15.S04 Hands-on Deep Learning

Image Captioning using CNNs and Attention Mechanism

Team Members: Abhranil Chakrabarti, Nikos Galanos, Arushi Jain, Xiaoyu (Chloe) Wu

Introduction

Over the past years, the abundance of available data and computing power has led to significant progress in the field of Computer Vision and Natural Language Processing. In particular, the areas of image understanding and analysis, as well as the ones of language generation, have experienced great advancements, bringing them closer to the human level. Among the various challenging tasks within this domain, image captioning has emerged as a captivating research problem, aiming to automatically generate human-like descriptions for images.

Image captioning is a deep learning task that involves generating textual descriptions, or captions, for images. It aims to develop models that can automatically understand and describe the visual content of an image in a way that is meaningful and relevant to humans. Image captioning has been successfully applied in various real-world scenarios, such as providing descriptions of images for visually impaired individuals, generating engaging captions for social media posts, and creating product descriptions on e-commerce platforms at scale.

The image captioning problem typically involves two main tasks:

- ➢ **Image understanding:** The model needs to analyze and understand the visual content of the image, including objects, scenes, and relationships between them. This may involve object detection, scene recognition, and visual feature extraction to capture relevant visual information from the image.
- ➢ **Natural language generation:** The model needs to generate a coherent and contextually relevant caption in natural language that describes the visual content of the image. This requires language modeling, and semantic and contextual understanding to generate meaningful and fluent captions that are understandable to humans.

Problem Description

Traditionally, image captioning systems relied heavily on handcrafted features and linguistic rules to generate captions. However, these approaches often suffered from limited generalization and lacked the ability to capture the diverse and intricate visual semantics of images. By leveraging the power of deep learning, we can effectively learn relevant image features and focus on specific regions of interest, resulting in more accurate and contextually relevant captions. Through this project, we seek to develop an efficient algorithm that will accurately understand images and enable machines to generate more human-like and contextually meaningful descriptions for them.

The main objective of our project is to develop an image captioning system that utilizes Convolutional Neural Networks, Gated Recurrent Units, and Attention Mechanisms to generate accurate and relevant image captions. In order to achieve that, we aim to address the following challenges so that our system not only produces captions closely aligned with human descriptions but also demonstrates a high degree of generalization:

- **Feature Extraction**: Utilizing CNN architectures and pre-trained models to extract meaningful image features that capture both low and high-level image details
- **Attention Mechanism:** Implementing an attention mechanism that enables our system to adaptively allocate attention to different image regions based on their significance

- **Language Generation:** Developing a language model that generates accurate captions describing the images
- **Evaluation:** Assessing the performance of our model through BLEU score

A potential stretch goal of this project is to test our image captioning model's robustness against adversarial attacks, which alter input data in small ways to deceive machine learning models and generate misleading captions.

Dataset

For the purposes of our project, we utilized the Flickr 8k[1] dataset. It is a popular image captioning dataset consisting of 8000 images from day-to-day scenarios such as people, landscapes, and objects sourced from the photo-sharing website Flickr. Each image in the dataset is accompanied by 5 human-created captions ensuring a diverse range of descriptions and interpretations which facilitate the exploration of different approaches to caption generation. The images of this dataset are already split into training and test sets (75-25 split).

Additionally, once we successfully created an image captioning model, we attacked the model by "generating" adversarial examples in the feature space to deceive the network into predicting wrong captions.

Methodology

Image Captioning

Caption generation is a challenging problem that requires methods both from the area of computer vision to understand the content of the image, and from natural language processing to generate a meaningful caption. Traditionally, image captioning models have an image encoder that uses a pre-trained Convolutional Neural Network to extract image features. These features are then decoded using a language model such as LSTM/ GRU that generates the caption word-by-word. The performance is evaluated using the BLEU score[2].

The problem with the traditional approach is that the captions are focused on local image features and cannot capture the "essence" of the image. To address this, we leveraged an attention mechanism that can identify the relevant parts of the image given some context in the form of partially generated captions which can be used to predict the next word in the caption.

Data Preprocessing

Images: To extract the features from images (e.g., Figure 1), we utilize the VGG-16 model which is pre-trained on the ImageNet dataset for classifying images.

