A Novel Actor Dual-Critic Model for Remote Sensing Image Captioning

Ruchika Chavhan [1] Biplab Banerjee [1] Xiao Xiang Zhu [2] Subhasis Chaudhuri [1]

[1] Indian Institute of Technology, Bombay, India
[2] Signal Processing in Earth Observation, Technical University of Munich, Germany

International Conference on Pattern Recognition (ICPR 2020)

Overview

1. Overview
2. Motivation
3. Problem Statement
4. Prior work
5. Proposed Methodology
6. Results
7. Discussion
8. References

Motivation

Problems with Remote Sensing Image Captioning Data

- Remote sensing images suffer from high inter-class similarity
- Identical reference sentences for multiple images
- Existing caption generators are supervised by these repetitive captions

Why use Reinforcement Learning (RL) ?

- RL based approaches are exploration based
- Captions can be enhanced by increasing exploration of the environment

Problem Statement

Supervised Learning

- Given an image I, a model is trained to maximize the likelihood $p(W|I)$ where $W = w_1, w_2, w_3, ...w_n$ where all w_i are words from a pre-defined vocabulary.
- All supervised learning methods aim to generate sentences that are exactly identical to ground truth

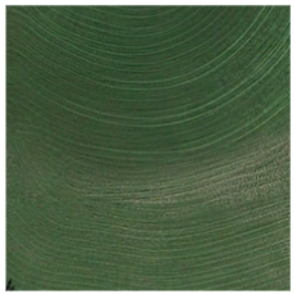
Reinforcement Learning: Actor-Critic Methods

- Given an image I and partially generated sentence $S = (w_1, w_2, ..., w_t)$, w_{t+1} is viewed as an action that a RL policy predicts
- The environment presents a reward for an action that the policy performs.
- A critic criticizes the actions performed by the actor to promote reward maximizing behaviour

Prior Work

Actor-critic sequence training for image captioning Zhang et.al, NIPS, 2017 [ZSL+17].

- Advantage-Actor Critic (A2C) setup based image captioning
- Semantically accurate captions on the COCO Dataset
- Captions generated by this method of remote sensing images provide no new information as compared to ground truth



Original: A tree is near a piece of green meadow
A2C: It is a piece of green meadow.

Original: Wrinkles can be seen in this bright yellow desert
A2C: It is a piece of yellow desert

Original: It is a densely arranged residential area where most houses are with red roofs
A2C: Many rectangular buildings and green trees are in a dense area

Figure: Captions generated by the A2C setup on the RSICD Dataset

Prior Work

Exploring models and data for remote sensing image caption generation , Lu et al, GRSL 2017 [LWZL]

- Generation of the RSICD Dataset consisting of 30 classes
- Various experiments on different kinds of CNNs, RNNs and LSTMs with soft and hard attention
- Cross dataset caption generation on the UCM-captions [QLTL16] (21 classes) and RSICD datasets

Caption: Many planes are parked in an airport

Caption: Five baseball fields are surrounded by green trees

Caption: Many cars are parked near a large building

Figure: Ground truth reference captions from the RSICD dataset

Shortcomings of the A2C setup on RSICD dataset

Captions Generated by A2C setup

- As seen in Figure 1, captions generated by the A2C setup are not semantically diverse and accurate
- Generated captions on remote sensing images are very similar to the ground truth
- This indicates that this type of RL setup is unable to explore the environment efficiently for remote sensing image captioning

Our solution to this problem

- We introduce a novel RL framework: Actor Dual-Critic where an additional critic is deployed in the form of a encoder-decoder LSTM
- This critic performs sentence-to-image translation to validate the predicted sentences and promote prediction of superior captions

