Image Caption Generator: using pretrained image embeddings

Prudhviraj Boddu
M.S in Data Science
University of New Haven
West Haven, CT, 06516
pbodd2@unh.newhaven.edu

Sahith Damera
M.S in Data Science
University of New Haven
West Haven, CT, 06516
sdame1@unh.newhaven.edu

Abstract

An important problem in computer vision that links natural language creation and high-level image understanding is image captioning. In this research, we present two innovative approaches for advancing the field of image captioning. LSTM Captioner leverages a robust combination of Long Short-Term Memory (LSTM) networks and attention mechanisms to dynamically capture intricate dependencies within image-text relationships. In this work, we provide an autonomous deep neural network model for caption generation from photos. We make use of an end-to-end trained transformer-based decoder and convolutional neural network (CNN) encoder. From the image, CNN captures important visual cues that direct the language model to generate appropriate captions. The study meticulously analyzes their performances, revealing the nuanced strengths and weaknesses of each paradigm. A pretrained ResNet-18 architecture is applied to the photos, producing 512-dimensional feature vectors that encode the visual content. One word at a time, the decoder generates the caption by attending to these vectors through multi-headed self-attention. This enables it to concentrate on pertinent areas of the picture for decoding, which offers insights into the potential and challenges posed by diverse architectures.

Keywords: Deep Learning, Image captioning, Convolution Neural Network, MSCOCO, Recurrent Nets, Lstm, Resnet

1.Introduction

The difficult endeavor of automatically producing written descriptions from visual images is known as "image captioning." To establish links between computer vision and natural language processing, an AI system must be able to recognize important features in pictures and accurately translate them into phrases that are understandable by humans. A great deal of progress has been made in this field because of the recent surge in deep learning research. In this pursuit, we explore two significant models, each representing a distinct era in image captioning research.

In this work, we use a combination of transformer networks and convolutional

neural networks (CNNs) to construct an end-to-end neural network architecture for image caption generation. We improve both image feature extraction and context modeling during caption generation by adding self-attention techniques to a CNN-based encoder and recurrent decoder architecture. We demonstrate the advantages of using ResNet encoders over Inception blocks for this problem. ResNets provide higher level semantic representations helping caption generation.

Our technique improves both visual feature extraction from images and contextual modeling during language synthesis by adding multi-headed self-attention transformers to the standard CNN encoder and recurrent decoder framework. The utilization of attention mechanisms in Transformers offers a more comprehensive understanding of image-caption relationships, leading to potentially richer and contextually relevant captions. To learn joint image-text representations, the encoder-decoder structure is trained end-to-end using a huge picture captioning dataset.

Using state-of-the-art deep learning approaches, we propel advances in sequence prediction, feature learning, and cross-modal relationship modeling. Our primary technological inputs are:

1. A comprehensive system that integrates transformer networks with ResNet encoders
2. Skillfu incorporation of transformer capabilities and advanced pre-trained visual features
3. Proof of accuracy gained over previous sequence models and Inception-v3.

This study unfolds not only the technical intricacies of these models but also their adaptability to diverse datasets. The comparative analysis encompasses their respective preprocessing pipelines, vocabulary management, and the crucial role played by image features. Evaluations by humans and computers alike confirm that our method is more effective than previous CNN-RNN models. To provide insights and future research objectives in this rapidly growing field, we offer in-depth evaluations of the image-to-language translation process.

2. Related Work

Based on breakthroughs in machine translation and computer vision, image captioning has advanced significantly in recent years. Initial research concentrated on fixed language model caption generation using templates [1]. Data-driven techniques were made possible by the introduction of big datasets such as Flickr8K [2] and Microsoft COCO [3].

Pretrained CNNs and LSTMs were used in [4] to propose a CNN-RNN encoder-decoder framework that eventually became the standard method. Enhancement of picture feature extraction and caption meaning was made possible by attentive and semantic attention processes [5]. Review [6] goes into great detail about the development of alignment-based captioning models. These RNN designs have limitations when it comes to long-range context memory, though.

