Image Captioning in Bangla Using Encoder-Decoder Architecture and Hybrid Dataset

Members

Kawsar Ahmed
170041021

Safayet Hossain Masum
170041050

Introduction

Image Captioning refers to the process of generating textual description of an image.

Significance:

Automatically generating captions of an image is a task very close to the heart of scene understanding — one of the primary goals of computer vision.

Amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language.

Problem Statement

Given an image as input, we want to come up with a model that will generate a caption of that image in Bangla with as much accuracy as possible.

Reference :

ভিনটি পাখি আছে।
There are three birds.

৩ টি চড়াই পাখি বসে আছে।
Three sparrows are sitting.

Predicted :

Greedy:

একজন পুরুষ বসে আছে।
A man is sitting.

BLEU-1 : 0.50 BLEU-2 : 0.41 ROUGE_L: 0.39 SPICE : 0.42

Beam:

একজন মানুষ বসে আছে।
A man is sitting.

BLEU-1 : 0.50 BLEU-2 : 0.41 ROUGE_L: 0.39 SPICE : 0.43

Literature Review

- Chittron: An Automatic Bangla Image Captioning System

  Publication: Procedia Computer Science (2019)

- Improved Bengali Image Captioning via deep convolutional neural network based encoder-decoder model

  Publication: International Joint Conference on Advances in Computational Intelligence 2021

- A Hybridized Deep Learning Method for Bengali Image Captioning

  Publication: International Journal of Advanced Computer Science and Applications, 2021

Dataset

BanglaLekhaImageCaptions Dataset

- 9,154 images with two captions for each image.
- Captions are generated by two native Bengali speakers.
- Has a considerable amount of human bias which hinders any model's ability to describe non-human subjects.
- 7154 images for training, 1000 images for validation and 1000 images for testing

Translated Flickr8k Dataset:

- 8000 Images with 5 captions for each of the image
- Translated to Bangla from English using google translate

Dataset: BanglaLekhaImageCaptions

- 'একজন পুরুষ ও একজন নারী বসে মাটির পাত্র বানাচ্ছে।'
- 'একটি লোক ও একটি মেয়ে হাড়ি বানাচ্ছে।'

- 'অনেকগুলো নারী ও শিশু আছে।'
- 'একটি ঘরের বারান্দায় দাড়িয়ে আছে অনেকগুলো নারী এবং বাচ্চা যার মাঝে একজন বিদেশি মহিলা একটি বাচ্চাকে কুলে নিয়ে আছে। '

Dataset: BanglaLekhaImageCaptions

Advantages

- Bangladeshi Landscape
- No Translation needed
- Variety of Bangla objects

Disadvantages

- Small in size
- Human Bias

Dataset: Translated Flickr8k

- Large brown dog running in the grass
- বড় বাদামী কুকুর ঘাসে দৌড়াচ্ছে
- A brown dog is playing with a garden hose
- একটি বাদামী কুকুর বাগানের পায়ের পাতার মোজাবিশেষ সঙ্গে খেলছে
- A brown dog chases the water from a sprinkler on a lawn
- একটি বাদামী কুকুর একটি লনে একটি স্প্রিংকলার থেকে জল তাড়া করছে
- A brown dog running on a lawn near a garden hose
- একটি বাদামী কুকুর একটি বাগান পায়ের পাতার মোজাবিশেষ কাছাকাছি একটি লনে ছুটছে
- A dog is playing with a hose
- একটি কুকুর একটি পায়ের পাতার মোজাবিশেষ সঙ্গে খেলছে

- Man and child in a yellow kayak
- একটি হলুদ কায়াক মানুষ এবং শিশু
- A man and a young boy ride a yellow kayak
- একটি লোক এবং একটি যুবক একটি হলুদ কায়াক চড়ে
- A man riding a kayak with a little boy
- একজন মানুষ একটি ছোট ছেলের সাথে কায়াক চড়ছেন
- A man and a baby are in a yellow kayak on water
- একটি মানুষ এবং একটি শিশু জলের উপর একটি হলুদ কায়াক আছে
- A man and child kayak through gentle waters
- মৃদু জলের মধ্য দিয়ে একজন মানুষ এবং শিশু কায়াক

