Multimodal Data Fusion for Image Classification

Stefano Ruggiero

Project Goal

- The goal of this project was to classify objects in the NuScenes dataset by combining data from two modalities: **LiDAR point clouds** and **camera images**.

Approach

- A **Vision Transformer (ViT)** to process image data.
- **PointNet** to encode 3D LiDAR data.
- An **early fusion strategy** that combines features from both modalities into a single representation.

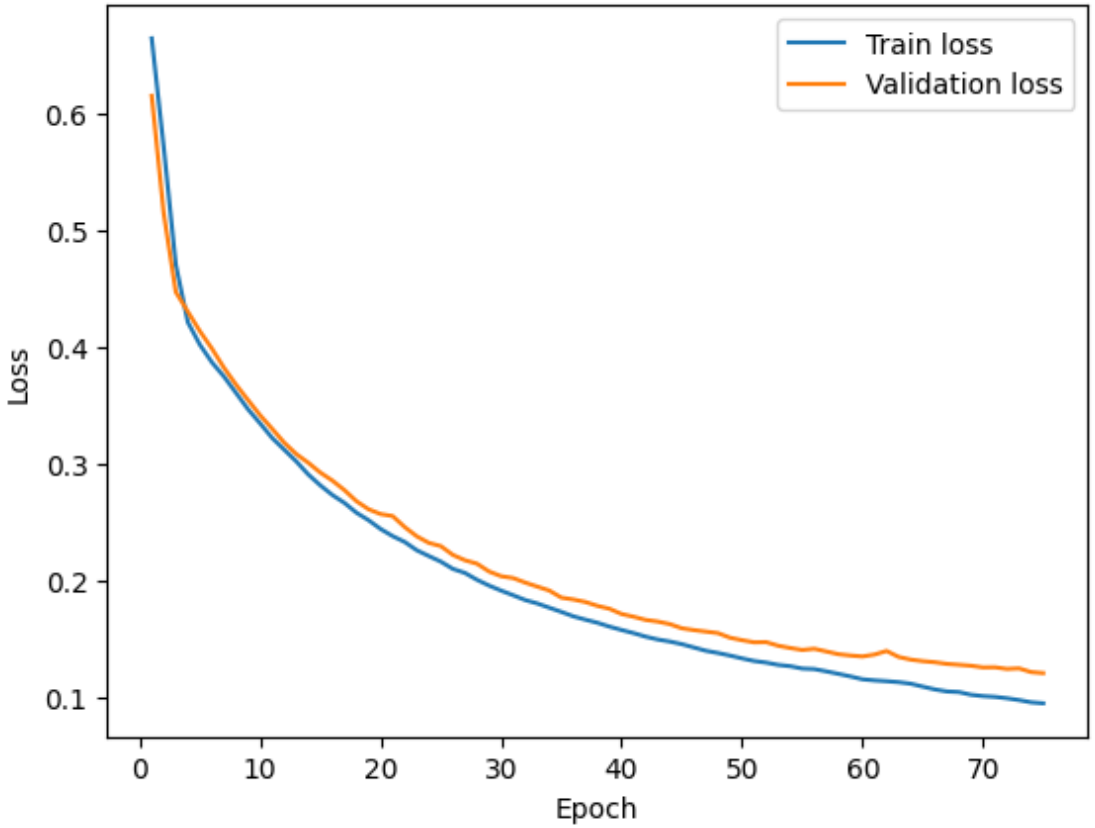
Run the code

- **Dataset Loading & Preprocessing: Feature Extraction**
- **Fusion**
- **Classification**
- **Evaluation (F1- score)**

Result

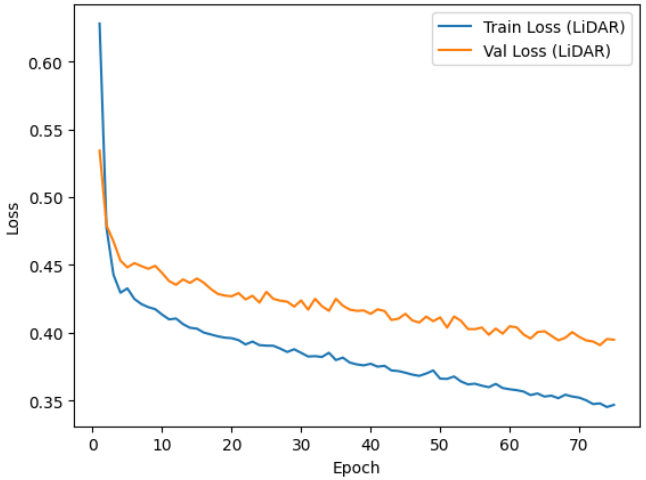
- **Fused embeddings**
  F1 Score: 0.9324

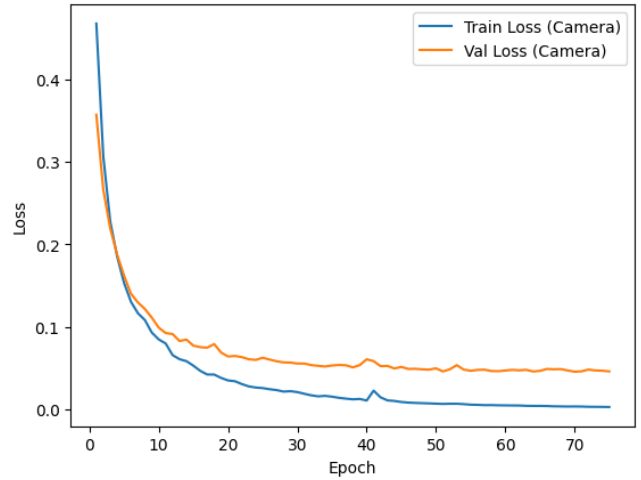
- Is good a result ?



Results

- **Camera embeddings** achieved the highest F1 score (0.9707), highlighting the effectiveness of ViT for image data.
- **Fused embeddings** performed well (F1 Score: 0.9324) but slightly underperformed compared to camera-only embeddings, likely due to noise or modality alignment issues.
- **LiDAR embeddings** scored the lowest (F1 Score: 0.7075), reflecting the challenges of encoding fine details from point clouds.





Challenges

- Aligning and synchronizing LiDAR and image data during fusion introduced potential noise.
- LiDAR data lacks the granularity of visual data, which limited its standalone performance.
- The computational cost of processing multimodal data, especially with a transformer-based architecture, required careful optimization.

Takeaways

- **Strength of Vision Transformers**: ViT is highly effective at capturing global spatial relationships in image data.
- **Challenges in Multimodal Fusion**: Effective fusion requires addressing modality alignment and noise issues, especially in early fusion strategies.
- **Importance of Representation**: While LiDAR data provides complementary information, it alone lacks the resolution for detailed object classification.
- **Evaluation of Fusion Strategies**: