Image Caption Generation With CLIP+GPT-2 Model

Siyuan Jing and Haonan Wu
Boston University
siyuan16@bu.edu and whn17@bu.edu

May 3, 2025

Abstract

This paper presents an image captioning system that combines CLIP with GPT-2. We implement a multi-head attention mechanism to effectively bridge between visual and textual modalities. Trained on the MS COCO dataset, our model achieves 94.4% of human-level consensus quality as measured by CIDEr scores. Our experiments show that the proposed architecture effectively captures both semantic and syntactic structure in generating natural language descriptions. Additionally, we develop a web-based deployment for real-time captioning The results highlight the potential of leveraging pre-trained vision and language models for efficient multimodal understanding with minimal architectural modifications. Codes are avalibale at `https://github.com/ChingSsuyuan/Clip_Pro_Caption.git`

1 Introduction

Image caption generation, the task of recognizing images to generate natural language descriptions, lies at the critical intersection of computer vision and natural language processing. It plays an important role in a variety of applications, including image retrieval, accessibility for the visually impaired, and automated image content processing.

Recent advances in vision-language models have enabled more accurate and fluent caption generation by leveraging large-scale pretraining on aligned image-text pairs . These models demonstrate remarkable capabilities in understanding visual content and translating it into coherent text descriptions.

Motivated by this progress, we focus our project on replicating a CLIP+GPT-2-based image captioning system, with inspriation from Nukrai et al.[1]. Through this project, we seek to explore and deepen our understanding of machine learning, as well as its broader applications in fields such as computer vision and natural language generation.

Our goal is to:

- Leverage the pretrained CLIP model to extract semantically rich image embeddings without the need for training a custom vision encoder.

- Utilize the generative capabilities of GPT-2 to produce fluent and coherent natural language captions.

- Bridge the gap between visual and textual modalities by introducing a projection layer that maps image embeddings into GPT-2's input space.

- Enable flexible and data-efficient image captioning, where the visual semantics guide the generation through prefix-based conditioning.

- Evaluate the quality of the generated captions using standard metrics such as BLEU and CIDEr, in order to quantitatively assess the model's accuracy and relevance.

2   Background and Related Work

2.1   Contrastive Language-Image Pretraining(CLIP)

CLIP from OpenAI is a visual-language model. Instead of relying on task-specific supervised learning, CLIP is trained on a dataset of 400 million image-text pairs collected from the internet using a contrastive loss function. CLIP consists of two separate encoders: a visual encoder (ResNet) for images, and a text encoder (Transformer) for captions. Its ability to generate rich, semantically meaningful image embeddings makes CLIP a powerful foundation for our systems and an ideal visual component in our CLIP+GPT-2 image captioning pipeline. For example, according to Mokady et al.[2], it mentioned that the visual encoding capability of CLIP can be used to embed and project the generated images into the input space of GPT-2 to generate prefixes, which helps the final caption generation of GPT-2. Inspired by this article, we decided to study the CLIP architecture and implement related deployments.

2.2   Generative Pre-trained Transformer 2 (GPT-2)

GPT-2 is a large-scale language model based on the Transformer decoder architecture proposed by Radford et al.[3]. According to the paper, Transformer completely replaces the traditional RNN or CNN structure with self-attention, which is more efficient and accurate when processing long sequence dependencies. Therefore, we consider implementing the transformer structure as our decoder of the whole pipeline. Unlike the traditional Transformer, which contains both encoder and decoder components, GPT-2 uses only a decoder. This design enables the model to predict the next token based solely on previously generated tokens, making the generated text semantically relevant and well suited for text generation tasks such as image captioning.

2.3   Multi-Head Attention Mechanism

Inspired by the *"Attention Is All You Need"* paper [4], we employ a multi-head attention mechanism as a cross-modal connector between the CLIP image embedding and the GPT-2 language model. Instead of using a linear layer to directly project the CLIP features into the GPT-2 embedding space, we employ an attention mechanism that enables the model to selectively focus on different aspects of the visual features when generating each token in the caption, making the generated tokens more consistent with the image information. Multi-head Attention network allows the model to learn how different regions in the image affect language production.

We compare these advanced approaches of image caption generation to understand its advantages and limitations, and implement these methods in the process of model training to get our own model.

