FINAL PROJECT PROPOSAL

Image Caption Generator

DATS 6312 - Natural Language Processing for Data Science

Professor: Dr. Amir Hossein Jafari, Ph.D.

Team Members
Amrutha Jayachandra, Vishal Fulsundar

Problem Selection and Motivation

In today's online shopping, millions of product listings need short, interesting, and accurate captions. Writing descriptions for each product by hand is hard work, not always accurate, and can lead to mistakes. to solve this problem, we suggest , a system for generating captions that uses both visual features from product images and structured metadata (attributes, brand, price) to create marketing copy that sounds like person. The primary NLP task is text generation (image captioning), which combines language modeling, controlled text generation, and multimodal grounding.

We will use below datasets for this project:

1. FAshion CAptioning Dataset (FACAD)
2. FashionGen
3. ABO

If we face issue to FACAD datasets or unavailable due to licensing issue, we will substitute with FashionGen or ABO, both of which contain product images, attributes, and descriptions

---

Dataset Description

We are going to use the FAshion CAptioning Dataset (FACAD) — a large-scale, publicly available multimodal dataset designed for fashion e-commerce research.
As well as we are looking to check with other dataset such as Amazon Berkeley Objects (ABO) Dataset, MEP-3M (Multi-Modal E-commerce Product Dataset) .

Fashion Captioning Dataset FACAD-
includes approximately 130K human-written fashion descriptions and 993K images, But we are going to use the subset of dataset which will be near of 30K images.

Structured attributes (color, material, fit, neckline, style, and brand) are included in every sample.

Compared to standard image-captioning datasets, the captions are longer and more focused on marketing.

The GLAMI-1M dataset, which offers multilingual product descriptions and extra metadata, is one example of an optional extension. These datasets are balanced and appropriate for deep learning models that integrate language and vision.

Network Architecture

We use a three-phase multimodal design for our architecture:

1) Image encoder: Each product image is analyzed by a pretrained Vision Transformer (ViT) network, which then transforms it into a collection of visual feature vectors that represent information about color, texture, and shape.

2) Text Encoder (for Metadata): Brand, material, style, color, and other structured product details are formatted into a brief text prompt (e.g., "Attributes: color=navy, brand=Levi's, material=denim"). The meaning of each attribute is represented by embeddings created from this text.

3) Language Decoder: Using the visual features and the metadata embeddings as input, a large language model (Flan-T5) creates a fluid, marketing-style caption that accurately describes the product.

Instead of using traditional statistical models, this project makes use of contemporary Transformer-based NLP techniques. We will customize the BLIP-2: Bootstrapping Language-Image Pre-training Models multimodal architecture by fine-tuning the googles Flan language model with parameter-efficient LoRA adapters and conditioning it on structured metadata. This enables the model to produce text based on attribute and image data.

The entire model is based on BLIP-2 (Bootstrapping Language-Image Pretraining), which uses a tiny bridging to connect the language and image components.

Implementation Framework

We will use PyTorch and Hugging Face Transformers for model implementation and fine-tuning, benefiting from its modular design and GPU support for efficient experimentation.
Streamlit for the interactive demo (“Upload Image → Generate Caption”).

Reference Materials and Background Sources

- **Research Papers**:
  - BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models
    https://arxiv.org/abs/2301.12597
  - Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards
    https://arxiv.org/abs/2008.02693

Performance Evaluation Metrics

Model performance will be assessed with both standard captioning metrics and custom task-specific metrics:

Metric	Purpose
BLEU-1/2	Measures n-gram precision of generated text.
ROUGE-1/2/L F1	Evaluates content recall against reference captions.
BERTScore	Semantic similarity based on contextual embeddings.
CIDEr	Consensus metric for captioning quality.

Project Timeline

- **Week 1:** Reproduce baseline BLIP-2 LoRA training on 20 K FACAD subset; collect baseline BLEU/ROUGE.
- **Week 2:** Fine-tune BLIP-2 Flan-T5-XL with LoRA on Q-Former + decoder using new training schedule.
- **week 3:** Deploy Streamlit demo (Upload Image + Enter Metadata → Caption Comparison).
- **Week 4 :** Finalize report, slides, and performance visualizations.