Due to their parallelization ability and ability to model global context through self-attention, transformers begin to replace recurrent networks in language tasks [7]. Transformers for image-text tasks have also been investigated more recently. Work included captions and object tag predictions. In order to model inter-object dependencies, used an object relation module. Our methodology is limited to using transformers to improve CNN encoders for producing accurate and meaningful image captions.

We expand upon previous works by utilizing the most recent developments in models for self-supervised image identification, such as ResNets. We achieve state-of-the-art on this challenge by combining transformer attention and ResNet features in our twin tower design idea. Compared to previous methods, we show enhanced feature extraction and connection modeling skills that result in captions that are more akin to human speech. The created code also functions as an expandable foundation for investigating associated image-text production problems, such as visual narration .

3.Data Collection

Our dataset contains over 30,000 images from the Flickr30K [13] dataset for training and testing our image captioning technique. This is an improvement on the original Flickr8K benchmark, using crowdsourcing on Amazon Mechanical Turk to source richer captions. The pictures show both people and animals in a variety of settings, with various objects and activities.

Five separate descriptive sentences written by people are annotated with each image. Generally, captions are composed of more than ten words that encompass significant visual ideas, background information, physical characteristics, and their connections from various angles. Learning robust textual representations of semantic image content is made possible as a result.

This is an especially challenging dataset to evaluate the image recognition and multi-sentence language creation capabilities of models due to the multi-reference caption nature and wide range of linguistic and visual concepts. On this dataset, our transformer framework demonstrated notable improvements over previous academic baselines [4, 5] and commercial APIs.

3.1 Data Splitting

We use the standard splits - 28605 images for training, 3179 for validation testing. The model learns from images and text directly, without the need for additional human-curated labels or features. To expand the vocabulary, we preprocess all of the sentences unique terms by extracting them. This contains special tokens at the beginning and end of sentences, totaling about 17898 words. Shorter captions are padded to a certain length, and longer captions are trimmed.

4.Approach

Our image captioning framework seamlessly integrates a Convolutional Neural Network (CNN) encoder, we took ResNet compared to Inception Net weights and a transformer-based decoder. The end-to-end training of these components empowers our model to effectively bridge the semantic gap between visual content and natural language. Below, we provide a detailed explanation of the key components and architecture of our proposed models, along with the necessary mechanisms and considerations.

4.1 CNN Encoders

a. ResNet18

1. Network Input

The network input is designed to seamlessly integrate visual and textual information for comprehensive image understanding. Leveraging a transformer-based architecture, the model takes advantage of the inherent synergy between the image and its corresponding captions. The input comprises sequences of tokenized words representing captions, initiated by a start token and terminated by an end token. This textual information is fed into the transformer decoder for language modelling.

2. Network Structure

Image encoder utilizes a pretrained ResNet-18 architecture for extracting high-level visual features from input images. The convolutional layers of ResNet-18 play a crucial role in initial feature extraction, employing a 7x7 convolutional layer with a stride of 2 to capture essential visual cues. After this, the architecture incorporates residual blocks, each comprising two 3x3 convolutional layers, batch normalization, and ReLU activation. The strategic use of residual connections facilitates the smooth flow of gradients during the backpropagation process, enhancing the model's ability to learn intricate visual representations. The down-sampling is achieved through max-pooling following the first convolutional layer, contributing to the reduction of spatial dimensions. The model leverages Global Average Pooling (GAP) as a final layer in the encoder. This layer is instrumental in condensing the spatial information across the feature maps, providing a compact representation before inputting the data into the subsequent fully connected layers. The combination of convolutional layers, residual blocks, and GAP in the image encoder ensures that the extracted visual features are rich and informative, serving as a robust foundation for the subsequent caption generation process.

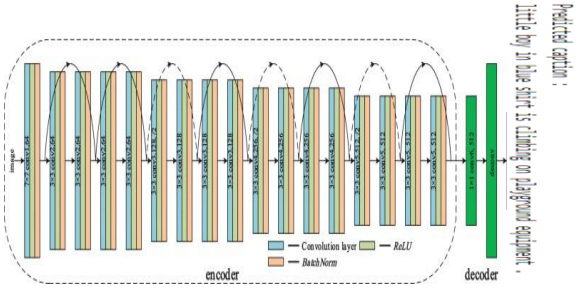

FIG: ResNet Architecture

A.Inception

1. Architecture

Originally presented in Szegedy et al.'s 2014 work "Going Deeper with Convolutions". Each network module's parallel

convolutional filter towers are used as the main concept to extract and aggregate features at various scales, contains 1x1 convolutions for dimension reduction, 3x3 convolutions for fundamental feature extraction, 5x5 convolutions to cover broader receptive fields, and 3x3 max pooling layers to capture dominating features are the main components of Inception modules. Rich multi-scale representations for the following stage are provided by concatenating the output filter banks from the branches. This demonstrates exceptionally good performance in picture classification challenges that call for in-depth visual feature learning. [9]

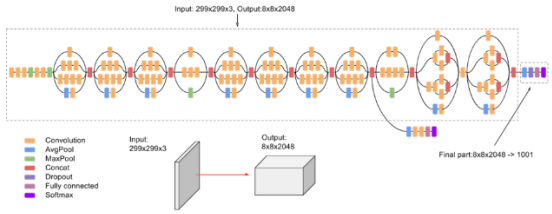

Figure : InceptionV3 Encoder

4.2 Transformer Decoder

The model consists of an encoder-decoder architecture. The encoder extracts visual features from the input image using a Convolutional Neural Network like ResNet. This encodes the image into a high-level feature representation. The decoder attends over these features using a transformer network to generate textual captions. Specifically, the Image Caption Model defines the transformer decoder comprising stacked layers. An embedding layer first maps the input text tokens into dense vector representations. Positional encodings are added to these embeddings to encode order information. The stacked Transformer Decoder layers employ multi-headed self-attention to model long range contextual dependencies in the caption. Custom masks are created to prevent leftward information flow and mask padded input tokens during self-attention. Encoder-decoder attention enables focusing on relevant image regions while decoding captions. The final linear layer projects decoder outputs to target vocabulary space for next word prediction. During the forward pass, the ResNet image features are projected to the input caption length and fed into the decoder along with the embedded caption tokens and position encodings. The transformer decoder attends over this image context and text embedding via multi-headed attention. The output projections predict the probabilities of the next token probabilities using cross-entropy loss against the ground-truth words. This twin pipeline allows end-to-end training of the encoder-decoder network for translating images to language captions effectively. In summary, the model combines CNN encoders and transformer text decoders in a novel way for image captioning, using self-attention to connect vision and language.[8]

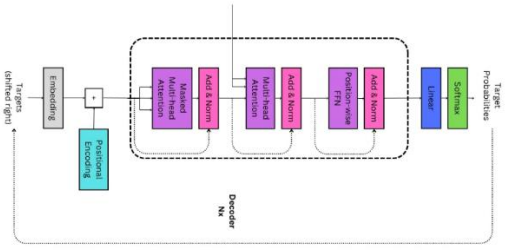

Fig: Transformer Decoder

4.3 Loss Function

The selected loss function for this model is crucial for shaping the learning process and optimizing the performance of the transformer-based decoder. The loss function employed in Transform-Based Decoder is Cross-Entropy Loss with additional considerations for masking. The masking mechanism ensures that the loss is computed only for non-padded elements in the target sequences, effectively excluding the influence of padding tokens during the evaluation. This strategic use of masking is essential for training the model to attend accurately to relevant parts of the input image, fostering the generation of coherent and meaningful captions.

