Describing Wildlife with Deep Learning: Techniques for Accurate Image Captioning

Shabbir Hussain
Lahore University of Management Sciences
Lahore, Punjab, Pakistan

Humaira Fasih Ahmed Hashmi
Lahore University of Management Sciences
Lahore, Punjab, Pakistan

ABSTRACT

Wildlife preservation is crucial for maintaining ecological balance and biodiversity. By observing animals in their natural habitat we can make informed decisions and improve conservation efforts. One common method to study animals is through the use of camera traps, which provide a non-invasive way to capture images and videos of wildlife without disturbing them. However, analysis and use of data generated by camera traps has largely been a laborious and manual activity. Much research effort has thus focused on automating this task. In this work, we build upon these efforts to investigate the possibility of captioning camera trap images, thus extracting more detailed and useful information from these images to aid in wildlife conservation efforts.

KEYWORDS

Image Captioning, Classification, Deep Neural Networks, Inception, LSTM

INTRODUCTION

Comprehensive animal statistics are needed to manage and conserve natural ecosystems. Over the past 20 years, motion-sensor cameras, also known as camera traps, have revolutionised wildlife ecology and conservation by collecting massive amounts of data without disrupting the ecosystem [2]. Camera trap images can identify rare and elusive species that traditional survey methods overlook. Ecologists use these devices to estimate population numbers and distributions [7], analyse animal habitat use [3], and prioritize conservation efforts.

Traditionally, experts or volunteers analyse camera trap data to find useful information and statistics. These camera traps generate an extensive amount of data, making manual analysis challenging, time-consuming, and expensive. Camera trap images can be redundant, blurry, or have shadows and poor lighting, making the task harder. Most current research is hence focusing on automated methods that use deep learning to extract relevant information from camera trap images.

Building on previous efforts, we propose an approach that utilizes standard natural language processing techniques and deep neural networks to generate accurate captions for camera trap images. The captions are comprehensive and encapsulate species identification, count, and activity attributes.

RELATED WORK

In recent years, several research efforts have focused on employing deep neural networks and assessing the efficacy of the performance of these networks. One prominent effort is by H. Nguyen et al. [5] reported in the paper titled **“Animal Recognition and Identification with Deep Convolutional Neural Networks for Automated Wildlife Monitoring"**, which proposes a framework for automated animal recognition in the wild using deep convolutional neural networks. The paper used a single-labeled dataset from the Wildlife Spotter Project consisting of 108,944 labeled images of wild animals and birds taken in South-central Victoria, Australia. The authors applied convolutional neural network architectures, such as VGGNet, ResNet, and Lite AlexNet which achieved up to 96.6% accuracy and 93.7% F1-score for detecting images containing animals. However, the classification task achieved only 90% accuracy for identifying the three most common species in the image indicating there is room for improvement. Another effort focused on specie identification is that of Gomez et al. [4] outlined in the paper titled **“Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks"** whose choice of dataset is identical to ours: Snapshot Serengeti. However, Gomez et al. use 26 specie classes from the SS dataset, and omit the last 22 classes which have less training samples. We use 32 classes, and our approach does not involve such systematic removal. Our model captions images with the correct specie even when there are less training samples for that specie class in the training dataset. The above two works focus mainly on identifying animals, and do not provide any additional information about their behavior or activities. The closest work to our approach is

Table 1: Frequency of Different Species in Dataset

Species Name	Dataset Frequency
Giraffe	444
Buffalo	432
Grant’s gazelle	370
Impala	308
Bird	274
Warthog	257
Hippopotamus	249
Elephant	244
Spotted hyena	232
Reedbuck	158
Guinea fowl	158
Dik dik	141
Topi	121
Baboon	55
Kori bustard	36
Secretarybird	33
Female lion	31
Vervet monkey	27
Jackal	26
Ostrich	24
Serval	21
Hare	17
Cheetah	14
Zebra	9
Bat-eared fox	6
Male lion	6
Waterbuck	3
Wildebeest	3
Eland	1

