# 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

## Introduction

Remote Sensing Image Captioning Data (RSICD) [1] suffers from high inter class similarity and identical reference sentences for multiple images.

Figure 1:Images from the RSICD Dataset that have the caption "It is a piece of yellow desert"

Supervised Learning methods generate sentences identical to these repetitive captions. In contrast, Reinforcement Learning methods are exploration based methods which can be leveraged to generate diverse sentences.

## Related Work

- *Actor-critic sequence training for image captioning* Zhang et.al, NIPS, 2017 [2].
  - Advantage-Actor Critic (A2C) setup based image captioning
  - Captions generated by this method of remote sensing images provide no new semantic information
- *Exploring models and data for remote sensing image caption generation* , Lu et al, GRSL 2017 [1]
  - Generation of the RSICD Dataset consisting of 30 classes
  - Various experiments on different kinds of CNNs, RNNs and LSTMs with soft and hard attention

## Methodology

- In the ADC setup, an actor generates a sentence given the image and the critics provide rewards based on the quality and relevance of this sentence. In contrast to [2], we utilize two critics: a **Value Function** $v_\theta^\pi(s_t)$ and an **Encoder-Decoder RNN critic** $D(S)$ to provide two rewards to update parameters of the actor using the REINFORCE Algorithm.
- **Learning the policy:** The actor/policy consists of a pre-trained feature extractor and a LSTM which is trained to generate a caption given these features. The actor in Figure 2 provides a measure of confidence $q_\pi(a_t|s_t)$ to predict the next action $\mathbf{a_t} = \mathbf{w_{t+1}} \in \mathbb{R}^d$. The total optimization objective for the policy is:

$$\min_\pi \sum_{t=0}^{T} \log(q_\pi(a_t|s_t)) \tag{1}$$

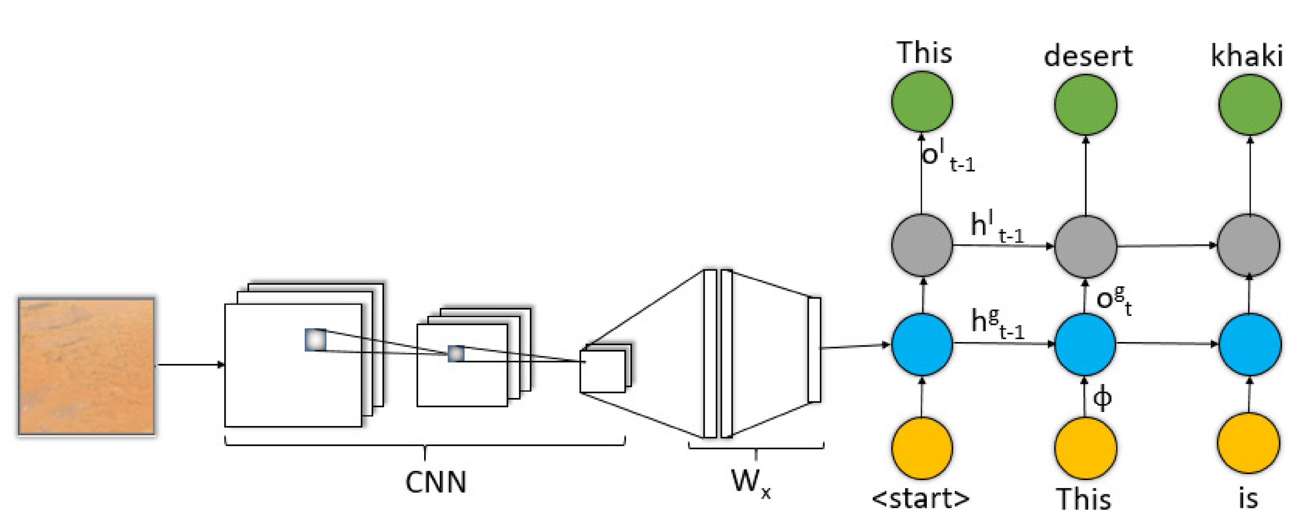


Figure 2:Working of the actor. The words are converted into the embedding space (not shown) before being fed into the GRU (denoted by blue) and LSTM (denoted by grey).