3 Methodology

3.1 Implementation Details

Detail of Implementation and pipeline: This is the pipeline of our entire training process, which mainly includes multiple steps: image input, feature extraction and encoding, feature processing, feature decoding, model training and error evaluation. We will also introduce the optimization we used.
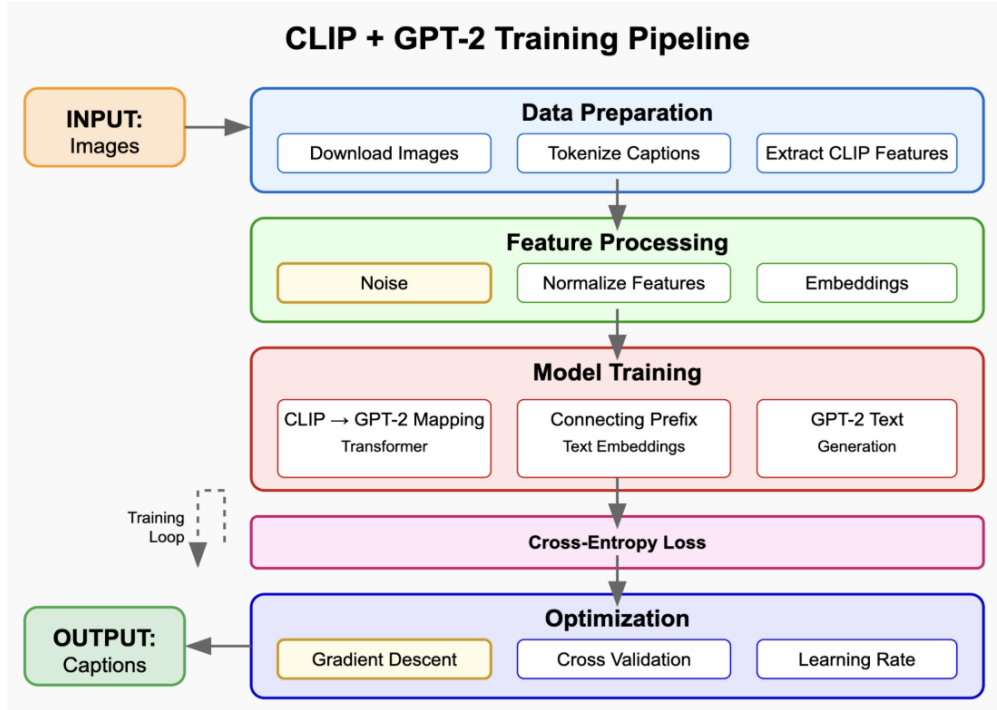

Figure 1: CLIP+GPT-2 Training Pipeline

3.1.1 DataSet

We utilize the MS COCO dataset [5]for training and evaluation, MS COCO (Microsoft Common Objects in Context) is a large-scale image recognition, segmentation, and annotation dataset that is widely used in research in the field of computer vision, especially in tasks such as image captioning, object detection, and semantic segmentation. It contains 330K images with 5 captions each. We will split our dataset into three parts: 80% training, 10% validation, 10% testing.

3.1.2 Data Preparation

After we get the original CLIP embedding, we will normalize it for better training performance. In the actual training process, we found that the trained model may be overly sensitive to some features of some pictures, so we add some noise to 10% of the feature information of each picture to improve generalization ability and reduce overfitting. These embeddings can then be projected into the GPT-2 space for training.

3.1.3 Feature Process

After we get the original CLIP embedding, we will normalize it for better training performance. In the actual training process, we found that the trained model may be overly sensitive to some features of some pictures, so we add some noise to 10% of the feature information of each picture to improve generalization ability and reduce overfitting. These embeddings can then be projected into the GPT-2 space for training.

3.1.4 Model training

Next, we use a multi-layer Transformer to map the CLIP embedding to the word embedding space of GPT-2 to create a prefix. This prefix is then concatenated with the tokenized caption, which is generated in the data preparation stage, to form the whole input sequence for GPT-2. After receiving these input sequences, the GPT-2 model is trained for a set rounds of training loop to generate the caption of the corresponding image based on the given prefix and token.

3.1.5 Optimization method

For the weight matrix obtained after each round of training, we use sub-gradient descent to adjust the weights. Additionally, during the entire training process, we use cross-validation to find the optimal solution for multiple hyperparameters in the model, and reasonably control the learning rate to obtain better training results.

