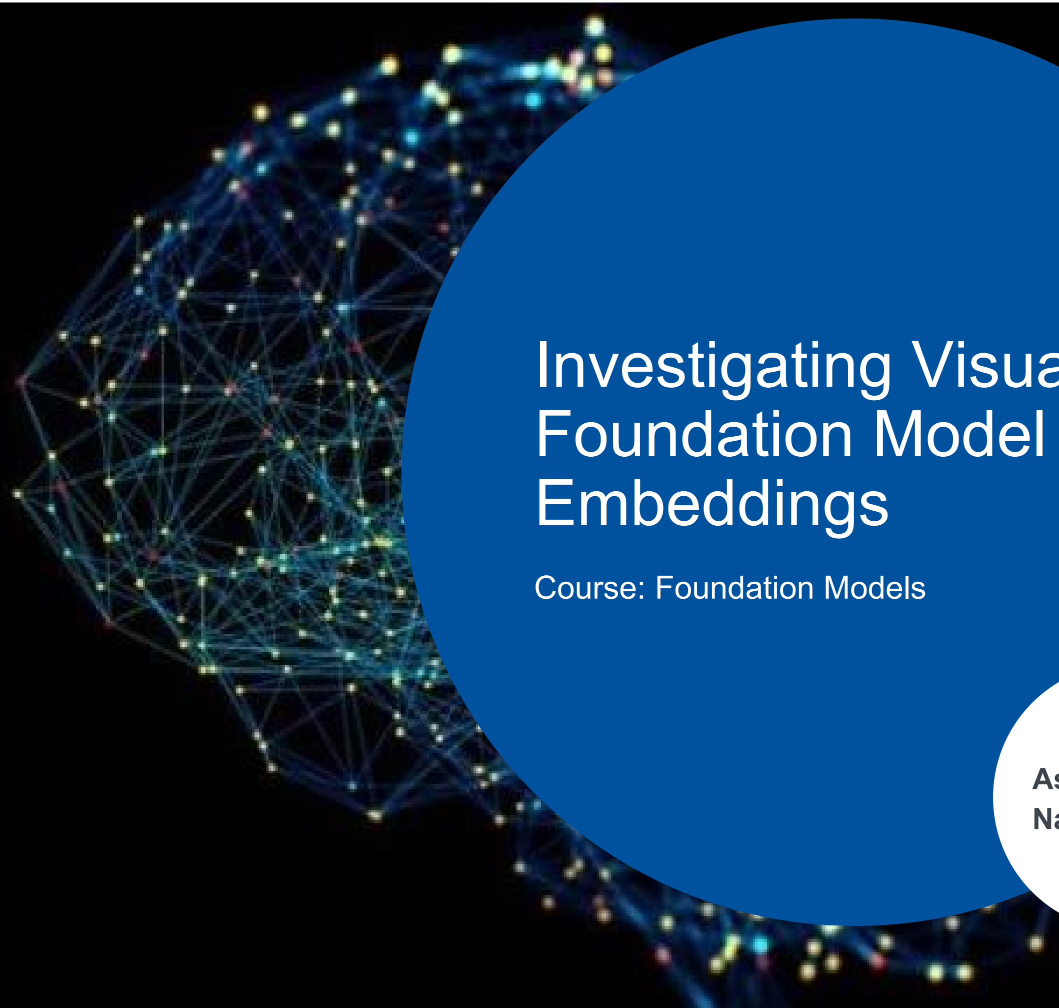
Universität Stuttgart



Investigating Visual Foundation Model Embeddings

Course: Foundation Models

Ashwin Murali
Namrata Jangid

Visual Foundation Models

Motivation

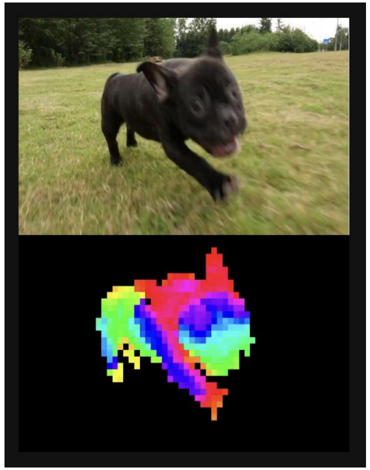
- In recent years, the field of computer vision has seen a huge surge in the development of advanced visual foundation models.
- However, the internal workings and representations of these models remain somewhat mysterious.
- Bridging the gap between the high-dimensional embedding representations and human understandable insights explains the inner workings of the model.
- Understanding the underlying features captured by their embeddings when subjected to dimensionally reduction is a way to interpret these models.
- We focus on CLIP, DINOv2, and SAM models for our analysis.

