Image Captioning using Flickr8K Dataset

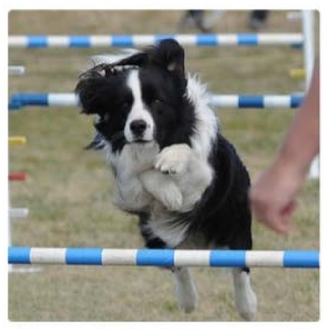
Aditya Kismatrao, Manas Mishra, Pratik Pandey, Yash Srivastava
Robert H. Smith School of Business
University of Maryland, College Park
{amkismat, manasm, pratikp, yashsri}@umd.edu

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Introduction

Image Captioning is a field of artificial intelligence that is swiftly becoming an important aspect of computer vision. Through our project, we aim to address issues in different domains like medicine, biodiversity, education, and defense by optimizing content-based image retrieval (CBIR) through image indexing. There is a pressing need for the creation of sophisticated solutions in all the below mentioned areas:

- Development of software that aid in fingerprint detection to reduce crime,
- Help generate real-time human-like captions for pictures circulating on social media,
- Creation of advanced digital libraries through artificial intelligence,
- Detection of new species of living organisms through the application of advanced modeling techniques,
- Help a visually impaired person to hear videos and socialize better.


"black and white dog jumps over bar."

Related Work

The recent after developments in deep learning led to drastic improvements from primitive visual recognizer and rule based methodology.

1. Template based method:
Uses fixed template (Action-verb-action); creates less natural caption but works well with out of box components.

2. Transfer based method:
Uses image retrieval; generates more human like captions but adds too many redundant words.

Flickr8K Dataset

The data set contains 8,092 images of people and animals performing some action handpicked from picture sharing website, Flickr.com. To provide conceptual descriptions of the images, we train our model using 5 captions provided by a few selected people for each image. The images and captions are handpicked to increase the diversity and randomness of the dataset.

Methodology

1. Generation of Image Features:

Image features are generated using the VGG16 Convolutional Neural Network (CNN) architecture pre-trained on the ImageNet dataset.

2. Captions for Supervised Learning:

The primary objective of the model is to output a caption for the image. Hence, the caption is used as a target variable. The way captions are converted is to make each word of the caption as a target variable appended with "startcap" and "endcap" tokens.

3. Combine Image and Text Features:

The image and text features are combined using the model architecture shown here containing LSTM, Embedding and Fully-Connected layers.

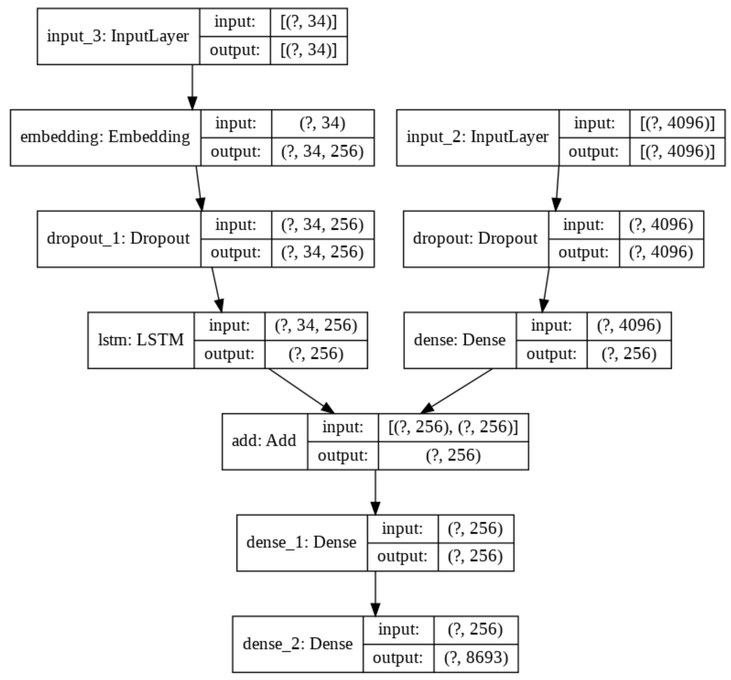


Figure 1: Model Architecture combining Image and Text features

Experimental Setup, Results and Observations

Parameters	Values
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Number of Epochs	10
Batch Size	32
Dropout rate	0.5

Model Settings

Loss	Value
Training	2.5230
Validation	4.1712

Loss Metrics

Predicted Caption: with a pierced necklace is smiling with a drink in the background and a crowd of people in the background is wearing a blitz hat and a red headband and a red mohawk

Actual Caption(s):
1. man in an orange hat staring at something
2. man wears an orange hat and glasses
3. man with gauges and glasses is wearing a blitz hat
4. man with glasses is wearing a beer can crochet hat
5. man with pierced ears is wearing glasses and an orange hat

Conclusion and References

Conclusion: Image captioning is one of the most sophisticated artificial intelligence tools today. Inspired by recent advances in deep learning, neural machine translation models have lately led to drastic improvements in image captioning domain. It has applications in multiple domains as listed in our report. We believe that the model we have built can apply to a wider gamut of problems, not limited to fields mentioned in the paper. When we compare our results to baseline architectures, our concerns about image captioning's reliability are alleviated.

References: Vinyals, Oriol Toshev, Alexander Bengio, Samy Erhan, Dumitru. (2015). Show and tell: A neural image caption generator. 3156-3164. 10.1109/CVPR.2015.7298935.