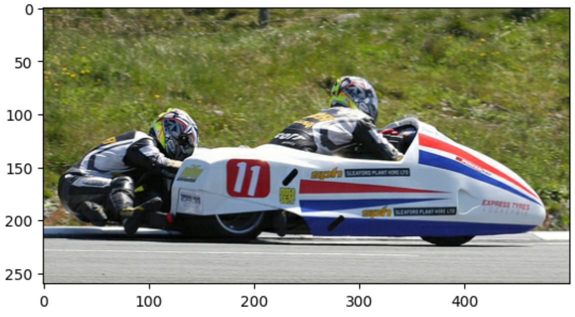

Figure 1: Original Image

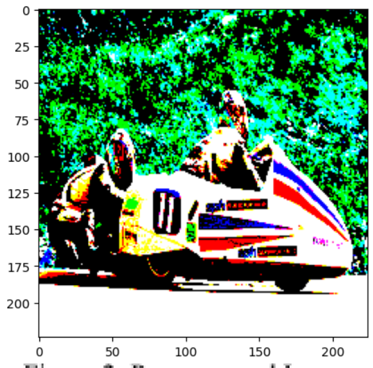

Figure 2: Preprocessed Image

To use this model, we applied some preprocessing to our images such as reshaping and changing from RGB to BGR before feeding them to the VGG-16 model (e.g., Figure 2).

Captions: We also perform some preprocessing to the captions provided in the Flickr dataset by

- Removing punctuations and common stopwords
- Creating a vocabulary of 5000 words from the training dataset and tokenizing the captions in the train and test dataset based on the training dataset vocabulary
- Padding the train and test sequences to keep the captions of a fixed length to be input to the GRU model (max_length = 24 tokens)

Original Caption: A man street racer armor examine tire another racer motorbike
Tokenized Caption Vector: [2, 4, 6, 52, 465, 3339, 1212, 323, 53, 465, 661,
3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Here, 2 is the start token, 3 is the end token and 0 is the pad token.

Model

Feature Extraction from VGG-16

We pass the preprocessed images through the VGG-16 model and extract the image features from the second last layer. We use transfer learning to make our training faster as we believe that the pre-trained weights could be a good starting point for our model and would give better performance since we are using a powerful model trained on a large dataset of day-to-day images.

Encoder

The next step is to define a VGG-16 encoder, through which the VGG-16 image features are passed to a fully connected layer followed by Dropout and the ReLU activation function.

Decoder with Attention Mechanism

The output of the encoder, which consists of the encoded features of each image and the previous hidden layer of the decoder (initially set to zeros) is used to compute an attention score.
We allow for 2 types of attention mechanisms:

- Bahdanau Attention[3]

$$score(h_t, h_s) = v_a * tanh(W_1 * h_t + W_2 * h_s)$$

- Luong Attention[4]

$$score(h_t, h_s) = h_t * W * h_s$$

Notation:
h_t: $Previous\ Decoder\ Hidden\ State$
h_s: $All\ Encoder\ Hidden\ States\ (total\ S)$
v_a, W_1, W_2: $Attention\ Weights$

Once we have the attention scores, we compute the attention weights (softmax):

$$\alpha_{ts} = exp(score(h_t, h_s))/ \sum_{all\ states\ s'\ in\ S} exp(score(h_t, h_{s'}))$$

Finally, we compute the contextualized vector for each feature from the encoder:

$$c_t = \sum_s \alpha_{ts} * h_s$$

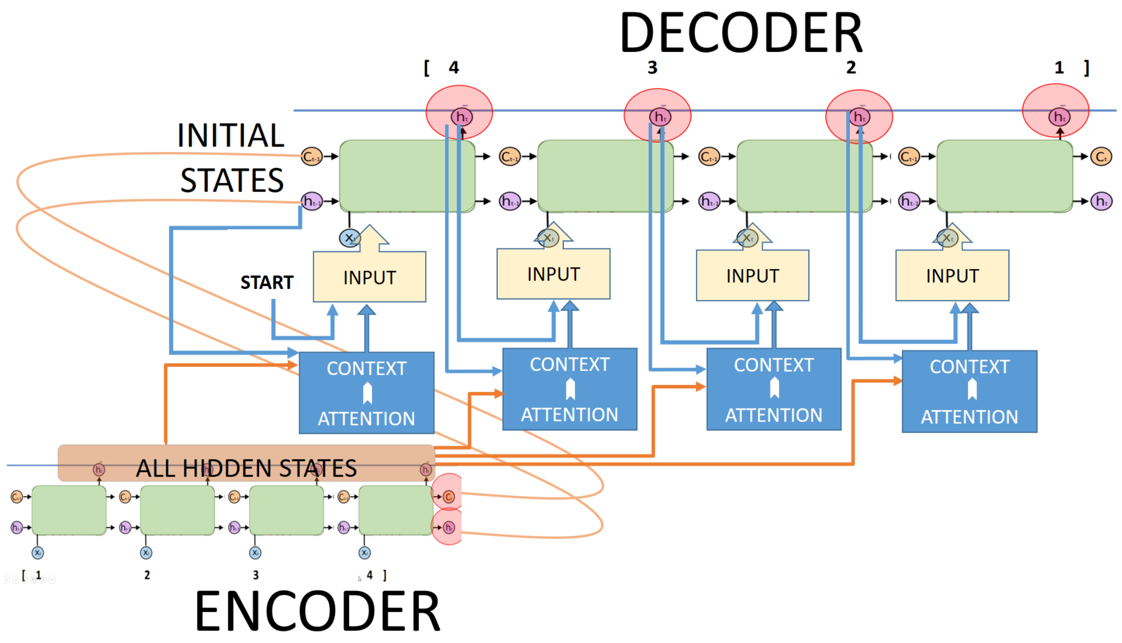

Figure 3: Illustrative Example of An Encoder-Decoder Architecture with Attention Mechanism

The output of the attention mechanism, which is the context vector, along with an input sequence, is now passed to the GRU Decoder. The input sequence is passed through an embedding layer to generate word embeddings. The context vector is then concatenated with the embeddings to generate the final input which is passed through the GRU layer and two fully connected layers with batch normalization and dropout. This produces a new hidden state and an output that consists of a list of the next words in the sequence with their respective probabilities.

Training

To train our model, we initiate the VGG-16 encoder, the GRU decoder and we choose Adam as our optimizer. We also need to initialize the hidden state for each batch as the captions are not related from image to image. We also choose categorical cross entropy as our loss function.

Prediction

To predict the caption based on the output probabilities, we employ different strategies:

- Greedy Approach: A naive strategy to generate the caption would be to predict the most probable word at each step.
- Beam Search: An alternate approach is to consider sequences of words rather than words in isolation. Beam search employs this approach wherein at each stage it considers the 'k' most probable sequences (k being the beam index). The rest of the possible sequences are discarded and the k-best sequences at each step are propagated further till an end token is encountered. The most probable sequence is then chosen as the prediction.

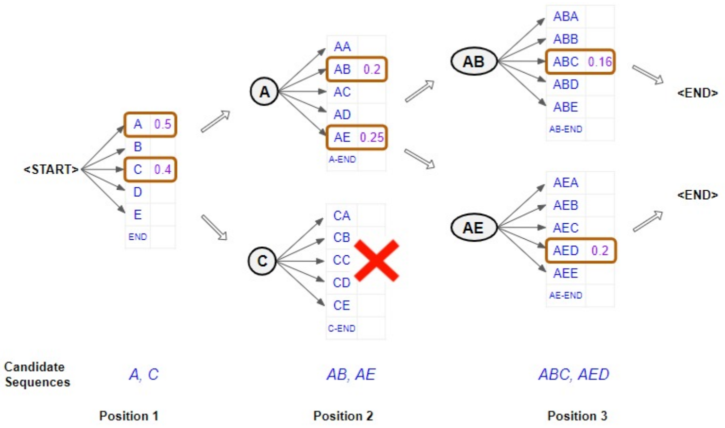

Figure 4: Illustrative Example of Beam Search

Adversarial Attacks

It is often possible to generate examples (images/image features) that are quite close to the original but can deceive the network into predicting incorrect outputs. When training a neural network, we use backpropagation to obtain the gradient of the loss with respect to the network parameters and use it to update the parameters that minimize the loss.

