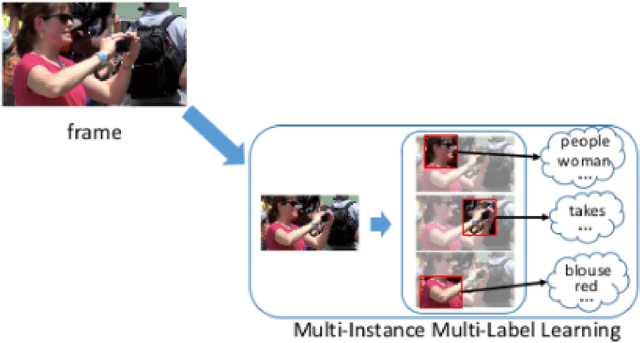

Video
the woman holds the child
golden brown cookies test tasting
a woman is showing the audience how to bake cookies

Approach

- visual sub-model - **Lexical FCN**
- discover region-sequence - **submodular maximization**
- language sub-model - **sequence-to-sequence**

Lexical FCN Model

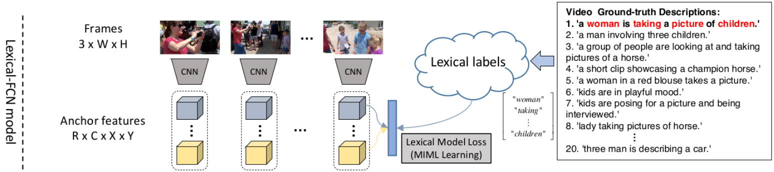
- learn good representation of each regions
- map frame regions to lexical labels



Lexical FCN Model

1. build a lexical vocabulary from training set
2. create a FCN model trained on ImageNet.
   - VGG-16 re-cast FC to Conv => $4x4x4096$
   - Resnet-50 delete final softmax layer => $4x4x2048$

   => 16 regions per frame, each having 4096/2048 chanels
3. sample frames, resize to 320 pixels and fine-tune using MIML loss



MIML loss

Bitdefender

- $L(\mathbf{X}, \mathbf{y}; \theta) = \frac{1}{N} \sum_{i=1}^{N} [\mathbf{y}_i \log \hat{\mathbf{p}}_i + (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{p}}_i)]$
- $p_{ij}^{w} = \sigma(w_w x_{ij} + b_w)$
  $\hat{p}_i^{w} = 1 - \prod_{x_{ij} \in \mathbf{X}_i} (1 - p_{ij}^{w})$

  $p^{w} = \max_i p_i^{w}$ N - number of frames
  θ - parameters
  X_i - i^{th} frame
  x_{ij} - last layer of FCN
  y_i - words from sentence
  $\hat{p}_i^{w}$ - probability of w word in frame i
  p_{ij}^{w} - probability vector of w word in region j of frame i
  p^{w} - probability of w word in region-sequence

Discover region-sequence

- **region-sequence**: a sequence of regions, one from each frame (16^{nr_frames} sequences)
- a sequence A_t is described by $f = [f_{inf}, f_{div}, f_{coh}]^T$, where:

Discover region-sequence

- **region-sequence**: a sequence of regions, one from each frame (16^{nr_frames} sequences)
- a sequence A_t is described by $f = [f_{inf}, f_{div}, f_{coh}]^T$, where:
  - $\mathbf{f}_{inf}$ measures the **informativeness** of the sequence

$$f_{inf}(x_v, A_t) = \sum_w (p^w);$$
$$p^w = \max_{i \in A_t} p_i^w$$

Discover region-sequence

- **region-sequence**: a sequence of regions, one from each frame (16^{nr_frames} sequences)
- a sequence A_t is described by $f = [f_{inf}, f_{div}, f_{coh}]^T$, where:
  - **$\mathbf{f}_{inf}$** measures the **informativeness** of the sequence
    $$f_{inf}(x_v, A_t) = \sum_w (p^w);$$
    $$p^w = \max_{i \in A_t} p_i^w$$
  - **$\mathbf{f}_{coh}$** ensures the temporal **coherence**. we select regions with the smallest changes temporally
    $$f_{coh} = \sum_{r_s \in A_{t-1}} < x_{r_t}, x_{r_s} >$$