Methodology

Actor

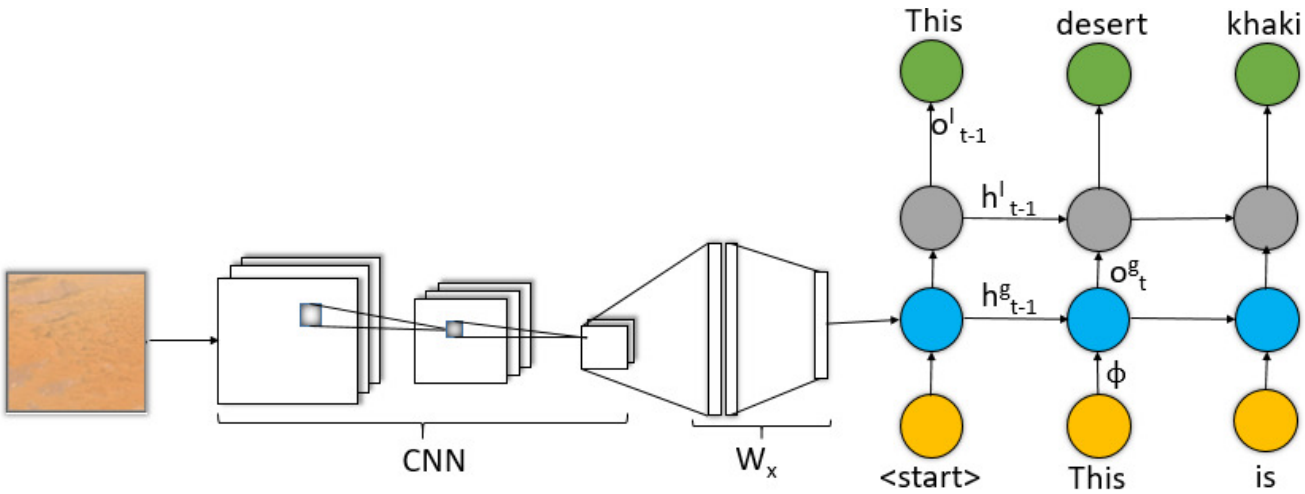


Figure: Working of the actor/policy

- The actor generates captions given the features extracted by a pre-trained CNN
- We observed substantial improvement in performance by employing AlexNet [KSH12]
- The actor provides a measure of confidence $q_\pi(a_t|s_t)$ to predict the next action $\mathbf{a_t} = \mathbf{w_{t+1}} \in \mathbb{R}^d$

Methodology

$$
\begin{aligned}
f &= W_x(\mathsf{CNN}(I)) \\
\phi_0 &= f \\
o_t^g, h_t^g &= \mathsf{GRU}(\phi_{t-1}, h_{t-1}^g) \\
o_t^l, h_t^l &= \mathsf{LSTM}(o_t^g, h_{t-1}^l) \\
q_\pi(a_t|s_t) &= \psi(o_t^l) \\
\phi_t &= \zeta(w_{t-1})
\end{aligned}
\tag{1}
$$

- Here, W_x is the weight of the linear embedding model of the CNN.
- Here, o_t^g and o_t^l are the outputs of the GRU and LSTM respectively at time step t
- $\psi : \mathbb{R}^n \mapsto \mathbb{R}^d$ is a non-linear function that transforms the output of the LSTM to dimension of word embedding model
- $\zeta : \mathbb{R}^d \mapsto \mathbb{R}^n$ denotes the word embedding model
- We denote the policy network by $\pi(a_t|s_{t-1})$.

Methodology

The total optimization objective for the policy is:

$$\min_{\pi} \sum_{t=0}^{T} \log(q_{\pi}(a_t|s_t)) \tag{2}$$

Value Network

- This critic outputs a value function

$$v_{\theta}^{\pi} = \mathbb{E}[\sum_{l=0}^{T-t-1} \gamma^{l} r_{t+l+1} | a_{t+1}, ..., a_T \sim \pi, I] \tag{3}$$

  given a caption $W = (w_1, w_2, ..., w_T)$, features f, and discount factor $\gamma \in [0, 1]$
- Gradients of the policy Parameters are updated using the REINFORCE Algorithm:

$$\mathbf{E}[\sum_{t=0}^{T} A^{\pi}(s_t|a_{t+1}) \nabla \log \pi(a_t|s_{t-1})] \tag{4}$$

