Expected Outcomes by End of Week 3:

- **Results on a small validation set, including sample captions for a few images.**
- **Optimized model architecture trained on the full dataset.**

Objectives:

- **Build a basic prototype combining CNN and transformer-based architectures.**
- **Optimize the model architecture for better accuracy and efficiency.**

Tasks and Activities:

1. Building a basic prototype and testing on a small dataset

Implement Self-Attention:

- Introduce basic attention between image features and textual embeddings.
- Ensure the model learns contextual relationships between visual and textual inputs.

Train a Minimal Prototype:

- Train on a subset of the data (e.g., 1,000 images, use a subset of dataset Flickr8k) for quick testing and debugging.
- Experiment with small batch sizes and reduced vocabulary size for faster training cycles.

Evaluate Results:

- Generate captions for a small validation set and analyze qualitative outputs.
- Identify potential bottlenecks in image feature extraction, text encoding, or fusion.

2. Enhance the Model Architecture :

- **Fine-Tune the CNN Backbone:**
  - Use a pre-trained CNN (e.g., EfficientNet, ResNet) with fine-tuning for image feature extraction.
  - Freeze initial layers and allow fine-tuning for higher layers to adapt to the dataset.
- **Improve the Text Decoder:**
  - Experiment with transformer-based decoders or sequence-to-sequence models.
  - Add positional encoding to better model sequential relationships in captions.
- **Integrate Advanced Attention Mechanisms:**
  - Include multi-head self-attention to improve context understanding.
  - Implement cross-attention between visual features and the textual embeddings.