Discover region-sequence

Bitdefender

- **region-sequence**: a sequence of regions, one from each frame (16^{nr_frames} sequences)
- a sequence A_t is described by $f = [f_{inf}, f_{div}, f_{coh}]^T$, where:
  - $\mathbf{f}_{inf}$ measures the **informativeness** of the sequence
    $$f_{inf}(x_v, A_t) = \sum_w (p^w);$$
    $$p^w = \max_{i \in A_t} p_i^w$$
  - $\mathbf{f}_{coh}$ ensures the temporal **coherence**. we select regions with the smallest changes temporally
    $$f_{coh} = \sum_{r_s \in A_{t-1}} < x_{r_t}, x_{r_s} >$$
  - $\mathbf{f}_{dif}$ measures the degree of **difference** between a candidate and all the existing region-sequences
    $$f_{div} = \sum_{i=1}^{N} \int_w p_i^w \log \frac{p_i^w}{q^w} dw$$

Discover region-sequence

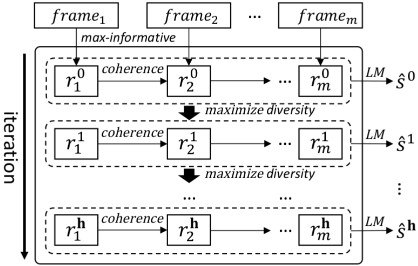


Figure 4: Illustration of region-sequence generation. r_i^j is the j-th region-sequence in i-th frame and '*LM*' denotes language model.

- f_{inf} measure the **informativeness** of the sequence
- f_{coh} ensure the temporal **coherence**. we select regions with the smallest changes temporally
- f_{dif} measure the degree of **difference** between a candidate and all the existing region-sequences

Discover region-sequence

- objective function to optimize:
  $$R(x_v, A) = w_v^T f(x_v, A)$$
  $$A^* = arg\max_{A \in S_v} R(x_v, A)$$
- There are 2 unknown elements:
  - parameter w_v
  - ground truth (region, sequence) pair

Discover region-sequence

- **Q:** How to find best A, given w_v?

Discover region-sequence

- **Q:** How to find best A, given w_v?
  A: Greedy

CELF optimization method

- Define marginal gain:
  $$L(w_v; r) = R(A_{t-1} \cup \{r\}) - R(A_{t-1})$$
- **CELF greedy algorithm**:
  1. $A_0 = \emptyset$
     $t = 1$
  2. $r_t = arg\max_{r \in S_t} L(w_v; r)$
     $A_t = A_{t-1} \cup \{r\}$
     $t = t + 1$
  3. repeat step 2 until the end of the video

Submodular maximization

- **Def:** Given a function f and arbitrary sets $A \subseteq B \subseteq S_v \setminus r$ f is **submodular** if it satisfies:
  $$f(A \cup \{r\}) - f(A) \geq f(B \cup \{r\}) - f(B)$$
- $[f_{inf}, f_{div}, f_{coh}]^T$ is a submodular function

- Submodular functions have many properties desirable for optimization
- A greedy algorithm yields a good aproximation of maximum solution (**CELF** - cost-effective lazy forward-selection method)

WTA

- **Q:** How to find best region from a set that match sentence s?
- **A:** WTA algorithm

- **WTA algorithm**:
  1. extract words from sentence
  2. compute probability of each word in each region-sequence:
     $p_i^w = \max_j p_{ij}^w$, where p_{ij}^w is the output of FCN
  3. threshold p_i^w with θ
  4. compute matching score: $f_i = \sum_{w \in V} p_i^w$
  5. $i^* = arg \max_i f_i$

Optimize w_v

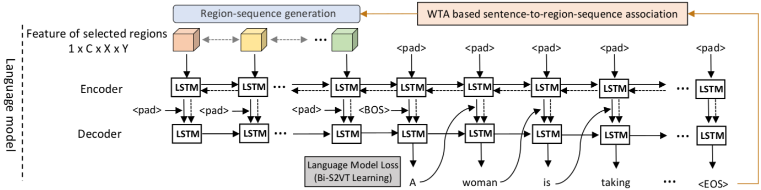
- **Q:** How to find best w_v, given N pairs (region, sentence)
- **A:** $\min_{w_v \geq 0} \frac{1}{N} \sum_{i=1}^{N} \max_{r \in r_i} L_i(w_v; r) + \frac{\lambda}{2} \|w_v\|^2$

Discover region-sequence

- But we do not know either w_v or ground truth pairs
- Use alternative optimization:
  1. initialize $w_v = \mathbf{1}$
  2. using w_v generate a sequences with submodular maximization
  3. associate sentences to sequences using WTA
  4. using pairs from step 3, optimize w_v
  5. repeat step 2-4 until w_v converge

