Comparative evaluation of CLIP-based models on Image Captioning Generation

Yuhan Wei   Xiaoyu Chen
Department of Computer Science, Rice University
{yw157, xc55}@rice.edu

Abstract

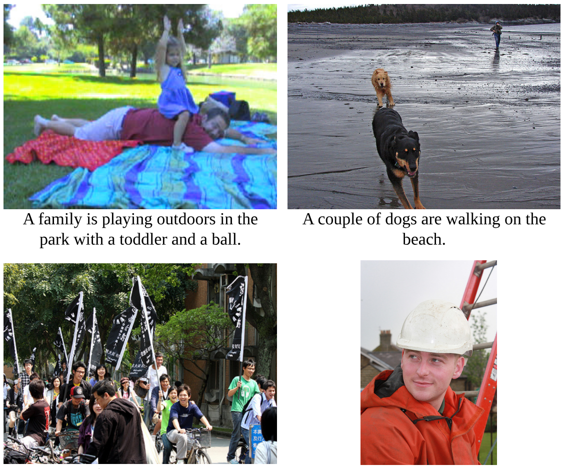
Image captioning is a challenging task that involves training deep learning models to understand the visual content of an image and generate a corresponding natural language description. One popular method is to use a CNN for image feature extraction and an RNN for sequence generation. In this paper, two new CLIP-based approaches for image captioning are proposed. The first approach is zero-shot prediction, which uses CLIP to detect objects in an image and then converts these objects into a caption using FLAN-T5. The second approach fine-tunes the pre-trained language model GPT-2 with the image feature, extracted from the CLIP image encoder, to generate a sequence of words. The results show that the CLIP-based methods perform well, with the CLIP+FLAN-T5 model producing comparable captions to the traditional CNN+LSTM model under no training step, and the CLIP+GPT-2 model generates better results with a faster training process.

1. Introduction

In simpler terms, human-readable textual descriptions for images are important for search engines and object recognition systems to understand and classify images. To achieve this, artificial intelligence combines computer vision with natural language processing. Most current methods use a model called CNN-LSTM.

CNN stands for Convolutional Neural Networks, which are designed to identify patterns and extract features from images. These features are then used to create vector representations of the images. LSTM, on the other hand, stands for Long-Short-Term Memory, which is a type of recurrent neural network that predicts words sequentially based on the image representations generated by the CNN [6]. Then the similarity between generated captions with the referenced captions is worked as the measurement.

Our work aims to address the limitations of the traditional image captioning model that can only capture dominant objects in the image, leading to missing fine details and reduced accuracy, as well as requiring long training times. To overcome these challenges, we propose using OpenAI's pre-trained CLIP model to extract objects or image features and feed them into FLAN-T5 (zero-shot prediction) and GPT-2 model (transfer learning) to generate image captions. Our main contributions are as follows:



A family is playing outdoors in the park with a toddler and a ball.

A couple of dogs are walking on the beach.

A group of people are riding their bikes towards the crowd at an event.

A construction worker with a helmet and vest is working outdoors.

Figure 1. Our image captioning models generate texts describing the input images. Here we show some samples from Flickr8K Dataset captioned by our zero-shot prediction model (CLIP + FLAN-T5).

- No training time is required for the zero-shot prediction approach that utilizes CLIP and FLAN-T5 for object detection and caption generation step.
- A comparable performance but faster model training process, when fine-tuning the GPT-2 model with the CLIP-extracted image features.

2. Related Work

CLIP is a powerful approach developed by Radford et al [10]. that can understand visual concepts from natural language descriptions. It uses a multi-modal transformer architecture that is pre-trained on a large dataset of 400 million (image, text) pairs to associate textual and visual information. This joint training approach makes CLIP a powerful tool for a wide range of applications, such as image captioning. In this paper, we propose two approaches for generating image captions. The first approach uses FLAN-T5 [1], a language model developed by Google that leverages pre-existing knowledge to generalize to new tasks using few-shot learning. This approach utilizes CLIP to detect objects in the input image and FLAN-T5 to generate a caption based on these objects. The second approach involves using CLIP to extract image features and feeding them into the GPT-2 language model [11], which is trained on a large corpus of text and can be fine-tuned to generate natural language text.