  where

$$A^{\pi}(s_t|a_{t+1}) = (\gamma^{T-t-1} r_T - v_{\theta}^{\pi}) \tag{5}$$

Methodology

Encoder-Decoder LSTM critic

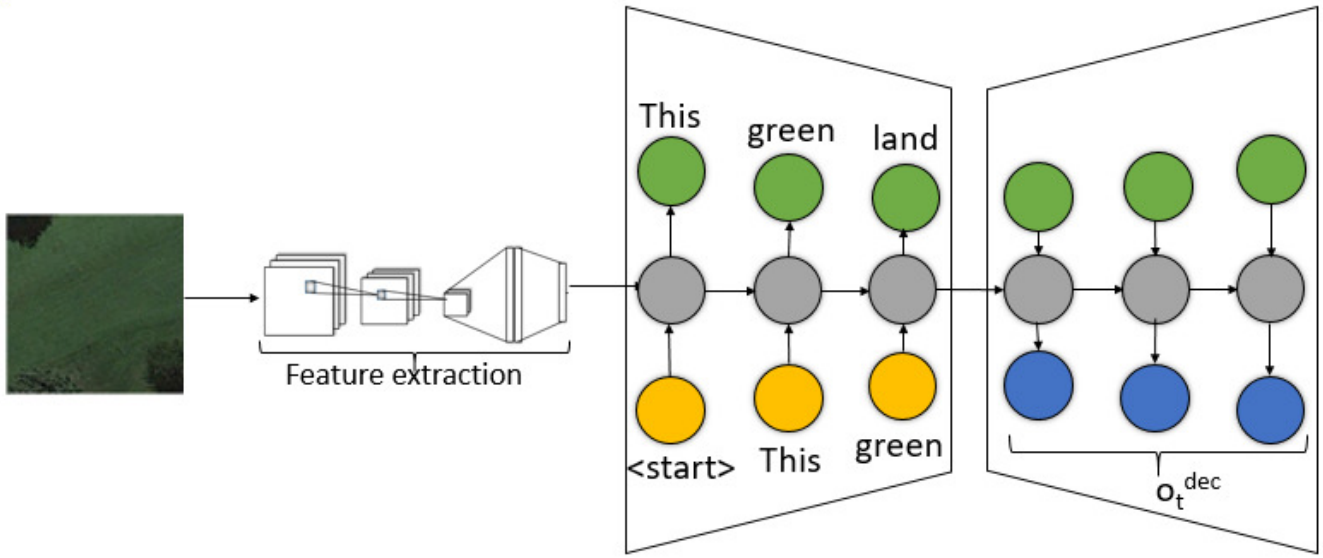


Figure: Working of the Encoder-Decoder LSTM critic

- Image captioning is defined as translation of images into sentences that aptly describe the images.
- This critic translates sentences back into image features to penalize the actor for irrelevant sentences
- It also promotes generation of diverse sentences by capturing semantic information

Methodology

$$
\begin{aligned}
h_0^{enc} &= W_x(\mathsf{CNN}(I)) \\
\eta_t &= \zeta(S) \\
o_t^{enc}, h_t^{enc} &= \mathsf{RNN}_{enc}(\eta_t, h_{t-1}^{enc}) \\
h_0^{dec} &= \psi_2(h_T^{enc}) \\
i_1^{dec} &= \psi_1(o_T^{enc}) \\
o_t^{dec}, h_t^{dec} &= \mathsf{RNN}_{dec}(i_t^{enc}, h_{t-1}^{dec})
\end{aligned}
\tag{6}
$$

- We denote this critic by $D(S)$
- Here, RNN_{enc} and RNN_{dec} are the encoder-decoder RNN respectively
- $S = (w_1, w_2, ..., w_T)$ denotes a natural language description of the image
- $\psi_1, \psi_2 : \mathbb{R}^n \mapsto \mathbb{R}^n$ are non-linear functions that map to word embedding space

