Name: Fahad Ahmed Pranto

Id: 1612607042

Paper name: **Nepali Image Captioning.**

Paper link: https://scihub.wikicn.top/https://ieeexplore.ieee.org/abstract/document/8947436

**Abstract:** Two areas in machine learning-computer vision and natural language processing-are merged in the image captioning project. For Nepali language with a complicated grammatical form, the difficulty is growing. Inserting an image and producing the definition of the image in Nepali with right grammar with necessarily complex grammar is difficult to obtain and also there is no standard dataset. There are many applications for defining the meaning of an image in Nepali. In particular, this work helps as a key to more complex issues such as a video descriptor framework and image search for web pages in the Nepalese. Two encoder-decoder architectures, one with visual attention and another without visual attention, are used in this research. Scientifically, they illustrate the lack and confusion of model outputs using multiple optimizers.

**Introduction:** In artificial intelligence, creating a caption that explains a picture in a realistic way is important to understanding the image. Even so, image captioning is the method of using an image-based method to capture consistent features of an image and explain the image using a language-based model in a natural language that uses the selected features by the image-based model. In contrast to languages like English, image captioning for a complicated language is tough and there are some difficulties. In aspects of grammatical form, Hindi and Bangla are some of the nearest languages to Nepali, among many other languages where picture captioning has been attempted. The researchers have manually developed Nepali captions from the MS-COCO data set to resolve the rich dataset-related problem. They assume that this study serves as a framework on which other researchers will construct and analyze for upgrades in the future.

**Literature review**: In comparison to languages like English, image captioning for a complicated language is challenging and there are some difficulties. There is a shortage of dataset and Nepali captions from the MS-COCO data set have been manually created by the researchers to fix the rich problem linked to the dataset. Using multiple optimizers, they demonstrate the failure and uncertainty of model outputs. A few translated captions found out to be incomprehensible and several pairs of image-captions were qualitatively processed and first the writers using an existing API to validate spelling The captions created are satisfactory and compatible with the images in particular and in the future leave scope for changes.

**Methods:** In this paper, they propose two models, model A and model B, for the Nepali image descriptor. Model A defines a common encoder-decoder structure where a classifier ResNet model is the encoder and a basic LSTM algorithm is the decoder. Model B is a Show, Attend, and Tell model with Visual attention architecture. A ResNet-50 system with pre - trained is the encoder used. Each one the image input is converted into a 2245222 tensor with RGB channels. That image has been rotated horizontally and compressed with a rate of 0.5. Then the attention of the encoder-decoder is to use those indirect vectors generated by the encoder after proper calculation, allowing them to pay closer attention to the most significant vector corresponding to the object of concern in the picture. After, while producing the caption, any intermediate element of the image model generated by the encoder will be used by the decoder feature vector.

**Datasets:** On top of the existing Microsoft COCO, the data set was created. There are 100,000 + image-caption combinations in the original data set. The training collection contains 82,783 image-caption pairs, 40,504 image-caption pairs in the validation set, and 40,775 in the test set.

**Results:** Strangely, model A performed better than model B without visual perception, with perception indicating that perception is not always the best image captioning process. Model A generates a loss of 0.02 with RAdam and a perplexity of 0.88, which is considerably better than the Adam or ASGD outcomes with the same model. The experiment demonstrates that model B produces a minimum loss of 0.06 with a minimum loss of 0.06 with 1,006 of Adam's perplexity. Model A's perplexity is lower than that of model B, which means that model A It makes assumptions to unseen samples better.

**Conclusion:** The purpose of this paper is to compile and customize the data collection. In order to enhance the analysis in these fields, this report has been publicly published. They demonstrated that, for a sophisticated language like Nepali, the visual-attention model does not always underperform a simple encoder-decoder without visual attention. Manually produced training with a comprehensive training data set Nepali captions and proper fine-tuning should allow researchers to come up with more conceptually accurate captions for the test samples to correlate with the benchmark provided in this paper.

.