Language model

- use sequence-to-sequence model S2VT to generate language:
  - encoder: bi-directional LSTM
  - decoder: LSTM



Results

Model	METEOR	BLEU@4	ROUGE-L	CIDEr
Mean-Pooling [49]	23.7	30.4	52.0	35.0
Soft-Attention [53]	25.0	28.5	53.3	37.1
S2VT [48]	25.7	31.4	55.9	35.2
ruc-uva [6]	27.5	39.4	60.0	48.0
VideoLAB [34]	27.7	39.5	61.0	44.2
Aalto [40]	27.7	41.1	59.6	46.4
v2t_navigator [15]	29.0	43.7	61.4	45.7
Ours w/o category	27.7	39.0	60.1	44.0
Ours category-wise	28.2	40.9	61.8	44.7
Ours + C3D + Audio	**29.4**	**44.2**	**62.6**	**50.5**

Table 3: Comparison with state of the arts on the *validation set* of MSR-VTT dataset. See texts for more explanations.

Model	METEOR	BLEU@4	ROUGE-L	CIDEr
ruc-uva [6]	26.9	38.7	58.7	45.9
VideoLAB [34]	27.7	39.1	60.6	44.1
Aalto [40]	26.9	39.8	59.8	45.7
v2t_navigator [15]	28.2	40.8	60.9	44.8
Ours	**28.3**	**41.4**	**61.1**	**48.9**

Table 4: Comparison with state of the arts on the *test set* of MSR-VTT dataset. See texts for more explanations.

Results

- diversity measure: $D_{div} = \frac{1}{N} \sum_{s^i, s^j \in S; i \neq j} (1 - < s_i, s_j >)$
- LSA representation

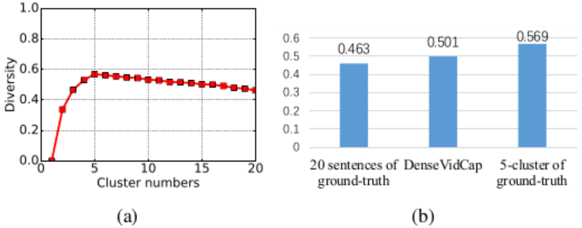


Figure 6: (a) Diversity score of clustered ground-truth captions under different cluster numbers; (b) Diversity score comparison of our automatic method (middle) and the ground-truth.

Results

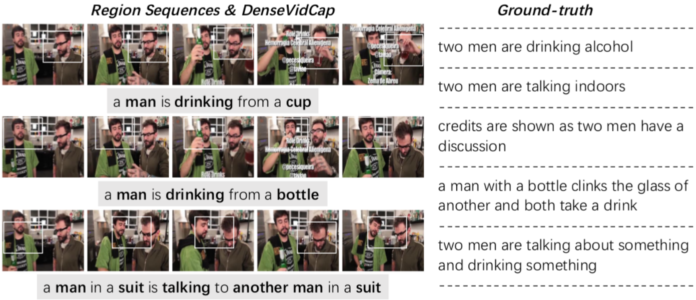


Figure 9: Left: Examples of dense sentences produced by our ***DenseVidCap*** method and corresponding ***region sequences***; Right: Ground-truth (video6974).

Bitdefender

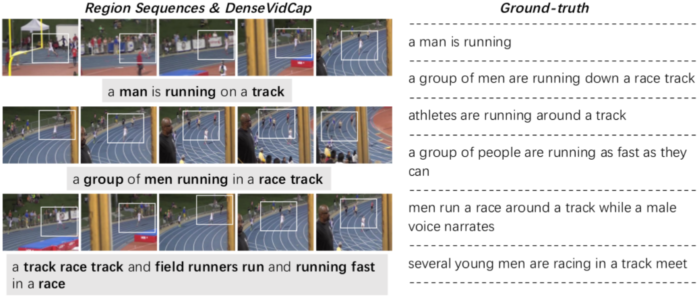


Figure 10: Left: Examples of dense sentences produced by our ***DenseVidCap*** method and corresponding ***region sequences***; Right: Ground-truth (video6967).

Questions?

References I

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. URL `http://arxiv.org/abs/1311.2524`.

A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014. URL `http://arxiv.org/abs/1412.2306`.

Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y. Jiang, and X. Xue. Weakly supervised dense video captioning. *CoRR*, abs/1704.01502, 2017. URL `http://arxiv.org/abs/1704.01502`.