Methodology

This loss function for training this critic is:

$$L = (\frac{\sum_{t=0}^{T} o_t^{dec}}{|S|} - f)^2 \tag{7}$$

Defining accuracy between the output of the critic and features extracted by encoder as:

$$A_{gen} = \frac{\frac{\sum_{t=0}^{T} o_t^{dec}}{|S|} f}{\|f\| \left\| \frac{\sum_{t=0}^{T} o_t^{dec}}{|S|} \right\|} \tag{8}$$

We defined an advantage factor for this critic to be:

$$A_{ed} = A_{gen} - \delta_t A_{orig} \tag{9}$$

Here, A_{gen} and A_{orig} are the accuracies of the network when captions generated by the actor and ground truth captions are fed into the encoder respectively.

Algorithm

Algorithm 1 Training Algorithm

Input: Pre-trained models $\pi(a_t|s_{t-1})$, $D(S)$ using the objectives given by the equations 2 and 7 respectively and $V(s_t)$ using the Huber Loss as done in [ZSL+17].

1: **for** $episode = 1$ to total episodes **do**
2: Given an Image I sample action $(a_1, a_2, ..., a_T)$ from the current policy using a multinomial distribution given by $q_\pi(s_t|a_t)$;
3: Calculate advantage factor A^π using the reward r_T for the value network;
4: Update the parameters of the policy using A^π by the REINFORCE Algorithm;
5: Update parameters of the critic by optimising the Huber Loss between r_T and v_θ^π;
6: Calculate advantage factor A_{ed} using the encoder-decoder critic;
7: Update the parameters of the policy using A_{ed} by the REINFORCE Algorithm;
8: Update parameters of the critic using A_{orig}.
9: **end for**

Results

Quantitative results of the Actor Dual-Critic Setup

Metric	B-1	B-2	B-3	B-4	METEOR [LA07]	ROUGE-L	CIDEr[VZP14]
MM [LWZL]	0.57905	0.41871	0.32628	0.26552	0.26103	0.51913	2.05261
SA [LWZL]	0.65638	0.51489	0.41764	0.34464	0.32924	0.61039	1.87415
HA [LWZL]	0.68968	0.5446	0.44396	0.36895	**0.33521**	0.62673	1.98312
A2C [ZSL$^+$17]	0.60157	0.41991	0.364516	0.28788	0.19382	0.63185	2.098
Ours	**0.73973**	**0.55259**	**0.46353**	**0.41016**	0.22126	**0.71311**	**2.243**

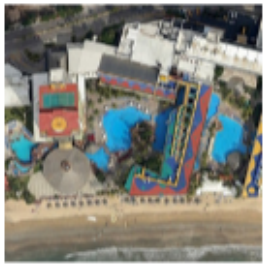
Table: Results of ADC setup on the RSICD dataset

Metric	B-1	B-2	B-3	B-4	METEOR [LA07]	ROUGE-L	CIDEr[VZP14]
MM [LWZL]	0.37066	0.32344	0.32346	0.23259	0.40476	0.4236	1.708
SA [LWZL]	0.79693	0.71345	0.6514	0.59895	0.74952	0.41676	2.12846
HA [LWZL]	0.78498	0.70929	0.65182	0.60167	0.77357	0.43058	2.19594
A2C [ZSL$^+$17]	0.373089	0.23776	0.15857	0.12222	0.39645	0.35989	2.381
Ours	**0.85330**	**0.75679**	**0.67854**	**0.61165**	**0.83242**	**0.80872**	**4.865**

Table: Results of ADC setup on the UCM dataset

Results

A2C: Green trees are in two sides of a curved river.
Ours: This s shaped green river with an island in it goes through this land devided into blocks of farms.



A2C: Some buildings and green trees are in a resort.
Ours: Several buildings with swimming pools and some green plants are near a beach.

