MENTOR
Brahmani Nutakki

TEAM
Bishal Goswami
VVSS Sameer Chakravarthy
Ramya Gurijala
Sandeep Tiwari

AUTOMATIC IMAGE CAPTIONING

Group 13

1.   Introduction

Image captioning is important for many reasons. Captions for every image can lead to faster and descriptively accurate images searches and indexing. *Image captioning* has various applications in fields such as biomedicine, commerce, web searching and military etc.

Generating captions for images is a vital task relevant to the area of both Computer Vision and Natural Language Processing. Mimicking the human ability of providing descriptions for images by a machine is itself a remarkable step along the line of Artificial Intelligence. The main challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language (like English). Traditionally, computer systems have been using pre-defined templates for generating text descriptions for images. However, this approach does not provide sufficient variety required for generating lexically rich text descriptions. This shortcoming has been suppressed with the increased efficiency of neural networks. Many state-of-the-art models use neural networks for generating captions by taking images as input and predicting the next lexical unit in the output sentence.

The image captioning models are usually categorized as [1]
- Template-based,
- Retrieval based
- Neural network-based models

The template-based models [2-4] first detect all the image attributes using image classification and object detection methods. These methods generate captions by filling in pre-defined templates from the identified objects. This approach produces too flexible captions that cannot correctly describe the relationships between attributes [5].

Retrieval-based models [6-8] create a pool of similar images in an image database and rank the retrieved images by measuring their similarities and then change the found image descriptions to create a new description for the queried image. The usefulness of this strategy is severely constrained when dealing with images that are not in the dataset and thus not classified, i.e., unseen.

The neural network-based models are inspired by the success of deep neural networks in machine learning tasks and use in an encoder–decoder architecture [9-20]. An encoder extracts image contents by a CNN, a module associates contents to words, and a decoder by an RNN is used for language modeling and creating image captions. Kiros et al. [11] proposed a multimodal language model that jointly learned the high-level image features and word representations. Their model can generate image captions without using any default template or structure, making the model more flexible. Nevertheless, their model could not learn latent representations of the interactions between the objects in the image. Moreover, they investigated a manual algorithm including multiple modules that cannot learn from each other during the training process.

1.1   Problem Statement

The goal of this project work is to generate descriptive captions for an input image. The captions generated should be meaningful and consistent with the image and closely relate to human

translation. This will be achieved by creating an image captioning deep neural network model and evaluate the performance metrics, finetune and improve the model captioning predictions.

1.2   Objective

To transform the image captioning problem into a modeling problem and apply deep learning techniques and performance evaluation metrics for generating appropriate captions for an input image.

2.   Methodologies

2.1   Data Pre-processing

We will use Flickr8K, Flickr30K image dataset and a set of 5 captions per image. We will need to do some data-preprocessing like transformation of images, cleaning of caption text and tokenization etc. before feeding the input to the model for training.

2.2   CNN-LSTM network model (Preliminary model)

Since we have two set of inputs (image and caption text). The image needs to be encoded into higher dimensional vector space first. For this purpose, we will use a pre-trained CNN model (transfer learning approach) to generate feature vectors from the image as part of encoding. The output from the CNN network will be used as an input along with the starting word from the related caption to a LSTM network. LSTM networks are great in generating sequences and will help predict the next word in the caption based on the inputs (image feature vector and input as shown in Fig.1.
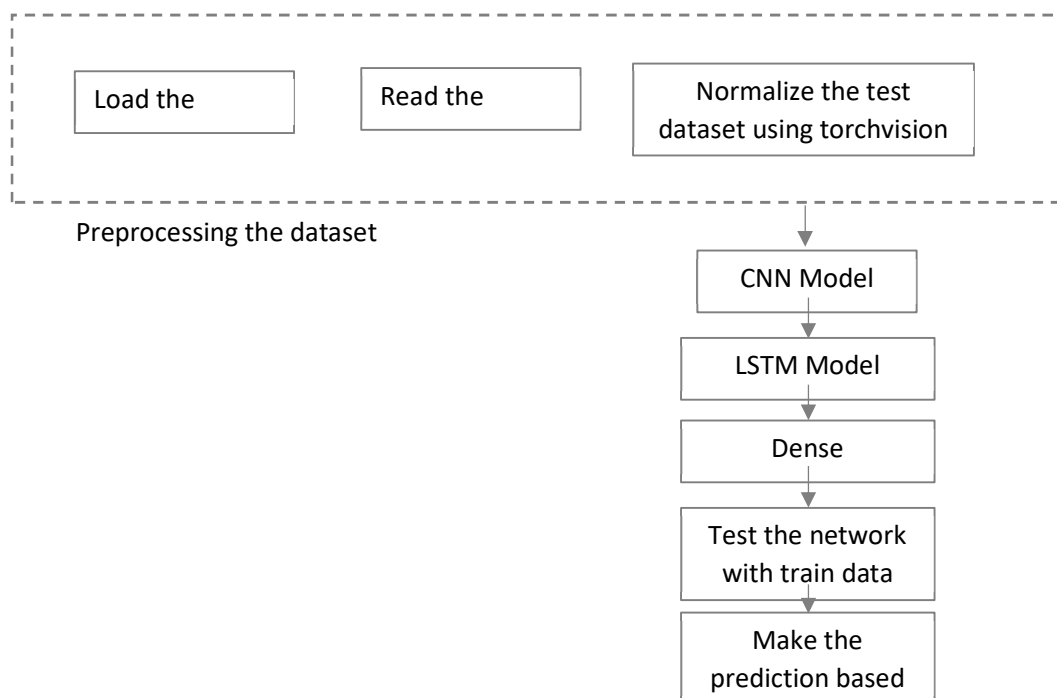

