Image Caption generation using Encoder- Decoder and Transformers

Aman Goyal

Venkata Krishna Kapardhi Dendukuri

Abstract

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. Starting from 2015 the task has generally been addressed with pipelines composed of a visual encoder and a language model for text generation. During these years, both components have evolved considerably through the exploitation of object regions, attributes, the introduction of multimodal connections, fully attentive approaches, and BERT-like early-fusion strategies. However, regardless of the impressive results, research in image captioning has not reached a conclusive answer yet. This work aims at providing a comprehensive overview of image captioning approaches, from visual encoding and text generation to training strategies, datasets, and evaluation metrics. In this respect, we quantitatively compare many relevant state-of-the-art approaches to identify the most impactful technical innovations in architectures and training strategies.

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.

1. Introduction

The challenge of captioning an image involves using a visual understanding system and a language model that can produce coherent and syntactically sound sentences to describe the visual content of the picture in natural language.Just recently has neuroscience research made the connection between human vision and language production clear.

Similar to this, artificial intelligence (AI) architectures that can analyse photos and produce words are a relatively new development. Finding the most efficient pipeline to analyse an input image, represent its content, and convert it into a sequence of words by creating links between visual and textual elements while keeping language fluency is the aim of these research endeavors.

The research community has made significant advancements in model design over the past few years, starting with the first deep learning-based proposals that adopted Recurrent Neural Networks (RNNs) fed with global image descriptors, and progressing through attentive approaches, reinforcement learning, and breakthroughs involving Transformers, self-attention, and single-stream BERT-like approaches.

The difficulty of developing appropriate assessment methods and metrics to compare findings with human-generated ground truths has been tackled by the Computer Vision and Natural Language Processing (NLP) communities at the same time. Nonetheless, despite the research and advancements made in these years, picture captioning is still far from being thought of as a problem that has been addressed.

2. Dataset

A number of datasets are used for training, testing, and evaluation of the image captioning methods. The datasets differ in various perspectives such as the number of images, the number of captions per image, format of the captions, and image size. Three datasets: Flickr8k, Flickr30k, and MS COCO Dataset are popularly used. In the Flickr8k dataset, each image is associated with five different captions that describe the entities and events depicted in the image that were collected. By associating each image with multiple, independently produced sentences, the dataset captures some of the linguistic variety that can be used to describe the same image. Flickr8k is a good starting dataset as it is small and can be trained easily on low-end laptops/desktops using a CPU.

Our Flick8k dataset structure contains the 8000 images, also Flick8k Text contains the image id along with the 5 captions, training image id's test image id's.

Word vectors map words to a vector space, where similar words are clustered together and different words are separated. The advantage of using Glove over Word2Vec is that GloVe does not just rely on the local context of words, but it incorporates global word co-occurrence to obtain word vectors.

The basic premise behind Glove is that we can derive semantic relationships between words from the co-occurrence matrix. For our model, we will map all the words in our 38-word long caption to a 200-dimension vector using Glove.

3. Literature 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.

4. Experiments

We will tackle this problem using an Encoder-Decoder model. Here our encoder model will combine both the encoded form of the image and the encoded form of the text caption and feed to the decoder.

Our model will treat CNN as the ‘image model’ and the RNN/LSTM as the ‘language model’ to encode the text sequences of varying length. The vectors resulting from both the encodings are then merged and processed by a Dense layer to make a final prediction.

We will create a merge architecture in order to keep the image out of the RNN/LSTM and thus be able to train the part of the neural network that handles images and the part that handles language separately, using images and sentences from separate training sets.

In our merge model, a different representation of the image can be combined with the final RNN state before each prediction.

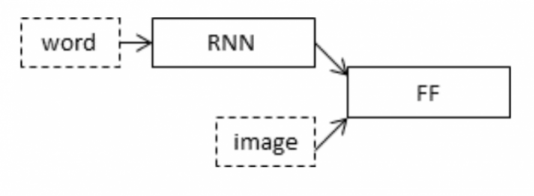
With an Attention mechanism, the image is first divided into n parts, and we compute an image representation of each When the RNN is generating a new word, the attention mechanism is focusing on the relevant part of the image, so the decoder only uses specific parts of the image.

The transformer network employs an encoder-decoder architecture similar to that of an RNN. The main difference is that transformers can receive the input sentence/sequence in parallel, i.e, there is no time step associated with the input, and all the words in the sentence can be passed simultaneously.

5. Methodology

We have implemented 3 different techniques to implement Image Captioning.

5.1. Merged Encoder - Decoder Architecture



The above diagram is a visual representation of our approach.

The merging of image features with text encodings to a later stage in the architecture is advantageous and can generate better quality captions with smaller layers than the traditional inject architecture (CNN as encoder and RNN as a decoder).

To encode our image features we will make use of transfer learning. There are a lot of models that we can use like VGG-16, InceptionV3, ResNet, etc. We will make use of the inceptionV3 model which has the least number of training parameters in comparison to the others and also outperforms them.

To encode our text sequence we will map every word to a 200-dimensional vector. For this will use a pre-trained Glove model. This mapping will be done in a separate layer after the input layer called the embedding layer.