We can also use backpropagation to obtain the gradient of the loss with respect to the input image/image features, given the network parameters. If we update the image features according to that gradient, we can generate features that maximize certain activations, or minimize the loss. If properly applied, small changes in the features can completely change the neural network's prediction and result in adversarial examples.

We implement this for a few train images by matching them against wrong captions, generating adversarial features which can deceive the network, demonstrating the need for more robust training procedures which remains an open challenge to this day.

Results

Image Captioning with Attention Mechanism

Figure 5 shows the loss for the training set against the test set for different epochs. Our result suggests that the model has a lower loss for the test set and has no sign of overfitting, which generalizes well on unseen data. We further introduce the BLEU (Bilingual Evaluation Understudy) score to examine the quality of our generated captions in the test set.

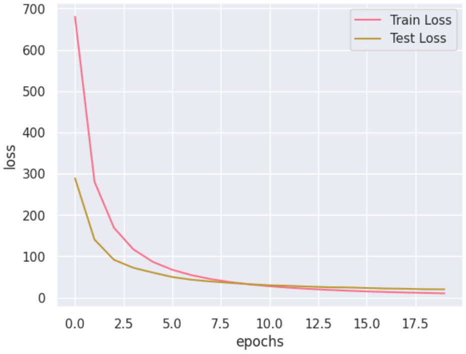

Figure 5: Train and Test Loss

The score suggests the fraction of n-grams in the predicted sentence that appear in the ground-truth caption, with 100 being very similar. BLEU score is chosen for this project because it is useful to compare the predicted sentence against multiple references, which is applicable to the Flickr 8k dataset, where one image corresponds to five captions.

BLEU Score	**Interpretation**
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High-quality translations
50 - 60	Very high-quality, adequate, and fluent translations
> 60	Quality is often better than human

Table 1: Interpretation of BLEU Score[4]

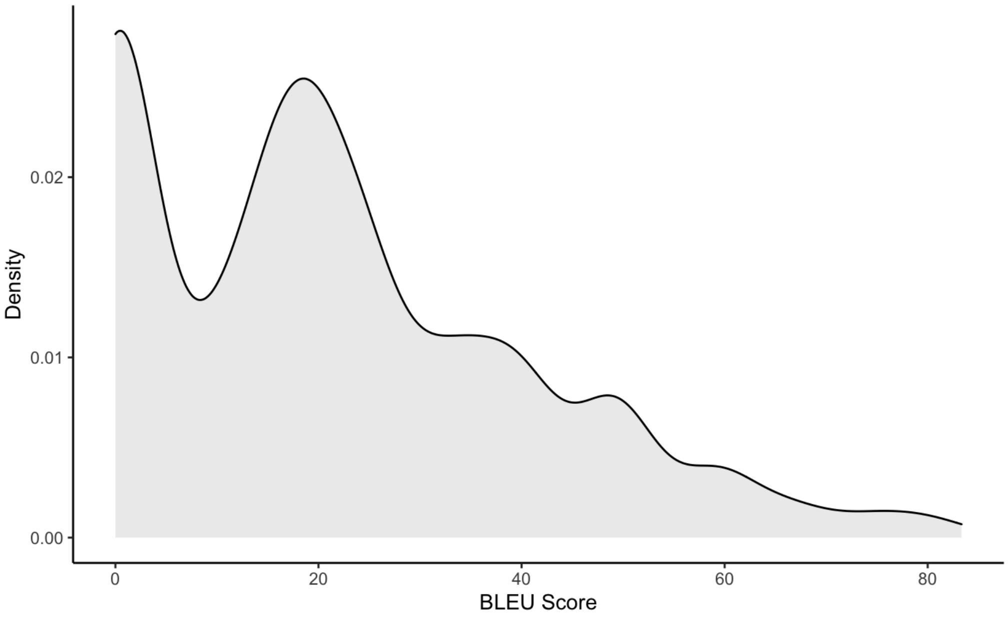

Figure 6: Distribution of BLEU Scores on Test Set

Figure 6 illustrates the distribution of BLEU scores for images in the test set. 46.6% have predicted captions with clear gist (BLUE score >= 20). These captions summarize the objects in the images well, with some grammatical errors that do not fundamentally alter the understanding of the image. However, only 29.4% of test images have good-quality captions with clear gist (BLUE score >=30) that meet the standard of public usage without major errors. Figure 7 shows selected images with high-quality predicted captions, such as “A small brown dog green grass” for the image on the left with a BLEU score of 83.3. These captions capture the main objects of the image, as well as details such as size and color.

