Final Presentation

Bangla Image Captioning

CSE472 : Machine Learning Sessional

1705092 - Asif Ajrof
1705100 - Farhana Khan

PROBLEM FORMULATION

Image Captioning in Bengali Language

- Image Captioning is the process of generating textual description of an image.
- It uses both Natural Language Processing and Computer Vision to generate the captions.

EXPLORED DATASETS

DATASET-I

- Khan Raqib Mahmud, Abul Kalam Al Azad. (2019). *Bangla Natural Language Image to Text (BNLIT)* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/629699

DATASET-I DESCRIPTION

- 8743 labeled images
- Resized
  - 224 x 224 pixels
  - 500 x 375 pixels
- PNG format

দুইটি ছেলে একটি গরু নিয়ে ব্রিজের উপর দাড়িয়ে আছে।

DATASET-I LIMITATION

- Too small compared to those the SOTA models have used
- 1 caption per image
- Most of the images contain human being

DATASET-II

- Nafees Mansoor, Abrar Hasin Kamal, Nabeel Mohammed, Sifat Momen, Md Matiur Rahman (2019). *BanglaLekhaImageCaptions* https://doi.org/10.17632/rxxch9vw59.2

DATASET-II DESCRIPTION

- 9154 labeled images
  - Two captions per image
- PNG format

1. একটি নৌকার উপর দুইজন মানুষ আছে।
2. একটি নৌকায় ২ জন মানুষ এবং পাশে জমি।

DATASET-II LIMITATION

- Too small compared to those the SOTA models have used
- Too long caption due to manual human annotation

DATASET-III

- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár (2014). *Microsoft COCO: Common Objects in Context.* https://doi.org/10.48550/arXiv.1405.0312

DATASET-III DESCRIPTION

- 330k images
- 200k labeled images
  - 5 captions per image

1. A black Honda motorcycle with a dark burgundy seat.
2. A black Honda motorcycle parked in front of a garage.
3. A Honda motorcycle parked in a grass driveway
4. Ma motorcycle parked on the gravel in front of a garage
5. A motorcycle with its brake extended standing outside

DATASET-III (BANGLA TRANSLATION)

- Due to the limited size of the Bengali datasets (BNLIT and BanglaLekhaImageCaptions), we increased our data resources by augmenting the COCO dataset (2017 Train(118k) and 2017 Validation(5k)) with Bengali language.
- We used Google Translator for the translation task.
- The task took a little over 30 hours on Google Colaboratory.
- Codes in the `MSCOCO_translate.ipynb` file.

DATASET-III (BANGLA TRANSLATION)

DATASET-III (BANGLA TRANSLATION)


```
{
  "id": 86220,
  "file_name": "000000086220.jpg",
  "captions": [
    "A car and a public transit vehicle on a road.",
    "a car in front of a train on train tracks",
    "A tram and a car make their way through town.",
    "A silver car in the street next to a metal railing.",
    "A white car and a white bus parked parallel from one another. "
  ]
}
```

```
{
  "id": 86220,
  "file_name": "000000086220.jpg",
  "captions": [
    "একটি রাস্তায় একটি গাড়ি এবং একটি পাবলিক ট্রানজিট যান।",
    "ট্রেনের ট্র্যাকে একটি ট্রেনের সামনে একটি গাড়ি",
    "একটি ট্রাম এবং একটি গাড়ি শহরের মধ্য দিয়ে যায়।",
    "ধাতব রেলিংয়ের পাশে রাস্তায় একটি সিলভার গাড়ি। ",
    "একটি সাদা গাড়ি এবং একটি সাদা বাস একে অপরের থেকে সমান্তরালভাবে দাঁড়িয়ে আছে।"
  ]
}
```

LITERATURE STUDY

STATE OF THE ART METHODS

- Image Captioning on COCO Captions

1. mPLUG
2. OFA

Rank	Model	BLEU-4	CIDER	METEOR	SPICE	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	Paper	Code	Result	Year	Tags
1	**mPLUG**	46.5	155.1	32.0	26.0					mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections			2022	
2	**OFA**	44.9	154.9	32.5	26.6					OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework			2022	

OLDER STATE OF THE ART METHOD

1. Show, Attend and Tell

MPLUG

- mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections
- VLP (Vision-Language Pre-training) framework for both cross-modal understanding and generation.
- Asymmetric vision-language architecture with novel cross-modal skip-connections
  - To address information and computation inefficiency in multi-modal fusion
- Pretrained on large-scale image-text pairs
- Demonstrates strong zero-shot transfer ability when directly applied to multiple video-language tasks.