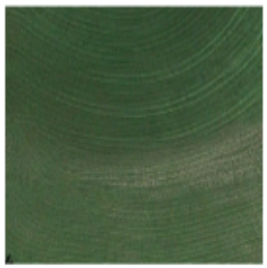
A2C: Many tall buildings are in a commercial area.
Ours: Three rows of skyscrapers stands at this prosperous commercial area.

A2C: It is a piece of yellow desert.
Ours: It is a rather flat desert stained with several black stains

A2C: Some planes are parked near an airport with a parking lot.
Ours: Several white planes are around a circle building with a parking lot

A2C: Many rectangular buildings and green trees are in a dense area.
Ours: Houses with red roofs on both sides of the road

A2C: Many white boats are in the port.
Ours: Two rows of white boats are in port



A2C: It is a piece of green meadow.
Ours: A dirt lines are in this meadow.

Figure: Qualitative results of the ADC setup on the RSICD dataset

Results

(a) **A2C:** There are many airports at the airport.
Ours: There is a red airplane with lots of cars.

(b) **A2C:**There are many airports at the airport.
Ours: There is a red airplane in the airport.

(c) **A2C:** There are many buildings.
Ours: There is one road next to many buildings.

(d) **A2C:** There are lots of cars with some buildings.
Ours: Lots of cars are rectangular and close to each other in the parking lot.

Figure: Qualitative results of the ADC setup on the UCM-captions dataset

Cross-captioning on RSICD dataset

- Model trained on UCM-captions dataset tested on the RSICD Dataset
- Gives an understanding of the model's ability to generalize and utilise it for real time predictions in the absence of labelled data.

Results

Metric	B-1	B-2	B-3	B-4	METEOR [LA07]	ROUGE-L	CIDEr[VZP14]
MM [LWZL]	0.19618	0.01481	0.00721	0.00445	0.07416	0.2457	0.08015
A2C [ZSL+17]	0.19405	0.04137	0.00714	0.00175	0.18855	0.1846	0.961
Ours	**0.38810**	**0.08643**	**0.019065**	**0.00608**	**0.23964**	**0.2888**	**2.013**

Table: Results of cross dataset captioning on the RSICD dataset

(a) **Original:** Many tall buildings are in a commercial area.
Ours: There is one road next to many buildings.

(b) **Original:** Many cars are parked in the parking lot.
Ours: Lots of cars are parked neatly in a parking lot.

(c) **Original:** Yellow beach is between green ocean and green trees.
Ours: This is a beach with blue sea and white sands.

