NEWS HEADLINE GENERATOR

(Text to Text Sequence Generetor)

Guide: Dr.Dipti Ghusse

MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

AUTHORS

Name:

- Sachin Jadhav, Shivanjali Jagtap ,Khushi Narad
- MIT Academy Of Engineering,Alandi,pune.

INTRODUCTION

- **Chain-of-Thought (CoT) prompting helps language models solve complex reasoning tasks by encouraging intermediate reasoning steps.**
- **However, real-world tasks often require multimodal inputs—such as images and text—not just language.**
- **This paper introduces Multimodal CoT Reasoning, extending CoT prompting to vision-language models (VLMs).**
- **The goal is to make language models not only understand text, but also reason step-by-step across both images and language.**

CHALLENGES

- Image-text alignment is crucial for generating correct reasoning chains.
- May require high compute for inference with large VLMs + LLMs.
- Dataset limitations—not all reasoning tasks have curated multimodal CoT annotations.

MOTIVATION

- Humans naturally combine visual and textual reasoning (e.g., interpreting diagrams in science questions).
- Standard large language models (LLMs) cannot process visual inputs.
- Vision-language models (like BLIP-2) can bridge this gap by generating textual representations of images.
- By combining image understanding + CoT prompting, we can enable better visual reasoning.

DATASETS

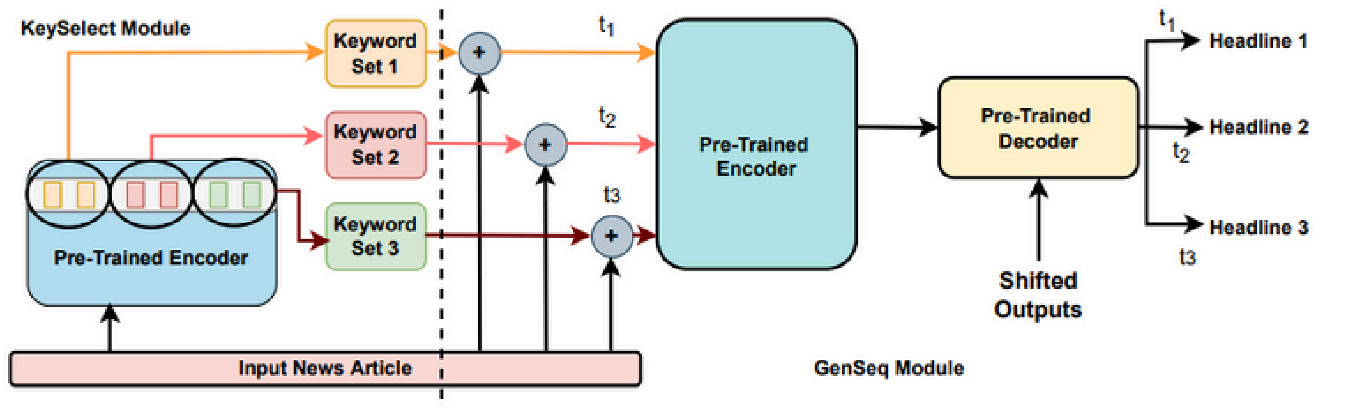
- **VQAv2 (Visual Question Answering)**
  - Image-based open-ended questions with text answers.
  - Requires both visual and reasoning skills.
- **ScienceQA**
  - Science and reasoning-based multiple-choice questions.
  - Includes text, images (diagrams), and explanations.

QUALITATIVE ANALYSIS

- **Generated rationales with images are more grounded and logically structured.**
- **Human evaluations show:**
  - Better faithfulness to visual content.
  - Easier to understand and verify reasoning steps.
  - Strong zero-shot generalization in unseen domains.

METHODOLOGY

1. **Multimodal Chain-of-Thought Prompting (MCoT)**
   a. **Objective:** Enhance reasoning by breaking down the problem using both visual and textual information.
   b. **Prompt Format:**
      i. Input: (Image + Question)
      ii. Example: Image of a plant cycle + "What process is shown?"
      iii. Intermediate Reasoning: The model explains the thought process in steps.
      iv. Final Answer: Based on the reasoning, the model generates the headline or answer.



2.SYSTEM ARCHITECTURE

a. **Image Encoder:**
   - → BLIP-2 extracts features from the image and generates descriptive text.

b. **Language Model:**
   - → GPT-3 / Flan-PaLM processes image text + question using Chain-of-Thought (CoT) reasoning.

c. **Prompting:**
   - → Few-shot: Uses 2–3 examples for guidance.
   - → Zero-shot: No examples; relies on model's prior knowledge.

3.Reasoning Enhancement

a. Incorporates step-by-step logic emulation similar to human reasoning.
b. Helps model handle complex queries, ambiguous visuals, and context-rich tasks like headline generation.

