Personalized AI Image Captioning

Lokesh Pandey 01FE21BEC195 , Santoshi Vajrangi 01FE21BCS298

Under the Guidance of
Prof. Ramesh Ashok Tabib
KLE Technological University, Hubballi, Karnataka, India.

Academic Year 2024-25

1. Introduction
2. Proposed Methodology
3. Equations and Algorithms
4. Results

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.
- Applications include:

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.
- Applications include:
  - Content-based search engines.

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.
- Applications include:
    - Content-based search engines.
    - Social media auto-captioning.

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.
- Applications include:
  - Content-based search engines.
  - Social media auto-captioning.
  - Personalized marketing.

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.
- Applications include:
  - Content-based search engines.
  - Social media auto-captioning.
  - Personalized marketing.
  - Accessibility tools for visually impaired individuals.

Introduction

- Image captioning bridges the gap between human language and visual perception by combining deep learning techniques from computer vision and NLP.
- Applications include:
  - Content-based search engines.
  - Social media auto-captioning.
  - Personalized marketing.
  - Accessibility tools for visually impaired individuals.
- The demand for personalization in captioning has risen to cater to diverse user contexts, preferences, and needs.

Literature Review

- **Advancements:**

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**
  - Dataset bias hinders model generalization.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**
  - Dataset bias hinders model generalization.
  - Lack of metrics for evaluating personalization.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**
  - Dataset bias hinders model generalization.
  - Lack of metrics for evaluating personalization.
  - Maintaining coherence in personalized captions.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**
  - Dataset bias hinders model generalization.
  - Lack of metrics for evaluating personalization.
  - Maintaining coherence in personalized captions.
- **Research Gaps:**

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**
  - Dataset bias hinders model generalization.
  - Lack of metrics for evaluating personalization.
  - Maintaining coherence in personalized captions.
- **Research Gaps:**
  - Focus on personalized captioning based on user preferences.

Literature Review

- **Advancements:**
  - Pre-trained CNNs boost feature extraction.
  - Transformer models (e.g., BLIP) excel in contextual learning.
  - Contextual tags enable personalized captions.
- **Challenges:**
  - Dataset bias hinders model generalization.
  - Lack of metrics for evaluating personalization.
  - Maintaining coherence in personalized captions.
- **Research Gaps:**
  - Focus on personalized captioning based on user preferences.
  - Use of advanced NLP for coherence and relevance.

Block Diagram for "Problem Statement"

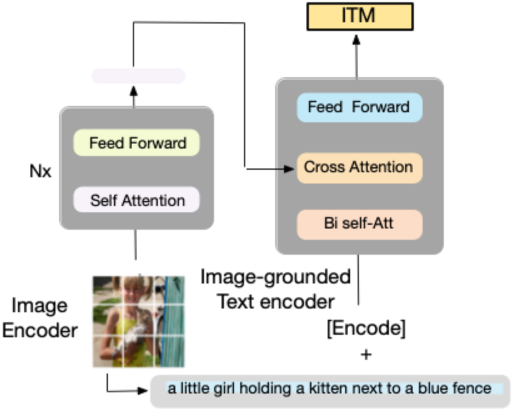

Figure 1: Block Diagram of BLIP [?]

Equations

Key Equation for Caption Generation

$$J(\theta) = \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \pi_{\theta}(a|s) Q^{\pi}(s,a) \tag{1}$$

Where:

- $J(\theta)$: Objective function.
- $d^{\pi}(s)$: Distribution of states under policy π.
- $Q^{\pi}(s,a)$: Action-value function.

Algorithms

Require: User image I, User preferences P
Ensure: Personalized caption C

1: Extract features from I using CNN $F = f(I)$
2: Encode preferences $E = g(P)$
3: Combine features and preferences $U = h(F, E)$
4: Generate caption $C = t(U)$
5: **return** C

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
  - Significant improvements in BLEU and CIDEr scores.

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
  - Significant improvements in BLEU and CIDEr scores.
  - Distinction between generic and personalized captions visible in visual analysis.

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
    - Significant improvements in BLEU and CIDEr scores.
    - Distinction between generic and personalized captions visible in visual analysis.
- Example Outputs:

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
  - Significant improvements in BLEU and CIDEr scores.
  - Distinction between generic and personalized captions visible in visual analysis.
- Example Outputs:
  - **Input 1:** A tropical beach at sunset.

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
  - Significant improvements in BLEU and CIDEr scores.
  - Distinction between generic and personalized captions visible in visual analysis.
- Example Outputs:
  - **Input 1:** A tropical beach at sunset.
  - **Output 1:** *"A serene and relaxing atmosphere with palm trees and a hammock."*

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
  - Significant improvements in BLEU and CIDEr scores.
  - Distinction between generic and personalized captions visible in visual analysis.
- Example Outputs:
  - **Input 1:** A tropical beach at sunset.
  - **Output 1:** *"A serene and relaxing atmosphere with palm trees and a hammock."*
  - **Input 2:** A young woman enjoying photography.

Results

- Performance evaluated using BLEU, CIDEr, and SPICE.
- Observations:
  - Significant improvements in BLEU and CIDEr scores.
  - Distinction between generic and personalized captions visible in visual analysis.
- Example Outputs:
  - **Input 1:** A tropical beach at sunset.
  - **Output 1:** *"A serene and relaxing atmosphere with palm trees and a hammock."*
  - **Input 2:** A young woman enjoying photography.
  - **Output 2:** *"Capturing the beauty around her with a camera."*

References I

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and Tell: A Neural Image Caption Generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Alexey Dosovitskiy, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations (ICLR)*, 2021.

[3] Md Shad Akhtar, Md Main Uddin Hossain, and others. "A Comprehensive Survey of Deep Learning for Image Captioning." *ACM Computing Surveys*, 2019.

[4] Alec Radford, et al. "Learning Transferable Visual Models From Natural Language Supervision." *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

References II

[5] Zehua Huang, et al.
"BLIP: Bootstrapped Language-Image Pre-training for Unified Vision-Language Understanding and Generation."
arXiv preprint arXiv:2201.12086, 2022.

[6] Andrej Karpathy and Li Fei-Fei.
"Deep Visual-Semantic Alignments for Generating Image Descriptions."
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[7] Aishwarya Agrawal, et al.
"VQA: Visual Question Answering."
International Journal of Computer Vision (IJCV), 2016.

References III

[8] Peng Wang, et al.
"Towards Personalized Image Captioning via Multimodal Memory Networks."
Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

Thank You