Attentive Visual Semantic Specialized Network for Video Captioning

Jesus Perez-Martin  Benjamin Bustos  Jorge Pérez
Problem: Video Captioning

- A kid pushes a stroller
- A girl is pushing a doll stroller
- A little girl is pushing a stroller through a grocery store
Model Overview
Model Overview
Model Overview
Model Overview
Model Overview
Model Overview

2D-CNN 3D-CNN

2D-CNN 3D-CNN

2D-CNN 3D-CNN

average pool

v-LSTM

TA

s-LSTM

concepts detector

Adaptive Gate

a man is holding a small animal
Length-Weighted Loss

Given a video $x$, and the ground-truth caption $y = (y_1, y_2, \ldots, y_L)$ of $x$

We minimize

$$
\mathcal{L}_\Theta = -\frac{1}{L^\beta} \sum_{t=1}^{L} \log p_\Theta(w_t|w_{<t})
$$
Experimental Evaluation: Datasets

**MSVD**
- 1,970 videos
  - 1,200 train
  - 100 validation
  - 670 test
  - ~ 40 captions per video
- Dictionary size: ~ 6K words

**MSR-VTT**
- 10,000 videos
  - 6,512 train
  - 498 validation
  - 2,990 test
  - ~ 20 captions per video
- Dictionary size: ~ 14,000 words
Results - Comparison with State of the Art

**MSVD**

- Wang et al., ICCV 2019
- Gao et al., TPAMI 2019
- Chen et al., FRAI 2020
- AVSSN (ours)
Results - Comparison with State of the Art

MSR-VTT

- Wang et al., ICCV 2019
- Gao et al., TPAMI 2019
- Chen et al., FRAI 2020
- AVSSN (ours)
Qualitative Analysis

Ours: a man *is holding* a small animal

GT1: a man *is holding* a pet animal
GT2: someone *holds* a baby skunk

w/o AAG: a person is petting a small animal
w/o s-LSTIM: a woman is petting a small animal
Qualitative Analysis

*Ours:* a girl is *pushing* a *stroller*

*GT1:* a little girl is *pushing* a *stroller* through a grocery store

*GT2:* a kid *pushes* a *stroller*

*w/o AAG:* a bayby is *playing*

*w/o s-LSTM:* a girl is *dancing*
Qualitative Analysis

Ours: a man is sprinkling spices on some bacon

GT1: a man is seasoning some bacon
GT2: a person seasons some bacon

w/o AAG: a man is seasoning some meat
w/o s-LSTM: a woman is adding spices to a bowl of meat
Attentive Visual Semantic Specialized Network for Video Captioning

**Poster:** PS T3.9 #1574

Code/Features/Models available on GitHub

https://github.com/jssprz/visual_syntactic_embedding_video_captioning

jeperez@dcc.uchile.cl

@jes_prz