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

**Figure**: Ground truth reference captions from the RSICD dataset

- **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
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
Proposed Methodology

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 \( a_t = w_{t+1} \in \mathbb{R}^d \)
Methodology

\[ f = W_x(CNN(I)) \]
\[ \phi_0 = f \]
\[ o_t^g, h_t^g = \text{GRU}(\phi_{t-1}, h_{t-1}^g) \]
\[ o_t^l, h_t^l = \text{LSTM}(o_{t}^g, h_{t-1}^l) \]
\[ q_\pi(a_t|s_{t}) = \psi(o_t^l) \]
\[ \phi_t = \zeta(w_{t-1}) \]

(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 the 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:

\[
\min_{\pi} \sum_{t=0}^{T} \log(q_\pi(a_t|s_t))
\]  

(2)

### Value Network

- This critic outputs a value function

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

(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:

\[
\mathbb{E}\left[ \sum_{t=0}^{T} A^\pi(s_t|a_{t+1}) \nabla \log \pi(a_t|s_{t-1}) \right]
\]  

(4)

where

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

(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.
Proposed Methodology

Methodology

\[ h_{\text{enc}} = W_x(\text{CNN}(I)) \]
\[ \eta_t = \zeta(S) \]
\[ o_{t,\text{enc}}, h_{t,\text{enc}} = \text{RNN}_{\text{enc}}(\eta_t, h_{t-1}^{\text{enc}}) \]
\[ h_{0,\text{dec}} = \psi_2(h_{T,\text{enc}}) \]
\[ i_{1,\text{dec}} = \psi_1(o_{T,\text{enc}}) \]
\[ o_{t,\text{dec}}, h_{t,\text{dec}} = \text{RNN}_{\text{dec}}(i_{t,\text{enc}}, h_{t-1}^{\text{dec}}) \]

(6)

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

**Methodology**

This loss function for training this critic is:

\[
L = \left( \sum_{t=0}^{T} o_{t}^{dec} \frac{|S|}{|S|} - f \right)^2
\]  

(7)

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

\[
A_{gen} = \frac{\sum_{t=0}^{T} o_{t}^{dec}}{|S|} \frac{f}{\|f\|} \frac{\|\sum_{t=0}^{T} o_{t}^{dec}\|}{\|\sum_{t=0}^{T} o_{t}^{dec}\|}
\]  

(8)

We defined an advantage factor for this critic to be:

\[
A_{ed} = A_{gen} - \delta_{t} A_{orig}
\]  

(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 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 $A^{\pi}$ using the reward $r_T$ for the value network;
4: Update the parameters of the policy using $A^{\pi}$ by the REINFORCE Algorithm;
5: Update parameters of the critic by optimising the Huber Loss between $r_T$ and $v^\pi_\theta$;
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

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

**Table:** Results of ADC setup on the RSICD dataset

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

**Table:** Results of ADC setup on the UCM dataset
Results

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

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 divided 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.
Results

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>METEOR [LA07]</th>
<th>ROUGE-L</th>
<th>CIDEr [VZP14]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>0.19618</td>
<td>0.01481</td>
<td>0.00721</td>
<td>0.00445</td>
<td>0.07416</td>
<td>0.2457</td>
<td>0.08015</td>
</tr>
<tr>
<td>A2C [ZSL\textsuperscript{+17}]</td>
<td>0.19405</td>
<td>0.04137</td>
<td>0.00714</td>
<td>0.00175</td>
<td>0.18855</td>
<td>0.1846</td>
<td>0.961</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.38810</strong></td>
<td><strong>0.08643</strong></td>
<td><strong>0.019065</strong></td>
<td><strong>0.00608</strong></td>
<td><strong>0.23964</strong></td>
<td><strong>0.2888</strong></td>
<td><strong>2.013</strong></td>
</tr>
</tbody>
</table>

**Table:** Results of cross dataset captioning on the RSICD dataset

![Image1](image1.png) ![Image2](image2.png) ![Image3](image3.png) ![Image4](image4.png)

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

**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


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