NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
School of Computer Science and Engineering
RISE Rapid-Rich Object Search Lab 博 云 搜 索 实 验 室

Unpaired Image Captioning by Language Pivoting

Jiuxiang Gu[1], Shafiq Joty[1], Jianfei Cai[1], Gang Wang[2]
[1] Nanyang Technological University, [2] Alibaba AI Labs

Ali A.I. Labs 阿里人工智能实验室
ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

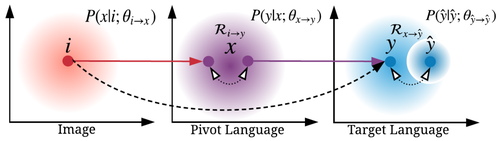
Overview

- We present an approach to this *unpaired image captioning* problem by *language pivoting*.
- Our method can effectively capture the characteristics of an image captioner from the *pivot language (Chinese)* and align it to the *target language (English)* using another pivot-target (Chinese-English) parallel corpus.
- Quantitative comparisons against several baseline approaches demonstrate the effectiveness of our method.

Challenges

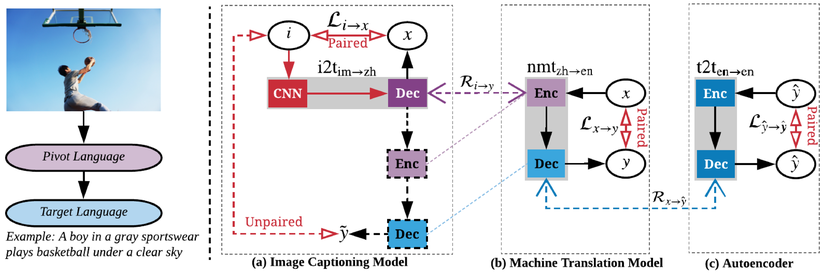
1. The task of image caption generation has so far been explored only in English since most available datasets are in English. *However, only 5.5% of the world population are native speakers of English (source from Wikipedia).*
2. In many applications and languages, the large-scale annotations are not readily available, and are expensive and slow to acquire.
3. Getting *image→caption* paired data for each target language is expensive, but, *language parallel data is relatively easier to get.*

Motivation



1. Our pivot-based captioning framework learns to:
   - Generate captions for *image→pivot* language in *domain A*.
   - Translate sentences for *pivot → target* language in *domain B*.
   - Match it to the target language with caption style in *domain C*.
   - The two proposed connection terms ($\mathcal{R}_{i\to y}$ and $\mathcal{R}_{x\to\hat{y}}$) bridge the gap between the pivot language in *domain A* and *domain B* and the gap between the target language in *domain B* and *domain C*, respectively.

Pivot-based image captioning framework



1. We use the pivot language x to learn the mapping: $i \xrightarrow{\theta_{i\to x}} x \xrightarrow{\theta_{x\to y}} y$.
2. Our pivot-based image captioning approach has an image captioning model $P(x|i;\theta_{i\to x})$ to generate a caption x in the pivot language from an image i and a machine translation model $P(y|x;\theta_{x\to y})$ to translate this caption into the target language.
3. In addition, we have an autoencoder in the target language $P(\hat{y}|\hat{y};\theta_{\hat{y}\to\hat{y}})$ that guides the target language decoder to produce caption-like sentences.
4. We train these components jointly so that they interact with each other.

Learning

1. We first generate the image description x from the global image feature v. We maximize the probability of the ground truth caption words given the image:

$$\tilde{\theta}_{i\to x} = \arg\max_{\theta_{i\to x}} \{\mathcal{L}_{i\to x}\} = \arg\max_{\theta_{i\to x}} \Big\{ \sum_{n_i=0}^{N_i-1} \sum_{t=0}^{M^{(n_i)}-1} \log P_x(x_t^{(n_i)} | x_{0:t-1}^{(n_i)}, i^{(n_i)}; \theta_{i\to x}) \Big\} \quad (1)$$

2. The second model is to learn to translate the Chinese sentence x to the English sentence y, where the maximum-likelihood training objective of the model can be expressed as:

$$\tilde{\theta}_{x\to y} = \arg\max_{\theta_{x\to y}} \{\mathcal{L}_{x\to y}\} = \arg\max_{\theta_{x\to y}} \Big\{ \sum_{n_x=0}^{N_x-1} \sum_{t=0}^{M^{(n_x)}-1} \log P_y(y_t^{(n_x)} | y_{0:t-1}^{(n_x)}; x^{(n_x)}; \theta_{x\to y}) \Big\} \quad (2)$$

3. Connecting *Image-to-Pivot* and *Pivot-to-Target*.

$$\mathcal{R}_{i\to y}(\theta_{i\to x}^{w_x}, \theta_{x\to y}^{w_y}) = - \sum_{w_x \in \mathcal{V}_{i\to x}^{x} \cap \mathcal{V}_{x\to y}^{x}} ||\theta_{i\to x}^{w_x} - \theta_{x\to y}^{w_x}||_2 \quad (3)$$

$$\mathcal{R}_{x\to\hat{y}}(\theta_{x\to y}^{w_y}, \theta_{\hat{y}\to\hat{y}}^{w_y}) = - \sum_{w_y \in \mathcal{V}_{x\to y}^{y} \cap \mathcal{V}_{\hat{y}\to\hat{y}}^{y}} ||\theta_{x\to y}^{w_y} - \theta_{\hat{y}\to\hat{y}}^{w_y}||_2 \quad (4)$$

