EECS 442 Final Project: Neural Image Captioning Model

Chloe Zeien
czeien

Cameron Nichols
camnic

Thomas He
mperform

1. Introduction

Our goal with this project is to create a model that can generate an appropriate caption for a given image. Object detection is a fundamental problem in computer vision, and we would like to take it even a step further by being able to identify and label the objects in an image. This project will allow us to combine and explore many different concepts from this class and apply them directly to a relevant problem. Image captioning is also a unique problem because it combines computer vision techniques with natural language processing techniques in order to generate captions for the images. In this project, we extend and build on the work of Vinyals et al. in *Show and Tell: A Neural Image Caption Generator*. [Vin+15] Recognizing the technological constraints of the time, we use a more advanced CNN architecture, ResNet-50, that has been pre-trained on the ImageNet dataset for feature extraction. We also utilize the METEOR score over the BLEU score for a more comprehensive evaluation of the final outputs.

2. Related Work

The problem of generating captions from images is one that has been widely studied in computer vision. Our project largely considers many techniques that are discussed in the paper *Show and Tell: A Neural Image Caption Generator* [Vin+15] . In this paper, the authors propose a hybrid model combining Convolutional Neural Networks and Recurrent Neural Networks for image caption generation. The CNNs extract the visual features from the images, while the RNNs generate contextually specific captions. The model is trained on the COCO dataset, which contains images paired with human-generated captions. This teaches the model to combine the visual content with the appropriate textual descriptions. This paper's greatest strength is the integration of both CNNs and RNNs in the Image Captioning Model, which allows the model to capture both visual and contextual information. This is a technique that we would like to recreate and implement in our own image captioning model. One of the major weaknesses of this paper that the authors mention is the lack of high quality large datasets available to them when they were first writing this report. The authors claim that they experienced many challenges with overfitting because they were limited in the amount of training data that they had available to them. We wanted to be able to improve upon this weakness by using the pre-trained ResNet-50 model in our encoder, which is trained on the extremely large ImageNet dataset. Another weakness of this paper is the use of the BLEU score in evaluating the model. BLEU contains limitations in focusing on precision without adequate consideration for recall and semantic understanding. It frequently overlooks the significance of including all words in the reference caption and fails to recognize semantic similarities between words (for example, "red" and "scarlet"). For this reason, we will abandon the use of BLEU score for model evaluation in favor of the METEOR score. METEOR provides a more balanced evaluation by taking into account both precision and recall, as well as recognizing synonyms, resulting in a more comprehensive assessment of the quality of generated captions. We hope that evaluating our model based on the METEOR score as opposed to the BLEU score, we will have a more accurate understanding of the outcome of our model.

3. Method

Our approach to image captioning is based on combining a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN), which gives us a complete system for translating visual data into de-

scriptive language. In our encoder, the ResNet-50 model performs image feature extraction. The features are then interpreted by an LSTM (Long Short-Term Memory) network, which generates relevant captions. ResNet-50 is CNN network with 50 layers which has been pre-trained on the ImageNet dataset, giving it a strong foundational understanding of various visual features. The final fully connected layer of ResNet-50 has been modified to meet the requirements of our captioning task. It generates a feature vector for each image, which serves as input to the RNN in the decoder.

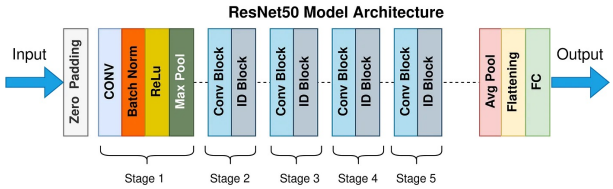

Figure 1. Architecture of the ResNet-50 Model[Muk22]

An LSTM is an RNN architecture widely used in deep learning. We use an LSTM in our decoder because of its ability to handle sequential data and maintain long-term dependencies, which is critical for generating coherent sentences. The LSTM receives the feature vector of each image from the ResNet-50 and generates words to form a caption. This procedure converts the high-dimensional feature space into a linguistic space of potential captions.