reported by Norouzzadeh et al. [6] in their paper **"Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning"** aimed to provide a cutting-edge deep learning approach for identifying, counting, and describing behaviors of wild animals in camera-trap images. This paper used the Snapshot Serengeti dataset too, and trained deep convolutional neural networks in emsembles to identify species, count animals, and predict their behaviors. The paper shows that deep learning can handle large amounts of visual data with high accuracy and efficiency, and that animal identification for 99.3% of the 3.2 million-image dataset can be automated with the same 96.6% accuracy as crowdsourced teams of human volunteers. This paper classifies, counts, and describes species, however, it does not generate natural language descriptions of images. Moreover, their activity attribute multilabel classification accuracy is 75%, which could be improved. Another effort in this domain is one reported in the paper **"Insights and approaches using deep learning to classify wildlife"**. The purpose of this paper is to evaluate the methods and visual features learned by deep learning algorithms for classifying wildlife from camera trap data. It presents the results of training a convolutional neural network (CNN) to classify 20 African wildlife species with an overall accuracy of 87.5% from a fully annotated dataset from Gorongosa National Park, Mozambique, containing 111,467 images. It uses Grad-CAM with VGG-16 and ResNet-50 architecture to extract the most salient pixels in the final convolution layer and interpret the image features with the strongest responses. It also utilizes hierarchical clustering of feature vectors has been to produce a visual similarity dendrogram of identified species, and Mutual Information (MI) to determine the importance of each output layer neuron across a species. While the efforts of the authors are to be lauded for providing insights into the methods used by deep learning algorithms to classify wildlife species, the reported accuracy of 87.5% in this paper is lower than the accuracies reported in the other two papers.

While previous studies have shown promising results using deep neural networks (DNNs) that are trained on larger and more balanced datasets than the one used in our approach, our method is still effective at identifying animal species even with unbalanced and low-frequency class samples (See frequencies of specie classes in Table 1). We intentionally did not balance the dataset to ensure that our model is more robust to capturing the real-life imbalance and bias in the dataset.

DATASET

To train deep neural networks for the identification and captioning of animal species, it is necessary to have image and caption data. In our research, we utilized one of the largest publicly available datasets which was generated by Snapshot Serengeti, a large-scale survey [2] that deployed 225 camera traps across a 1,125 km2 area in the Serengeti National Park in Tanzania[8]. This data contained 3.2 million images of 48 different species. The labelling of this data (species type, count, and behavior) was done by 28,000 registered and approximately 40,000 unregistered volunteer citizen-scientists through a website monitored by the project members. Each image set was circulated to multiple volunteers to improve the accuracy of the data. A simple plurality algorithm was then applied to produce a ‘consensus dataset’ of final classifications for each image set. The consensus classifications were validated against 4,149 ‘gold-standard’ image-sets that had been classified by experts, resulting in 96.6% accuracy for species identifications and 90% accuracy for species counts. [1]

To generate captions for our work, we used the labelled annotations provided on Lila BC's official website and developed a script. Some limitations in the generation of captions are as follows:

- We used only the most prominent species per image for the species type even in the presence of multiple animals in our captions, resulting in an inherent bias towards the most visible animal species in the captions.
- We used categories based on a range of values instead of exact counts. For example, if the count of animals was between 3-6, we categorized them as 'Few'; if it was between 7-10, we classified them as 'A group,' and for higher values, we used 'Many.'
- The annotations included proportion of annotator who chose each behavioural activity or presence of young [1], and we included the attribute in our caption if more than 50% of the annotators agreed upon it.

Due to computational resource and time restrictions, we trained our model on only 3834 images containing 32 species from the SS dataset. Despite these limitations, our model still produced excellent results.

METHODOLOGY

Our baseline model is a simple encoder-decoder architecture where the encoder is used to extract useful features from the image and project them into a latent dimension which is defined as the embedding size. The decoder mainly consists of an LSTM layer which is a RNN and a linear layer that projects the result into vocabulary dimension. During training, the model receives an image, and its objective is to generate a corresponding caption for the image. To ensure that the model generates accurate captions, we employ the approach of teacher-forcing and feed the correct captions to our model as ground truth (Fig. 1). This means that at each time step, the model receives the correct word as input, rather than using its own predicted word from the previous time step. This approach helps eliminate accumulation of any errors made by the model in the initial words while generating captions which could lead to further errors later on. By providing the correct words as input, we force the model to generate accurate captions from the beginning of the sequence.

Our initial approach was to use ResNet50 pretrained on the ImageNet dataset to extract image features as it is light weight and has shown to work well on image classification tasks. We removed the classification layer from ResNet50 and replaced it with a linear layer to project the features into our desired embedding dimension. The weights of all layers except the last one were frozen to stop our model from doing unnecessary backpropagation. Even though this basic architecture performed well on larger datasets, on our dataset it did not. The loss decreased rapidly at the start but then it became stagnant when the model memorized a few tokens that helped achieve small loss over the entire dataset instead of extracting useful features to distinguish between species, their count, and other attributes. The resultant captions were poorly structured and did not convey any meaning.

We tried using multi-headed attention to resolve the above issue. The idea was that the model would be able to better understand sentence structure and relate them to image features. Upon training the network after implementing this strategy, we observed that the loss decreased, and the model started producing results that were grammatically correct. However, since our captions were simple and the vocabulary size was limited to 74, the probabilities of each word generated by the LSTM lied very close to each other for every image. Consequently, the model outputted same captions for different images.