Fig.1 : Block diagram

2.3   Transformer Network model

We have used the GIT transformer model which is an ideal model for image/video captioning and visual question answering.

2.4   Performance evaluation

The performance of our trained model can be evaluated by measuring how well is the generated caption for an input image compared to the reference caption for that image. For this purpose, we will be exploring evaluation metric BLEU. Based on the score we will be finetuning our model.

BLEU stands for Bilingual Evaluation Understudy. It is an algorithm, which has been used for evaluating the quality of machine translated text. BLEU can be used to check the quality of the generated caption. It is easy to compute.

Bleu measures precision: how much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries.

3.   Literature Review

Based on literature review, we would like to call out two approaches from the paper Conceptual Captions [21] to accomplish the project objective described above. One uses RNNs with LSTM cells to implement the encoder and decoder functions, corresponding to the Show-And Tell model. The other uses Transformer self-attention networks to implement the encoder and decoder functions. Both the models use Inception-ResNet-v2 as the CNN component. The other paper[22] uses Faster R-CNN for image recognition and LSTM for captioning. Faster R-CNN is an object detection model designed to identify instances of objects belonging to certain classes and localize them with bounding boxes. Another recent published work [23] uses a Swin Transformer in the encoder, instead of a traditional CNN based architecture and a LSTM combined with an attention module in the decoder. A recent approach called ViLBERT extends the popular BERT architecture (which has proven to be effective for transfer learning to multiple natural language processing tasks) to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers [24].

Gurari et al [25] developed a system that can help the blind to take images and get the description of the photograph in terms of captions. They have developed the dateset titled VizWiz and made it publicly available to give a scope for more challenges. Benchmarking of nine algorithms based on three modern image captioning algorithms known as Up-Down, SGAE, and AoANet, is presented for the MSCOCOCaptions challenge. UpDown, OSCAR, VIVO, Meta Learning and a model that uses conditional generative adversarial nets are the well explored methodologies for captioning problem [26]. Although the general adversarial network (GAN) based model achieves the highest score, UpDown represents an important basis for image captioning and Object-Semantics Aligned Pre-training (OSCAR) [27] and VIsual VOcabulary pretraining (VIVO) [28] are more useful as they use novel object captioning. The GAN-based model is the most performant, UpDown has the most impact and OSCAR and VIVO are the more useful however, this varies with the dataset considered.

Show and Tell [29, 30] is a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model resulted in a BLEU-1 score of 59 which is competitively performing with the human intellect which scores around 69 on the Pascal data set. The proposed model reported a BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Similarly, the model performed well on the COCO dataset, achieving a BLEU-4 score of 27.7, benchmarking the current state-of-the-art.