## The Actor Dual-Critic Setup

We introduce an additional critic to the A2C setup in the form of an Encoder-Decoder LSTM that jointly encodes the sentences and images and encourages prediction of semantically more precise sentences. The critic provides a significant boost in performance in the following ways:

- **Sentence to image translation:** This critic translates back into images to generate a quantity which closely resembles the features extracted from images, thus validating the contextual accuracy of the predicted captions.
- **More exploration of environment:** The policy successfully investigates the environment consisting of images and captions and gains more knowledge due to this critic's extra upgrade step in the optimization of policy objective.

We demonstrate the efficiency of our proposed method on two datasets: RSICD and UCM-captions and perform interesting experiments to validate the functionality of the critic

## Proposed Critic

This loss function for training the critic in Figure 3 is given by

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

Here $S$ denotes the reference caption and $f$ denotes the features extracted by a pre-trained CNN. We define accuracy $A$ between the output of the critic and features extracted by encoder as cosine distance between $\frac{\sum_{t=0}^{T} o_t^{dec}}{|S|}$ and features $f$

The advantage factor for this critic is defined as: $A_{ed} = A_{gen} - \delta_t A_{orig}$. 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 this critic.

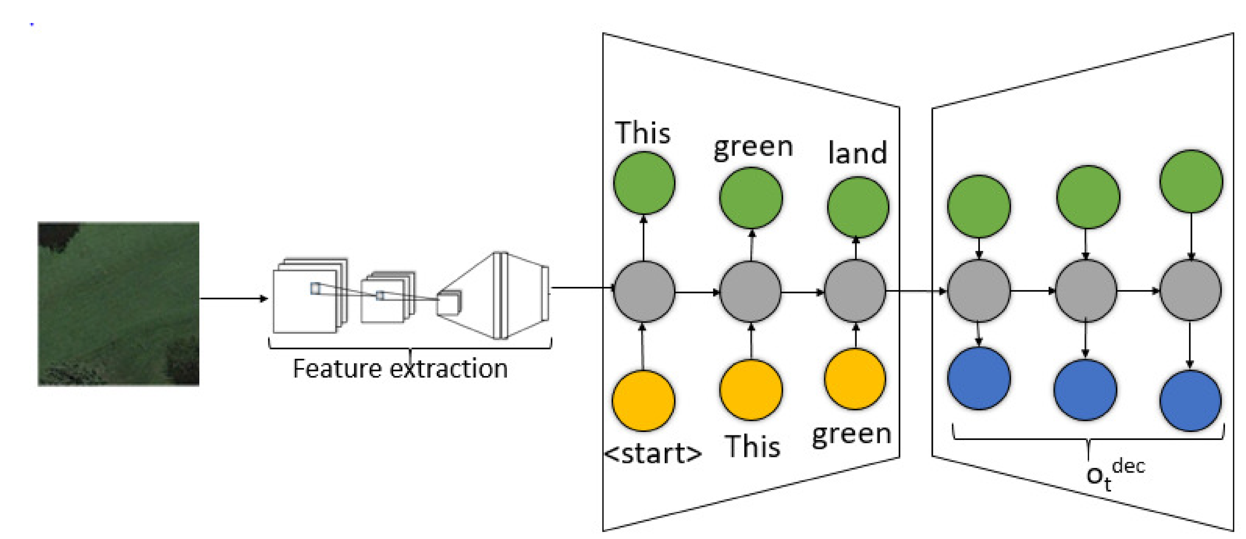


Figure 3:Working of the proposed Encoder-Decoder RNN critic.

