Image2tweet: Datasets in Hindi and English for Generating Tweets from Images

Rishabh Jha[1] Varshith Kaki [1] Varuna Krishna Kolla[1] Shubham Bhagat[1]
Parth Patwa[2] Amitava Das[3,4] Santanu Pal[3]
[1]Indian Institute of Information Technology Sri City, India
[2]University of California Los Angeles, USA
[3]Wipro AI Labs, India [4]AI Institute, University of South Carolina, USA
[1]{rishabh.j19, vishnusaivarshith.k19, varunakrishna.k19, shubham.b18}@iiits.in
[2]parthpatwa@g.ucla.edu
[3]{amitava.das2, santanu.pal2 }@wipro.com

Abstract

Image Captioning as a task that has seen major updates over time. In recent methods, visual-linguistic grounding of the image-text pair is leveraged. This includes either generating the textual description of the objects and entities present within the image in constrained manner, or generating detailed description of these entities as a paragraph. But there is still a long way to go towards being able to generate text that is not only semantically richer, but also contains real world knowledge in it. This is the motivation behind exploring image2tweet generation through the lens of existing image-captioning approaches. At the same time, there is little research in image captioning in Indian languages like Hindi. In this paper, we release Hindi and English datasets for the task of tweet generation given an image. The aim is to generate a specialized text like a tweet, that is not a direct result of visual-linguistic grounding that is usually leveraged in similar tasks, but conveys a message that factors-in not only the visual content of the image, but also additional real world contextual information associated with the event described within the image as closely as possible. Further, We provide baseline DL models on our data and invite researchers to build more sophisticated systems for the problem.

1 Introduction

Generating a textual description of an image is called image captioning. It can be an easy process for most adults, but for a machine to generate a rich and vivid description is a difficult task. Image captioning requires to recognize the important objects, their attributes and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. This task involves the knowledge of both computer vision and natural language processing.

Image Captioning has been a very popular research area since the last decade. Even before the boom of neural network based techniques people tried various hand crafted features such as Local Binary Patterns (LBP) (Ojala et al., 2000), Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), the Histogram of Oriented Gradients (HOG) (De Marneffe et al., 2006) along with classical ML methods like SVM for Image Captioning. On the other hand, while using neural network based techniques, features are learned automatically from training data and they can handle a large and diverse set of images (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Xu et al., 2015). Moreover, the availability of large and new datasets has made the learning-based image captioning an interesting research area. The popular datasets for English Image Captioning are - Flickr30K Dataset (Young et al., 2014), MS COCO (Lin et al., 2014), and Google Conceptual Caption dataset (Sharma et al., 2018). However, there is almost no research of image captioning in Hindi and/or Indian languages.

Image captioning is important for many reasons. For example, they can be used for automatic image indexing. Image indexing is important for Content-Based Image Retrieval (CBIR) and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching.

Image2Tweet takes one step ahead of regular image captioning task. It involves generating captions that are not only semantically rich but also contain some real world knowledge

Proceedings of the 18th International Conference on Natural Language Processing, pages 670–676
Silchar, India. December 16 - 19, 2021.

(Sharma, 2020). The task is that given an image, the machine has to generate a tweet from it. An example is provided in figure 1. Generating this level of detailed tweets requires person identification (Sachin Tendulkar), Object detection (BJP logo) etc.

In this paper, we describe the image2tweet task and release a new dataset for the task and also release a novel hindi dataset to ignite the Image Captioning research for Indian languages.

Figure 1:
COCO Style: A man in front of a crowd.
Conceptual Caption: A closeup of a mid-aged man, and a parade.
Expected Image2Tweet: Sachin Tendulkar and BJP parade.

2 Related Work

There are quite a few popular image captioning datasets. Flickr30k (Young et al., 2014) consists of 30K images and each image has 5 captions. COCO (Lin et al., 2014) dataset consists of 330K images and each image has 5 captions. Google Conceptual Caption (Sharma et al., 2018) has approximately 3.3 million images and each image has only one caption. However, such datasets use commonly found images over the web and couple the images with alt-text descriptions. Most of the descriptions use proper nouns (such as *characters, places, locations, organizations, etc.*). Such proper nouns pose some problems because a image captioning model is difficult to learn such fine-grained proper noun inference from the input image pixels. At the same time, there is very little research done on Hindi image captioning. To the best of our knowledge, ours is the first dataset to generate tweet from images and to release a Hindi dataset.

Deep learning methods are the most popular to solve the image captioning task. Jiang et al. (2018) proposed novel Recurrent Fusion Network (RFNet), which exploits complementary information from multiple encoders to tackle image captioning. Xu et al. (2015) propose an encoder-decoder method which incorporate spatial attention mechanism to help the model to determine which regions to focus in an image. Yang et al. (2016) propose a framework called ReviewNet. Zhou et al. (2020) proposed a Unified Vision-Language Pre-Training for Image Captioning which can be easily fine tuned.