Due to the autoregressive nature of the existing methods, the computational complexity increases linearly as the length of the generated captions grows. Hence, a non-autoregressive image captioning approach is proposed in [31] that can generate captions in a length-irrelevant complexity. The non-autoregressive model outperformed the autoregressive baselines in terms of controllability and diversity, and also significantly improves the decoding efficiency for long captions. The GIT (Generative image-to-text Transformer for Vision and Language) is extremely helpful to unify vision-language tasks such as image/video captioning and question answering. GIT simplifies the architecture as one image encoder and one text decoder under a single language modeling task. We also scale up the pre-training data and the model size to boost the model performance [32]

## 4. Model Summary

4.1 CNN-LSTM model

The feature extractor model uses a VGG16 pretrained output with three dense layers and activation function as relu. The decoder is a LSTM + Dense model with an embedding layer with input dimensions as the size of the vocabulary and output dimension as 256 units. The two LSTM layers are of 512 and 256 units and the two dense layers use the relu activation as shown in Fig.2. The output layer uses softmax as activation function. The loss function is categorical cross entropy and optimizer used is Adam.
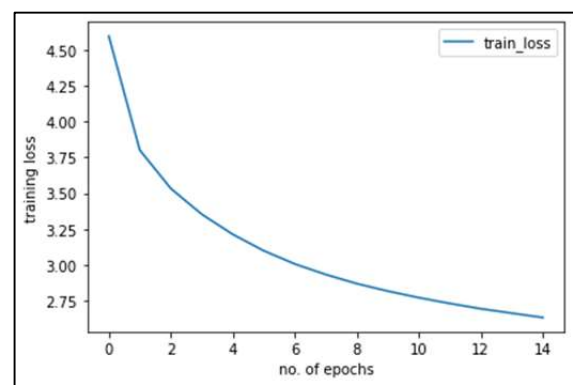
```
Model: "model_2"
_____
Layer (type)                    Output Shape          Param #       Connected to
=========================================================================================
input_image_features (InputLay  [(None, 4096)]        0             []
er)

input_text_sequence (InputLaye  [(None, 74)]          0             []
r)

image_features_layer1 (Dense)   (None, 1024)          4195328       ['input_image_features[0][0]']

embedding_1 (Embedding)         (None, 74, 256)       5052928       ['input_text_sequence[0][0]']

dropout_2 (Dropout)             (None, 1024)          0             ['image_features_layer1[0][0]']

LSTM_1 (LSTM)                   (None, 74, 512)       1574912       ['embedding_1[0][0]']

image_features_layer3 (Dense)   (None, 512)           524800        ['dropout_2[0][0]']

dropout_3 (Dropout)             (None, 74, 512)       0             ['LSTM_1[0][0]']

image_features_layer4 (Dense)   (None, 256)           131328        ['image_features_layer3[0][0]']

LSTM_2 (LSTM)                   (None, 256)           787456        ['dropout_3[0][0]']

add_1 (Add)                     (None, 256)           0             ['image_features_layer4[0][0]',
                                                                     'LSTM_2[0][0]']

dense_3 (Dense)                 (None, 256)           65792         ['add_1[0][0]']

dense_4 (Dense)                 (None, 512)           131584        ['dense_3[0][0]']

dense_5 (Dense)                 (None, 19738)         10125594      ['dense_4[0][0]']

=========================================================================================
Total params: 22,589,722
Trainable params: 22,589,722
Non-trainable params: 0
_____
```