A common practice for image captioning models is to convert an image into a set of vectors using CNNs, which capture the visual features and information present in the image. The generated vectors are then passed into an RNN, which generates a sequence of words that form the image description. It utilizes the probability distribution function for the next word generation. Jia et al [9] incorporates Long Short-Term Memory(LSTM) to construct long sentences. Donahue et al [2] optimizes pre-trained CNN models as an image encoder and LSTM as an image decoder, an end-to-end model. The word embedding and extracted image feature vectors are merged and then fed to the decoder. The attention mechanism is involved in stemming distinctive features from tedious texts. Sulabh et al [4] integrate the Attention mechanism with dozens of CNN-LSTM models and conclude that without the attention mechanism, the similarity metrics have better performance.

3. Model

3.1. CNN + LSTM Model

The approach for the common image caption model is to extract image vectors via CNN(DenseNet201), which could be set as an initial hidden state for the LSTM decoder. Then output the last two layers from the pre-trained models as the feature extractor. Through data cleaning and word embedding, textual vectors concatenate with image vectors feed LSTM. The LSTM successively updates hidden states and units.

3.2. Zero-shot Prediction: CLIP + FLAN-T5 Model

Our approach in the zero-shot prediction phase involves a combination of object detection and caption generation using CLIP and FLAN-T5, respectively. To identify objects in an image, we extract nouns and objectives from the ground truth captions in the training dataset and use them to create a prompt that describes the image. The prompt and the image are then inputted into the CLIP model, which generates multiple text descriptions of the image. We assume that the objects mentioned in the top selected texts are present in the image, which we refer to as object detection.

Using the top selected objects, we generate another prompt for FLAN-T5 and use beam search to produce a set of five image captions. We then reuse the CLIP model and the original image to select the best caption from the set based on the highest CLIP score. This zero-shot method, which is illustrated in Figure 2, eliminates the need for additional training data and enables the generation of high-quality image captions. Our approach has shown promising results in accurately describing the visual content of images.

3.3. Fine-tuning: CLIP + GPT2 Model

Our approach to transfer learning consists of two steps. Firstly, we extract visual information from input images and use an adapter layer to transform the image vectors to the same dimension as the word embedding vectors, acquired from the GPT2 Tokenizer. Secondly, we feed the image vectors and word embeddings together into the GPT-2 language model and generate image captions in an auto-regressive manner. The cross-entropy loss is used to train the adapter layer and fine-tune the GPT-2 model. Our approach (shown in Figure 3) has a similar structure to the traditional CNN-LSTM model, as described previously. The key difference is that we use the image encoder from CLIP to extract visual information, and the GPT-2 model to generate the word sequence. This allows us to take advantage of the pre-trained image encoder from CLIP and achieve better performance in generating image captions.

4. Experiments and Results

Dataset and Platform. We conduct our experiments using the deep learning library *PyTorch* and *TensorFlow* on Google Colab to accelerate the process. The Flickr8K [1] dataset is used for the experiments, which consists of 8,000 images collected from the Flickr website, and each image is paired with five different human-generated captions, resulting in a total of 40,000 captions. The images in the dataset represent a wide range of scenes, objects, and people, making it ideal for evaluating the performance of image captioning models.

We have split the entire dataset into training and test data, with 85% of the data being allocated for training and 15% for testing. Within the training data, we further split into a training set, consisting of 70% of the data, and a validation set, consisting of 15% of the data.

