CS 584 Project Proposal

Image Caption generation using Encoder-Decoder and Transformers

Team Members:

Venkata Krishna Kapardhi Dendukuri. A20482375
Aman Goyal A20482265

Description

In recent years, neural networks have fueled dramatic advances in image captioning. Researchers are looking for more challenging applications for computer vision and Sequence to Sequence modeling systems. They seek to describe the world in human terms. This task is significantly harder in comparison to the image classification or object recognition tasks that have been well researched.

The biggest challenge is most definitely being able to create a description that must capture not only the objects contained in an image, but also express how these objects relate to each other.

Survey

There have been several attempts to provide solutions to this problem, including template-based solutions using image classification. However, more recent research has focused on recurrent neural networks. RNNs are already very popular for several natural language processing tasks such as: B. A machine translation that produces a sequence of words. The Image Caption Generator extends the same application by generating word-by-word descriptions of images.

Computer vision reads an image and sees it as a two-dimensional array. Venugopalan thus described captioning as a language translation problem. Previously, language translation was complex and involved several different tasks, but recent research shows that recurrent neural networks can be used to accomplish the task in a highly efficient manner. shown. However, his regular RNN suffers from the vanishing gradient problem, which was important in our application case. The solution to this problem is to use LSTM and GRU. They contain internal mechanisms and logic gates that store information for a long time and pass only useful information.

One of the biggest challenges we faced was choosing a suitable model for our subtitling network. In their work, Tanti categorized generative models into two types: injection architectures and merge architectures. In the former he puts both the tokenized label and the image vector into the RNN block, whereas in the latter he puts the label into the RNN block and merges the output with the image. Experiments show that there is no significant difference in the accuracy of the two models, but we chose the merge architecture due to the simplicity of the design. This reduces hidden states and speeds up training. Images are also not repeatedly passed through her RNN network, which makes better use of RNN memory.

Preliminary Project Milestones:

- Data collection and Preprocessing:-
    - Flickr8k
    - Glove6B
- Building State of the art models :-
    - Merged Encoder-Decoder
    - Attention Mechanism
    - Transformers
- Model evaluation using SOTA algorithms:-
    - Greedy Search
    - Beam Search
    - BLEU

References

- [1] ”Every Picture Tells a Story: Generating Sentences from Images.” Computer Vision ECCV (2016) by Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hocken-maier, and David Forsyth

- [2] Automatic Caption Generation for News Images by Yansong Feng, and Mirella Lapata, IEEE (2013).

- [3] Image Caption Generator Based on Deep Neural Networks by Jianhui Chen, Wenqiang Dong and Minchen Li, ACM (2014).

- [4] Show and Tell: A Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).

- [5] Image2Text: A Multimodal Caption Generator by Chang Liu, Changhu Wang, Fuchun Sun, Yong Rui, ACM (2016).