Image Data, Video Data and Both in VTT Model Training
Video-to-Text Task in TRECVID 2019

Jorma Laaksonen, PicSOM Team

Department of Computer Science
Aalto University School of Science
Espoo, Finland

November 13th, 2019
Contents

Background
Motivation
Approach
Results
Analysis
Conclusions
People

- Jorma Laaksonen
- Héctor Laria Mantecón
- (Danny Francis & Benoit Huet of EURECOM)
Lessons from TRECVID 2018

- We used only cross-entropy training, others did better with reinforcement learning
- Validation with VTT 2016 data was not able to select the best models
- Training with COCO image dataset gave equally good results as with video datasets
- We could move from old Theano-based code to new PyTorch-based
Development of scores

METEOR scores by submission

- PicSOM pre experiments
- PicSOM submissions
- Other submissions
- PicSOM post experiments
Work between TRECVID 2018 and 2019

- Implemented self-critical reinforcement learning
- Studied methods to combine image and video datasets and features
- Also wanted to study optimal combination of different video datasets
TGIF and COCO datasets

Statistics:

- TGIF: 125,713 videos with 125,713 captions
- COCO: 123,287 images with 616,767 captions

Which approach would be the best:

- 125,713 video feature vectors and 125,713 captions
- 123,287 image feature vectors and 616,767 captions
- 249,000 image feature vectors and 742,480 captions
- 249,000 image and video feature vectors and 742,480 captions
Videos to image features and vice versa

- Image features can be extracted from videos in multiple ways, e.g.
  - use only the middle frame
  - max or mean pool features of multiple or all frames
- Genuine video features such as I3D cannot be extracted from still images
  - we used fake video features for COCO images
  - average of all video features in TGIF was used assigned to all COCO images
- The final feature vector was concatenation of

<table>
<thead>
<tr>
<th>TGIF videos:</th>
<th>I3D video feature</th>
<th>ResNet image feature of middle frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO images:</td>
<td>constant average I3D feature</td>
<td>ResNet image feature</td>
</tr>
</tbody>
</table>
Contents

Background

Motivation

Approach

Results

Analysis

Conclusions
Methodology

- COCO image and TGIF video datasets in training
- model validation and early stopping with VTT 2018 dataset
- ResNet-152 CNN image and I3D video features
- fake I3D video features for COCO images
- “DeepCaption” LSTM language model decoder in PyTorch
- cross-entropy loss training in the beginning
- self-critical reinforcement learning learning in the end
We submitted four runs:

- **PIC SOM.1-M EMAD.PRIMARY**: uses ResNet and I3D features for initialising the LSTM generator, and is trained on MS COCO + TGIF using self-critical loss,
- **PIC SOM.2-M EMAD**: uses I3D features as initialisation, and is trained on TGIF using self-critical loss,
- **PIC SOM.3**: uses ResNet features as initialisation, and is trained on MS COCO + TGIF using self-critical loss,
- **PIC SOM.4**: is the same as **PIC SOM.1-M EMAD.PRIMARY** except that the loss function used is cross-entropy,
Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>p-18-s2</td>
<td>Ice rn+fr C+M</td>
<td>0.1541</td>
<td>0.1657</td>
<td>0.0476</td>
<td>0.0091</td>
<td>0.1773</td>
<td>0.1858</td>
<td>0.0722</td>
<td>0.0207</td>
<td>–</td>
</tr>
<tr>
<td>p-19-s1</td>
<td>Ice rn+i3d C+T</td>
<td>0.2055</td>
<td>0.3025</td>
<td>0.1157</td>
<td>0.0294</td>
<td>0.2285</td>
<td>0.3277</td>
<td>0.1615</td>
<td>0.0385</td>
<td>0.4168</td>
</tr>
<tr>
<td>p-19-s2</td>
<td>Ice i3d T</td>
<td>0.1958</td>
<td>0.2718</td>
<td>0.0949</td>
<td>0.0348</td>
<td>0.2139</td>
<td>0.2773</td>
<td>0.1245</td>
<td>0.0379</td>
<td>0.4169</td>
</tr>
<tr>
<td>p-19-s3</td>
<td>Ice i3d C+T</td>
<td>0.2007</td>
<td>0.2777</td>
<td>0.1074</td>
<td>0.0301</td>
<td>0.2254</td>
<td>0.3130</td>
<td>0.1569</td>
<td>0.0345</td>
<td>0.4282</td>
</tr>
<tr>
<td>p-19-s4</td>
<td>Ice i3d C+T</td>
<td>0.1850</td>
<td>0.2190</td>
<td>0.0822</td>
<td>0.0213</td>
<td>0.2049</td>
<td>0.2348</td>
<td>0.1147</td>
<td>0.0319</td>
<td>0.4057</td>
</tr>
</tbody>
</table>