Baseline and Model Configurations. In our experiments, we used the CNN-LSTM model as our baseline model. This model was trained using a batch size of 64 and a learning

[1] https://www.kaggle.com/datasets/adityajn105/flickr8k

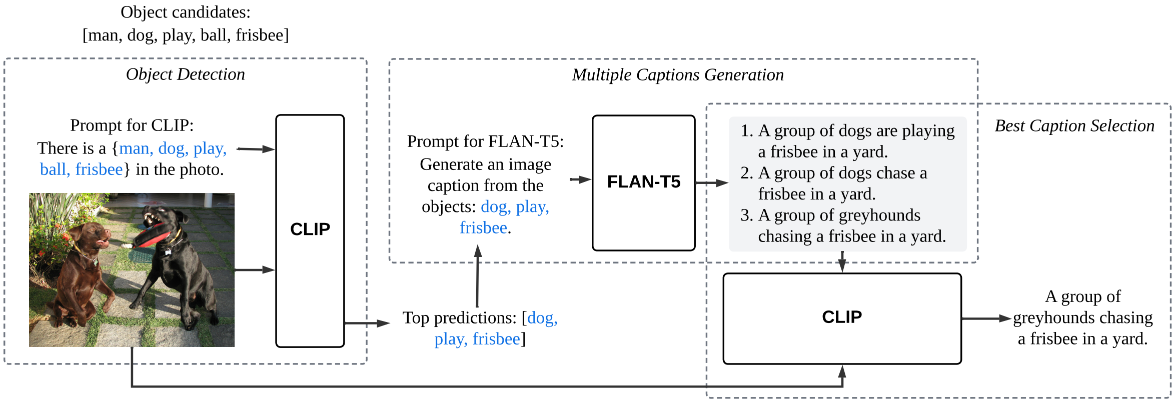

Object candidates:
[man, dog, play, ball, frisbee]
Object Detection
Prompt for CLIP:
There is a {man, dog, play, ball, frisbee} in the photo.
CLIP
Multiple Captions Generation
Prompt for FLAN-T5:
Generate an image caption from the objects: dog, play, frisbee.
FLAN-T5
Top predictions: [dog, play, frisbee]
Best Caption Selection
1. A group of dogs are playing a frisbee in a yard.
2. A group of dogs chase a frisbee in a yard.
3. A group of greyhounds chasing a frisbee in a yard.
CLIP
A group of greyhounds chasing a frisbee in a yard.

Figure 2. Here we show a simplified example of our zero-shot CLIP and FLAN-T5 model. We begin by constructing five input prompts for the CLIP model, which detects ”dog”, ”frisbee” and ”grass” in the input image. Next, we create another prompt for FLAN-T5 to generate several image captions and then reuse CLIP to select the best one.

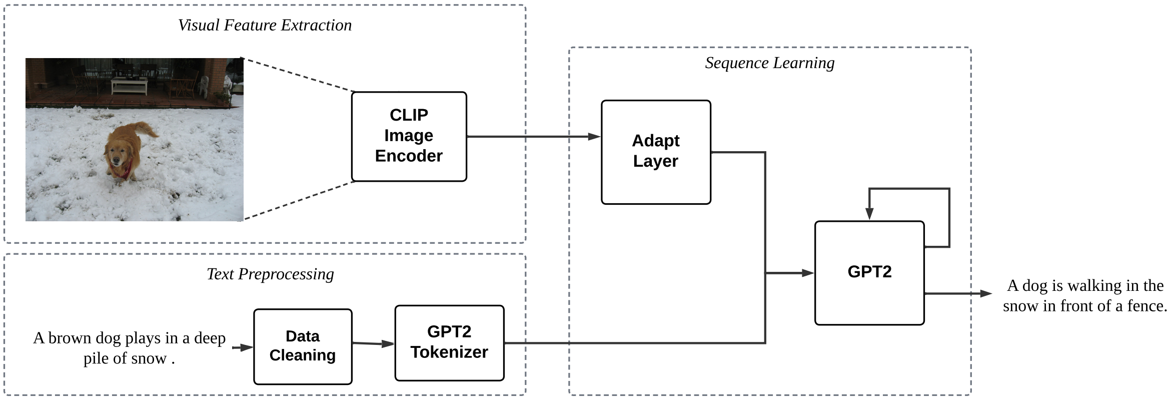

Visual Feature Extraction
CLIP Image Encoder
Sequence Learning
Adapt Layer
GPT2
A dog is walking in the snow in front of a fence.
Text Preprocessing
A brown dog plays in a deep pile of snow .
Data Cleaning
GPT2 Tokenizer

Figure 3. Here is the architecture of the CLIP+GPT-2 model. The CLIP-extracted image features are first passed through an adapter, which is a multilayer perceptron that transforms the features into a format suitable for the GPT-2 model. Next, the text embedding and transformed image features are concatenated and fed into the GPT-2 model to generate image captions.