Our next approach to fix our model was to use a different encoder network. After experimenting with several models, we found that InceptionV3 produced the best results. Unlike our previous attempt, we did not freeze the weights of the model. Our reasoning for this was that camera trap images are different from images in the ImageNet dataset. ImageNet contains high-quality images of daily life activities and objects, and there is a lot of variety in the dataset. However, camera trap images are all very similar, and it is possible that the model was unable to extract useful information without learning the true representation of our dataset. This reasoning proved to be correct as the model was able to generate captions that closely resembled the actual captions.

Another problem we observed with our network was that the attention mechanism was slowing down the training process. Whenever we used the attention mechanism, our model tended to learn a single sentence and repeatedly generate it. As a result, we decided to remove the attention mechanism altogether. Finally, we identified a problem with our model's ability to identify species that appeared less frequently in our dataset. To address this issue, we added a classification layer to our encoder and included the classification loss in the overall loss. This approach ensured that the model learned features specifically useful for classifying the species of the animal or bird in the image, rather than random features. As a result, our model not only improved its ability to classify less frequent species but also demonstrated better generalization and avoided overfitting to the training dataset.

RESULTS

The overall performance of our model has been reported using BLEU (bilingual evaluation understudy) score in Table 2 which is a commonly used metric for evaluating the quality of machine-generated translations or captions by comparing

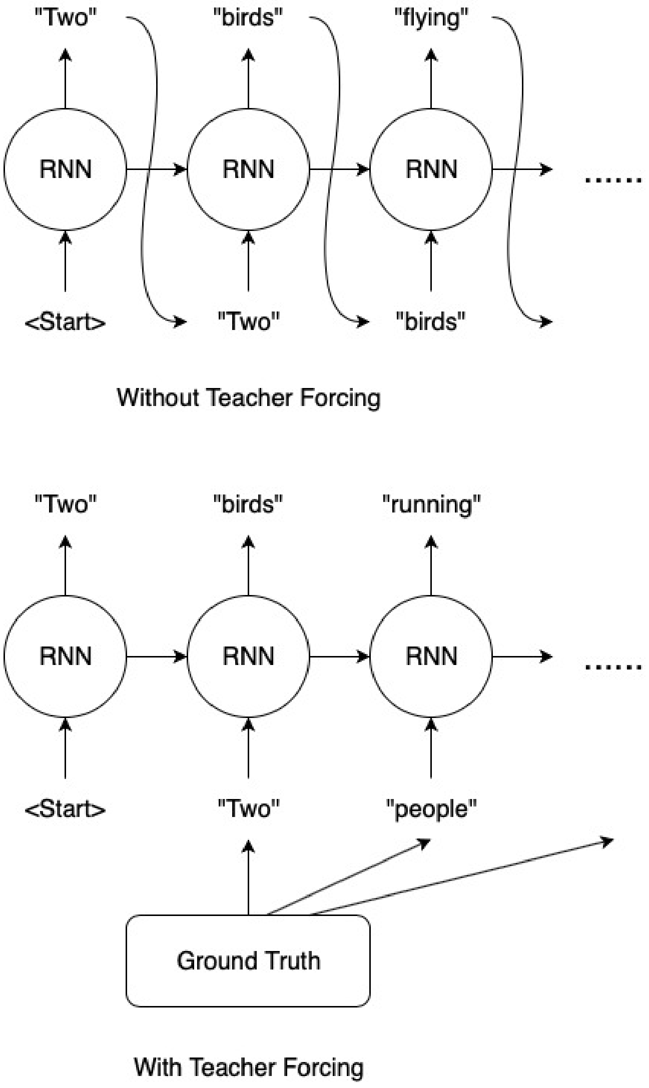


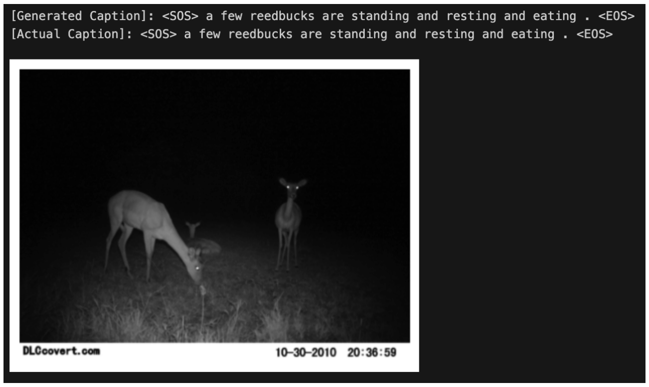
Figure 1: Example of RNN with and without teacher-forcing

them with human-written reference translations or captions. We generated captions for all 3800+ images in our dataset and compared them with the actual captions. We've reported BLEU-1, BLEU-2, BLEU-3, and BLEU-4 for the entire dataset which measure the overlap between model generated caption and actual caption at different n-gram level (from 1 to 4 respectively).

