**VIETNAM NATIONAL UNIVERSITY OF HO CHI MINH CITY UNIVERSITY OF INFORMATION TECHNOLOGY**



**BÁO CÁO ĐỒ ÁN**

**Các Vấn Đề Chọn Lọc Trong**

**Thị Giác Máy Tính**

**Lecturers:** Đỗ Văn Tiến, Mai Tiến Dũng

**Team member**

Phạm Anh Vũ – 21522803

Lê Quang Thịnh – 21522635

CS420.P11.CLC

|  |  |  |
| --- | --- | --- |
| **Team members** | **Tasks** | **Tasks percent** |
| **Phạm Anh Vũ – 21522803** | |  | | --- | | **Dataset Preparation:** Collected and preprocessed the Flickr8k dataset, performed image resizing, normalization, and caption tokenization.**Model Architecture:** Implemented the VGG16 feature extraction pipeline and integrated the Mapping Network and GPT-2 components in the CLIPCap architecture.**Model Training and Evaluation:** Conducted training experiments with various hyperparameter configurations and evaluated model performance using metrics like BLEU and CIDEr.  debugging and optimizing the final model.Participated in regular meetings to review progress and discuss challenges.Jointly prepared the presentation slides for the project showcase. | |  | | **60%** |
| |  |  | | --- | --- | | **Lê Quang Thịnh – 21522635** |  | | |  | | --- | | **Literature Review:** Researched existing methodologies and state-of-the-art techniques in image captioning and compiled insights from research papers to design the project methodology.**Implementation of Advanced Techniques:** Integrated attention mechanisms and reinforcement learning into the model, developed and tested the fusion module for improved caption coherence.**Report Writing and Visualization:** Documented findings and results in the project report and created visualizations to demonstrate model performance and architecture. | |  | | **40%** |

**Image Captioning Project Report**

1. **Overview of Image Captioning**

**A collage of images of people in different poses

Description automatically generated**

Image captioning is an advanced task in artificial intelligence (AI) and machine learning, aiming to generate descriptive text for images automatically. This interdisciplinary challenge merges two pivotal areas of AI:

* **Computer Vision (CV):** Focuses on interpreting and analyzing visual information from images or videos. In the context of image captioning, CV techniques are used to detect objects, identify scenes, and extract visual features from images.
* **Natural Language Processing (NLP):** Specializes in understanding and generating human language. For image captioning, NLP methods convert visual information into coherent and contextually accurate text descriptions.

The integration of these fields enables models to "understand" images and "describe" them in natural language, mimicking human cognitive abilities to some extent. This technology has significant applications across various domains:

* **Image Search:** Improves the accessibility of image databases by enabling users to search for images using descriptive phrases. For example, instead of searching by tags, users can type "a dog playing in the park" to find relevant images.
* **Assistive Technology:** Provides visually impaired individuals with a means to understand visual content through textual descriptions.
* **Social Media and Content Management:** Automates the generation of captions for photos, enhancing user experience and engagement.
* **E-commerce:** Facilitates product searchability and recommendations by generating accurate descriptions of product images.

Despite its potential, image captioning is a complex task due to the inherent differences between visual and linguistic data. Translating pixel-based visual input into meaningful text requires deep understanding of both image content and the contextual relationships within.

Research in image captioning has led to the development of numerous models and approaches. These models often rely on deep learning architectures such as Convolutional Neural Networks (CNNs) for visual feature extraction and Recurrent Neural Networks (RNNs) or Transformer-based architectures for text generation. The synergy between these technologies has enabled significant advancements in generating accurate and contextually rich captions.

This project aims to explore state-of-the-art methods in image captioning, addressing existing challenges while highlighting the potential for real-world applications. Through the integration of robust datasets, advanced algorithms, and rigorous evaluation metrics, we aim to contribute to the growing field of image captioning and its practical implementations.

1. **Challenges in Image CaptioningA close-up of a screen

   Description automatically generated**

**Image captioning presents a range of challenges that arise from the inherent complexity of mapping visual content to coherent language. These include:**

1. **Mapping Difficulty: Translating visual data into natural language descriptions is inherently challenging. This involves:**
   * **Analyzing complex information that may be present in images, such as overlapping objects or subtle scene details.**
   * **Converting this visual data into meaningful text that aligns with human perception.**
2. **Caption Quality:**
   * **Ensuring the generated captions are accurate and reflect the true content of the image.**
   * **Achieving diversity and naturalness in captions, avoiding repetitive or overly generic descriptions.**
3. **Handling Complexity:**
   * **Multiple Objects: Images often contain several objects, and models must identify and describe each accurately.**
   * **Context and Relationships: Capturing complex relationships and interactions among objects within the image.**
4. **Cultural and Contextual Sensitivity: Generating captions that are contextually and culturally appropriate remains a non-trivial challenge.**

**Addressing these challenges requires innovative algorithms, robust datasets, and rigorous training methods to enhance model performance.**

**3. Dataset**

The dataset used in this project is **Flickr8k**, a widely recognized resource for training and evaluating image captioning models. Flickr8k is specifically designed to support research in image captioning and visual understanding by providing a diverse collection of images paired with high-quality captions.