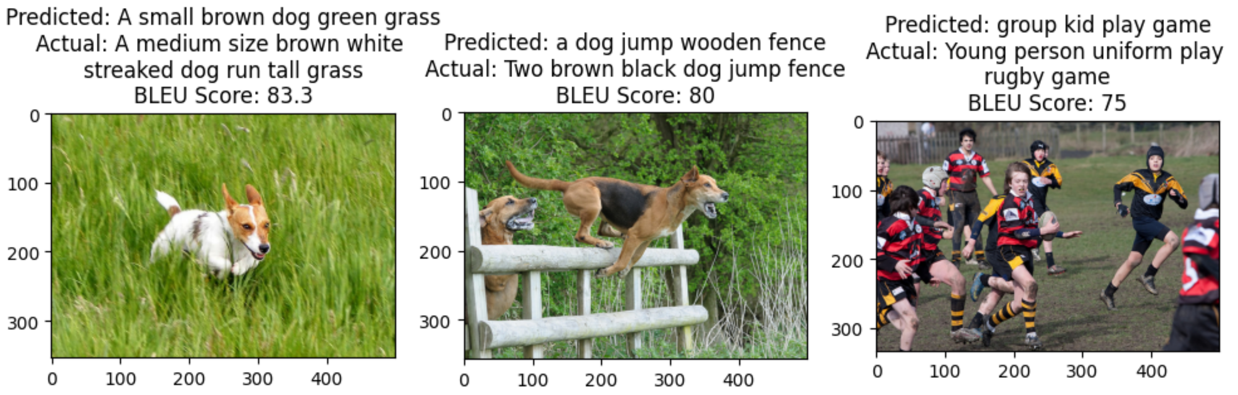
Predicted: A small brown dog green grass
Actual: A medium size brown white streaked dog run tall grass
BLEU Score: 83.3
Predicted: a dog jump wooden fence
Actual: Two brown black dog jump fence
BLEU Score: 80
Predicted: group kid play game
Actual: Young person uniform play rugby game
BLEU Score: 75
0
100
200
300
400

Figure 7: High-Quality Predicted Captions

For captions with lower quality, one major limitation of our model is that it fails to capture the macro elements in the image background. In Figure 8, both images predict "A dog" without any information on their surroundings, namely beach and snow ground. It makes sense as image captioning often focuses on the main objects and their relationship, rather than the surroundings.

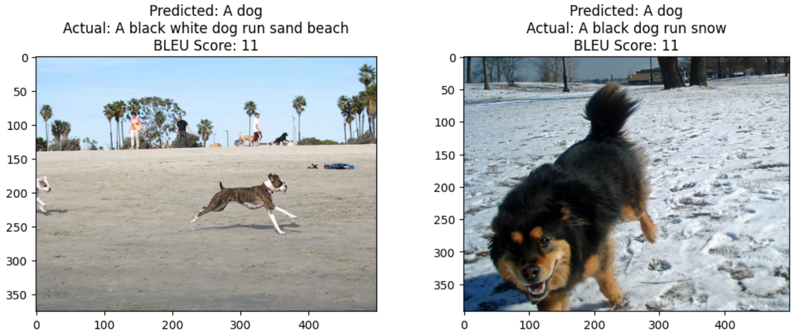

Figure 8: Model Ignorance of Surroundings

Our model is also limited in predicting the number of objects. In the two images in Figure 9, our model predicts the caption "A dog" and "a boy white shirt play game," with a BLEU score of 11 and 33.3 respectively, but they fail to identify there are two dogs and two boys in the images. The model also fails to capture the macro elements in the image background.