Attached are examples that demonstrate the accuracy and structure of the generated captions on different specie classes,including the less frequent ones.

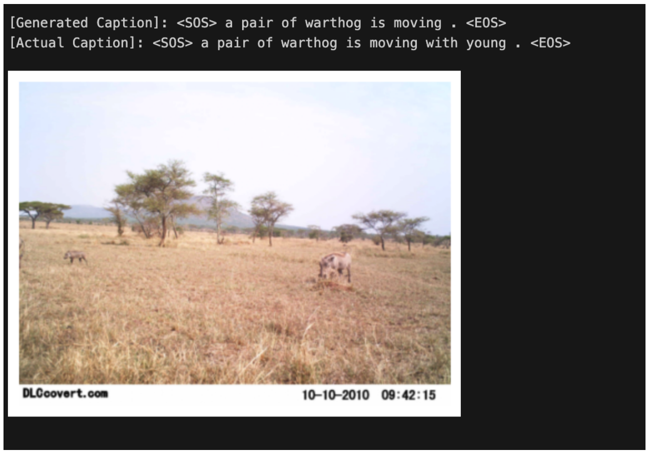
Table 2: BLEU scores over the entire dataset

Type	Score
BLEU-1	0.9313777663096084
BLEU-2	0.9004411297229316
BLEU-3	0.8658755091948893
BLEU-4	0.8372741890588865



```
[Generated Caption]: <SOS> a few reedbucks are standing and resting and eating . <EOS>
[Actual Caption]: <SOS> a few reedbucks are standing and resting and eating . <EOS>
```

Figure 2: Result on an image containing reedbucks



```
[Generated Caption]: <SOS> a pair of warthog is moving . <EOS>
[Actual Caption]: <SOS> a pair of warthog is moving with young . <EOS>
```

Figure 3: Result on an image containing warthogs

FUTURE WORK & IMPROVEMENTS

To improve the applicability of the current model, it should be trained on a larger dataset that includes a broader range of species beyond the 32 classes. Additionally, there is a need for better captions as the current ones are simplistic and only contain relevant classification, count, and activity information extracted from publicly available volunteer-labelled dataset annotations. To enhance the quality of the captions, more detailed information such as whether the activity took place during nighttime or daytime can be incorporated. These improvements can increase the accuracy and usefulness of the model in wildlife conservation efforts.

CONCLUSION

In this work, we combined language processing techniques with a popular state-of-the-art deep neural network - InceptionV3 - to extract relevant information from camera trap data and describe wildlife via captions. We explored the feasibility of image captioning on camera trap data with limited, unbalanced training data with significant success.

REFERENCES

[1] 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data* 2, 1 (2015), 150026–150026. https://doi.org/10.1038/SDATA.2015.26

[2] O'Connell Allan F., Nichols James D., and Karanth K. Ullas. 2011. *Camera Traps in Animal Ecology : Methods and Analyses.* Springer. https://search.ebscohost.com/login.aspx?direct=true&db=edsebk&AN=371104&site=eds-live

[3] Andrew Bowkett, Francesco Rovero, and Andrew Marshall. 2007. The use of camera-trap data to model habitat use by antelope species in the Udzungwa Mountain forests, Tanzania. *African Journal of Ecology* 46 (11 2007), 479 – 487. https://doi.org/10.1111/j.1365-2028.2007.00881.x

[4] Alexander Gómez, Augusto Salazar, and Jesús Francisco Vargas-Bonilla. 2016. Towards Automatic Wild Animal Monitoring: Identification of Animal Species in Camera-trap Images using Very Deep Convolutional Neural Networks. *CoRR* abs/1603.06169 (2016). arXiv:1603.06169 http://arxiv.org/abs/1603.06169

[5] Hung Nguyen, Sarah J. Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G. Ritchie, and Dinh Phung. 2017. Animal Recognition and Identification with Deep Convolutional Neural Networks for Automated Wildlife Monitoring. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 40–49. https://doi.org/10.1109/DSAA.2017.31

[6] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Ali Swanson, Craig Packer, and Jeff Clune. 2017. Automatically identifying wild animals in camera trap images with deep learning. *CoRR* abs/1703.05830 (2017). arXiv:1703.05830 http://arxiv.org/abs/1703.05830

[7] Leandro Silveira, Anah T.A. Jácomo, and José Alexandre F. Diniz-Filho. 2003. Camera trap, line transect census and track surveys: a comparative evaluation. *Biological Conservation* 114, 3 (2003), 351–355. https://doi.org/10.1016/S0006-3207(03)00063-6

[8] Alison Surridge, Robert Timmins, Godfrey Hewitt, and Diana Bell. 1999. Striped rabbits in Southeast Asia. *Nature* 400 (08 1999). https://doi.org/10.1038/23393