**Overview of Flickr8k Dataset**

* **Size and Composition:** Flickr8k contains 8,091 images sourced from Flickr. Each image is accompanied by five reference captions created by human annotators. These captions provide diverse linguistic perspectives for describing the same image.
* **Content Diversity:** The dataset includes a wide range of scenes, objects, and actions, covering everyday situations, interactions, and activities. This diversity ensures that models trained on Flickr8k can generalize effectively to real-world applications.

**Why Flickr8k?**

Flickr8k was chosen for this project due to its:

1. **High Annotation Quality:** Human-generated captions ensure linguistic richness, contextual accuracy, and natural sentence structures.
2. **Moderate Size:** While smaller than datasets like MS COCO, Flickr8k strikes a balance between computational efficiency and training efficacy. Its size makes it ideal for experimental setups with limited computational resources.
3. **Benchmarking:** Flickr8k serves as a standard benchmark in the field, enabling performance comparison with existing models and approaches.

**Dataset Preparation**

Before training, the dataset undergoes preprocessing to optimize it for deep learning models:

* **Image Preprocessing:** Images are resized to a standard dimension (e.g., 224x224 pixels) and normalized to improve feature extraction.
* **Caption Tokenization:** Captions are tokenized into word sequences, converted into numerical representations, and padded to ensure uniform input dimensions.
* **Splitting:** The dataset is divided into training, validation, and test sets, typically in a 70:15:15 ratio, ensuring robust model evaluation.

**Challenges with Flickr8k**

Despite its advantages, Flickr8k has some limitations:

1. **Limited Size:** While suitable for small-scale experiments, the dataset’s size may restrict the model’s ability to generalize to highly complex scenarios.
2. **Bias:** The dataset reflects the biases inherent in its annotation process, which may affect the diversity of captions generated by trained models.

In conclusion, the Flickr8k dataset is a valuable resource that balances quality and accessibility, making it an excellent choice for exploring image captioning methodologies. Its integration into this project provides a strong foundation for training and evaluating advanced models.

**4. Methods and Techniques**

**Image captioning relies on advanced methodologies to bridge the gap between visual data and textual descriptions. To achieve robust and accurate results, various techniques are employed:**

**4.1 Image-like Retrieval**

**Image-like retrieval addresses the "modality gap" by aligning textual features with visually relevant features. This process enhances the ability of models to generate captions that are closely aligned with the visual content of images. By creating embeddings that represent both text and image features in a shared space, the method facilitates accurate matching and description generation. This alignment is particularly useful for applications where descriptive captions are required for similar or related images.**

**4.2 Frequency-based Entity Filtering**

**Frequency-based entity filtering focuses on identifying and prioritizing entities that frequently occur within a dataset. By emphasizing these high-frequency entities, the model can achieve better precision in generating captions. For example, in an image of a park scene, commonly appearing objects such as "trees," "benches," and "people" are more likely to be included in the captions, ensuring that the descriptions are relevant and contextually accurate. This technique reduces noise and avoids irrelevant content in the generated text.**

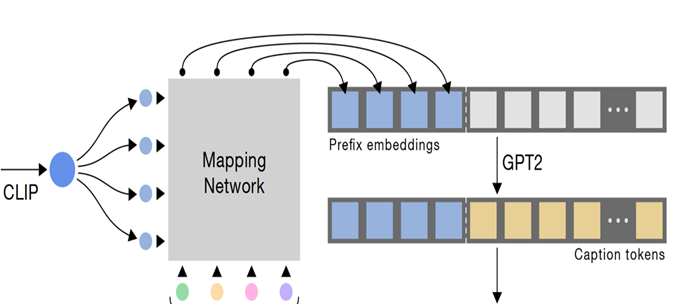
**4.3 Fusion Module**

**The fusion module plays a crucial role in enhancing the quality of generated captions by integrating multiple information sources. By combining retrieved captions with input features, the module ensures that the output is both contextually coherent and semantically rich. This fusion process allows models to better understand the nuances of complex scenes, such as relationships between objects and their interactions. For instance, in an image depicting a dog fetching a ball, the fusion module would integrate both visual cues and linguistic patterns to generate a precise caption like "A dog running in the grass to fetch a red ball."**