Fig.2: Features

The model has been trained using both Flickr8K_Dataset and Flickr30K_Dataset. The challenges included handling a lot of in memory data as the model accepts image feature vector per sequence token in a caption. The high RAM usage issue was mitigated by running the data in batches by implementing generators which helped in processing a subset of data in each batch. The training summary for Flickr8K dataset are: input captions/images = 32364 and final training loss: 1.7302 (Fig.3). The training summary for the Flickr30K dataset were captions/images = 127132 and final training loss: 2.6310 (Fig.3).
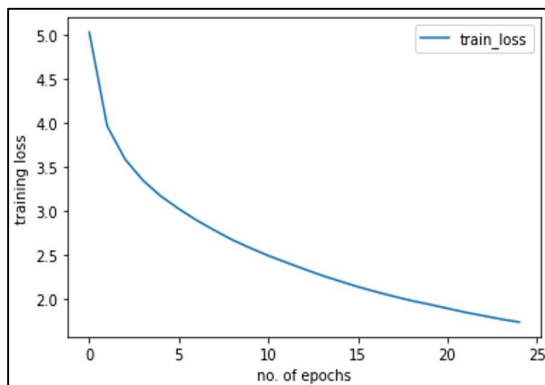
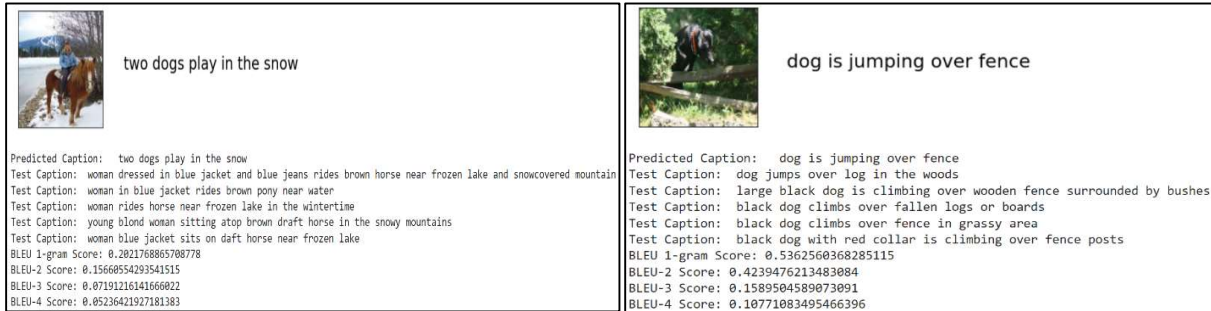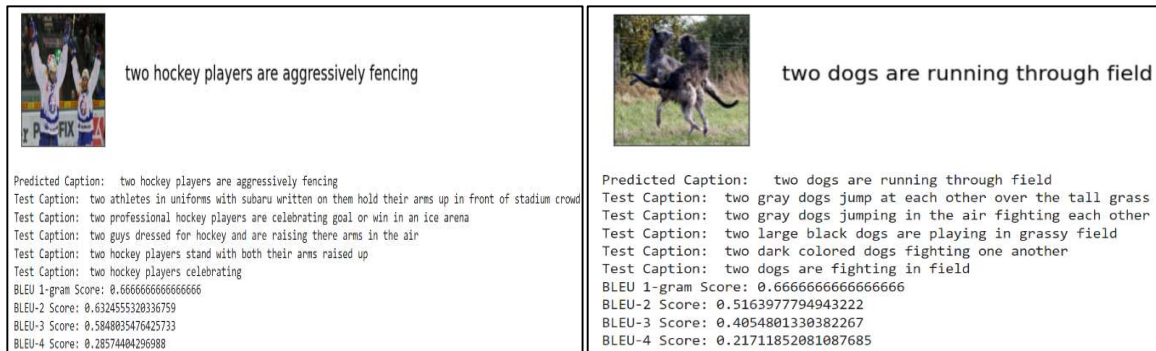(a)   (b)

Fig.3: Training loss summary (a) 8K Dataset (b) 30K Dataset

We also implemented multiple captions generation for a single image by sampling top three results from the next predicted sequence and choosing randomly any one of the sequence.

The performance evaluation of the model using the BLEU scores were as given in Fig.4.

two dogs play in the snow

Predicted Caption: two dogs play in the snow
Test Caption: woman dressed in blue jacket and blue jeans rides brown horse near frozen lake and snowcovered mountain
Test Caption: woman in blue jacket rides brown pony near water
Test Caption: woman rides horse near frozen lake in the wintertime
Test Caption: young blond woman sitting atop brown draft horse in the snowy mountains
Test Caption: woman blue jacket sits on daft horse near frozen lake
BLEU 1-gram Score: 0.2021768865708778
BLEU-2 Score: 0.15660554293541515
BLEU-3 Score: 0.07191216141666022
BLEU-4 Score: 0.05236421927181383

dog is jumping over fence

Predicted Caption: dog is jumping over fence
Test Caption: dog jumps over log in the woods
Test Caption: large black dog is climbing over wooden fence surrounded by bushes
Test Caption: black dog climbs over fallen logs or boards
Test Caption: black dog climbs over fence in grassy area
Test Caption: black dog with red collar is climbing over fence posts
BLEU 1-gram Score: 0.5362560368285115
BLEU-2 Score: 0.4239476213483084
BLEU-3 Score: 0.1589504589073091
BLEU-4 Score: 0.10771083495466396