3.2 Loss and objective function for training process

We use the cross-entropy loss function to measure the difference between the caption generated by the model and the ground-truth captions. The loss function is defined as:

The Cross-entropy Loss Function:

$$\text{Loss} = -\sum_{t=1}^{T} \log P_{\text{GPT-2}}(w_t \mid w_{<t}, \text{prefix}) \tag{1}$$

Where:

- T — the total number of tokens (words) in the caption

- w_t — the actual word at position t in the ground-truth caption

- $w_{<t}$ — the sequence of all previous words before time step t

- prefix — the mapped image feature vector used to condition the generation

- $P_{\text{GPT-2}}(w_t \mid w_{<t}, \text{prefix})$ — the probability assigned by GPT-2 to word w_t given previous words and the image prefix

This loss encourages the model to assign a higher probability to the correct next word at each time step. Minimizing the cross-entropy loss over the training data helps the model generate captions that are more fluent and accurate.

The Objective Function in Maximum Likelihood Estimation:

$$\hat{\theta} = \arg\min_{\theta} - \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P_\theta(w_t^{(i)} \mid w_{<t}^{(i)}, \text{prefix}^{(i)}) \tag{2}$$

Where:

- θ — model parameters, including GPT-2 weights and the parameters of the prefix mapping network

- N — the total number of training samples (image-caption pairs)

- T_i — the number of tokens in the i-th caption

- $w_t^{(i)}$ — the t-th word in the i-th ground-truth caption

- $w_{<t}^{(i)}$ — all previous words before position t in the i-th caption

- $\text{prefix}^{(i)}$ — the mapped image features for the i-th image

This objective function aims to find the parameter set $\hat{\theta}$ that minimizes the total negative log-likelihood across the training dataset. During training, the model updates θ using optimization techniques such as gradient descent, guided by cross-validation to avoid overfitting and improve generalization.
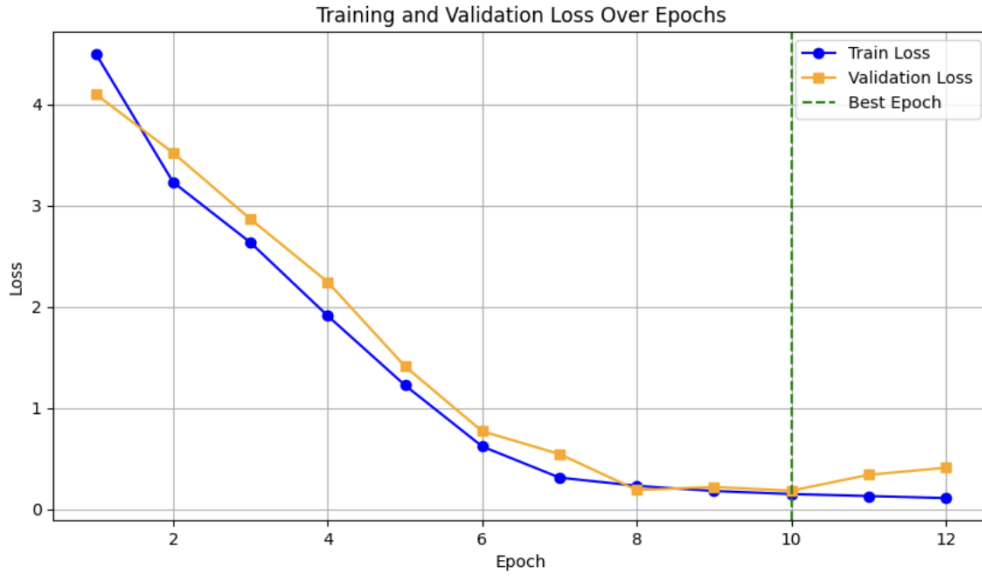

Figure 2: Training and validation loss over epochs

4 Experiments and Results

4.1 Training Details

We train our model using these hyperparameter after cross-validation:

- Prefix length: 40

- Transformer layers: 8

- Number of epochs: 12

- Noise rate: 10%

- Optimizer: Adam with a learning rate of 5×10^{-5}

We conduct experiments on:

- **COCO Image Dataset:**

  Training set: 80, 000 images with 5 captions per image.