Dataset: Translated Flickr8k

Advantages

- More captions
- Less bias

Disadvantages

- Lacks Bangladeshi objects
- Landscape mismatch

Hybrid Dataset

Size:

- 14000+ total images
- 27000+ captions
- 9000+ images from BanglaLekha
- 5000+ imagesfrom Flickr8k

Captions:

- 2 per image
- 10783 distinct words

Training:

- ~6000 from BanglaLekhaImageCaptions
- ~3000 from Translated Flickr8k

Validation:

- ~2000 from BanglaLekhaImageCaptions
- ~1000 from Translated Flickr8k

Testing:

- ~1000 from BanglaLekhaImageCaptions

Hybrid Dataset

- Translation using Google translate API

  Code: Hybrid Dataset Translation.ipynb

- Data Filtering

  Code: Hybrid Dataset Filtering.ipynb

- Concatenation of Flicker8k and BanglaLekhaImageCaptions dataset

  Code: Hybrid Dataset Concat.ipynb

	filename	caption
0	1.png	তিন জন মেয়ে মানুষ আছে। এক জন দাড়িয়ে আছে আর দুই...
1	1.png	একটি হলুদ জামা পায়জামা পরা মহিলা দাড়িয়ে হাতে এ...
2	2.png	অনেক মেয়ে মানুষ বসে আছে।
3	2.png	একটি নীল জামা পরা মহিলা একটি নীল ল্যাপটপ এর দি...
4	3.png	অনেক মানুষ একসাথে বসে কাজ করছে।

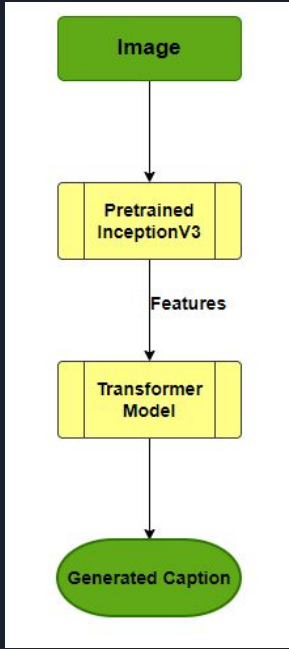
```
['<start> তিন জন মেয়ে মানুষ আছে। এক জন দাড়িয়ে আছে আর দুই জন বসে আছে। <end>',
 '<start> একটি হলুদ জামা পায়জামা পরা মহিলা দাড়িয়ে হাতে একটি বেত নিয়ে পিটানোর ভাব দেখাচ্ছে আর ছোট একটি মেয়ে পিছনে ব্যাগ নিয়ে বসে কাঁদছে। <end>',
 '<start> অনেক মেয়ে মানুষ বসে আছে। <end>',
 '<start> একটি নীল জামা পরা মহিলা একটি নীল ল্যাপটপ এর দিকে তাকিয়ে আছে এবং পিছনে তার দিকে বসে শারি পরে তাকিয়ে আছে অনেকগুলো মহিলা। <end>',
 '<start> অনেক মানুষ একসাথে বসে কাজ করছে। <end>',
 '<start> ২ টি ছোট ছেলে একজন শার্ট প্যান্ট দাড়িয়ে চেয়ে আছে আরেকজন বসে গার্মেন্টস এ কাজ করছে নীল শার্ট পরে তাদের পিছনে অনেকগুলো মহিলা বসে দাড়িয়ে কাজ করছে। <end>',
 '<start> ছয় জন মানুষ দাড়িয়ে আছে। <end>',
 '<start> ৬ জন মানুষ এলোমেলো দাড়িয়ে আছে, তাদের মাঝে ২ জন ছেলে ৪ জন পুরুষ, তাদের একজন লুঙ্গী পরে দাড়িয়ে আছে। <end>',
 '<start> এক জন মেয়ে মানুষ মাথায় ঘোমটা দিয়ে কাজ করছে। মাটিতে বিভিন্ন রঙের মসলা আছে। <end>',
 '<start> একটি মহিলা হালকা পানির উপরে দাড়িয়ে আছে শারি পরে, মহিলার মুখ ডানদিকে ঘুরানো, পানির রং হলুদ দেখাচ্ছে। <end>']
```