(a)

two hockey players are aggressively fencing

Predicted Caption: two hockey players are aggressively fencing
Test Caption: two athletes in uniforms with subaru written on them hold their arms up in front of stadium crowd
Test Caption: two professional hockey players are celebrating goal or win in an ice arena
Test Caption: two guys dressed for hockey and are raising there arms in the air
Test Caption: two hockey players stand with both their arms raised up
Test Caption: two hockey players celebrating
BLEU 1-gram Score: 0.6666666666666666
BLEU-2 Score: 0.6324555320336759
BLEU-3 Score: 0.5840035476425733
BLEU-4 Score: 0.28574404296988

two dogs are running through field

Predicted Caption: two dogs are running through field
Test Caption: two gray dogs jump at each other over the tall grass
Test Caption: two gray dogs jumping in the air fighting each other
Test Caption: two large black dogs are playing in grassy field
Test Caption: two dark colored dogs fighting one another
Test Caption: two dogs are fighting in field
BLEU 1-gram Score: 0.6666666666666666
BLEU-2 Score: 0.5163977794943222
BLEU-3 Score: 0.4054801330382267
BLEU-4 Score: 0.21711852081087685

(b)

Fig.4 : Performance evaluation of BLEU (a) 8K Dataset (b) 30K Dataset

4.2   GIT Transformer model

The GIT Transformer [21] based model is ideal for image/video captioning and visual question answering. The architecture of the same is as follows:
- An image encoder whose output is a compact 2D feature map, which is flattened into a list of features
- A text decoder which is a transformer module to predict the text description
- The transformer module consists of multiple transformer blocks, each of which is composed of one self-attention layer and one feed-forward layer
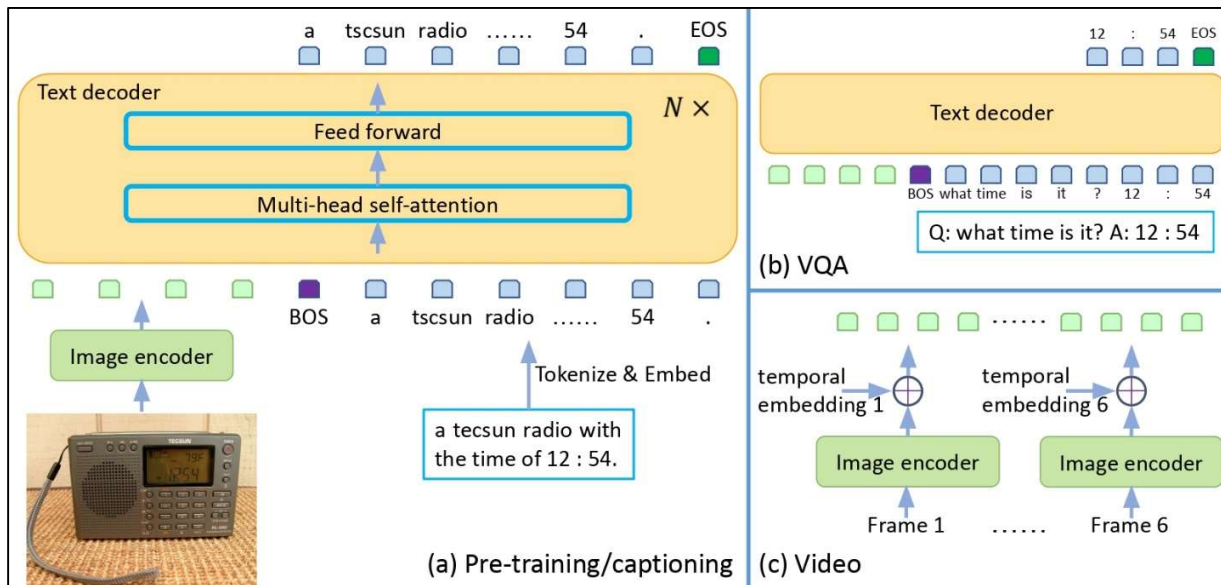
Fig.5 : (a) Pre-training/captioning, (b) VQA and (c) Video