  Validation set: 2,000 images.

4.2 Evaluation Metrics

After completing model training, we need to evaluate the quality of generated captions to assess the effectiveness of our approach. Therefore, we adopt two widely used automatic evaluation metrics in the field of image captioning: BLEU and CIDEr.

- **Bilingual Evaluation Understudy**

  BLEU is a baseline evaluation metric of our model that measures the accuracy of n-gram overlap between predicted captions and one or more real captions[6]. It emphasizes syntactic similarity and is particularly sensitive to exact word matching and word order. However, this criterion lacks semantic awareness and often penalizes correct words due to different word position information.

- **Consensus-based Image Description Evaluation** To address the shortcomings of the above evaluation method, we also chose the CIDEr (Consensus-based Image Description Evaluation) [7] evaluation metric, which is designed specifically for image captioning. It uses a TF-IDF weighted n-gram model to evaluate how well the generated captions match the manually generated captions. This design increases the weights of rare but informative words (e.g., nouns, verbs, adjectives, etc., which often contain unique image information) while reducing the weights of function words such as articles (a, the), quantifiers (some, many), and prepositions (in, on, at).

4.3 Generation Results

Prediction Scores:

We compute BLEU and CIDEr on our test set and compare the original captions' scores with our outputs:
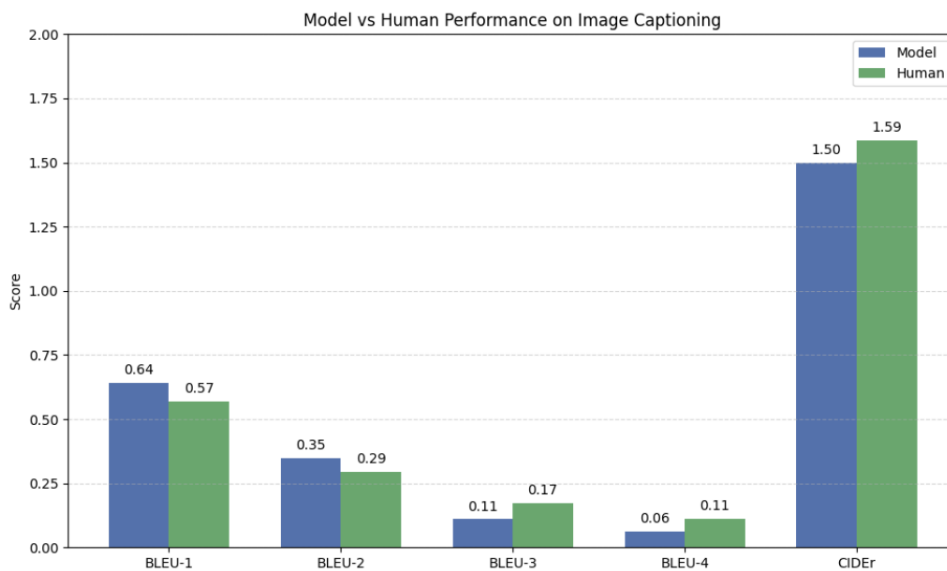

Figure 3: Model vs Human Performance on Image Captioning

- Human captions show better performance on longer word sequences, indicating more natural grammatical structure.

- The CIDEr score shows our model can achieve 94.4% of human-level consensus quality.