4. Training/Inference Pipeline

a. **Preprocessing:** Resize, normalize images; clean and tokenize text.
b. **Multimodal Fusion:**
   - Combine visual and text embeddings.
   - Use attention mechanisms to align relevant parts of the image with question context.
c. **Inference:**
   - Prompted with CoT format.
   - Outputs both a rationale and the final headline or answer.

6. Training/Inference Pipeline

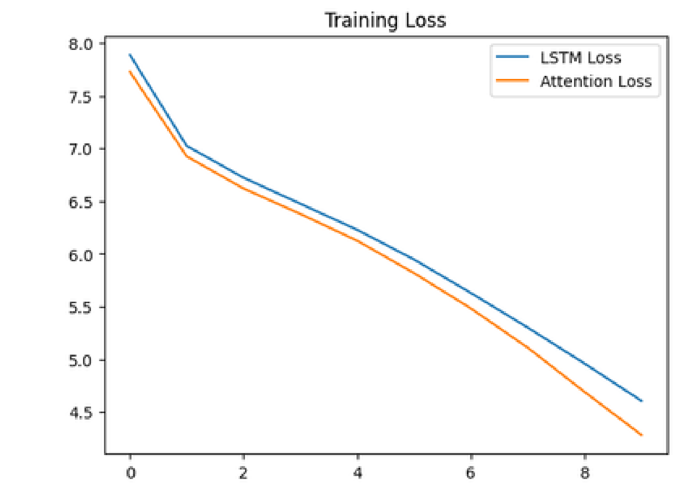
a. **Preprocessing:** Resize, normalize images; clean and tokenize text.
b. **Multimodal Fusion:**
   - Combine visual and text embeddings.
   - Use attention mechanisms to align relevant parts of the image with question context.
c. **Inference:**
   - Prompted with CoT format.
   - Outputs both a rationale and the final headline or answer.

EXPERIMENTAL RESULTS

We evaluated different methods on VQAv2 (Visual Question Answering) and ScienceQA datasets. Results show that our Multimodal Chain-of-Thought (MCoT) approach achieves the highest accuracy across tasks.

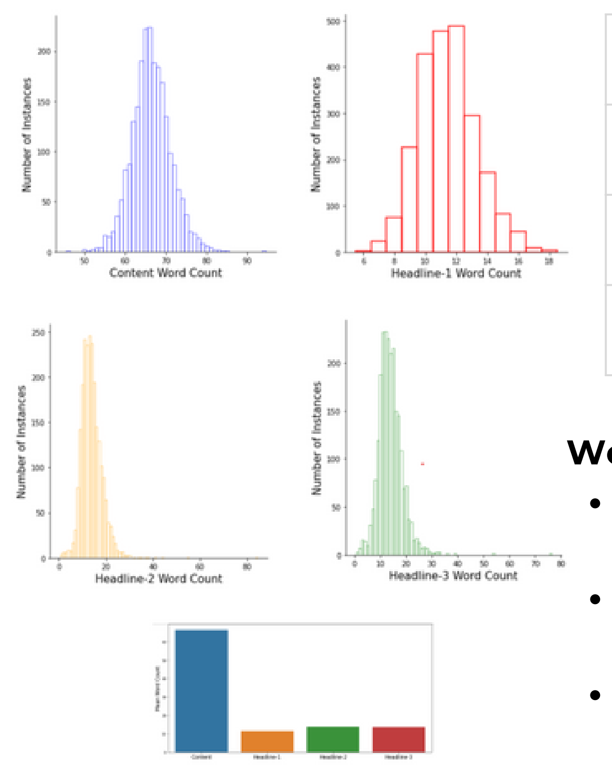
Key Findings:

- Text-only CoT lacks visual grounding → Lower performance.
- BLIP-2 without reasoning performs better, but lacks step-by-step logic.
- Multimodal CoT (Ours) significantly improves accuracy and interpretability.



Performance Comparison Table

Method	VQAv2 Accuracy	ScienceQA Accuracy
Text-Only CoT	Lower	Moderate
BLIP-2 w/o CoT	Moderate	Higher
Multimodal CoT (Ours)	Highest	Highest



(a) Mean word count of content and all headlines

Word Count Analysis

- Content has an average word count of ~68 words with a normal distribution.
- Headline-1 averages around 11 words, showing concise summaries.
- Headline-2 and Headline-3 are mostly under 20 words and right-skewed.
- The model maintains consistent content length and generates shorter, controlled-length headlines.

CONCLUSION

- This research introduces Multimodal Chain-of-Thought prompting, enabling LLMs to reason with both text and images.
- Combines the strength of LLMs in logical reasoning with the visual understanding of VLMs.
- Delivers state-of-the-art results on VQAv2 and ScienceQA benchmarks.
- Paves the way for interpretable multimodal AI systems in real-world applications.

REFERENCES

- Zhang et al., ACL 2023. Multimodal Chain-of-Thought Reasoning in Language Models.
- BLIP-2: Bootstrapped Language-Image Pretraining.
- ScienceQA Dataset: Lu et al., 2022.
- GPT-3, Flan-PaLM: LLMs for reasoning tasks.