Unsupervised Image Captioning

Yang Feng et al

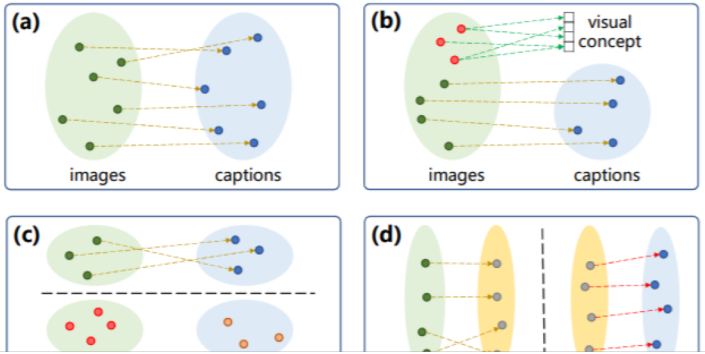
Winter 2019

Unsupervised Image Captioning

Yang Feng[‡*] Lin Ma[†] Wei Liu[†] Jiebo Luo[‡]
[†]Tencent AI Lab [‡]Univeristy of Rochester
{yfeng23,jluo}@cs.rochester.edu forest.linma@gmail.com wl2223@columbia.edu

Abstract

Deep neural networks have achieved great successes on the image captioning task. However, most of the existing models depend heavily on paired image-sentence datasets, which are very expensive to acquire. In this paper, we make the first attempt to train an image captioning model in an unsupervised manner. Instead of relying on manually labeled image-sentence pairs, our proposed model merely requires an image set, a sentence corpus, and an existing vi-

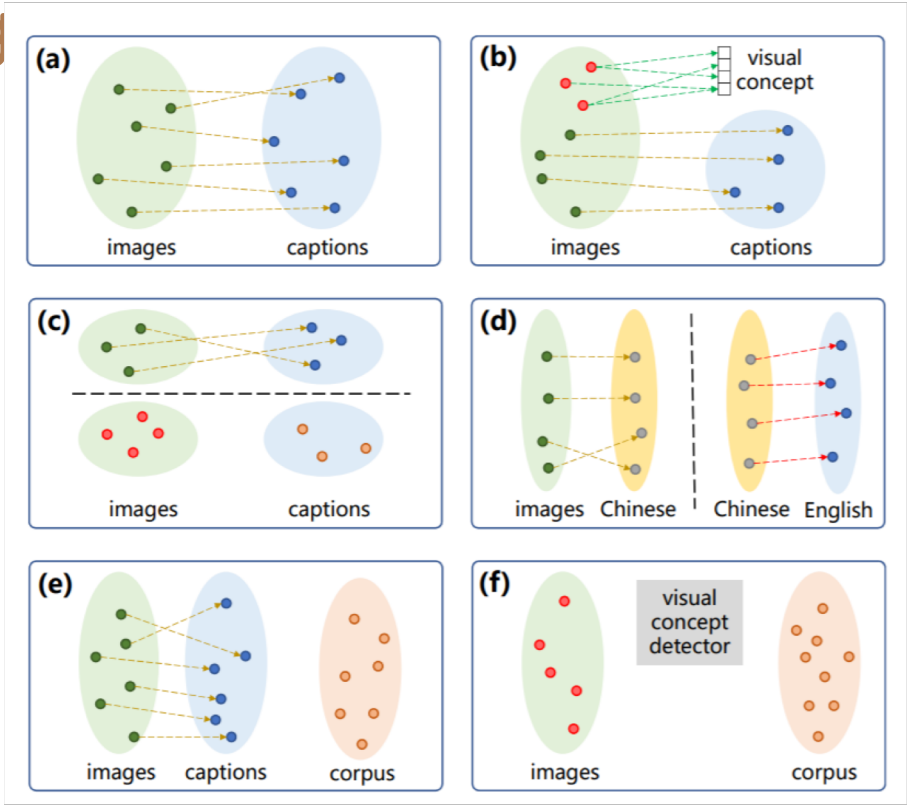


Introduction

- Similar to unsupervised machine translation

Different Captioning Models

a) Supervised manner
b) Visual Concepts from the paired dataset
c) New Domain Generalization
d) Double translate
e) Semi-Supervised approach
f) Unsupervised Image Captioning