4. where $\theta_{i\to x}^{w_x}$ is already a learned model and kept fixed during adaptation.
5. Our goal is to find a set of source-to-target model parameters that maximizes the training objective:

$$\mathcal{J}_{i\to x,x\to y,y\to\hat{y}} = \mathcal{L}_{i\to x} + \mathcal{L}_{x\to y} + \mathcal{L}_{\hat{y}\to\hat{y}} + \lambda\mathcal{R}_{i\to x,x\to y,y\to\hat{y}} \quad (5)$$

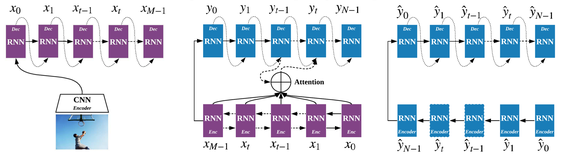
$$\mathcal{R}_{i\to x,x\to y,y\to\hat{y}} = \mathcal{R}_{i\to y}(\theta_{i\to x}^{w_x}, \theta_{x\to y}^{w_x}) + \mathcal{R}_{x\to\hat{y}}(\theta_{x\to y}^{w_y}, \theta_{\hat{y}\to\hat{y}}^{w_y}) \quad (6)$$

6. During inference, given an unseen image i to be described, we use the joint decoder:

$$y \sim \arg\max_{y} \{P(y|i;\theta_{i\to x},\theta_{x\to y})\} \quad (7)$$

Implementation

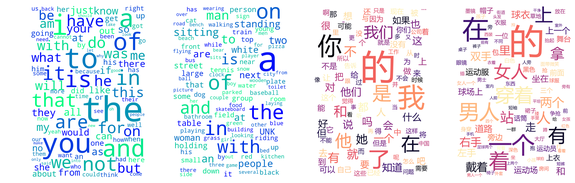
The three models used in our pivot-based image captioning framework:



We choose the two independent datasets used from AI Challenger (AIC): AIC Image Chinese Captioning (AIC-ICC) and AIC Chinese-English Machine Translation (AIC-MT), as the training datasets, while using MSCOCO and Flickr30K as the test datasets.

	Dataset	Lang.	Source		Target	
			# Image/Sent.	Vocab. Size	# Sent.	Vocab. Size
Training	AIC-ICC	im → zh	240K	–	1,200K	4,461
	AIC-MT	zh → en	10,000K	50,004	10,000K	50,004
Testing	MSCOCO	im → en	123K	–	615K	9,487
	Flickr30K	im → en	30K	–	150K	7,000

Word clouds of datasets



(a) AIC-MT (English) (b) MSCOCO (English) (c) AIC-MT (Chinese) (d) AIC-ICC (Chinese)

Results

Results of unpaired image-to-English captioning on MSCOCO 5K test split.

Approach	Lang.	B@1	B@2	B@3	B@4	M	CIDEr
MSCOCO							
$\text{i2t}_{\text{im}\to\text{en}}$ (Upper bound, XE Loss)	en	73.2	56.3	42.0	31.2	25.3	95.1
FC-2K [42] (ResNet101, XE Loss)	en	–	–	–	29.6	25.2	94.0
$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ ($\mathcal{R}_{i\to x,x\to y,y\to\hat{y}}$)	en	**46.2**	**24.0**	**11.2**	**5.4**	13.2	**17.7**
$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ ($\mathcal{R}_{i\to x,x\to y}$)	en	45.5	23.6	11.0	5.3	13.1	17.3
$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ (Lower bound)	en	42.0	20.6	9.5	3.9	12.0	12.3
$\text{i2t}_{\text{im}\to\text{zh}}$ + Online Google Translation	en	42.2	21.8	10.7	5.3	**14.5**	17.0
Flickr30K							
$\text{i2t}_{\text{im}\to\text{en}}$ (Upper bound, XE Loss)	en	63.1	43.8	30.2	20.7	17.7	40.1
$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ ($\mathcal{R}_{i\to x,x\to y,y\to\hat{y}}$)	en	**49.7**	**27.8**	**14.8**	**7.9**	13.6	**16.2**
$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ ($\mathcal{R}_{i\to x,x\to y}$)	en	48.7	26.1	12.8	6.4	13.0	14.9
$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ (Lower bound)	en	45.9	25.2	13.1	6.9	12.5	13.9
$\text{i2t}_{\text{im}\to\text{zh}}$ + Online Google Translation	en	46.2	25.4	13.9	7.7	**14.4**	15.8

Evaluation results of user assessment on MSCOCO 1.2K test split.

Approach	$\text{i2t}_{\text{im}\to\text{zh}}$ + $\text{nmt}_{\text{zh}\to\text{en}}$ ($\mathcal{R}_{i\to x,x\to y,y\to\hat{y}}$)	Upper Bound	Ground-Truth
Relevant	3.81	3.99	4.68
Resemble	3.78	4.05	4.48