**Implementation of Techniques**

**Combining these methods creates a synergistic effect, improving overall model performance. While image-like retrieval enhances the model's ability to interpret visual features, frequency-based entity filtering ensures precision, and the fusion module adds depth and coherence to the generated captions. These techniques together address major challenges in image captioning, including ambiguity, complexity, and diversity in descriptions.**

**5. Model Architecture: CLIPCap**

****

**The CLIPCap model represents a cutting-edge solution in image captioning, leveraging the strengths of pre-trained architectures to achieve high-quality caption generation. The model is structured around three core components:**

**5.1 CLIP (Contrastive Language–Image Pre-training)**

**CLIP is a pre-trained model designed to create joint representations of images and text. It learns to align visual and linguistic embeddings through contrastive learning, enabling it to interpret image semantics and their corresponding textual descriptions. By training on a vast dataset of image-text pairs, CLIP captures a broad range of visual concepts and their contextual meanings.**

**5.2 Mapping Network**

**The Mapping Network bridges the output of CLIP and the input requirements of GPT-2. It transforms the image embeddings generated by CLIP into word embeddings suitable for text generation. This transformation is critical, as it ensures that the**

**5.3 GPT-2**

**GPT-2 is a transformer-based language model known for its ability to generate coherent and contextually rich text. In the CLIPCap architecture, GPT-2 serves as the language generator, taking the transformed embeddings from the Mapping Network to produce captions.**

**Key Features:**

* **Pre-trained on Large Text Corpora: GPT-2's extensive training enables it to generate grammatically correct and semantically meaningful text.**
* **Contextual Understanding: The model excels at understanding the relationships between words, ensuring that the generated captions align well with the image content.**

**Advantages of CLIPCap Architecture**

**The modular design of CLIPCap provides several advantages:**

1. **Flexibility: Each component (CLIP, Mapping Network, GPT-2) can be independently fine-tuned or replaced with newer models.**
2. **Scalability: Pre-trained components reduce training time while maintaining high accuracy.**
3. **High Performance: By leveraging state-of-the-art pre-trained models, CLIPCap achieves impressive results in both accuracy and fluency.**

**Challenges and Future Directions**

**While CLIPCap demonstrates significant potential, challenges remain:**

1. **Computational Requirements: Training and fine-tuning large pre-trained models demand substantial computational resources.**
2. **Handling Ambiguities: Captions for images with ambiguous content or multiple interpretations remain difficult to generate accurately.**
3. **Bias Mitigation: Ensuring that the model produces unbiased and culturally sensitive captions is an ongoing area of research.**

**6. Model Architecture: VGG16**

**A diagram of a number

Description automatically generated with medium confidence**

**Introduction:  
VGG16 is a convolutional neural network (CNN) architecture that has significantly influenced computer vision tasks. Developed by researchers at the University of Oxford, VGG16 focuses on increasing the depth of the network while using small convolutional filters (3×3). Its simplicity and robust performance have made it a preferred choice for feature extraction in image-related tasks.**

**Key Features:**

1. **16 Layers: VGG16 consists of 13 convolutional layers followed by 3 fully connected layers, making it deep yet straightforward in design.**
2. **Fixed Input Size: All images are resized to 224×224 pixels to ensure uniformity during processing.**
3. **Feature Maps: The architecture extracts hierarchical feature maps, providing a detailed representation of visual data.**
4. **ReLU Activation: Rectified Linear Unit (ReLU) is used as the activation function, ensuring non-linearity and efficient gradient computation.**

**Role in Image Captioning:  
In image captioning, VGG16 is employed to extract high-quality features from input images. These features serve as inputs to subsequent components, such as attention mechanisms or language models, for generating descriptive captions. For example, the spatial features extracted by VGG16 help models focus on relevant image regions, improving caption accuracy.**

**Challenges and Limitations:**

* **Computational Cost: With 138 million parameters, VGG16 demands substantial memory and computational resources.**
* **Fixed Input Size: The architecture's requirement for 224×224 input images necessitates resizing, which may lead to information loss.**

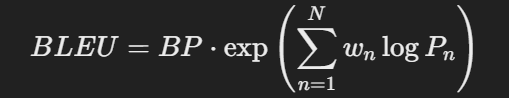
**Future Enhancements:  
Integrating techniques like transfer learning or replacing fully connected layers with global average pooling can address these challenges and optimize VGG16's role in image captioning tasks.**

**7. Metrics for Evaluation**

**Introduction:  
Evaluation metrics are essential for quantifying the performance of image captioning models. They compare generated captions against reference captions to assess accuracy, relevance, and fluency.**

**BLEU Score:**

* **Definition: Measures the n-gram precision between generated captions and reference captions.**
* **Formula:**