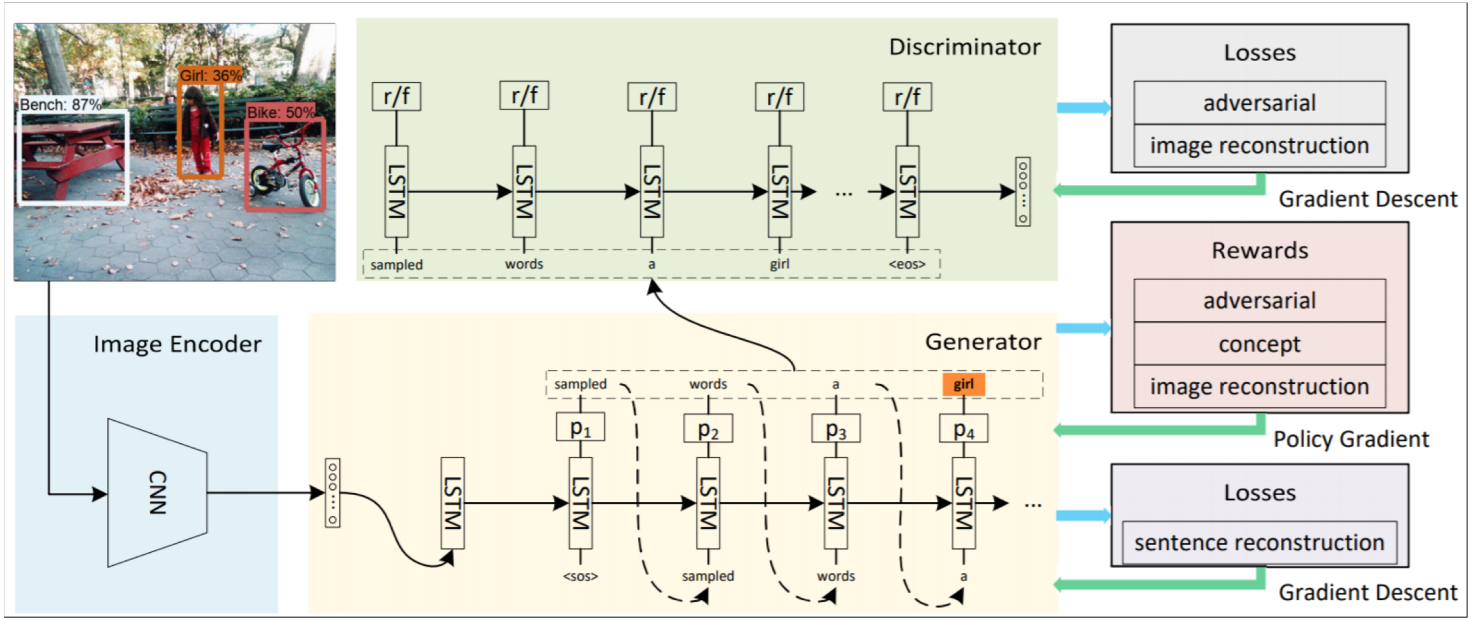
Contributions

- Not relying on any labeled image-sentence pairs
- Propose 3 objectives to train the model
  - Adversarial training
  - Visual Concept Detector
  - Bi-directional reconstruction
- A novel model initialization pipeline
- Provide a new image description corpus

Model

a) Image Encoder
b) Sentence Generator
c) Sentence Discriminator

Model


Bench: 87%
Girl: 36%
Bike: 50%
Discriminator
r/f
LSTM
sampled
words
a
girl
<eos>
Losses
adversarial
image reconstruction
Gradient Descent
Rewards
adversarial
concept
image reconstruction
Policy Gradient
Losses
sentence reconstruction
Gradient Descent
Image Encoder
CNN
Generator
p1
p2
p3
p4
<sos>
sampled
words
a
girl

Training

- **Adversarial Caption Generation**
- **Visual Concept Distillation**
- **Bi-directional Image-Sentence Reconstruction**

Adversarial Caption Generation

Goal: human plausible generated sentences

Adversarial Reward for generator:

$$r_t^{adv} = \log(q_t).$$

Adversarial loss for Discriminator

$$\mathcal{L}_{adv} = -\left[\frac{1}{l}\sum_{t=1}^{l}\log(\hat{q}_t) + \frac{1}{n}\sum_{t=1}^{n}\log(1-q_t)\right].$$

Limitation: Reward only encourages the model to generate sentences following grammar rules

Visual Concept Distillation

Concept Reward:

Specifically, when the image captioning model generates a word whose corresponding visual concept is detected in the input image, we give a reward to the generated word.