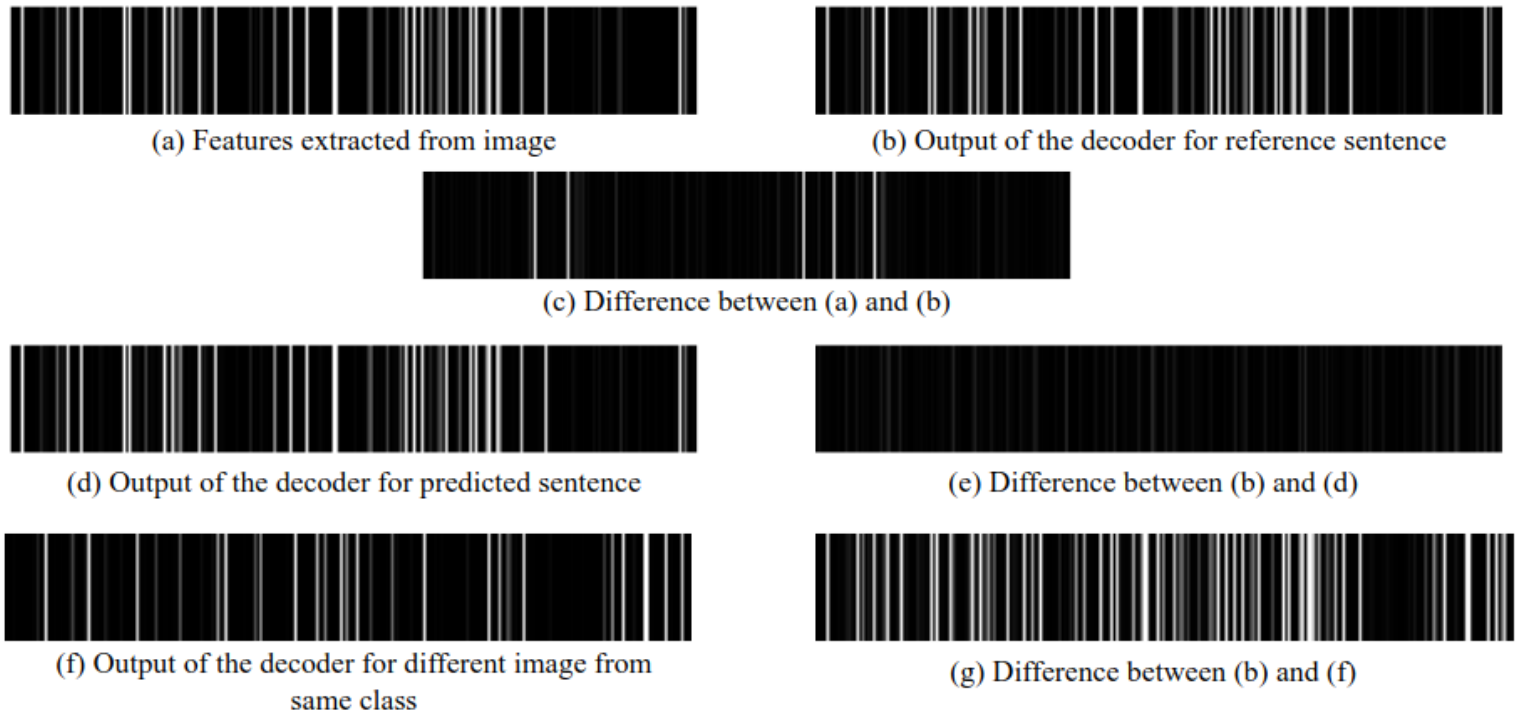
(d) **Original:**On the ground, there are two spherical storage tank.
Ours: Two small storage tanks are on the ground.

Figure: Qualitative comparison of results of cross captioning the ADC setup on the RSICD dataset.

Results

Demonstrating the validity of the proposed critic

- To validate if the critic alleviates high inter class similarity, we pass a different image from the same class with identical reference sentence as the test input



(a) Features extracted from image

(b) Output of the decoder for reference sentence

(c) Difference between (a) and (b)

(d) Output of the decoder for predicted sentence

(e) Difference between (b) and (d)

(f) Output of the decoder for different image from same class

(g) Difference between (b) and (f)

Figure: Qualitative results of the experiment demonstrating the validity of the critic

Discussion

- We proposed an Actor Dual-critic (ADC) method for Image Captioning for the Remote Sensing Image Captioning Dataset
- We introduced another critic to the A2C training setup to encourage the prediction of sentences capturing relevant details along with sentence diversity
- We prove that the policy has gained more knowledge compared to previous works due to this critic's extra upgrade step in the optimization of policy objective.
- The sentences generated by our model provide a highly accurate semantic explanation of the nature and localization of objects in the scene.

References I

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in Neural Information Processing Systems 25 (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2012, pp. 1097–1105.

Alon Lavie and Abhaya Agarwal, *Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments*, Proceedings of the Second Workshop on Statistical Machine Translation (USA), StatMT '07, Association for Computational Linguistics, 2007, p. 228–231.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li, *Exploring models and data for remote sensing image caption generation*, IEEE Transactions on Geoscience and Remote Sensing **56**, no. 4, 2183–2195.

B. Qu, X. Li, D. Tao, and X. Lu, *Deep semantic understanding of high resolution remote sensing image*, 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), 2016, pp. 1–5.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh, *Cider: Consensus-based image description evaluation*, CoRR **abs/1411.5726** (2014).

References II

Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy Hospedales, *Actor-critic sequence training for image captioning*.

The End