Similar to caption generator meme generation has also been a eye-catching task for researchers. the task is to generate memes based on the image. unlike captioning, here in meme generation it has to generate text for multiple persons, if multiple persons are involved in meme image. Kurochkin (2020) released a dataset consisting of 650K meme instances. They applied GPT-2(Radford et al., 2019) model for meme generation and observed that machine generated meme text's are not that engaging as human generated.

3 Task Description

Image Captioning for English is well studied paradigm and researchers have tried various methods like hand crafted features (Ojala et al., 2000; Lowe, 2004; De Marneffe et al., 2006) along with classical ML methods like SVM. During the last decade numerous of Big datasets have been released and quite a few efforts can be noticed but there is a still shortage of works in Indic Languages.

Image2Tweet is a shared task where we move a step forward from image captioning. The task is to generate a tweet like a human/news reporter given an image. We release datasets for two languages - English and Hindi. Figures 2 and 3 show an instance from the English and Hindi data respectively.

3.1 Evaluation Metric

For Image Captioning, most used metrics are n-gram based matching metrics such as BLEU, ROUGE, METEOR, and CIDEr.

Popular Image Captioning datasets like Flickr30k (Young et al., 2014), COCO (Lin et al., 2014), and Google Conceptual Caption

Figure 2: **Tweet**: Finance Minister Nirmala Sitharaman presents the full Budget of the second term of the Narendra Modi government #BudgetSession2020 #BudgetWithTimes #UnionBudget2020

Figure 3: **Tweet**: पिंकसिटी में सुबह से हो रही झमाझम बारिश किसी के लिए राहत तो कहीं आफत #jaipur #Monsoon2017.

(Sharma et al., 2018) provide multiple captions per image, as the same image can be described in many different ways. So, in these datasets, while evaluating they calculate the score between the system generated caption and all the reference captions in the gold data. Now, in our task having multiple tweets for a given image is difficult to collect, and having only one reference tweet will affect the evaluation score.

Since having multiple tweets for an image would be difficult, we assume that similar images may have similar tweets. With this in mind we apply content based similarity match on the collected data and keep all the similar images in one cluster. The released data is pre-processed accordingly, and all the clusters are marked along with image ids. For evaluation, we use CIDEr, where the score will be calculated between system generated tweet vs. all the tweets belong to the similar image cluster provided in the dataset.

4 Dataset

The data consists of image-tweet pairs. We provide 2 dataset - English and Hindi. The Hindi data is collected by crawling tweets from two well known Hindi Newspapers - Dainik Bhaskar and Dainik Jagran. The English data is crawled from the twitter handle of Times of India. We use Twitter API[1] to crawl the tweets. We collect total 70k Tweets for English Image2Tweet and 51K for Hindi Image2Tweet. Table 1 gives the data statistics.

Dataset	English	Hindi
Training	48792	35701
Validation	10209	7652
Test	10411	7652
Total	69412	51005

Table 1: Train, Validation and Test data split for the English and Hindi datasets.

Figures 4 and 5 show the word clouds of Hindi and English tweets respectively. We observe that most of the words are related to politics and Covid-19.

Clustering is the necessary part of making

[1] https://developer.twitter.com/en/docs/twitter-api

Figure 4: Word cloud of English dataset. Most of the words are related to Politics and Covid-19.

Figure 5: Word cloud of Hindi dataset. Most of the words are related to Politics and Covid-19.

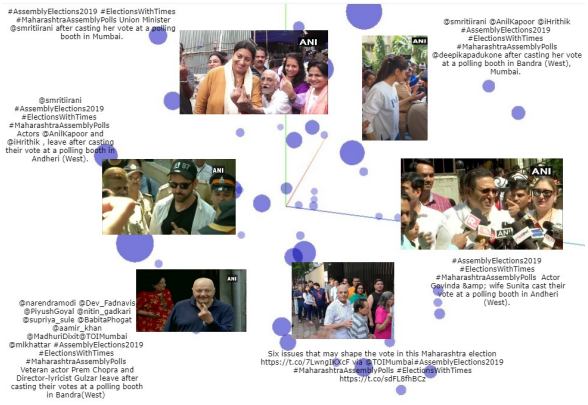

Figure 6: An example of a cluster from English data. All the tweet objects are related to political elections in India.

Figure 7: An example of a cluster from English data. All the tweet objects are related to cricket.

of the dataset, as mentioned in the previous section. For clustering we first remove unnecessary links, symbols and numbers. However keeping the hashtags and mentions can help in the process of clustering similar tweets (without symbols '@' and hashtags). In the next step we remove the words which doesn't add meaning to sentence, stopwords.