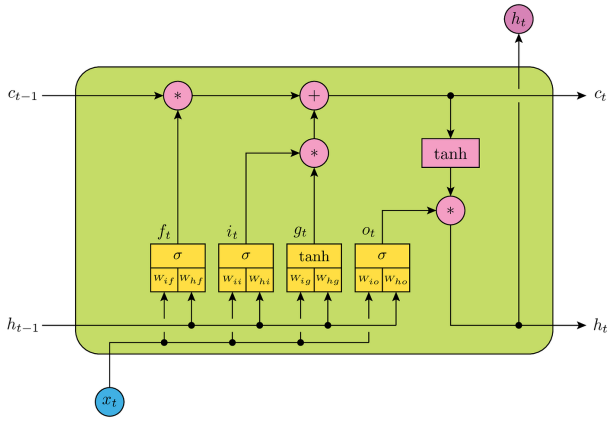

Figure 2. Architecture of the LSTM Model[Hol17]

4. Experiments

We trained and tested our model using the COCO 2017 dataset. This dataset is well-known for its diversity and size, with over 130,000 images each with five different captions, providing a diverse source of visual and textual data. Because of the limited computing power of Google Colab, we trained our model on a subset of this dataset consisting of 10,000 images. In the pre-processing steps, images were resized to size 224 x 244, and then normalized. The training set was tokenized, and a vocabulary was built from it. For sequence processing, special tokens for start, end, and padding were included. The ResNet-50 model has a predetermined number of layers implemented in the pytorch library. The LSTM network was set up with a predetermined number of hidden layers and units from the pytorch library as well. To avoid overfitting, we chose a learning rate of .001. We used a batch size of 32 and trained the model for 50 epochs.

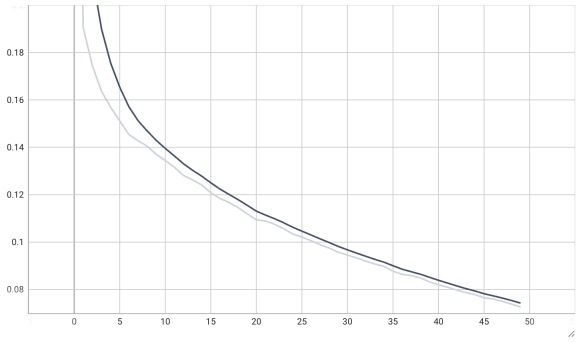

Figure 3. Training Loss over 50 Epochs

The METEOR score was chosen to evaluate the model based upon its balanced evaluation of precision and recall, as well as its ability to recognize synonyms, making it a more comprehensive measure of caption quality. The METEOR score is monitored after the testing phase of the process in order to track model output. The model described in *Show and Tell: A Neural Image Caption Generator*[Vin+15] served as the primary baseline for our research. This allowed for a direct comparison of the progress made by using ResNet-50.

In the end, our model was only able to obtain a final meteor score of 2.3. As we can see, this score is quite low compared the the meteor scores we are comparing to in Figure 4. This suboptimal score is one that we wish we had more time and resources to dig into. One setback that may have led to this result is the limited computing power of Google Colab, as we were only able to train our model on 10,000 images, though

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Table 1. Scores on the MSCOCO development set.

Figure 4. Performance of the Model from *Show and Tell: A Neural Image Caption Generator*[Vin+15]

the COCO dataset we used contained 130,000 images. We were also limited by time during this project as well. If we had had more time, we would have liked to dig deeper into the architecture and the hyperparameters of our model to tune these to optimal performance. We still think that the use of a ResNet-50 and the METEOR score are great improvements to the Image Captioning Model, and would have liked to see these features work strongly in conjunction with one another.

5. Conclusions

In contrast to the method from *Show and Tell: A Neural Image Caption Generator* [Vin+15], we hoped the use of ResNet-50, a more complex and deep architecture, would result in better feature extraction from images, resulting in more detailed and contextually rich captions. In addition, the use of the METEOR score instead of focusing on BLEU for caption evaluation has allowed for a more nuanced evaluation that values semantic richness and contextual accuracy. This combination of advanced architecture and a more comprehensive evaluation metric shows a clear improvement over baseline methods. We would have liked to have more time and resources to see the optimal improvement these changes could have had.

Looking ahead, the project offers several opportunities for additional research and development. Exploration of newer neural network architectures such as Transformers could improve the model's performance and robustness even further. We could also broaden the model's scope to include video data, making it more versatile and applicable in a variety of real-world scenarios. Another path to explore would be training the model on other large image datasets to see how the results compare. There are several options for this such as the CIFAR datasets or the Flickr 30k dataset.

References

[Vin+15] Oriol Vinyals et al. "Show and Tell: A Neural Image Caption Generator". In: *Cornell University* (2015).

[Hol17] Andre Holzner. *LSTM cells in Pytorch*. Tech. rep. Medium, 2017.

[Muk22] Suvaditya Mukherjee. *The Annotated ResNet-50*. Tech. rep. Towards Data Science, 2022.