Suggestions for Final Project Report and Presentation

Understanding and Improving the Project

During the implementation of the ImageCLEFmed Captioning Project, several key improvements and ideas emerged that should be highlighted in the final report and presentation to demonstrate a deep understanding of the project beyond basic execution.

Complete Project Workflow Overview:

1. **Dataset Preparation:**
    - The ROCOv2-radiology dataset was downloaded from Huggingface.
    - The dataset included medical images along with corresponding captions and UMLS concepts.
2. **Feature Extraction:**
    - Features were extracted from the images using a pre-trained ResNet50 model.
    - The extracted features were stored in a file called features.npy, and the captions were saved in captions.json.
    - Images were resized, normalized, and processed into feature vectors (2048-dimensional arrays).
3. **Baseline Model Training:**
    - Decision Tree classifiers were trained on the extracted features to predict caption categories.
    - Random Forest classifiers were also used to enhance the performance.
    - Different strategies such as clustering captions into 10 or 30 groups were tested.
    - The best baseline accuracy achieved using Random Forest was approximately 82%.
4. **Advanced Method Exploration:**
    - The team discussed implementing more advanced vision-language models.
    - Experiments began with models like CLIP and LLaVA, aiming for direct image-to-caption generation.
5. **Team Division for Subtasks:**

- The project was divided into two subtasks:
    - Concept Detection: using classifiers like Decision Trees, Random Forest, CNNs.
    - Caption Prediction: requiring the use of advanced models like Transformers (e.g., CLIP, BLIP).

6. **Identification of Gaps:**
    - Although CLIP and LLaVA were explored, BLIP or Encoder-Decoder models had not yet been implemented.
    - Fine-tuning directly on images was not performed initially, which could further improve results.

Main Improvements and Ideas to Include:

1. **Fine-tuning Directly on Images:**
    - Instead of only using pre-extracted features (like those from ResNet50), applying direct fine-tuning on raw images with models such as BLIP can lead to better performance.
    - This approach allows the model to learn image-specific patterns that might be lost during feature extraction.
2. **Data Augmentation:**
    - Using data augmentation techniques (e.g., random rotations, flips, brightness adjustments) can make the model more robust.
    - It helps the model generalize better, especially when working with limited or similar-looking medical images.
3. **Multimodal Embeddings:**
    - Combining visual features with textual metadata (such as UMLS concepts) could enhance the model's understanding of the medical context.
    - This multimodal approach could result in more accurate and medically relevant caption generation.
4. **Validation Set Usage:**
    - Regular evaluation on a validation set (separate from training and testing) is critical.

- Monitoring validation loss during training prevents overfitting and ensures the model generalizes well to unseen data.

5. **Exploring Advanced Models:**
    - In addition to using BLIP, consider investigating BLIP-2 or larger models if time and resources allow.
    - These newer models are specifically designed for vision-language tasks and might achieve even higher performance.