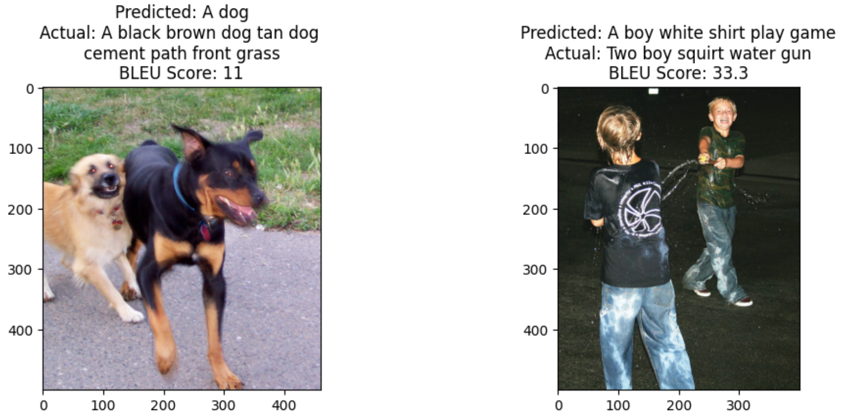
Predicted: A dog
Actual: A black brown dog tan dog cement path front grass
BLEU Score: 11
Predicted: A boy white shirt play game
Actual: Two boy squirt water gun
BLEU Score: 33.3
0
100
200
300
400

Figure 9: Model Ignorance of Quantity

The predicted caption is further limited by the quality of the image provided and whether the main objects stand out from its background. The swing in the left image in Figure 10 blends in with the tree shadows, and our model fails to identify the swing and the related motions and predicts the action as a “jump.” The girl in the right image is also covered by the mist, and our model only captures the facial expression “smile”.

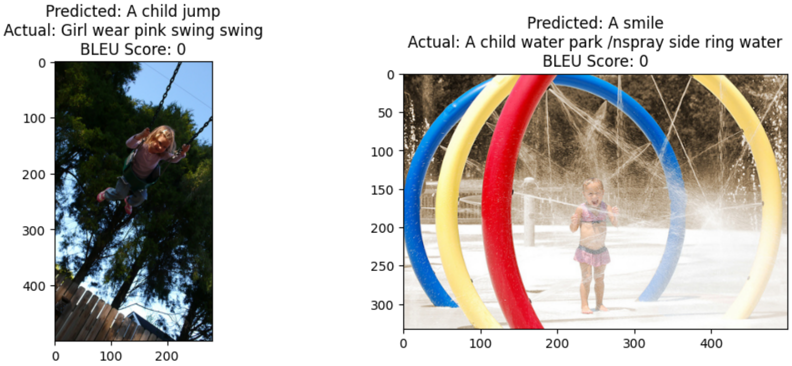

Figure 10: Model Performance and Image Quality

Image Captioning under Adversarial Attack

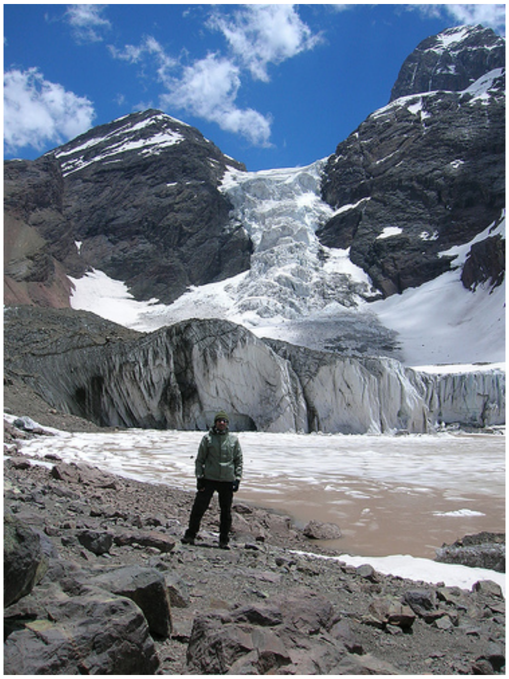

Figure 11: Image used for Adversarial Attack

Original Prediction: “A man stand front rise rock formation background”, BLEU Score = 62.5

Adversarial Caption used for Attack: “A man street racer armor examine tire another racer motorbike”, BLEU Score = 10

Prediction after Adversarial Attack: “A small blue shirt hold large blue shirt hold large”, BLEU Score = 4.34

We tried to re-generate the original image from the modified features. However, that is not possible since we are using the pre-trained VGG-16 model and hence, cannot use gradient descent to regenerate the image.
For comparison purposes to see how close are the adversarial features to the original, we **computed the average feature distance = 9.24.** This means that the original image could be relatively close to the adversarial one.