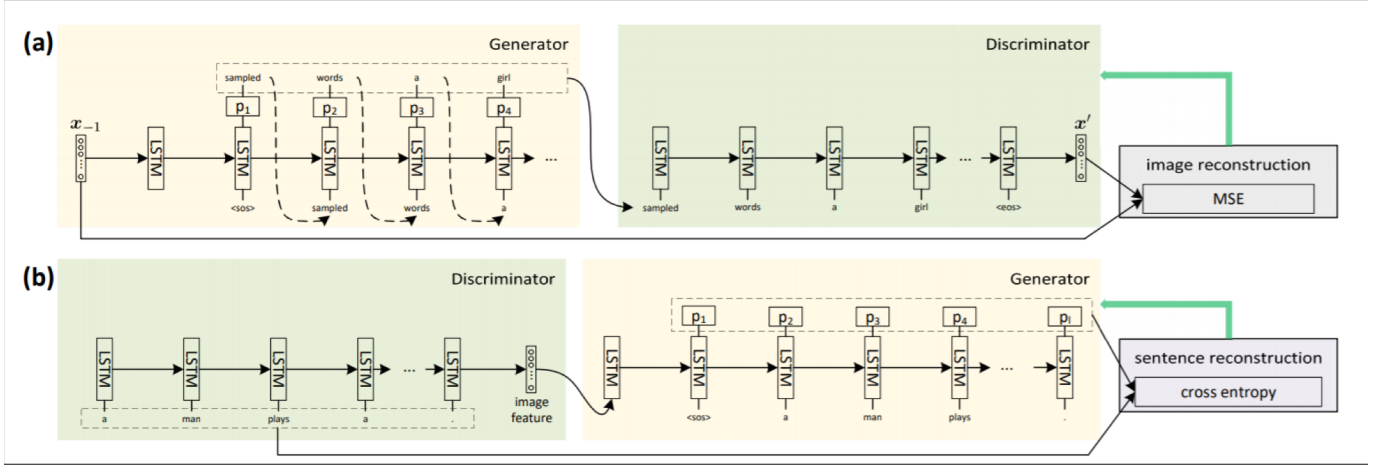
$$r_t^c = \sum_{i=1}^{N_c} \mathrm{I}(s_t = c_i) * v_i,$$

Limitations:

- The captioning quality would be largely determined by the visual concept detector because it is the only bridge between images and sentences.
- The existing visual concept detectors can only reliably detect a limited number of object concepts.

Bi-directional Image-Sentence Reconstruction

Project images and sentences into a common latent space such that they can reconstruct each other.



Integration

- For generator, the word sampling operation is not differentiable:

  Use policy gradient with respect to trainable parameters given the joint reward.
- Joint reward:
  - a) adversarial reward
  - b) concept reward
  - c) image reconstruction reward

Integration

Let θ denote the trainable parameters in the generator:

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}\left[\sum_{t=1}^{n}\left(\sum_{s=t}^{n}\gamma^s(\underbrace{r_s^{adv}}_{\text{adversarial}} + \underbrace{\lambda_c r_s^c}_{\text{concept}}\right.\right.$$
$$\left.\left. + \underbrace{\lambda_{im} r_s^{im}}_{\text{image reconstruction}} - b_t\right)\nabla_\theta \log(\boldsymbol{s}_t^\top \boldsymbol{p}_t)\right] + \underbrace{\lambda_{sen}\nabla_\theta \mathcal{L}_{sen}(\theta)}_{\text{sentence reconstruction}}, \quad (11)$$

Discriminator:

$$\mathcal{L}_D = \mathcal{L}_{adv} + \lambda_{im}\mathcal{L}_{im}.$$

Initialization

It's hard to train even with the three objectives and unpaired data → Propose an initialization pipeline

Generate a pseudo caption for each image to feed into generator:

a) build a concept dictionary consisting of the object classes in the OpenImages dataset
b) we train a concept-to-sentence (con2sen) model using the sentence corpus only.
c) detect concepts using visual concept detector
d) generate pseudo captions using concepts and concept-to-sentence model. Such model is called feat2sen and used to initialize the generator.

Experiments

- Shutterstock Image Description Corpus: online stock photography website
- Use 80 object categories from MSCOCO as the searching keywords
- Top 1000 results for each keyword
- 100 images per page resulting in 100,000 descriptions for each object category
- Remove sentences with less than eight words --> 2, 322, 628 distinct image descriptions in total.
- NLTK as tokenizer. Build vocabulary with words with frequency > 40
- 18, 667 words in our vocabulary, including special SOS, EOS, and an Unkown token.
- λc, λim, and λsen are set to be 10, 0.2, and 1, respectively. γ is set to be 0.9.
- Adam optimizer [26] with a learning rate of 0.0001.e
- BLEU [32], METEOR [13], ROUGE [31], CIDEr [39], and SPICE [1] scores computed with the cococaption code.

Experiments

"adv" alone leads to much worse results

Method	B1	B2	B3	B4	M	R	C	S
Ours w/o init	35.3	18.2	8.6	4.4	10.5	24.8	20.9	6.1
Ours	39.4	21.1	10.0	4.8	11.4	27.2	23.3	7.0
con2sen	35.9	18.7	8.7	4.1	11.5	26.3	17.6	7.0
feat2sen	38.0	20.4	9.6	4.7	11.6	27.2	19.5	6.6
adv	34.8	16.6	6.9	3.3	9.1	24.5	12.5	3.9
adv + con	36.6	18.4	8.3	3.9	10.7	25.5	19.7	6.3
adv + con + im	35.5	17.4	8.0	3.9	10.6	25.4	19.9	6.3

Experiments

concepts	bookcase, clothing, desk, person, table
con2sen	back to school concept . back to school concept . back to school . back to school concept
feat2sen	back view of a man in a clothing and a laptop . rear view people collection . backside
adv	young woman working on laptop in office .
adv + con	young woman working on laptop at desk in cafe
adv + con + im	young man working on laptop at home with laptop and drink
Ours w/o init	young woman working on laptop computer at home
Ours	**young man working on laptop at home office**

concepts	vehicle
con2sen	bangkok , thailand - june <UNK> : vehicle on the road in bangkok , thailand .
feat2sen	beautiful landscape with tree in the forest .
adv	young woman sitting on a bench in park on sunny day
adv + con	two wooden boat in the sea at sunset .
adv + con + im	a small fishing boat in the middle of the sea
Ours w/o init	small fishing boat tied to a tree in the sea
Ours	**a boat on the coast of the sea**

concepts	bowl, cat, plate, tableware
con2sen	a cat in a white plate with a bowl of tableware
feat2sen	the cat is sleeping on the floor .
adv	white wine glass isolated on white background with clipping path
adv + con	white wine in a glass on dark background
adv + con + im	a plate of red wine on a dark background
Ours w/o init	the cat is sleeping in the garden .
Ours	**a black and white cat on a wooden background**

concepts	bowl, food, hat
con2sen	food in a bowl with a hat on a white background
feat2sen	portrait of a happy young couple in santa hat
adv	happy young mother and her daughter sleeping in bed
adv + con	fresh orange juice in a wicker basket on a white background
adv + con + im	composition of fresh carrot on a plate , food
Ours w/o init	fresh organic vegetable on wooden background . healthy food
Ours	**top view of a bowl of healthy food**

Experiments

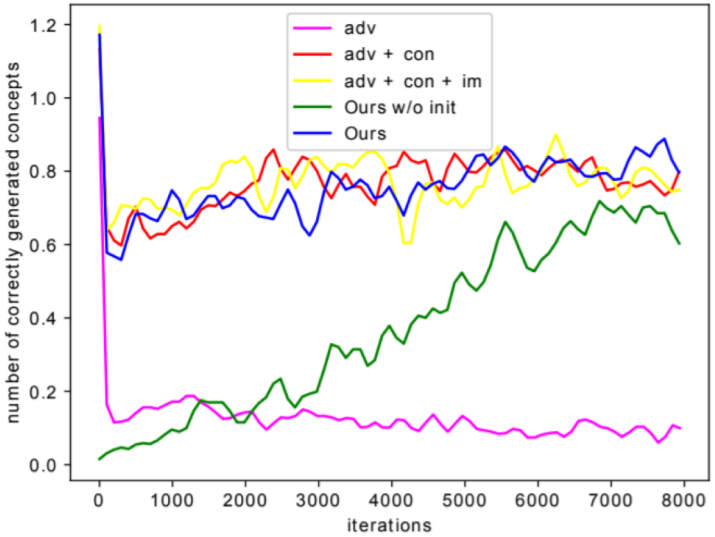


Figure 6. The average number of correct concept words in each sentence generated during the training process.

Experiments

Table 2. Performance comparisons on the test split [24] of the MSCOCO dataset *under the unpaired setting*.

Method	B1	B2	B3	B4	M	R	C	S
Pivoting [17]	46.2	24.0	11.2	5.4	13.2	-	17.7	-
Ours w/o init	53.8	35.5	23.1	15.6	16.6	39.9	46.7	9.6
Ours	58.9	40.3	27.0	18.6	17.9	43.1	54.9	11.1
con2sen	50.6	30.8	18.2	11.3	15.7	37.9	33.9	9.1
feat2sen	51.3	31.3	18.7	11.8	15.3	38.1	35.4	8.8
adv	55.6	35.5	23.1	15.7	17.0	40.8	45.8	10.1
adv + con	56.2	37.2	24.2	16.2	17.3	41.5	48.8	10.5
adv + con + im	56.4	37.5	24.5	16.5	17.4	41.6	49.0	10.5