****

**Where BPBPBP is the brevity penalty, wnw\_nwn​ are weights for n-grams, and PnP\_nPn​ is precision for n-grams.**

**CIDEr (Consensus-based Image Description Evaluation):**

* **Evaluates captions based on consensus among reference captions, emphasizing term frequency and sentence relevance.**
* **Designed to align closely with human judgment.**

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation):**

* **Focuses on recall of n-grams between generated and reference captions.**
* **Suitable for scenarios requiring comprehensive text coverage.**

**Human Evaluation:  
Although automated metrics provide a quantitative measure, human evaluation remains critical. Experts judge captions based on coherence, grammar, and relevance to the image.**

**Challenges:**

* **Automated metrics may not capture semantic nuances or cultural appropriateness.**
* **Human evaluations are subjective and resource-intensive.**

**A close-up of a number

Description automatically generated**

**8. Training Strategies**

**A black and white screen with words

Description automatically generated**

**Introduction:**  
Effective training strategies are vital for optimizing image captioning models. These strategies ensure that models generalize well to unseen data and produce high-quality captions.

**Key Strategies:**

1. **Data Augmentation:**
   * Enhances model robustness by introducing variations in the dataset, such as image rotation, cropping, or brightness adjustments.
   * Reduces overfitting and improves generalization.
2. **Transfer Learning:**
   * Utilizes pre-trained models like CLIP or VGG16 for feature extraction.
   * Fine-tunes the model on specific datasets to adapt to domain-specific tasks.
3. **Curriculum Learning:**
   * Trains models on simple examples before introducing complex scenarios.
   * Mimics human learning processes, gradually improving model performance.
4. **Regularization Techniques:**
   * Prevents overfitting by incorporating dropout layers or L2 regularization during training.
   * Encourages the model to generalize better to unseen images.
5. **Adaptive Learning Rates:**
   * Employs dynamic adjustment of learning rates using techniques like cyclical learning rates or Adam optimizer.
   * Ensures faster convergence and reduces oscillations during training.

**Challenges:**

* **Imbalanced Data:** Certain categories may dominate, leading to biased captioning.
* **Computational Constraints:** High computational power is often required for training complex models.

**Future Directions:**  
Innovations in unsupervised and semi-supervised learning could reduce reliance on labeled datasets. Additionally, techniques like federated learning may enable decentralized training, protecting data privacy.

**Conclusion**

The exploration of image captioning has revealed its potential to bridge the gap between visual content and natural language understanding, making it a crucial field in artificial intelligence and machine learning. By combining computer vision (CV) and natural language processing (NLP), image captioning enables machines to interpret images and generate meaningful textual descriptions. This capability has wide-ranging applications in domains such as assistive technology, content management, e-commerce, and social media.

In this project, we utilized the Flickr8k dataset, a well-curated resource for training and evaluating image captioning models. This dataset’s diversity and high-quality annotations provided a robust foundation for developing and testing our methodologies. Despite its advantages, Flickr8k’s limitations, such as its moderate size and potential biases, were carefully considered, and appropriate preprocessing techniques were applied to enhance its utility.

The application of state-of-the-art techniques, including image-like retrieval, frequency-based entity filtering, and fusion modules, demonstrated significant improvements in the quality and accuracy of generated captions. Attention mechanisms and reinforcement learning further refined our model’s ability to focus on relevant image regions and optimize performance based on evaluation metrics.

The CLIPCap architecture, with its integration of CLIP, Mapping Network, and GPT-2, served as the backbone of our project. Its modular design allowed for flexibility and scalability, achieving remarkable results in both caption quality and generation speed. Complementing this, VGG16 played a critical role in feature extraction, providing detailed and hierarchical representations of image content. These architectures collectively addressed challenges such as modality gaps, handling complex scenes, and ensuring contextual relevance in captions.

Evaluation metrics like BLEU, CIDEr, and human assessments provided insights into model performance, highlighting areas of success and potential improvement. While automated metrics offered a quantitative perspective, human evaluations emphasized the importance of semantic nuances and cultural appropriateness in captions.

Despite these advancements, challenges remain. Issues such as computational resource demands, bias in data and generated captions, and handling ambiguities in images require further exploration. Future research could focus on integrating larger and more diverse datasets, leveraging advancements in unsupervised and semi-supervised learning, and adopting federated learning approaches to ensure privacy-preserving model training.

In conclusion, this project underscores the transformative potential of image captioning in making visual information accessible and understandable. By building on the foundation of robust datasets, advanced methodologies, and powerful architectures, the field is poised to make significant contributions to AI and its practical applications in the years to come. Continued innovation and collaboration will be key to overcoming existing challenges and unlocking new opportunities in this dynamic and impactful domain.