Hybrid Dataset

Complete Dataset Available at Mendeley Data: [HybridDataset]

Experiments

- Image captioning using transformer architecture on Flickr8k
- Code : [Image_Captioning_with_Transformer_Flickr8k.ipynb]



Experiments

- Image captioning using transformer architecture on BanglaLekhaImageCaptions
- Code: [Image_Captioning_with_Transformer_BanglaLekha.ipynb]

একটি রাস্তার পাশে একটি কুকুর বসে আছে

একজন নারী ও একটি শিশু আছে

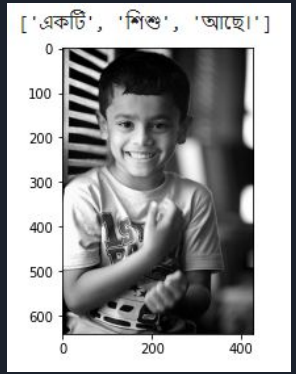
```
Blue Scores[1, 2, 3, 4] = 35.07546244891761 32.584083967091935 24.92793900867118 19.899133433889478
```

Experiment Setup 1

- Image Captioning in Bangla Using Encoder-Decoder Architecture and Hybrid Dataset
- Code: [Image_Captioning_with_Transformer_Hybrid_1.1]

Setup:

- Vocabulary_Size = 6400
- BATCH_SIZE = 128
- Batches = 200
- epochs = 20




```
Blue Scores[1, 2, 3, 4] = 47.574365074204984 37.940302020472785 29.144868433586772 20.889435278378983
```

Experiment Setup 2

- Image Captioning in Bangla Using Encoder-Decoder Architecture and Hybrid Dataset
- Code: [Image_Captioning_with_Transformer_Hybrid_1.2]

Setup:

- Vocabulary_Size = 8196
- BATCH_SIZE = 64
- Batches = 400
- epochs = 30


```
Blue Scores[1, 2, 3, 4] = 48.982585938476184 39.28958076400858 31.631547483290255 23.849581495241672
```

Results

Metric	Transformer on Banglalekha	Transformer on Flickr8k Translated	Transformer on HybridDataset setup-1	Transformer on HybridDataset setup-2	State of the art
BLEU-1	0.391	0.351	0.476	0.490	0.667
BLEU-2	0.314	0.326	0.379	0.393	0.552
BLEU-3	0.268	0.249	0.291	0.316	0.479
BLEU-4	0.221	0.199	0.209	0.238	0.413

Result Analysis

- BanglaLekha Dataset vs Flickr8k
- Flickr8k vs Hybrid
- BanglaLekha vs Hybrid

'শাড়ি পরা একজন মহিলা একটা গাছের নিচা বসে চালনি দিয়ে চাল চালছে।'

'একজন নারী বসে কাজ করছে।'

Findings:

- Less bias, More accuracy
- BLEU is not a good metric for Bengali Image Captioning

Future Works

- More Experiment with the current dataset
- Hyperparameter tuning
- Curating a larger, more varied data set with multiple captions per image

Conclusion

The current bangla image captioning models lack accuracy and we explored the opportunity of incorporating encoder-decoder architecture in Bangla image captioning.

Even though our model did not beat the state-of-the-art model we showed that our idea of reducing human bias from testset increased accuracy in all the evaluation metrics.

References

- Rahman, M., Mohammed, N., Mansoor, N. and Momen, S., 2019. Chittron: An Automatic Bangla Image Captioning System.
- Khan, M., Shifath, S. and Islam, M., 2021. Improved Bengali Image Captioning via deep convolutional neural network based encoder-decoder model.
- Humaira, M., Paul, S., Abidur, M., Saha, A. and Muhammad, F., 2021. A Hybridized Deep Learning Method for Bengali Image Captioning.

Project Drive Link: Image_Captioning_with_Transformer_Hybrid

THANK YOU