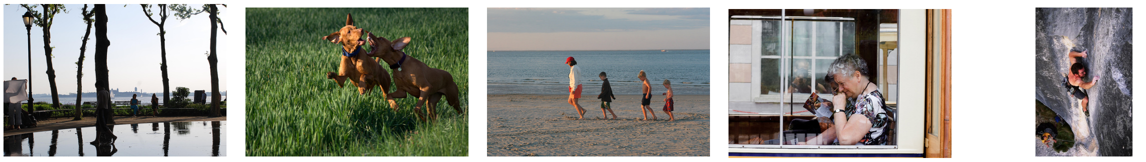
rate of $4e^{-5}$. We implemented early stopping, which led to the model stopping training at epoch 10.

Our CLIP+FLAN-T5 model does not require a training process, as we extract a set of object candidates from the ground truth captions in the training dataset and use them for zero-shot prediction on the test dataset. Due to limited computing power, we restrict the candidate objects to the top 200 frequently occurring objects for our CLIP+FLAN-T5 model in the ground truth captions, and we use no more than 20 different objects detected by the CLIP model from each image, to generate the image caption. On the other hand, our CLIP+GPT-2 model (which is modified from a publicly available implementation [7]) involves training the model using a multilayer perceptron with 2 hidden layers as the adapter layer. The adapter and GPT-2 are trained using a batch size of 64, a learning rate of e^{-5}, and trained for a total of 5 epochs.

For our CLIP-based models, we utilized the CLIP-ViT-B/32 model. we used FLAN-T5-Small and FLAN-T5-XL models in the CLIP+FLAN-T5 model to generate the image captions. In our CLIP+GPT-2 model, we compared the performance of the GPT-2 model with greedy search and beam search in generating image captions.

Evaluation metrics. The quantitative measures in our experiments include BLEU [8], ROUGE [5], CIDER, CLIPScore, and RefCLIPScore. The CLIPScore is a free-reference image text metric that could be applied without collections of captions. The RefCLIPScore is tightly focused on image text and is a tool for complementary tool for reference-based metrics. This method is to measure the cosine similarities between texts.

We also report our training times for each model. Shorter training times can enable faster model creation and facilitate the development of model ensembles. The number of trainable parameters is also considered to assess the feasibility and scalability of a model. Intermediate sized-models tend to be faster to train and may generalize better to new data [3].



Ground Truth	Person walking across a large puddle of water in a park .	Two dogs are playing in the grass.	A woman walks along the beach as three children follow her in a line.	A woman reads a book in front of a window.	One man is climbing a rock wall.
CNN+LSTM	Man is standing on the edge of the water.	Brown dog is running through the grass.	Two people are standing on beach.	Man is sitting on the floor with his arms outstretched.	Man in blue jacket is climbing up rock.
CLIP+FLAN-T5	Person walking towards the ocean in the park.	Couple of dogs playing catch in the grass.	Group of people walking on the beach with a toddler and a boy playing in the sand.	A woman is sitting in a window with a book.	A guy is working on a rope while a person is jumping off a rock.
CLIP+GPT-2	A couple of people sitting on top of a wooden bench.	Two dogs running in a grassy field.	A group of people walking along a beach.	A woman reading a book while sitting on a bench.	A man on a rock climbing up a cliff.

Figure 4. We demonstrate the comparison between ground truth captions, CNN+LSTM model-generated captions, and our CLIP-based model-generated captions. Overall, our CLIP+FLAN-T5 model can generate more generalized image descriptions including the key components ”Who, What, Where”, and our CLIP+GPT-2 model is capable to generate high-quality captions compare to the ground truth.