Using photos and phrases from different training sets, we will build a merge architecture to keep the image out of the RNN/LSTM and enable us to train the neural network's parts that deal with images and language independently.

A distinct picture representation can be blended with the final RNN state prior to each prediction in our merging model.

The merging of image features with text encodings to a later stage in the architecture is advantageous and can generate better quality captions with smaller layers than the traditional inject architecture (CNN as encoder and RNN as a decoder).

To encode our image features we will make use of transfer learning. There are a lot of models that we can use like VGG-16, InceptionV3, ResNet, etc. We will make use of the inceptionV3 model which has the least number of training parameters in comparison to the others and also outperforms them.

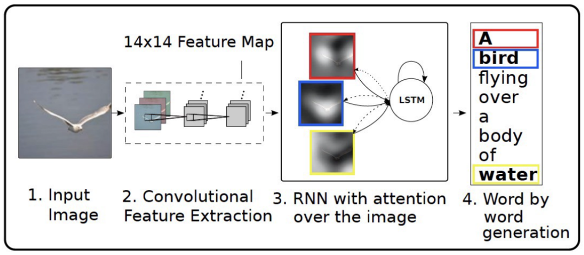
To encode our text sequence we will map every word to a 200-dimensional vector. For this will use a pre-trained Glove model. This mapping will be done in a separate layer after the input layer called the embedding layer.

5.2. Attention Mechanism

Rather than compressing an entire image into a static representation, the Attention mechanism allows for salient features to dynamically come to the forefront as and when needed. This is especially important when there is a lot of clutter in an image.

RNNs tend to be computationally expensive to train and evaluate, so in practice, memory is limited to just a few elements. Attention models can help address this problem by selecting the most relevant elements from an input image.

With an Attention mechanism, the image is first divided into n parts, and we compute an image representation of each When the RNN is generating a new word, the attention mechanism is focusing on the relevant part of the image, so the decoder only uses specific parts of the image.


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

In Bahdanau or Local attention, attention is placed only on a few source positions. As Global attention focuses on all source side words for all target words, it is computationally very expensive. To overcome this deficiency local attention chooses to focus only on a small subset of the hidden states of the encoder per target word.

Local attention first finds an alignment position and then calculates the attention weight in the left and right windows where its position is located and finally weights the context vector. The main advantage of local attention is to reduce the cost of the attention mechanism calculation.

In the calculation, the local attention is not to consider all the words on the source language side, but to predict the position of the source language end to be aligned at the current decoding according to a prediction function and then navigate through the context window, considering only the words within the window.

All hidden states of the encoder and the decoder are used to generate the context vector. The attention mechanism aligns the input and output sequences, with an alignment score parameterized by a feed-forward network. It helps to pay attention to the most relevant information in the source sequence. The model predicts a target word based on the context vectors associated with the source position and the previously generated target words.

To evaluate our captions in reference to the original caption we make use of an evaluation method called BLEU. It is the most widely used evaluation indicator. It is used to analyze the correlation of n-gram between the translation statement to be evaluated and the reference translation statement.

In this article, multiple images are equivalent to multiple source language sentences in the translation. The advantage of BLEU is that the granularity it considers is an n-gram rather than a word, considering longer matching information. The disadvantage of BLEU is that no matter what kind of n-gram is matched, it will be treated the same.

6. Evaluation Techniques

We have implemented 2 Evaluation Techniques for our Image Caption Generator.

6.1. Greedy Search

It is a local search algorithm that always chooses the best available option at each step, without considering the long-term consequences of that choice. The algorithm works by starting with an initial solution and then iteratively improving it by choosing the best possible option at each step. Greedy search is easy to implement and computationally efficient, making it a popular choice for solving a wide range of problems in various fields, including computer science, mathematics, and engineering.

It can get stuck in local optima, where a suboptimal solution is chosen early on and then further improvements become impossible. Greedy search also does not guarantee that the final solution will be globally optimal, as it does not explore all possible options before making a decision.

As the model generates a 1660 long vector with a probability distribution across all the words in the vocabulary we greedily pick the word with the highest probability to get the next word prediction.

6.2. Beam Search

This algorithm is an extension of the greedy search algorithm and works by maintaining a set of the most promising candidate solutions, called the ”beam,” at each step of the search. Unlike greedy search, beam search considers a limited set of possible solutions at each step, which helps to reduce the computational complexity of the algorithm. This also allows beam search to explore a larger portion of the search space, potentially leading to a better global solution.

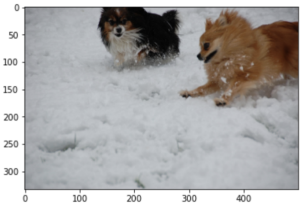
Beam Search is where we take top k predictions, feed them again in the model and then sort them using the probabilities returned by the model. So, the list will always contain the top k predictions and we take the one with the highest probability and go through it till we encounter ‘endseq’ or reach the maximum caption length.