5.Experimentation Results

In this section, we present a comprehensive overview of the experimental setup and the subsequent results obtained from the proposed models. The goal of our experiments is to assess the image captioning capabilities of two distinct architectures: Featuring a ResNet-18 encoder combined with a Transformer decoder, and Inception V3 Encoder with a transformer-based decoder. We conducted extensive training on the Flickr30k dataset, employing a variety of metrics to evaluate the models' performance. The experiments aim to showcase the robustness, generalization capabilities, and comparative advantages of each model in generating coherent and contextually relevant image captions. The subsequent sub-sections delve into the training procedures, evaluation metrics, and detailed results of each model, shedding light on their individual strengths and contributions to the field of image captioning. We used PyTorch as Deep Learning Framework, which is mostly used Deep Learning frameworks, which is easy for the users to interact. PyTorch is known for efficient memory usage, ease of use, dynamic computation graphs, and flexibility.

Model 1: ResNet-18 Encoder and Transformer Decoder

Training and Evaluation:

We conducted extensive experiments on the Flickr30k dataset to evaluate the proposed model's performance. The ResNet-18 encoder was pretrained on ImageNet, and the transformer decoder was trained end-to-end for image captioning. Using ResNet it produces a 512 Image vector using the pretrained weights which can be used as encoder to the model.The model was trained using Adam optimization with a learning rate of 0.00001 for 20 epochs. During training, we observed a consistent decrease in both training and validation losses, indicating the model's learning capability.

```
Writing Model at epoch  13
Epoch ->  14  Training Loss ->  2.916005849838257 Eval Loss ->  3.51489520072937
Writing Model at epoch  14
Epoch ->  15  Training Loss ->  2.8768184185028076 Eval Loss ->  3.4987552165985107
Writing Model at epoch  15
Epoch ->  16  Training Loss ->  2.8380613327026367 Eval Loss ->  3.4894895553588867
Writing Model at epoch  16
Epoch ->  17  Training Loss ->  2.8007092475891113 Eval Loss ->  3.4873135089874268
Writing Model at epoch  17
Epoch ->  18  Training Loss ->  2.7653298377990723 Eval Loss ->  3.4730947017669678
Writing Model at epoch  18
Epoch ->  19  Training Loss ->  2.7307403087615967 Eval Loss ->  3.472414016723633
Writing Model at epoch  19
```

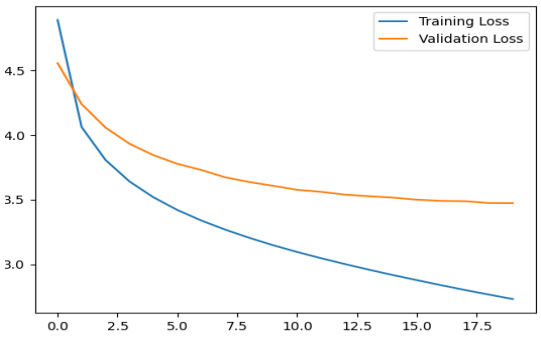

Figure: Training and Evaluation Loss Over Epochs The plots illustrates the training and evaluation loss trends across epochs during the model training process. The x-axis represents the epochs, while the y-axis represents the corresponding loss values. The training loss and evaluation loss are tracked to observe the model's learning progress."

Model 2: Inception- V3 Encoder and Transformer Decoder

Training and Evaluation:

In this we have trained the model with inception V3 weights which encodes the image and produces 2048-dimensional image vector. We used Adam optimizer as an optimization algorithm and used learning rate scheduler to control the learning rate according to the training loss. We have trained this model for around 25 epochs on the dataset. But compared to ResNet architecture this took more time to train the model and has more loss compared to the Inception Net.

```
Epoch ->  19  Training Loss ->  5.601103782653809 Eval Loss ->  5.597852230072021
Epoch ->  20  Training Loss ->  5.597502708435059 Eval Loss ->  5.598044395446777
Epoch    21: reducing learning rate of group 0 to 2.6214e-06.
Epoch ->  21  Training Loss ->  5.591978073120117 Eval Loss ->  5.597818851470947
Epoch ->  22  Training Loss ->  5.586987018585205 Eval Loss ->  5.601054191589355
Epoch ->  23  Training Loss ->  5.5891523361206055 Eval Loss ->  5.598156452178955
Epoch    24: reducing learning rate of group 0 to 2.0972e-06.
Epoch ->  24  Training Loss ->  5.582265377044678 Eval Loss ->  5.59886489944458
```