The model has been trained using Flickr8K_Dataset. The challenges included huge compute resources required to train the model and complex decoder architecture. It was handled by paid GPUs. The training results for 8K dataset were captions/images = 6480 and final training loss was 0.00129 (Fig.6).
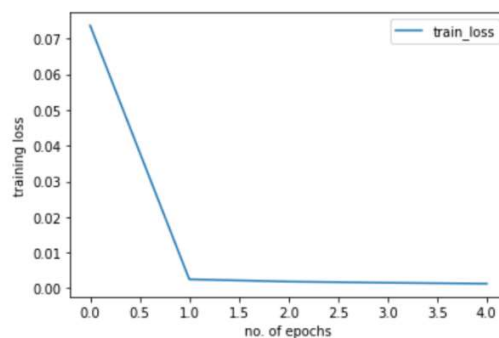

Fig.6: Training loss summary for Transformer model

The performance evaluation of the model using the BLEU scores are presented in Fig.7.
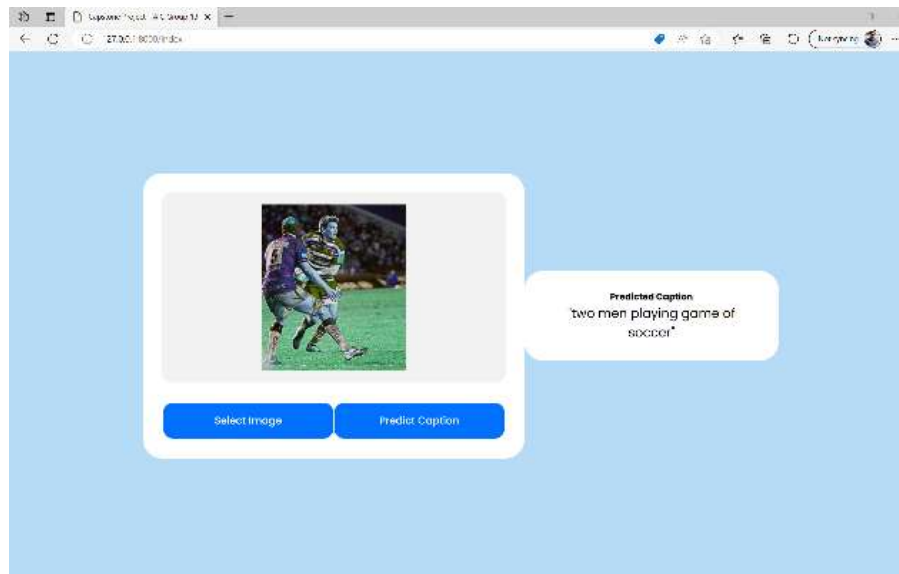

Fig.7 : Performance evaluation BLEU

5.   Model Deployment

The model was deployed using FastAPI + HTML responsive UI front-end with an interface for uploading image and displaying output as given in Fig.8. The upload screen facilitates uploading the image which needs captioning.
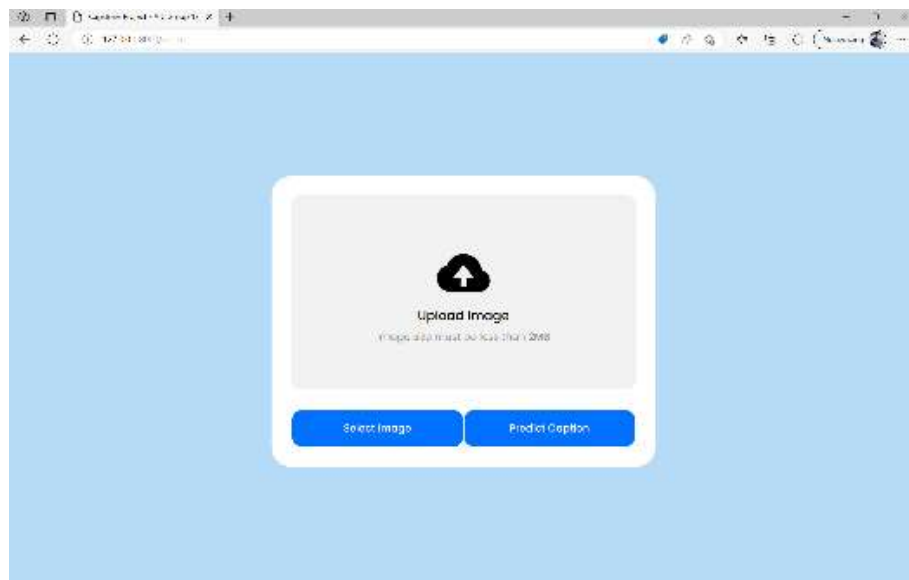
(a)

(b)

Fig.8: GUI (a) Upload image and (b) Output screen

REFERENCES :

[1]. Javanmardi, S., Latif, A. M., Sadeghi, M. T., Jahanbanifard, M., Bonsangue, M., & Verbeek, F. J. (2022). Caps captioning: a modern image captioning approach based on improved capsule network. *Sensors*, *22*(21), 8376.