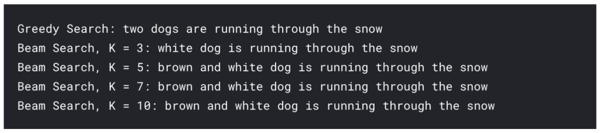
7. Results

7.1. Merge Encoder - Decoder

We will test our model on the merge encoder - decoder architecture and view its performance.

First, we will take a look at the example dog image. The caption for the image is ‘A black dog and a brown dog in the snow’. Let’s see how our model compares.

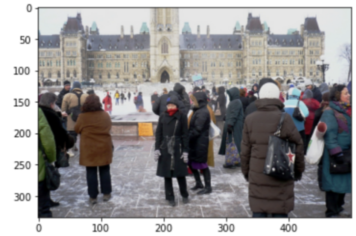


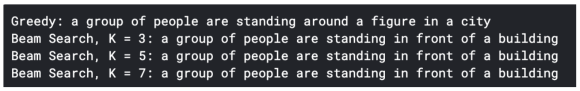


```
Greedy Search: two dogs are running through the snow
Beam Search, K = 3: white dog is running through the snow
Beam Search, K = 5: brown and white dog is running through the snow
Beam Search, K = 7: brown and white dog is running through the snow
Beam Search, K = 10: brown and white dog is running through the snow
```

You can see that our model was able to identify two dogs in the snow. But at the same time, it misclassified the black dog as a white dog. Nevertheless, it was able to form a proper sentence to describe the image as a human would.

Next, we will take a look at the another example image.Let’s see how our model compares.



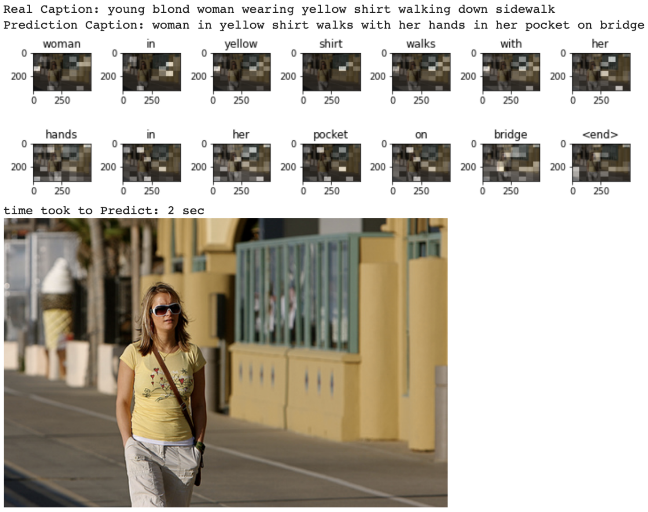


```
Greedy: a group of people are standing around a figure in a city
Beam Search, K = 3: a group of people are standing in front of a building
Beam Search, K = 5: a group of people are standing in front of a building
Beam Search, K = 7: a group of people are standing in front of a building
```

Here we can see that we accurately described what was happening in the image. You will also notice the captions generated are much better using Beam Search than Greedy Search.

7.2. Attention Mechanism

We will test our model on the attention mechanism architecture and view its performance.



```
Real Caption: young blond woman wearing yellow shirt walking down sidewalk
Prediction Caption: woman in yellow shirt walks with her hands in her pocket on bridge
```

```
time took to Predict: 2 sec
```

You can see even though our caption is quite different from the real caption, it is still very accurate. It was able to identify the yellow shirt of the woman and her hands in the pocket.

8. Conclusion

In conclusion, the image caption generator using a merged encoder-decoder and attention mechanism is a powerful technique for generating natural language descriptions of images. By combining the strengths of both encoder-decoder models and attention mechanisms, this approach is able to accurately capture the key features of an image and produce a semantically coherent and grammatically correct caption. Furthermore, the attention mechanism allows the model to focus on the most relevant parts of the image when generating the caption, leading to more accurate and descriptive captions.

While this approach has demonstrated impressive results, there is still room for improvement. One area for future research is exploring different ways of merging the encoder and decoder models, such as using a hierarchical structure or incorporating additional information sources. Additionally, improvements can be made to the attention mechanism, such as incorporating visual or contextual cues to guide the attention process. Overall, the image caption generator using merged encoder-decoder and attention is a promising approach that has the potential to revolutionize the field of computer vision and natural language processing.

9. Future Work

In the future, one potential direction for further research in the field of natural language processing is the implementation of transformer-based models and the use of the Bilingual Evaluation Understudy (BLEU) score for evaluation. Transformers have shown impressive performance on a wide range of NLP tasks, and incorporating them into image captioning models could potentially lead to even more accurate and natural language descriptions of images. Additionally, using the BLEU score as an evaluation metric would provide a more objective measure of the quality of the generated captions, allowing for more precise comparisons between different models. However, implementing transformers and BLEU in image captioning models poses several challenges, such as the high computational cost of training large transformer-based models and the difficulty of ensuring that the generated captions are both accurate and linguistically diverse. Addressing these challenges will require further research and development, but has the potential to significantly advance the field of image captioning.

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).