Mining for meaning: from vision to language through multiple networks consensus

Iulia Duţă1,2, Andrei Nicolicioiu1,3, Vlad Bogolin1, Marius Leordeanu1,3,4
iduta@bitdefender.com, anicolicioiu@bitdefender.com, vladbogolin@gmail.com, marius.leordeanu@imar.ro
1Bitdefender, Romania 2University of Bucharest, Romania 3University Politehnica of Bucharest, Romania 4Institute of Mathematics of the Romanian Academy

1. Overview

Video captioning: describe videos in natural language

Our approach:
• obtain a diverse pool of generated sentences by:
  • varying the video encoder (TCN)
  • use sparse intermediate representations (Two-Stage)
  • leverage learning on additional tasks (Two-Stage, Two-Wings)
• use a selection method based on:
  • consensus among whole pool of sentences for a video
  • pairwise comparisons between sentences

Main Contributions:
• propose a method for selecting a sentence that best describes a video
• propose two novel architectures and perform extensive tests with many others adapted from the literature
• achieve state of the art results on the MSR-VTT dataset

2. Two-Wings Network

• goal: improve vocabulary of generated sentences
• improve language decoder by also learning a language reconstruction task
• use a separate branch (shared decoder) for optimizing on raw text - Wikipedia

3. TCN

• goal: obtain a different video encoding
• use temporal convolution to aggregate features from neighbouring time steps
• encode the information - hierarchy of dilated convolutional layers

4. Two-Stage Network

• goal: use sparse representation of the video
• learn two stages of the model separately then fine-tune them jointly
• fist stage: learn to predict set of labels from video
• second stage: learn to construct sentences from a set of labels

5. Consensus

Agreement score:
• select the sentences that agree most with the others
• agreement score: for each generated sentence compute its CIDEr score against the others
• choose the top C sentences

Oracle Network:
• train a network to choose between 2 sentences given a video
• pairwise comparisons between each sentence from top C and all the others from the pool
• final caption is the one with most wins

6. Features

• each additional set of features bring improvement compared to single model
• consensus brings substantial improvements regardless of features used

7. Results

<table>
<thead>
<tr>
<th>CIDEr</th>
<th>Meteor</th>
<th>Rouge</th>
<th>Bleu 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>v2 navig [1]</td>
<td>44.8</td>
<td>25.2</td>
<td>60.9</td>
</tr>
<tr>
<td>MT-Ent [2]</td>
<td>47.1</td>
<td>28.8</td>
<td>60.2</td>
</tr>
<tr>
<td>HRL [3]</td>
<td>48.0</td>
<td>28.7</td>
<td>61.7</td>
</tr>
<tr>
<td>dense [4]</td>
<td>48.9</td>
<td>28.3</td>
<td>61.1</td>
</tr>
<tr>
<td>CIDEnt-RL [5]</td>
<td>51.7</td>
<td>28.4</td>
<td>61.4</td>
</tr>
<tr>
<td>TGM [6]</td>
<td>52.9</td>
<td>29.7</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>53.8</td>
<td>29.7</td>
<td>63.0</td>
</tr>
</tbody>
</table>

We obtain state of the art results on three evaluation metrics on MSR-VTT 2016 test set.

8. Qualitative Results

Top captions:
• a girl is knocking on a wall and texting
• a girl laying in bed and knocking on the wall
• a girl is laying in bed and knocking on the wall
• a girl is knocking on a wall and texting

Human annotations:
• a girl in bed
• a girl lays in bed and uses her phone

9. References