Web-based Deployment for Real-time Captioning

We design and deploy a website to realize our model function. The website is able to upload images and load the pre-trained model to generate captions within 10 seconds.
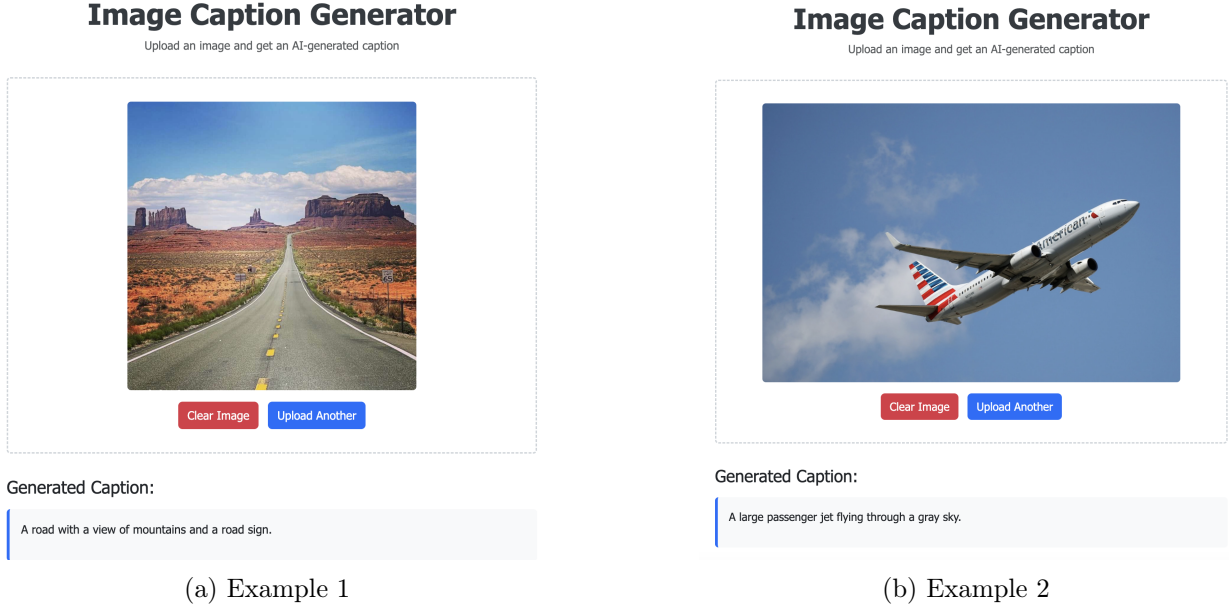

(a) Example 1

(b) Example 2

Figure 4: Web demo of our deployed model: From image upload to automatic caption generation

5 Discussion

5.1 Key Findings

- Automated organization of visual content:

  Our model demonstrates strong automatic organization capabilities for visual content, especially when dealing with a large number of unlabeled images. By generating relevant and informative captions, the system can group images without manual annotation.

- Effective CLIP-to-GPT-2 Mapping:

  The key design of our entire model is the multi-head attention mechanism used to connect CLIP visual embedding and GPT-2 language model input. Unlike simple linear projection, the multi-head attention mechanism enables the model to learn multiple semantic alignment patterns between image features and different dimensions of generated captions. Test results show that the implementation of this mechanism can generate more accurate captions that reflect the details of image semantics.

- Noise Injection Boosts Generalization: We observe that adding small Gaussian noise into CLIP embeddings during training improves the model's generalization across different image inputs. This regularization technique helps the mapping module avoid overfitting to specific embedding patterns and encourages the model to generate captions that are both accurate and diverse. Result shows that noise improves both BLEU and CIDEr scores, indicating improved robustness and semantic coverage of caption generation.

5.2 Future Work

In future work, we plan to expand the test dataset to cover more diverse and challenging images to more robustly evaluate the generalization ability of our model. In addition, we can try to study and combine more powerful visual language models (such as CLIP ViT-L or GPT-3 variants) to achieve stronger semantic alignment capabilities.

6 Conclusion

In this project, we implemented an image captioning system by combining the powerful visual encoding capabilities of CLIP and the generative language capabilities of GPT-2. By introducing a mapping module based on a multi-head attention mechanism, we effectively reduced the gap between visual and textual modalities, enabling the generation of fluent and semantically relevant captions. We further enhanced the generalization ability of the model by adding noise during training, and selected appropriate loss functions and objective functions for training. Finally, evaluations using BLEU and CIDEr metrics confirmed the effectiveness of our model, especially in capturing human-written semantics, with the CIDEr score reflecting its excellent performance.

References

[1] Nukrai, D., Mokady, R., & Globerson, A. (2022). Text-Only Training for Image Captioning using Noise-Injected CLIP. https://doi.org/10.48448/n7sq-p557

[2] Mokady, R., Hertz, A., & Bermano, A. H. (2021). ClipCap: CLIP Prefix for Image Captioning. http://arxiv.org/abs/2111.09734

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). "Language Models are Unsupervised Multitask Learners." OpenAI Technical Report.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention is all you need." Advances in Neural Information Processing Systems, 30.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[6] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation." ACL 2002.

[7] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). "CIDEr: Consensus-based image description evaluation." CVPR 2015.