Final Report 507: BLIP Model Learning

Jiayi Guo
Department of Applied Statistics
University of Michigan
Ann Arbor, United States
jiayiguo@umich.edu

***Abstract*—The project is adapted and based on the provided sample project 2 from hugging face of Fine-tune BLIP model [1].**

I. Introduction

A. Background and Motivation

The objective of this project is to develop a robust image recognition model capable of generating descriptive text based on images of animals. This will involve training the model on a real-world dataset of animal images, with the goal of adapting the performance from a dummy dataset used as a baseline in the sample project. The motivation behind this project comes from both a personal interest in animals and the challenge of bridging the gap between visual data and natural language processing. By focusing on real-world images, this project not only aligns with my personal interests but also provides an opportunity to explore the intersection of image classification and text generation, and the use of BLIP model which is a compelling area of research. The successful implementation of this model inspires in various domains, such as automated content generation.

B. Project Goal

The primary goals of this project are as follows:

To evaluate the adaptability and performance of a model initially trained on a dummy dataset of football players, with a focus on how it can be modified and fine-tuned to accurately recognize and describe images from a significantly different domain of animals.

To assess the quality of the descriptive text generated by the model, specifically evaluating the accuracy, relevance, and coherence of the generated descriptions in relation to the input images.

C. Literature Review

The *Show and Tell* [2] introduced a CNN-RNN architecture for image captioning, where the CNN extracts image features and the RNN generates captions. Recent advancements have significantly improved upon this approach, with the introduction of transformer-based models like *ViLT* and *Oscar*, which capture better cross-modal relationships between images and text. Attention mechanisms, such as spatial attention, have enhanced caption precision, while pre-training large models on diverse datasets (like CLIP or BLIP) and fine-tuning them on specific domains further improve caption quality. These advancements, including the use of multi-model fusion and specialized language models, offer more fluent, accurate, and context-aware captions for domain-specific tasks.

II. Methods

A. Problem formulation

The task is to generate captions for images in a dataset cats and dogs images. This is a vision-to-text task takes input images and output a descriptive caption summarizing the content. The input to the model consists of image data in the form of RGB images, and the output is a sequence of tokens corresponding to a caption that accurately describes the image content. The model learns to associate visual features from the image with textual descriptions, aiming to predict the correct caption given the input image.

B. Methodologies

Image Preprocessing and Transformation

The images are preprocessed before being fed into the model. This involves resizing the images to a fixed size to standardize the input dimensions. This transformation ensures that the images are ready for the BLIP model and prevents overfitting by exposing the model to different variations of the same images.

BLIP Model and Preprocessing

The BLIP model (Bootstrapping Language-Image Pretraining) is used for this vision-to-text task. The processor is used to handle the image-to-text transformation and ensure the input images are appropriately tokenized and prepared for the model's inference process.

Training with Backpropagation

The input images are paired during the training process with their corresponding captions, and the model generates predictions based on the image inputs. The loss function is used to compute the difference between the predicted and actual values. This loss is then backpropagated through the model, updating the model's weights using the existing *AdamW* optimizer. It helps to adjust the model's parameters to minimize the loss, enabling the model to improve its performance over time. The training process is over multiple epochs, and save the learned parameters after each epoch to track progress.

DataLoader and Batch Processing

The DataLoader is used to efficiently load and process the images in batches instead of feeding the model one image at a

time. It enables the model to process more data in parallel to speed up the training process and stabilizing gradient updates. The collate_fn function ensures that the images and captions are organized into batches, making it easier to process multiple samples simultaneously.

Final Saving

During the training process, model checkpoints are at the end of each epoch to better see the output and continue process. After training is complete, the final model is saved for later use.

III. Results

A. Data Pipeline and Model Set Up

Since the limitation of computer capability and the time consume for BLIP model, decreased the sample size and training loop.

The images are processed to generate pixel values, and their corresponding captions are tokenized by the model. During training, the BLIP model is fine-tuned using these inputs, optimizing with the *AdamW* to generate captions that accurately describe the content of each image.

B. Figures present numerical simulation results

The figures for epochs 1, 2, and 3 are images along with the captions generated by the model. Each figure includes six images with corresponding generated captions, providing a visual representation of the model's performance at different stages of training.

