Stefano Ruggiero – project 3

Multimodal Data Fusion for Image Classification

Objective

The project aims to classify objects in the NuScenes dataset using a multimodal approach. The data fusion leverages embeddings from LiDAR point clouds and images captured by the front-facing camera. This pipeline incorporates:

- Vision Transformer (ViT) for image data processing.
- PointNet for encoding 3D LiDAR data.
- A fully connected fusion network to combine multimodal features.
- Classification using fused, camera, and LiDAR embeddings.

The model's performance is evaluated using the F1 score, which measures the balance between precision and recall in multi-label classification.

Pipeline Overview

1. **Dataset and Preprocessing**

    - **NuScenes Mini Dataset**: A subset of NuScenes (v1.0-mini) provides synchronized LiDAR and camera data.
    - **Preprocessing**: Extracted image and LiDAR data for each sample. Images are resized and normalized for ViT processing, while LiDAR points are converted into tensors and structured as graphs for PointNet encoding.

2. **Image Encoder**

    - A pre-trained Vision Transformer (ViT) generates embeddings for images. The last hidden state corresponding to the [CLS] token is used as a condensed feature vector for classification.

3. **LiDAR Encoder**

    - **PointNet**: Encodes 3D point cloud data. The network uses:
        - Graph-based neighborhood calculations using k-nearest neighbors (k=5).
        - Two PointNetConv layers followed by global max-pooling for feature extraction.
        - Final linear transformation to produce a fixed 256-dimensional embedding.

4. **Fusion and Classification**

   The project uses **early fusion**, specifically **concatenation-based joint representation**, where features from image embeddings and LiDAR point clouds are combined at the feature level. After feature extraction, the embeddings from both modalities are concatenated into a unified vector, processed through a fully connected network for joint representation.

   - **Fusion Layer**: Combines 256-dimensional LiDAR embeddings with 768-dimensional image embeddings from ViT. A fully connected network reduces the concatenated 1024-dimensional feature vector to 128 dimensions.
   - **Classifier**: A three-layer fully connected network maps the 128-dimensional fused features to class probabilities. The model is trained with BCEWithLogitsLoss for multi-label classification.

5. **Training and Evaluation**

   - Data is split into training, validation, and test sets (80/20 ratio).
   - Separate models are trained for:
     - **Fused embeddings** (combined LiDAR and camera data).
     - **LiDAR embeddings** only.
     - **Camera embeddings** only.
   - Results are evaluated using the **averaged F1 score**.

Results

The classification performance is summarized as follows:

- **Fused Embeddings**: F1 Score: 0.9324
- **Camera Embeddings**: F1 Score: 0.9707
- **LiDAR Embeddings**: F1 Score: 0.7075

Analysis:

- The highest performance is achieved using only **camera embeddings**. This reflects the richness of visual data captured by the Vision Transformer.
- **Fused embeddings** slightly underperform compared to camera-only embeddings, possibly due to noise or misalignment in the LiDAR and camera data fusion process.
- **LiDAR embeddings** perform the worst, as the PointNet encoder may lose finer details compared to image features in this task. LiDAR may provide complementary information, but on its own, it lacks the granularity for detailed classification.

Conclusion

This project successfully demonstrates multimodal data fusion for object classification in a real-world autonomous driving dataset. While fused embeddings offer competitive performance, the camera-based embeddings dominate due to the effectiveness of Vision Transformers for image data.