Figure 6 and 7 show and example of English and Hindi cluster respectively. We can see that the tweet objects within a cluster are related/similar to each other. Our aim is to do multimodal clustering, Image+Text, hence we implement an algorithm in which similarity score between tweet object are calculated with every other tweet object and stored in the form of 2D dictionary, each row sorted in reverse order. i^{th} row contains the similarity score of i^{th} tweet object with every other object in reverse order. As for every pair of tweet data there are two entries (dict[i][j] & dict[j][i]), we eliminate that entry which is in lower relative position among those two rows. After that for every row we consider at most 5 element with highest similarity score and combine them to make a cluster group. Hence. each cluster has at most 6 tweet objects (1 tweet object and its 5 neighbors). The formula to calculate similarity between 2 tweet objects is :

$$Sim(i,j) = W1 * textSim(i,j) + W2 * imgSim(i,j)$$

textSim(i,j) function calculates the similarity between the textual part of the tweet object using the weighted average of overlap of unigrams, bigrams and trigrams. imgSim(i,j) function calculates the cosine similarity between

the feature vectors of the images extracted using DenseNet (Huang et al., 2017). Overall similarity is just the weighted average of both the similarity, where w1 + w2 = 1 and (w1, w2) ϵ [0,1].

The datasets are available at https://competitions.codalab.org/competitions/35702.

5 Baseline

We develop our baseline using BERT (Devlin et al., 2018) and VGG-19(Simonyan and Zisserman, 2014). The BERT model is pre-trained on whole English Wikipedia and Brown corpus for next sentence prediction objective.

We design is a two branch model (refer figure 8). While training, the image embedding obtained from VGG-19 is passed to one branch and the text is passed to the other branch. In the first branch the image embedding is passed to dense layer. In the second branch the text is sent to BERT tokenizer and its output passed to the pre-trained BERT. Then the output from the last layer of BERT is passed to an LSTM layer which is given as an input to max pooling and to and average polling. The output vectors of max pooling and average pooling are concatenated. After this, we concatenate the outputs from both the branches, and give the concatenated vector as input to an LSTM followed by a dense layer. The output vector of this dense layer is used to generate the words in Tweet.

For training we use Adam optimiser, and train it with a mini-batch size of 32. The learning rate is set to 1e -5. The max caption length is set to 34. While testing, we pass the image vector and the sequence of words generated so far and will predict the next word. Likewise we go on until the end token appears. We use greedy search method to generate the whole tweet.

The baseline code is available at https://github.com/git-rishabh-jha/Image2Tweet.

6 Results

Table 2 shows the results of the baseline system. The results are poor since we use a relatively simple approach to establish the baseline. There is a huge scope of improvement in the results, for which we encourage more innovative approaches.

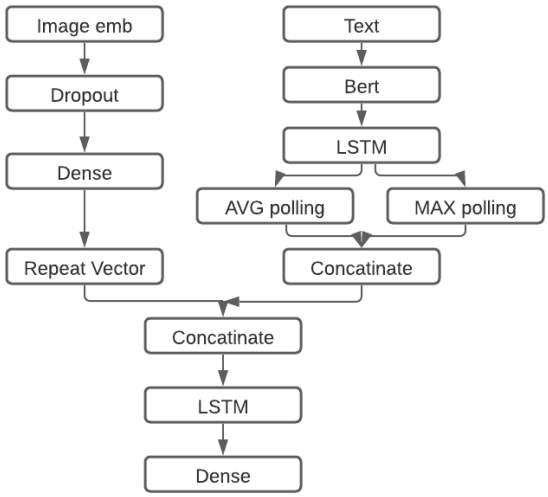

Figure 8: Architecture diagram of baseline model.

Table2 shows the results of the baseline on the image2tweet datasets. The results are poor since we use a relatively simple approach to establish the baseline. There is a huge scope of improvement in the results, for which We encourage more innovative approaches.

Table 3 shows the result of the baseline trained and tested on popular image captioning datasets. The results are much better than on our image2tweet datasets, which shows that image2tweet is a unique and more difficult task than image captioning.

7 Conclusion

In this paper we define the task image2tweet and release datasets in Hindi and English for the task. The English and Hindi datasets consists of 70k and 51k image-tweet pairs respectively. We cluster the similar tweets in our dataset for better evaluation of the system generated tweets. Generated tweets are evaluated using Cider. Further, we provide VGG19 + BERT based baseline systems for our data.

Image2tweet is more difficult than traditional image captioning and we believe it needs further research attention. Future work includes collecting data for more languages, building more complex systems for the task etc.

System	CIDEr	BLEU-4	METEOR	ROUGE
Baseline-English	0.0003	0.02	0.00013	0.00013
Baseline-Hindi	0.0004	0.03	0.00023	0.00023

Table 2: Results of baseline systems on Hindi and English datasets

Dataset	BLEU-4
Flickr30K	23.6
COCO	21.3
Conceptual Captions	20.3

Table 3: Results of baseline systems on popular image captioning datasets.

References

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Andrew Kurochkin. 2020. Meme generation for social media audience engagement.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, pages 404–420. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Shivam Sharma. 2020. Generating tweet-like text from images. where we are…and where we need to be.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Russ R Salakhutdinov. 2016. Review networks for caption generation. *Advances in neural information processing systems*, 29:2361–2369.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Uni-

fied vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.