Ahmedabad University

CSE623 - Group XYZ

Deploying Person Retreival System on Edge Devices

Meet Rathi
AU2240106
B.Tech CSE

Aditya Agrawal
AU2240153
B.Tech CSE

Harsh Panchal
AU2240160
B.Tech CSE

Problem Statement

Deploying a person retrieval system based on CLIP code, which will be optimized with Low-Rank Adaptation (LoRA) on the edge device, which is Nvidia Jetson Orin AGX for assessing its real-time performance in resource-constrained environment.

Key challenges:

- Computational Efficiency
- Latency
- Model Optimization
- Hardware Compatibility
- Real-Time Performance

Literature Review

Research Paper	Method Used	Key Findings	Limitations
Learning Transferable Visual Models From Natural Language Supervision (CLIP) [1] (2021)	Large-scale Pre-training with Image-Text Pairs	- Learns transferable visual representations from image-text pairs. - Excels in zero-shot tasks with ResNet and ViT variants.	- High computational cost for larger models. - Needs extensive datasets for effective pre-training.
Ultra Low-Power Deep Learning Applications at the Edge with Jetson Orin AGX Hardware [2] (2023)	Low-Power Vision Models on Edge Devices	- Deploys lightweight models (e.g., YOLOv4-tiny) efficiently on Jetson Orin AGX. - Enables real-time object detection with low power usage.	- Limited by edge device capacity, restricting model complexity. - Requires model optimization for balanced performance and power.
CLIP-Based Multi-level Alignment for Text-based Person Search [3] (2024)	CLIP + Fine-Grained Feature Extraction	- Enhances person search with multi-level vision-text alignment. - Auxiliary segmentation improves retrieval accuracy.	- Increased inference time due to multi-level alignment. - May introduce noise in feature alignment with diverse datasets.

Flickr Dataset

The Flickr Image Dataset is a large-scale collection of images sourced from Flickr, a popular photo-sharing platform. It is widely used in computer vision and machine learning research for tasks such as image classification, object detection, and image captioning.

- Images: 31,783
- Annotations: overall 158,915 User-generated tags, labels, captions

Dataset Features

- Diverse Images
- Metadata
- User-generated content

ICFG-PDES PRS Dataset

The ICFG-PDES (Image and Contextual Feature Graph - Person Detection and Retrieval System) dataset, focusing exclusively on person images, is designed for tasks related to person retrieval and re-identification. This dataset contains images of individuals captured in various environments, making it ideal for developing and evaluating algorithms for person detection, tracking, and retrieval.

- Images: 21000+ images
- Annotations: split, person detection id and captions.

Dataset Features

- Focus on Person Images
- Contextual Information
- Multiple Instances of Individuals

Methodology

Preprocessing:

- Each image is assigned a unique identifier (ID), with multiple captions grouped together.
- Cleaned and processed data is stored in a structured format with image filenames, captions, and unique IDs.
- A dataset class handles image loading, text tokenization, and transformations.

Model Architecture:

- Image Encoder: ResNet50 (2048-dimensional embeddings).
- Text Encoder: DistilBERT (768-dimensional embeddings).
- Projection Head: Maps embeddings to a shared 256-dimensional space.
- CLIP Model: Combines image and text encoders for similarity analysis.

Methodology

Key Training Details:

- **Optimizer:** AdamW with adaptive learning rates
- **Loss:** Contrastive Cross-Entropy with cosine similarity
- **Scheduler:** ReduceLROnPlateau (auto LR reduction)
- **Model:** ResNet-50 (Image) + DistilBERT (Text) + Projection Head
- **Training:** Batch 64, 3 epochs, saves best model

