Show, Translate, and Tell

Dheeraj Peri, Shagan Sah, Raymond Ptucha
{dp1248, sxs337, rwpeec}@rit.edu
Rochester Institute of Technology, Rochester, NY

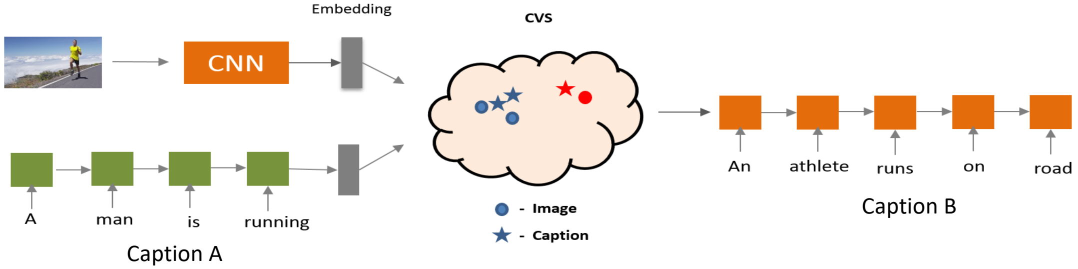
Introduction

We propose a unified model which jointly trains on images and captions and learns to generate new captions given either an image or a caption query. We evaluate our model on three different tasks-

- **Cross-modal retrieval**
- **Image captioning**
- **Sentence paraphrasing**

Model

- Deep convolutional network is jointly trained with sequence to sequence which generates a unified embedding space for image and text modalities. Semantically close images and captions are mapped close in this latent space.
- We build an attention mechanism that aligns individual regions in the image to the words in the sentence by computing cosine similarity between regions and words



- Multi-task weighted loss function to optimize the model-

$$\mathbf{L} = \lambda_1 \mathbf{L}_{\mathbf{sim}} + \lambda_2 \mathbf{L}_{\mathbf{IC}} + \lambda_3 \mathbf{L}_{\mathbf{SP}}$$

- $\mathbf{L}_{\mathbf{sim}}$ - Similarity loss, $\mathbf{L}_{\mathbf{IC}}$ - Captioning loss, $\mathbf{L}_{\mathbf{SP}}$ – Paraphrasing loss
- Joint training on three tasks along with attention improves the performance of the model especially on retrieval and captioning.

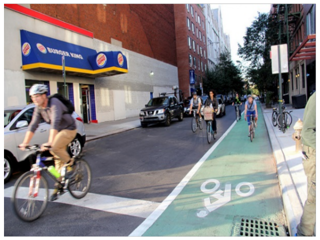
Comparison of STT on multiple tasks

Task	Attention	Metric	Score
Retrieval	No	I-T, T-I Recall@1	55.1, 41.0
	Yes		64.9, 49.8
Captioning	No	Bleu-1, Bleu-2	0.683, 0.506
	Yes		0.706, 0.530
Paraphrasing	No		0.744, 0.578
	Yes		0.747, 0.581

Comparison of STT with related works

Method	Sentence Retrieval		Image Retrieval	
	R@1	R@10	R@1	R@10
CSE	56.39	91.5	45.7	90.6
VSE++	58.3	93.3	43.6	87.8
STT w/ att	64.9	96.8	49.8	91.6
SCO	69.9	97.5	56.7	94.8

Representative Result



Captioning : a group of people riding bikes down a street

Paraphrasing
- a group of people riding bikes down a street
- a man riding a bike down a street next to a traffic light

Top 3 Retrieved captions
- bike riders passing Burger King in city street
- A group of bicyclists are riding in the bike lane .
- Bicyclists on a city street , most not using the bike lane

Groundtruth captions
- people on bicycles ride down a busy street
- A group of people are riding bikes down the street in a bike lane
- bike riders passing Burger King in city street
- A group of bicyclists are riding in the bike lane .
- Bicyclists on a city street , most not using the bike lane

- The performance improvement of STT is consistent on MSCOCO and Flickr-30K dataset..
- Paper link: https://arxiv.org/abs/1903.06275
  Code: https://github.com/peri044/STT