Method	BLEU-2	BLEU-4	ROUGE	CIDER	RefCLIP Score	CLIP Score	Training time
DenseNet201+LSTM	0.2906	**0.0904**	**0.3231**	0.3147	0.6474	0.6933	30 min
CLIP+FLAN-T5-Small	**0.0370**	0.0054	0.1035	0.0508	0.6365	0.6670	0
CLIP+FLAN-T5-XL	0.1935	0.0447	**0.2698**	0.1144	0.6526	0.6805	0
CLIP+GPT-2 (Greedy)	0.2860	**0.0991**	0.3213	0.2332	0.6783	0.7076	8 min
CLIP+GPT-2 (Beam)	0.2779	**0.0973**	0.3180	0.2339	**0.6812**	**0.7094**	8 min

Table 1. Here are quantitative measurements on the Flickr 8K dataset. According to the results, our CLIP+FLAN-T5 model is able to generate slightly worse captions compared to the baseline without training. Our CLIP+GPT-2 model achieves generate similar results as the baseline but with less training time.

Results. In Figure 4, we present examples of image captions generated by both our baseline models and the CLIP-based models. Overall, our CLIP+FLAN-T5 model can generate more generalized image descriptions including the key components ”Who, What, Where”, and our CLIP+GPT-2 model is capable to generate high-quality image captions which are much similar to the ground truth label.

Based on the results presented in Table 1, it is evident that the FLAN-T5-Small model for zero-shot prediction performs poorly compared to the other models, with a low BLEU-2 score indicating that the generated captions do not match well with the reference captions even for two-word sequences. However, the FLAN-T5-XL model shows a relatively better performance compared to the FLAN-T5-Small model, with a ROUGE score of 0.27. Although this score is slightly worse compared to the baseline model, which has a ROUGE score of 0.32, it still indicates that the captions generated by the FLAN-T5-XL model are similar to the reference captions and capture the same content without a training process.

The CLIP+GPT-2 model with both greedy search and beam search performed similarly on our dataset. Both approaches generated captions with comparable results across all metrics, with the added benefit of requiring only one-third of the training time compared to the baseline model. In addition, the CLIP+GPT-2 models achieved better BLEU-4 scores of 0.099 (greedy search) and 0.097 (beam search) compared to the baseline model with DenseNet201+LSTM, which had a BLEU-4 score of 0.09. This indicates that the CLIP+GPT-2 models are capable of generating captions that better match the reference captions in terms of four-word sequences.

Since our proposed models use CLIP in the image caption generation process, it is expected that all models achieve high CLIPScore and RefCLIPScore. According to the results presented, the captions generated by zero-shot prediction from the FLAN-T5-XL model achieve similar CLIPScore (0.68) and RefCLIPScore (0.65) compared to the baseline model, which has CLIPScore of 0.69 and RefCLIPScore of 0.65 respectively. The fine-tuned GPT-2 model achieves slightly better CLIPScore (0.71) and RefCLIPScore (0.68), indicating that it is able to generate captions that more closely match the visual content of the input images, and are more comparable in quality to human-created captions.

The results presented demonstrate the effectiveness of our proposed CLIP-based models for generating accurate and descriptive captions for images. In particular, the CLIP+FLAN-T5 model is able to generate captions that are comparatively good without requiring a training process. Overall, these results highlight the potential of CLIP-based models for image captioning tasks.

5. Code & Video

This is the link to our codes and video: `https://drive.google.com/drive/folders/1vwaZmjr6neC6jiljkF9rA48VbdROaNc8`.

References

[1] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[2] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[3] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

[4] S. Katiyar and S. K. Borgohain. Comparative evaluation of cnn architectures for image caption generation. *arXiv preprint arXiv:2102.11506*, 2021.

[5] C.-Y. Lin and F. Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*, 2004.

[6] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[7] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[9] Y. Pathak, K. Arya, and S. Tiwari. Feature selection for image steganalysis using levy flight-based grey wolf optimization. volume 78, pages 1473–1494. Springer, 2019.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.