[2]. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every picture tells a story: Generating sentences from images. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 15–29. Sensors **2022**, 22, 8376 19 of 20

[3]. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. BabyTalk: Understanding and Generating Simple Image Descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **2013**, 35, 2891–2903. [CrossRef] [PubMed]

[4]. Li, S.; Kulkarni, G.; Berg, T.; Berg, A.; Choi, Y. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland, OR, USA, 23–24 June 2011; pp. 220–228.

[5]. Jin, J.; Fu, K.; Cui, R.; Sha, F.; Zhang, C. Aligning where to see and what to tell: Image caption with region-based attention and scene factorization. arXiv **2015**, arXiv:1506.06272.

[6]. Kuznetsova, P.; Ordonez, V.; Berg, T.L.; Choi, Y. TreeTalk: Composition and Compression of Trees for Image Descriptions. Trans. Assoc. Comput. Linguist. **2014**, 2, 351–362. [CrossRef]

[7]. Kuznetsova, P.; Ordonez, V.; Berg, A.; Berg, T.; Choi, Y. Generalizing image captions for image-text parallel corpus. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 790–796.

[8]. Ordonez, V.; Kulkarni, G.; Berg, T. Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process Syst. **2011**, 24, 1143–1151.

[9]. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; van den Hengel, A. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. IEEE Trans. Pattern Anal. Mach. Intell. **2018**, 40, 1367–1381. [CrossRef]

[10]. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2017; pp. 7008–7024.

[11]. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal neural language models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 595–603.

[12]. Mason, R.; Charniak, E. Nonparametric method for data-driven image captioning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 592–598.

[13]. Devlin, J.; Gupta, S.; Girshick, R.; Mitchell, M.; Zitnick, C.L. Exploring nearest neighbor approaches for image captioning. arXiv **2015**, arXiv:1505.04467.

[14]. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE

[15]. Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.

[16]. Lebret, R.; Pinheiro, P.; Collobert, R. Phrase-based image captioning. Int. Conf. Mach. Learn. **2015**, 37, 2085–2094.

[17]. You, Q.; Jin, H.;Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2016; pp. 4651–4659.

[18]. Johnson, J.; Karpathy, A.; Fei-Fei, L. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2016; pp. 4565–4574.

[19]. Yang, Z.; Liu, Q. ATT-BM-SOM: A Framework of Effectively Choosing Image Information and Optimizing Syntax for Image Captioning. IEEE Access **2020**, 8, 50565–50573. [CrossRef]

[20]. Martens, D.; Provost, F. Pseudo-Social Network Targeting from Consumer Transaction Data; University of Antwerp: Antwerp, Belgium, 2011.

[21]. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning - Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut

[22]. Vietnamese Image Captioning for Healthcare Domain using Swin Transformer and Attention-based LSTM-Thanh Tin Nguyen et al.

[23]. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks Jiasen Lu Et al.

[24]. Gurari, D., Zhao, Y., Zhang, M., & Bhattacharya, N. (2020, August). Captioning images taken by people who are blind. In *European Conference on Computer Vision* (pp. 417-434). Springer, Cham.

[25]. Elhagry, A., & Kadaoui, K. (2021). A thorough review on recent deep learning methodologies for image captioning. arXiv preprint arXiv:2107.13114.

[26]. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020, August). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision* (pp. 121-137). Springer, Cham.

[27]. Hu, X., Yin, X., Lin, K., Wang, L., Zhang, L., Gao, J., & Liu, Z. (2020). Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training.

[28]. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[29]. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[30]. Peng, Z., Dai, Y., Tang, Q., Cui, X., & Guo, S. (2019). Show and tell: A neural image caption generator.

[31]. Deng, C., Ding, N., Tan, M., & Wu, Q. (2020, August). Length-controllable image captioning. In European Conference on Computer Vision (pp. 712-729). Springer, Cham.

[32]. GIT: A Generative Image-to-text Transformer for Vision and Language Jianfeng Wang et al.