MPLUG

- Text Encoder - initialized using first 6 layers of $\text{BERT}_{\text{base}}$.
- Visual Encoder - intialized using CLIP-ViT
- Base architecture - ViT-L/14

OFA

- OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework
- Task agnostic
- Modality agnostic
- Sequence to sequence learning framework

OFA

- Text Processing - BART.
- Image Processing - ResNet
- Base architecture - Transformer. Initialized using BART.

SHOW, ATTEND AND TELL

- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
- Encoder: image feature extraction with a CNN (VGG architecture)
- Decoder: caption generation with a RNN (LSTM architecture) and Attention Mechanism

SHOW, ATTEND AND TELL

- Two attention-based image caption generators under a common framework

1. “Soft” deterministic attention mechanism trainable by standard back-propagation methods
2. “Hard” stochastic attention mechanism trainable by maximizing an approximate variational lower bound

EXPLORED MODELS

APPROACH 1

- A warm-up exercise → how to work with image captioning models
- To analyze the usability of the COCO dataset we implemented an architecture using pretrained VGG16 encoder for image encoding and a simple RNN decoder for image captioning following this notebook.
- This was done with english captions

APPROACH 2 - RUNNING MPLUG

- We explored the model of mPLUG thoroughly
- Executed using MSCOCO 2014 image captioning dataset (english captions)
- The codebase was too complicated to modify for bangla captions within the allocated time for this project, so we decided to back out from this approach.

APPROACH 3 - ON BANGLA DATASET

- We ran the previously mentioned simple model modifying (different data pre-processing, BanglaBERT tokenizer) it a bit on the Bangla COCO Dataset.
- Codes in the `Bn_Image_Captioning_simple.ipynb` file.
- The produced result for a limited sample (10k) did not seem to be promising.
- So we decided to change our model again.

APPROACH 3 - ON BANGLA DATASET

- Example Result:

ছুরির ছুরির ছুরির ছুরির ছুরির ছুরির ছুরির

একটি মেরু ভালুক একটি গাছের উপর দাঁড়িয়ে আছে

একটি ছোট বিমান একটি নীল এবং সাদা

APPROACH 4 - THE FINAL MODEL

APPROACH 4 - THE FINAL MODEL

- We chose a model quite similar to the paper of Show, Attend and Tell as our final model.
- The code of the model was followed from this notebook, modifying it for our Bangla COCO Dataset.
- Codes in the `Bn_Image_Captioning_final.ipynb` file.