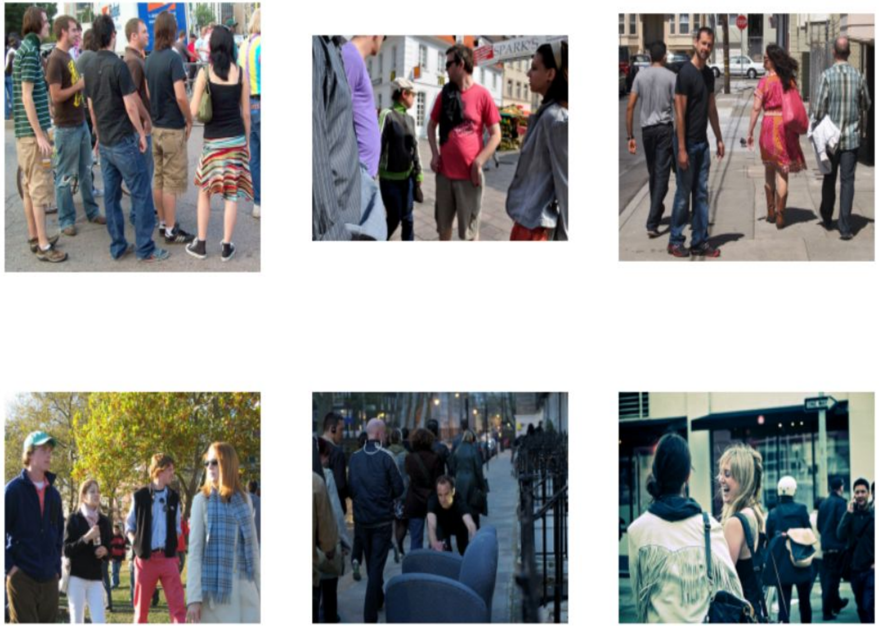
Inference:

- Get Image Embeddings: Generate 256-dimensional embeddings for validation set images.
- Find Matches: Compute similarity using dot product:

  similarity=text_embeddings · image_embeddings^T

- Saving Retrieval Results: Output file contains image file names and similarity scores.

Results



Query: A group of friends on a walk

Query: A single person performing
A ritual

Future work

1. **Fine-Tune Model Hyperparameters for Efficiency:**
   - Perform minor tweaks to optimize model parameters, reducing inference time while maintaining accuracy.
2. **Code Optimization for Jetson Compatibility:**
   - Refactor and optimize the code for smoother deployment on Jetson Orin AGX to minimize latency.
3. **Test and Evaluate Model Performance on Jetson:**
   - Run comprehensive evaluations on Jetson hardware to validate performance, focusing on speed, power consumption, and accuracy.
4. **Implement Resource-Efficient Scheduling on Jetson:**
   - Optimize task scheduling to balance CPU and GPU workloads effectively during inference.
5. **Address Potential Bottlenecks in Edge Deployment:**
   - Identify and resolve bottlenecks related to I/O operations and memory bandwidth when running on Jetson.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, ”Learning Transferable Visual Models From Natural Language Supervision.”

[2] Z. Wu and S. Ma, ”CLIP-Based Multi-level Alignment for Text-based Person Search.”

[3] M. Barnell and D. Isereau, ”Ultra Low-Power Deep Learning Applications at the Edge with Jetson Orin AGX Hardware,” in Proc. 2022 IEEE High Performance Extreme Computing Conference (HPEC), 2022, doi: 10.1109/HPEC55821.2022.9926369.

[4] M. Ye, J. Shen, G. Lin, et al., ”Deep Learning for Person Re-Identification: A Survey and Outlook,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 2872-2893, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, et al., ”Learning Transferable Visual Models From Natural Language Supervision,” in Proceedings of the International Conference on Machine Learning (ICML), 2021.

[6] J. Devlin, M. W. Chang, K. Lee, et al., ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, pp. 4171-4186.

[7] M. Barnell, C. Raymond, M. Wilson, D. Isereau, and C. Cicotta, ”Target Classification in Synthetic Aperture Radar and Optical Imagery Using Loihi Neuromorphic Hardware,” in Proc. 2020 IEEE High Performance Extreme Computing Conference (HPEC), 2020, pp. 1-6, doi: 10.1109/HPEC43674.2020.9286246.

[8] D. Isereau, C. Capraro, E. Cote, M. Barnell, and C. Raymond, ”Utilizing High-Performance Embedded Computing, Agile Condor, for Intelligent Processing: An Artificial Intelligence Platform for Remotely Piloted Aircraft,” in Proc. 2017 IEEE Intelligent Systems Conference (IntelliSys), 2017, pp. 1155-1159, doi: 10.1109/IntelliSys.2017.8324277.

Thank You