- p-18-s2 is our best submission in TRECVID 2018
- p-18-a3 is our best TRECVID 2018 post-conference result
- p-19-s* are our TRECVID 2019 submissions
Comparison: METEOR 2018

METEOR scores by submission

- PicSOM pre experiments
- PicSOM submissions
- Other submissions
- PicSOM post experiments
Comparison: METEOR

METEOR scores by submission

PicSOM 2018 models
PicSOM submissions
Other submissions
Comparison: CIDEr

CIDEr scores by submission

- PicSOM 2018 models
- PicSOM submissions
- Other submissions
Comparison: CIDEr-D

CIDErD scores by submission

- PicSOM 2018 models
- PicSOM submissions
- Other submissions

CIDErD scores by submission

- s1s3
- s2s8-a3
- 18-s2
Comparison: BLEU-4

BLEU scores by submission

PicSOM 2018 models
PicSOM submissions
Other submissions
Comparison: STS

STS scores by submission

PicSOM submissions
Other submissions
Comparison

- s4 run is always the worst — reinforcement learning is beneficial
- s1 run is almost always the best — combining image and video features is good
- s3 run wins s2 with 4–1 — COCO image features better than TGIF video features
Run types

In TRECVID VTT 2019 all submissions had to be tagged with their run type:

- Run type 'I': Only image captioning datasets were used for training
- Run type 'V': Only video captioning datasets were used for training
- Run type 'B': Both image and video captioning datasets were used for training
<table>
<thead>
<tr>
<th>team</th>
<th>image</th>
<th>video</th>
<th>both</th>
</tr>
</thead>
<tbody>
<tr>
<td>EURECOM</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDU</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMFD_IMPRESSEE</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insight_DCU</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>KU_ISPL</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KsLab</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PicSOM</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>RUCMM</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUC_AIM3</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UTS_ISA</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 teams</td>
<td>1</td>
<td>26</td>
<td>3</td>
</tr>
</tbody>
</table>
### Training datasets used per team

<table>
<thead>
<tr>
<th>team</th>
<th>COCO</th>
<th>TGIF</th>
<th>MSR-VTT</th>
<th>MSVD</th>
<th>VTT</th>
<th>VATEX</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>EURECOM</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>0+3</td>
</tr>
<tr>
<td>FDU</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0+1</td>
</tr>
<tr>
<td>IMFD_IMPRESEE</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>0+1</td>
</tr>
<tr>
<td>Insight_DCU</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0+1</td>
</tr>
<tr>
<td>KsLab</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>0+2</td>
</tr>
<tr>
<td>PicSOM</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1+1</td>
</tr>
<tr>
<td>RUCMM</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>0+3</td>
</tr>
<tr>
<td>RUC_AIM3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0+4</td>
</tr>
<tr>
<td>UTS_ISA</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0+4</td>
</tr>
<tr>
<td><strong>9 teams</strong></td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td><strong>0+0</strong></td>
</tr>
</tbody>
</table>

**Note:** The counts are the number of teams using each dataset, with a '+' indicating datasets exclusively used by each team.
# Statistics of the training datasets

<table>
<thead>
<tr>
<th>dataset</th>
<th>items</th>
<th>captions</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO</td>
<td>123,287</td>
<td>616,767</td>
</tr>
<tr>
<td>TGIF</td>
<td>125,713</td>
<td>125,713</td>
</tr>
<tr>
<td>MSR-VTT</td>
<td>6,513</td>
<td>130,260</td>
</tr>
<tr>
<td>MSVD</td>
<td>1,969</td>
<td>80,800</td>
</tr>
<tr>
<td>VTT</td>
<td>3,753</td>
<td>9,020</td>
</tr>
<tr>
<td>VATEX</td>
<td>41,300</td>
<td>826,000</td>
</tr>
<tr>
<td>LSMDC</td>
<td>108,536</td>
<td>108,536</td>
</tr>
</tbody>
</table>
### Video features used per team

<table>
<thead>
<tr>
<th>team</th>
<th>I3D</th>
<th>C3D</th>
<th>CNN+pool</th>
<th>CNN+seq</th>
<th>audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>EURECOM</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDU</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>IMFD_IMPRESEE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insight_DCU</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>KsLab</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>PicSOM</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUCMM</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUC_AIM3</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>UTS_ISA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>9 teams</strong></td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Contents

Background

Motivation

Approach

Results

Analysis

Conclusions
Conclusions

- In the PicSOM experiments the use of also the COCO dataset proved to be beneficial.
- Naïve use of fake video features for images was better than not to use images at all.
- This conclusion might be different if:
  - our overall result level were higher
  - we used more video data than just TGIF
  - we used better video features than I3D
  - we used pooling or RNN based aggregation of framewise features
  - our implementation of self-critical training were better
- Model performance was very stable from validation with 2018 data to 2019 test data.
- No other team used COCO dataset anymore.
- Our results we clearly behind those of the best teams.
- Specifying the run types in the way it was done now might be discontinued.