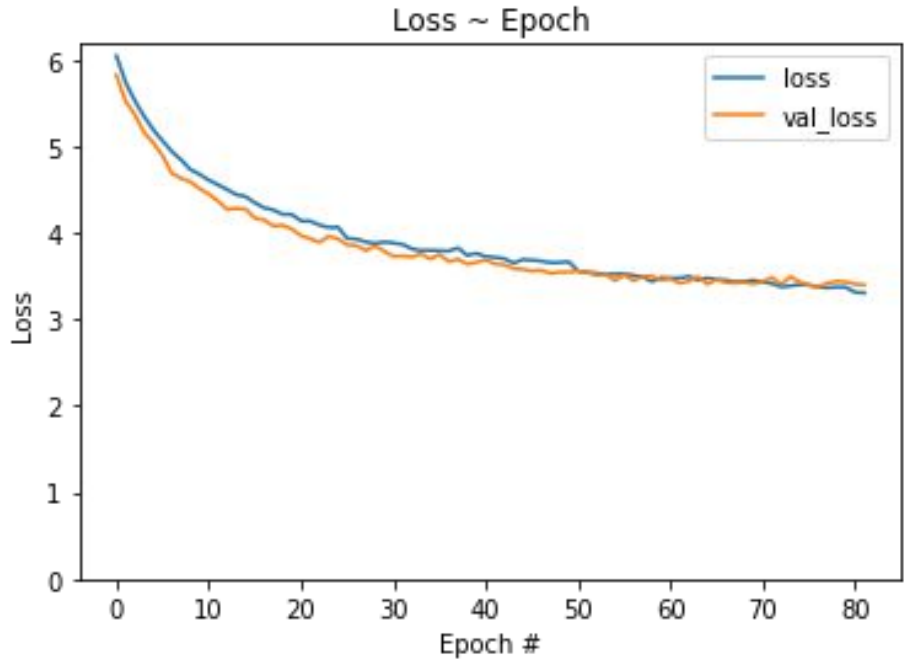
APPROACH 4 - THE FINAL MODEL

- The encoder is MobileNetV3, a CNN.
- The decoder is a transformer.
  - Input - The token embedding and positional encoding (*SeqEmbedding*).
  - Decoder - A stack of transformer decoder layers (*DecoderLayer*) where each contains:
    i. A causal self attention later (*CausalSelfAttention*), where each output location can attend to the output so far.
    ii. A cross attention layer (*CrossAttention*) where each output location can attend to the input image.
    iii. A feed forward network (*FeedForward*) layer which further processes each output location independently.
  - Output - A multiclass-classification over the output vocabulary.

APPROACH 4 - THE FINAL MODEL

- Train images: 32000 (more than that was not possible due to resource constraints)
- Validation images: 5000
- Batch size: 64
- Image feature extraction time: 00:04:30 (hh:mm:ss)
- Adam optimizer
- Learning rate: 1e-4
- Epoch: 82 (model converged after that)
- Training time: 00:35:00 (hh:mm:ss)
- RAM usage: 7.3 GB
- GPU usage: 1.9 GB

APPROACH 4 - THE FINAL MODEL



APPROACH 4 - THE FINAL MODEL

- Example Result:
  - ❏ Randomness in prediction enables different result on the same image on different test of the same model run.

লাল দেয়াল সহ একটি সাদা বিছানা এবং একটি ঘর

একটি বিছানা এবং একটি বাতি সহ একটি বিছানা ঘর

APPROACH 4 - THE FINAL MODEL

- Example Result:

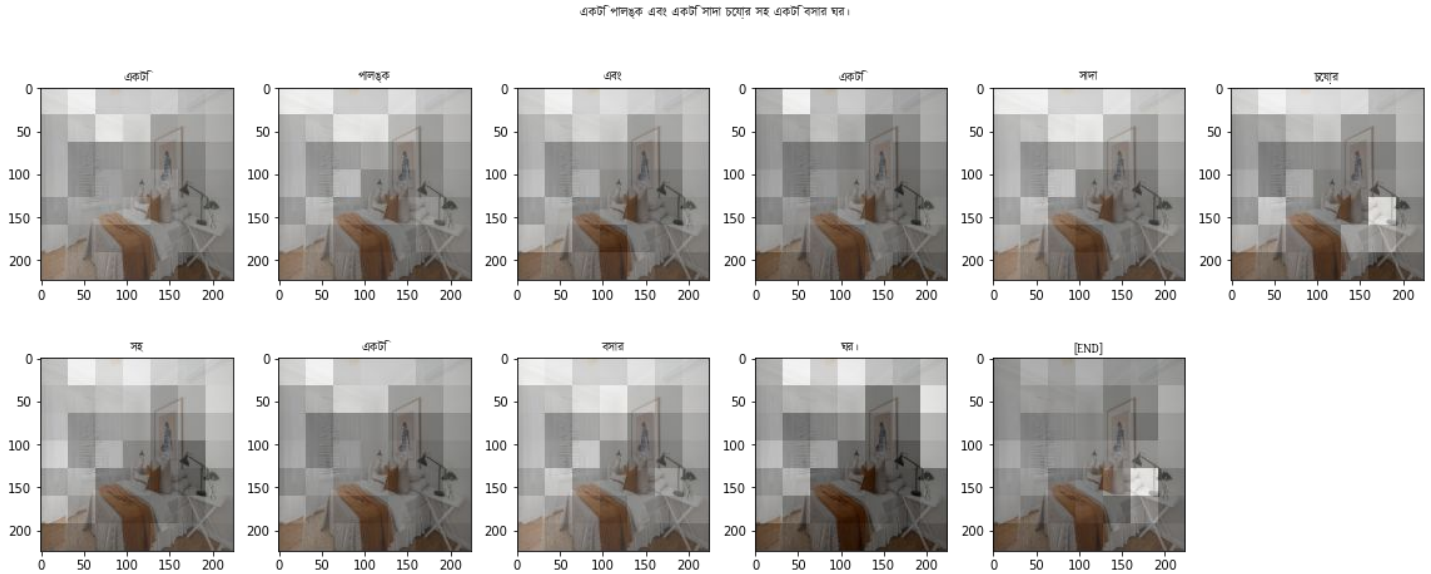
একটি নৌকায় বেশ কিছু লোক বসে আছে

একটি বড় ভবন

চশমা পরা একজন লোক একটি টাই এবং একটি টুপি পরে আছে

APPROACH 4 - THE FINAL MODEL

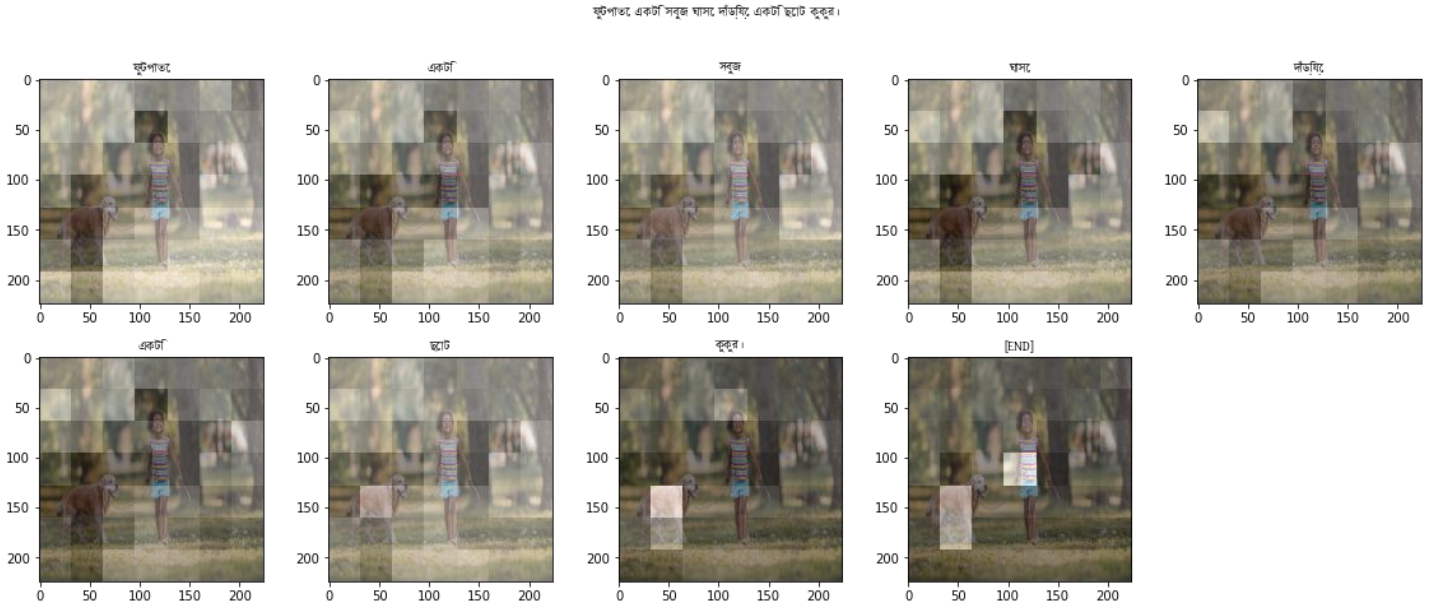
- Example Result with attention map:



একটি পালঙ্ক এবং একটি সাদা চেয়ার সহ একটি বসার ঘর

APPROACH 4 - THE FINAL MODEL

- Example Result with attention map:



ফুটপাতে একটি সবুজ ঘাসে দাঁড়িয়ে একটি ছোট কুকুর

DISCUSSION

- As the model was trained on MS COCO Dataset, it has bias towards some English words, e.g. Frisbee, Surfing etc.
- The CIDEr score was calculated for the BanglaLekhaImageCaptions dataset. But the dataset itself contains captions that are far from what they actually are. So CIDEr was not presented here. Rather qualitative measuring has been done with manual checking of the predicted captions.
- Additional results and remarks in the submitted notebook files.

THE END