Lessons Learned

Image Captioning

Image captioning is a challenging task involving translating between different mediums, from images to texts. Our project addresses this challenge by utilizing an encoder-decoder architecture with attention mechanism. The CNN-based encoder converts image input into encoded features that capture important information from the image, and the decoder, based on a GRU language model, takes the input and generates one text element at a time.

In a typical image-to-text task, the input sequence from the image is very lengthy, and the fixed-length context sequence that feeds into the decoder will not be able to capture all the information. The decoder structure also has limited retention of this information during later time steps. To address this problem, we incorporated an attention mechanism that considers all the hidden states of the encoder when deriving the context vector and reweigh them during each time step. We then use beam search to evaluate the model output based on the probabilities of the generated sequence.

We learned how to implement the sequential encoder-decoder architecture from scratch. Additionally, we studied the various types of global attention mechanisms (Bahdanau and Luong) and implemented them on our encoder features and decoder hidden states which was quite a challenging task. Furthermore, to generate captions, we studied how to improve upon the traditional greedy approach of caption generation which generates the most probable word at each time step using Beam Search algorithm.

However, in spite of the attention mechanism, our approach suffers from a major challenge for the model to consider the surroundings of the main objects. Another limitation that comes with the emphasis on objects is its failure to identify the number of objects. Furthermore, the quality of the input image can impact the model's performance, where the main objects blur into the background. To reduce the impact of noise introduced by the quality of the image, the model can be trained on a larger dataset such as MS-COCO or Flickr30 for better generalization and to improve the quality of generated captions. We can also use data augmentation to address this.

Given that only half of the generated captions are informative, the model is useful for generating titles for e-commerce products at a large scale, such as for medium to large businesses on Amazon. However, the grammatical errors in tense, omission of capitalization and punctuation, and the lack of personalization such as humor narrow down its usage in a social setting. Yoshida et al. proposed a potential solution to incorporate humor in the image captioning model[5]. The study uses the BoketeDB database that collects funny captions users posted on the Bokete website and incorporates a funny score into its loss function to maximize the humorous elements in the generated texts. This pre-trained model can adapt the image captioning model for more use cases.

Adversarial Attacks

Adversarial attacks are a significant challenge for computer vision systems, as they can significantly reduce the reliability and security of these systems. Our experiment highlights the vulnerability present in any machine learning model and despite recent progress, making a model robust to such attacks is a significant challenge. These attacks can be exploited by malicious actors and are a serious threat to the security of any system.

We were able to implement a basic adversarial attack by reverse engineering gradient descent on the image features, keeping the model weights fixed with respect to incorrect captions. This generated features that were still relatively close to the original features but drastically reduced the BLEU score. However, this technique requires more rigorous testing on various training examples and incorrect captions. Additionally, we were unable to generate images from features since we were not training our pre-trained VGG-16 model.

Augmenting the dataset with adversarial examples, smoothing, or using model ensembles are some of the techniques used to defend against such attacks.

References

1. “Framing image description as a ranking task: data, models and evaluation metrics” by Micah Hodosh, Peter Young, and Julia Hockenmaier (https://paperswithcode.com/dataset/flickr-8k)
2. “Evaluate a model - AutoML Translate” by Google Cloud (https://cloud.google.com/translate/automl/docs/evaluate)
3. "Neural Machine Translation by Jointly Learning to Align and Translate" by Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio (https://arxiv.org/abs/1409.0473)
4. "Effective Approaches to Attention-based Neural Machine Translation" by Minh-Thang Luong, Hieu Pham, Christopher D. Manning (https://arxiv.org/abs/1508.04025)
5. “Neural Joking Machine: Humorous image captioning” by Kota Yoshida, Munetaka Minoguchi, Kenichiro Wani, Akio Nakamura, and Hirokatsu Kataoka (https://www.researchgate.net/publication/325463963_Neural_Joking_Machine_Humorous_image_captioning)

Link to Colab

Image Captioning with Attention Mechanism and Adversarial Attacks