**Algorithm 1** Training Algorithm

**Input:** Pre-trained models $\pi(a_t|s_{t-1})$, $D(S)$ using the objectives given by the equations 1 and 2 respectively and $V(s_t)$ using the Huber Loss.

```
for episode = 1 to total episodes do
    Given an Image I sample action (a_1, a_2, ..., a_T) from the current policy using a multinomial distribution given by q_π(s_t|a_t);
    Calculate advantage factor A^π using the reward r_T for the value network;
    Update the parameters of the policy using A^π by the REINFORCE Algorithm;
    Update parameters of the critic by optimising the Huber Loss between r_T and v_θ^π;
    Calculate advantage factor A_ed using the encoder-decoder critic;
    Update the parameters of the policy using A_ed by the REINFORCE Algorithm;
    Update parameters of the critic using A_orig.
end for
```

## Results

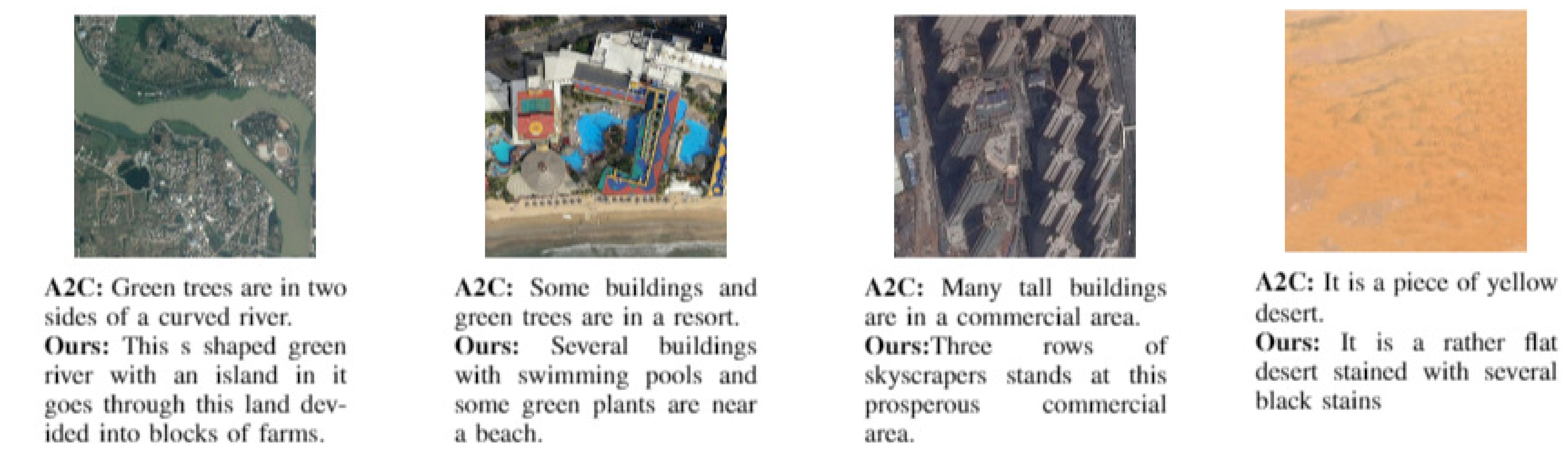


Figure 4:Captions generated by the ADC method on RSICD Dataset

| Dataset | RSCID | | | UCM-captions | | |
|---|---|---|---|---|---|---|
| Metric | B-4 | ROUGE-L | CIDEr | B-4 | ROUGE-L | CIDEr |
| MM [1] | 0.2655 | 0.51913 | 2.05261 | 0.2325 | 0.4236 | 1.708 |
| SA [1] | 0.3446 | 0.61039 | 1.87415 | 0.5989 | 0.4167 | 2.128 |
| HA [1] | 0.3689 | 0.62673 | 1.98312 | 0.6016 | 0.4305 | 2.195 |
| A2C [2] | 0.2878 | 0.63185 | 2.098 | 0.1222 | 0.3598 | 2.381 |
| Ours | **0.4101** | **0.71311** | **2.243** | **0.6116** | **0.8087** | **4.865** |

Table 1:Results of ADC setup on the RSICD and UCM-captions dataset

## An experiment demonstrating the validity of the 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 5:Qualitative results of the experiment demonstrating the validity of the critic

## Remarks

- 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

## References

[1] 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(4):2183--2195.

[2] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy Hospedales.
Actor-critic sequence training for image captioning.
06 2017.

Paper