Mini Project On

Multimodal Learning

Overview and code Analysis

Project Overview and Objectives

Objective :-

The primary objective of this project is to develop a multimodal learning model that can effectively process and integrate data from multiple modalities, specifically images and text, for tasks like image captioning and image-text matching

Methodology :-

1. Data Preprocessing: Clean and structure the multimodal data (images, text).
2. Fusion Strategies: Experiment with early and late fusion, cross-attention mechanisms.
3. Model Training: Train image and text encoders, and fuse them for multimodal understanding.
4. Evaluation: Assess the model's performance on tasks like image captioning and image-text retrieval.

Dataset Collection and Preprocessing

Dataset Selection:

- **Dataset:** The project uses the MS COCO dataset, which contains images paired with textual descriptions (captions). MS COCO is widely used for tasks like image captioning, image-text retrieval, and object detection.

Data Preprocessing:

- **Images:**
  - Resizing: Resize images to a fixed resolution (e.g., 224x224 or 256x256) to ensure consistency.
  - Normalization: Normalize pixel values to a standard range (e.g., [0, 1] or [-1, 1]).
  - Data Augmentation: Apply augmentation techniques (e.g., random flips, rotations, and brightness adjustments) to increase the diversity of the training data and improve model generalization.
- **Text:**
  - Tokenization: Convert textual descriptions into tokens using tokenizers like the WordPiece tokenizer for BERT or the Byte-Pair Encoding (BPE) for GPT.
  - Padding: Ensure text sequences are of a consistent length by padding or truncating longer text sequences.
  - Vectorization: Convert tokens into numerical representations using pre-trained word embeddings or transformer-based models (e.g., BERT embeddings).
- **Alignment:**
  - Image-Text Pairing: Ensure each image is correctly paired with its corresponding textual description for supervised learning.

Model Design and Fusion

Encoder Models:

- Image Encoder:
    - Use a pre-trained Convolutional Neural Network (CNN) such as ResNet or ViT (Vision Transformer) for image feature extraction.
    - The CNN model extracts high-level features from the images, which are then passed through a fully connected layer to produce fixed-size embeddings.
- Text Encoder:
    - Use a pre-trained Transformer-based model such as BERT or GPT for extracting text features. The model encodes the textual description into high-dimensional embeddings.

Multimodal Fusion:

- Early Fusion:
    - Combine image and text embeddings early in the network, typically by concatenating them at the input level or the initial layers of the model.
- Late Fusion:
    - Process images and text separately through their respective encoders, and merge them at a later stage (e.g., after feature extraction) using concatenation, bilinear pooling, or attention mechanisms.
- Cross-Attention Mechanisms:
    - Implement cross-attention to allow the model to focus on relevant parts of the image based on the words in the text and vice versa. This can be done by incorporating a multimodal attention layer that learns to attend to both modalities simultaneously.

Model Training

Training Setup:

- Loss Function: Use a suitable loss function based on the task. For image captioning, this may involve a cross-entropy loss between predicted and true captions. For image-text retrieval, use a contrastive loss to align matching image-text pairs and push apart non-matching pairs.
- Optimization: Use standard optimizers like Adam or AdamW to minimize the loss function.
- Batching: For efficiency, use batch processing with a batch size appropriate to the available computational resources.

Hyperparameter Tuning:

- Experiment with hyperparameters such as learning rate, batch size, and the number of layers to optimize model performance.

Regularization:

- Apply techniques like dropout or early stopping to prevent overfitting, especially in deep networks.

Evaluation

Evaluation Metrics:

- **Image Captioning:**
  - BLEU Score: Measures the overlap between predicted and true captions. A higher score indicates more accurate captions.
  - CIDEr Score: Measures the consensus between human-annotated captions and generated captions.
- **Image-Text Retrieval:**
  - Mean Average Precision (mAP): Measures the precision of retrieving relevant images/texts given a query.
- **Cross-modal Retrieval:**
  - Evaluate how well the model can retrieve the correct image for a given text query and vice versa.

Debugging and Model Optimization

Embedding Mismatch:

- Embedding Alignment: Ensure that the image and text embeddings are aligned in terms of dimensionality and meaningfulness before fusion. Resolve any issues with embedding mismatch, such as inconsistent sizes or improper normalization.

Optimization:

- Fine-Tuning: Fine-tune both the image and text encoders using transfer learning from pre-trained models (e.g., CLIP, VisualBERT).
- Pruning and Quantization: For deployment, apply techniques like model pruning and quantization to reduce model size and inference time.

Scalability:

- Ensure that the system can scale to handle larger datasets or real-time inference, optimizing both speed and accuracy.

.

Conclusion

The multimodal learning project demonstrated the power of combining data from multiple modalities—specifically images and text—to solve complex tasks like image captioning and image-text retrieval. By leveraging advanced models like ResNet for image processing and BERT for text encoding, we were able to design an architecture that effectively captures the relationship between visual and textual information.

Future Work

- Incorporating Additional Modalities: Future models can incorporate additional modalities like audio or video to improve the multimodal understanding of the content.
- Scalability: Work to make the model more scalable and suitable for real-time applications, reducing inference time and optimizing for large datasets.
- Improved Attention Mechanisms: Experiment with advanced attention mechanisms or transformers like Cross-Modal Transformers to better handle complex multimodal interactions.

Our Team Members

Yashika Tahaliyani
UGMR20230016@ihub-data.iiit.ac.in

Ujwal Dhomne
UGMR20230041@ihub-data.iiit.ac.in

Manasvi Namdev Harde
UGMR20230010@ihub-data.iiit.ac.in

Sujal wankhede
UGMR20230007@ihub-data.iiit.ac.in

Thank you