C. Results Interpretation

The results from the first three epochs indicate that while the model generates captions that are generally relevant to the images, there are inaccuracies and areas for improvement. For instance, of the first picture in the epoch 1, the model misinterpreted the tongue of the dog to a ball, which is reasonable since the dataset used for training is too small. The loss values during training suggest that the model is learning and improving its performance, but further fine-tuning and possibly a larger training set could enhance accuracy. The generated captions reflect a foundational understanding, but more precise and detailed captions will require continued training and optimization.

Also challenges remain for the model, particularly in generating highly accurate captions for complex or less common scenes. The model may still struggle with nuances such as identifying specific breeds or subtle actions within images. Future work would be focus on fine-tuning the model with a larger, more diverse dataset, exploring advanced image understanding techniques, and experimenting with different model architectures to enhance performance. The output images with generated captions illustrate the current effectiveness of the BLIP model in image captioning tasks, showcasing both its strengths and areas for potential improvement.

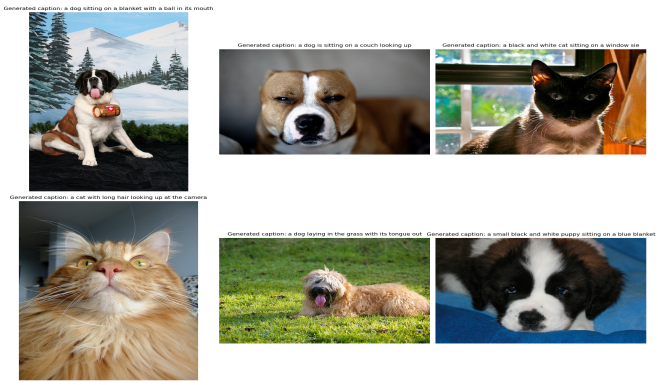

Fig. 1. Example of figure caption. Epoch1. Loss: 13.2992219924926 76.

Fig. 2. Example of figure caption. Epoch2. Loss: 10.378581047058105

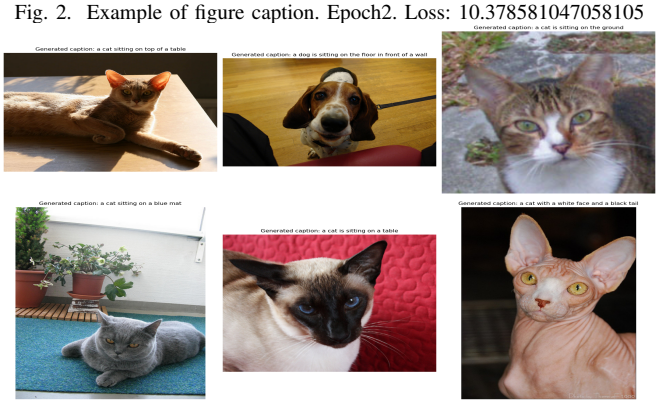

Fig. 3. Example of figure caption. Epoch3. Loss: 10.320682525634766

IV. Conclusion

The project explored the application of the BLIP model for generating image captions from a dataset and the fine-tune of model for animal images. The results demonstrated that the model was capable of generating captions that were generally relevant to the images, though some inaccuracies were observed, particularly in cases of ambiguous or complex scenes. The loss values from the training process indicated that the model was learning at a relatively slow pace due to the smaller dataset used. These results highlight the potential of BLIP for image captioning tasks, but also underscore the challenges that arise when working with limited data. Moving forward, increasing the training dataset, reduce the

time cost for testing, and further fine-tuning the model will be crucial steps to enhance the quality and accuracy of the generated captions. Future work would also focuse on more diverse datasets, experimenting with different architectures, and optimizing the model for more specialized or detailed captioning tasks.

Overall, the project serves as a foundational exploration into the capabilities of the BLIP model and its application in real-world image-to-text tasks.

REFERENCES

[1] Hugging Face, "Image Captioning with BLIP," *Google Colab notebook*, 2024. [Online]. Available: https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/image_captioning_blip.ipynb.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," *arXiv preprint arXiv:1411.4555*, 2014. [Online]. Available: https://arxiv.org/abs/1411.4555.