Project Proposal

Project Title: Automated Image Caption Generation using Deep Learning

Project Team:

The project team consists of the following members:

- Jaswanth Naidu Grorripati (jgorripa@kent.edu)
- Vikitha Ganta (vganta@kent.edu)
- Namrutha Ganga Pallerla (npallerl@kent.edu)

Problem statement:

In the field of computer vision, producing automated yet precise and context-sensitive captions for images is still a major challenge. Many existing systems primarily use either CNNs or RNNs, which each have their own drawbacks in terms of scalability, speed, and real-time performance. Additionally, most of these systems are not publicly accessible, making them less practical for everyday use. As images frequently contain complex scenes and a diverse range of objects interacting in complex ways, there is a pressing need for a more advanced system that can dynamically adapt, efficiently scale, and generate accurate captions on the fly.

Project Overview:

This project aims to create an advanced automated image captioning system that employs a blend of Transformer architectures and Convolutional Neural Networks. Utilizing Python as the programming language, the system will leverage TensorFlow and Keras libraries to build and train deep learning models.

The CNN component will serve as the encoder to understand and encode the visual context of images. A Transformer-based Encoder-Decoder framework will act as the caption generator, offering advantages in scalability and performance over traditional RNN-based methods. The model will be trained and validated using a specified dataset, while metrics like loss, accuracy, and BLEU scores will be closely monitored. We also plan to deploy the model as a public website where users can upload images to generate captions in real-time.

Project Goals:

The goals of this project:

1. **Automated Caption Generation**: Build a system capable of generating accurate and contextually relevant captions for a variety of images.
2. **Architecture Comparison**: Evaluate and compare the effectiveness of different deep learning architectures in the task of image captioning.
3. **Performance Visualization**: Use Matplotlib and other visualization tools to represent the system's performance metrics.
4. **Public Accessibility**: Deploy the finalized system as a user-friendly web application.

Project Timeline:

- **Phase 1:** Data Collection and Preprocessing
  - Collecting datasets
  - Image and text preprocessing
  - Data augmentation techniques
- **Phase 2:** Model Training and Evaluation
  - Building various deep learning models
  - Hyperparameter tuning
  - Model validation and comparison
- **Phase 3:** System Deployment and Evaluation
  - Backend and frontend development
  - Model deployment
  - User acceptance testing

Datasets :

- COCO:
  - The Common Objects in Context (COCO) dataset is a large-scale, diverse collection of images annotated for multiple vision tasks.
  - It includes over 200,000 labeled images, 1.5 million object instances, and 80 object categories.
  - COCO provides rich contextual information, multiple object annotations, and 5 captions per image.
  - It is ideal for training robust and nuanced image captioning models.
- Flickr30k:
  - The Flickr30k dataset complements our project by adding another layer of diversity and context to our training data.
  - It consists of 30,000 images, each annotated with five different captions.

    - This dataset is especially valuable for its focus on everyday situations and a broad array of scenes.

Use Cases

- Social Media Platforms: Automated captioning can assist in accessibility, helping visually impaired users understand image content.
- E-commerce Websites: Automated tagging and captioning can enhance product discoverability.
- Search Engines: Improve image search accuracy by considering the generated captions.

Previous Works

Several systems have been proposed for image captioning using RNNs, LSTMs, and GRUs. CNN-LSTM architectures have been frequently employed but are generally limited by computational resources and don't scale well for larger datasets.

Our Unique implementation strategies:

Enhanced Accuracy and Performance

Technique: Advanced Data Augmentation (Random Cropping, Flipping, Color Jittering)

We will use data augmentation techniques like random cropping, flipping, and color jittering during the training phase to enhance the model's ability to generalize to new, unseen data.

Technique: Multi-Head Self-Attention Mechanism

The Transformer model we'll be using comes with multi-head self-attention mechanisms, which are excellent for capturing long-range dependencies in data. This feature should allow our model to better understand the context of each image, leading to more accurate captions.

Real-time Predictions

Technique: Real-Time Inference API

We aim to offer real-time captioning to individual users, a feature largely absent in existing solutions. To achieve this, the trained model will be exposed as a RESTful API, capable of receiving image data and returning generated captions in real-time.

Modularity

Technique: Microservices Architecture & API-based Component Integration

To achieve high levels of modularity, we'll structure our application following a Microservices Architecture. Each component, such as data preprocessing, the CNN encoder, the Transformer model, and the web interface, will be developed as an individual service, Enabling any component to be replaced or upgraded independently without disrupting the entire system.

Conclusion:

This project aims to develop a state-of-the-art automated image captioning system, leveraging CNNs and Transformer architectures. Utilizing specified datasets like COCO and Flickr30k, the model will undergo rigorous training and validation, with key metrics such as loss, accuracy, and BLEU scores being monitored. Upon completion, the model will be deployed as a public web application for real-time captioning. The team is optimistic that this endeavor will contribute meaningfully to image captioning research and practical applications. Your consideration is highly appreciated.