Stefano Ruggiero – Project 3

Multimodal Data Fusion for Image Classification

Objective

The project aims to classify objects in the NuScenes dataset (mini version with 400 samples and 18 categories) by combining data from two modalities: LiDAR point clouds and camera images. This multimodal approach incorporates:

- Vision Transformer (ViT): A model pre-trained on ImageNet-21k and fine-tuned on ImageNet-2012 to process image data.
- PointNetConv: A graph-based encoder to process 3D LiDAR data using k-nearest neighbors (k-NN) to build a local graph structure.
- Fusion Network: Combines features from the two modalities via early fusion (concatenation).
- Classification: Neural network for object classification based on fused, camera-only, and LiDAR-only embeddings.

Performance is evaluated using the F1 score, a metric balancing precision and recall in multi-label classification tasks.

Pipeline Overview

1. Dataset and Preprocessing:
   - Dataset: NuScenes Mini Dataset (v1.0-mini), containing synchronized LiDAR and camera data for 400 samples.
   - Preprocessing:
     - Images: Resized and normalized the images for ViT input.
     - LiDAR: Converted into tensors and structured as graphs (via k-NN) for PointNetConv encoding.
2. Image Encoder:
   - Model: Vision Transformer (ViT, google/vit-base-patch16-224).
   - Processing:
     - Images are split into 16×1616 \times 16 patches, which are linearly embedded.
     - A [CLS] token is prepended to the sequence for classification tasks.
     - Absolute positional embeddings are added to the sequence before feeding it into the Transformer encoder.
   - Output: The [CLS] token from the last hidden state is used as the condensed image embedding.
3. LiDAR Encoder:
   - Model: PointNetConv, a graph-based network for 3D point cloud data.
   - Processing:
     - Graph construction using k-nearest neighbors (k=22) to compute edges.
     - Two PointNetConv layers extract local features, followed by global max-pooling to aggregate information.
     - A final linear transformation reduces the embedding to a fixed 256-dimensional representation.
4. Fusion and Classification:
   - Fusion Strategy: Early fusion via concatenation, combining the 768-dimensional image embedding with the 256-dimensional LiDAR embedding into a unified 1024-dimensional vector.
   - Fusion Layer:
     - A fully connected network reduces the 1024-dimensional vector to 128 dimensions.

- Classifier:
    - A three-layer fully connected network maps the fused embedding to class probabilities.
    - Training is performed using BCEWithLogitsLoss for multi-label classification.

5. Training and Evaluation:
- Data Splitting: Training, validation, and test sets are created using an 80/20 split.
- Models Trained:
    - Fused embeddings (combined LiDAR and camera data).
    - Camera embeddings only.
    - LiDAR embeddings only.
- Evaluation: Average F1 score over 1000 runs with different random splits.

Results

The classification performance (average over 1000 runs) is summarized as follows:
- Fused Embeddings: F1 Score: 0.914
- Camera Embeddings: F1 Score: 0.96
- LiDAR Embeddings: F1 Score: 0.74

Analysis

- Camera Embeddings: Achieved the highest F1 score, highlighting the effectiveness of Vision Transformers for extracting rich visual features.
- Fused Embeddings: Performed well but slightly underperformed compared to camera-only embeddings, likely due to:
    - Noise or misalignment in data fusion.
    - Limited complementary information provided by LiDAR in this task.
- LiDAR Embeddings: Scored the lowest due to:
    - Lack of granularity in standalone LiDAR data.
    - Possible underfitting of the PointNetConv encoder.
    - Computational constraints and limited dataset size.

Challenges

1. LiDAR limitations:
    - Lower granularity compared to visual data.
    - Challenges in encoding fine details due to graph sparsity or suboptimal kk.
2. Computational cost: Processing multimodal data significantly increases computational requirements.

Conclusion

This project demonstrates the potential of multimodal data fusion for object classification in autonomous driving datasets. While fused embeddings offer competitive performance (F1=0.914), the superior results of camera embeddings (F1=0.96) underscore the effectiveness of Vision Transformers for visual tasks. LiDAR embeddings alone (F1=0.74) highlight the limitations of standalone LiDAR for detailed classification, though it provides complementary information in a fused pipeline.