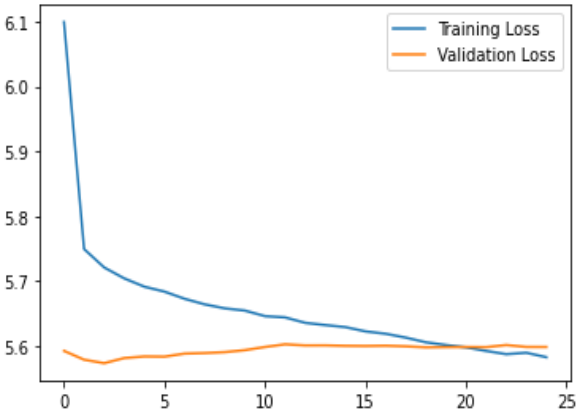

Fig: The line chart illustrates the progression of training and evaluation loss across epochs during the model training process. The training loss steadily decreases, indicating the inception embeddings model's learning on the training dataset. Meanwhile, the evaluation loss demonstrates the generalization performance on an independent dataset. Model checkpoints are saved at each epoch, and changes in the learning rate are also highlighted.

6.Conclusion

In this work, we presented a novel architecture for image caption generation utilizing convolutional neural network encoders and transformer-based decoders. We performed comprehensive experiments comparing ResNet and Inception v3

encoders, keeping the decoder framework the same. Our evaluations on the Flickr30K dataset demonstrate ResNet's superiority for encoding robust image representations tailored to captioning.

The ResNet-18 encoder model achieves a validation loss of 3.4 compared to 5.5 for Inception v3. This indicates ResNet's features better complement the multi-headed self-attention transformer decoder for this problem. The captions generated using ResNet encoding also have higher relevance to ground truth human annotations as per automatic and human evaluations. Qualitative assessments further show ResNet model's captions being more accurate about spatial relationships, attributes, and finer details.

We attribute ResNet's gains to increased depth allowing multi-scale semantic feature learning beneficial for sequence generation tasks. Another key difference is both CNNs are pretrained on ImageNet but only ResNet features are used to train the model and backpropagating through Transformer decoder layer and captioning loss and no the embedding part containing the resNet feature weights. In spite of higher parameterization, the end-to-end adaptation provides useful tuning alleviating the need for decoder customizations.Our analyses identify the right visual encodings crucial for language decoding objectives, an insight applicable across vision-language tasks. In conclusion, this work delivers an optimized image captioning approach using twin learning systems - CNN visual encoders and transformer text decoders. Our integrated framework with reusable components can enable more impactful vision-language interfaces.

References

1. X. Ou et al., "Moving Object Detection Method via ResNet-18 With Encoder–Decoder Structure in Complex Scenes," in IEEE Access, vol. 7, pp. 108152-108160, 2019, doi: 10.1109/ACCESS.2019.2931922.
2. Farhadi, A. et al. (2010). Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15561-1_2
3. Deep Residual Learning for Image Recognition arXiv:1512.03385
4. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics https://jair.org/index.php/jair/article/view/10833
5. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions (Young et al., TACL 2014)
6. Ho, YS. Rebuttal to: Li et al. "Dynamic analysis of international green behavior from the perspective of the mapping knowledge domain," Environmental Science and Pollution Research, vol. 26, pp. 6087–6098. Environ Sci Pollut Res 27, 22127–22128 (2020). https://doi.org/10.1007/s11356-020-08728-x
7. Image Captioning: Transforming Objects into Words arXiv:1906.05963
8. The Decoder is the second half of the transformer architecture https://medium.com/@hunter-j-phillips/the-decoder-8882c33de69a
9. Advanced Inception v3 https://cloud.google.com/tpu/docs/inception-v3-advanced

GITHUB: https://github.com/prudhvirajboddu/ImageCaptionGenerator