Background

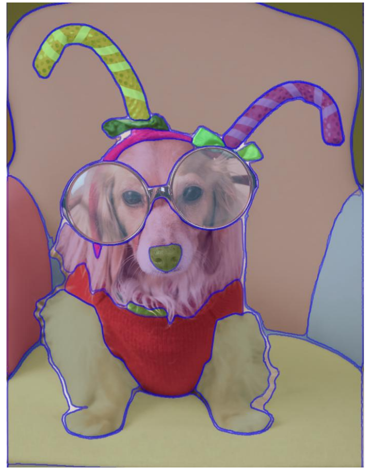
CLIP, DINOv2, SAM

- **CLIP** (Contrastive Language-Image Pretraining) learns a joint representation of images and corresponding textual descriptions.
- **DINOv2** can discover and segment objects in an image or video with no supervision and a segmentation-targeted objective.
- **SAM** (Segment Anything Model) is a promptable segmentation system with zero-shot generalisation to unfamiliar objects and images.

CLIP



DINOv2



SAM

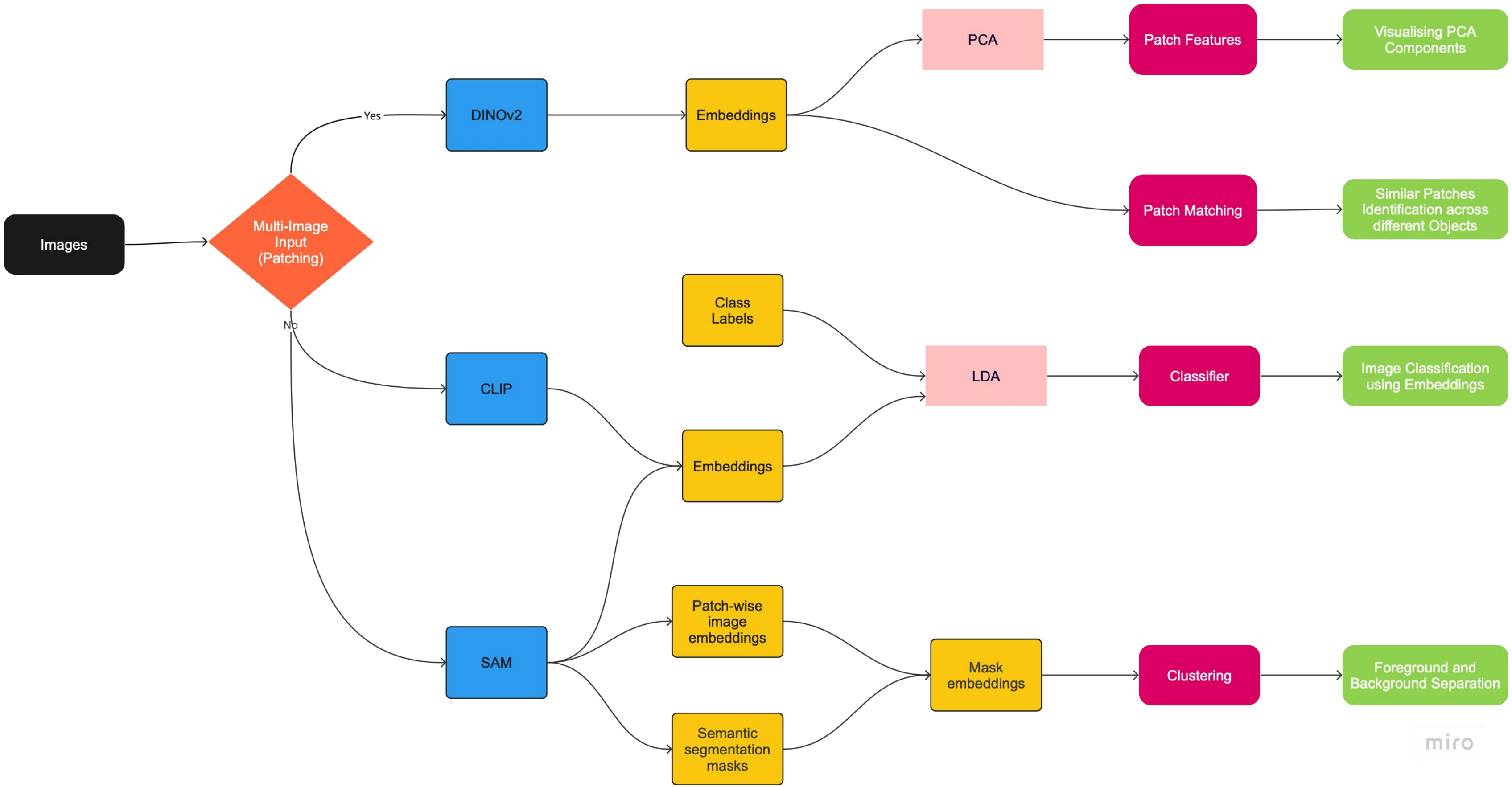
Background

Model Card

Aspect	CLIP	DINOv2	SAM
Architecture Variant	ViT-B/32	ViT-g14	MAE ViT-H/16
Input Image Shape	>= 224 x 224	>= 224 x 224	=1024 x 1024
Patch Input	No	Yes	No
Embedding Shape	512	1536	256

Methodology

Architecture Diagram



Methodology

Datasets

- CIFAR-10, Caltech-101 and KITTI were the three datasets taken for the analysis.

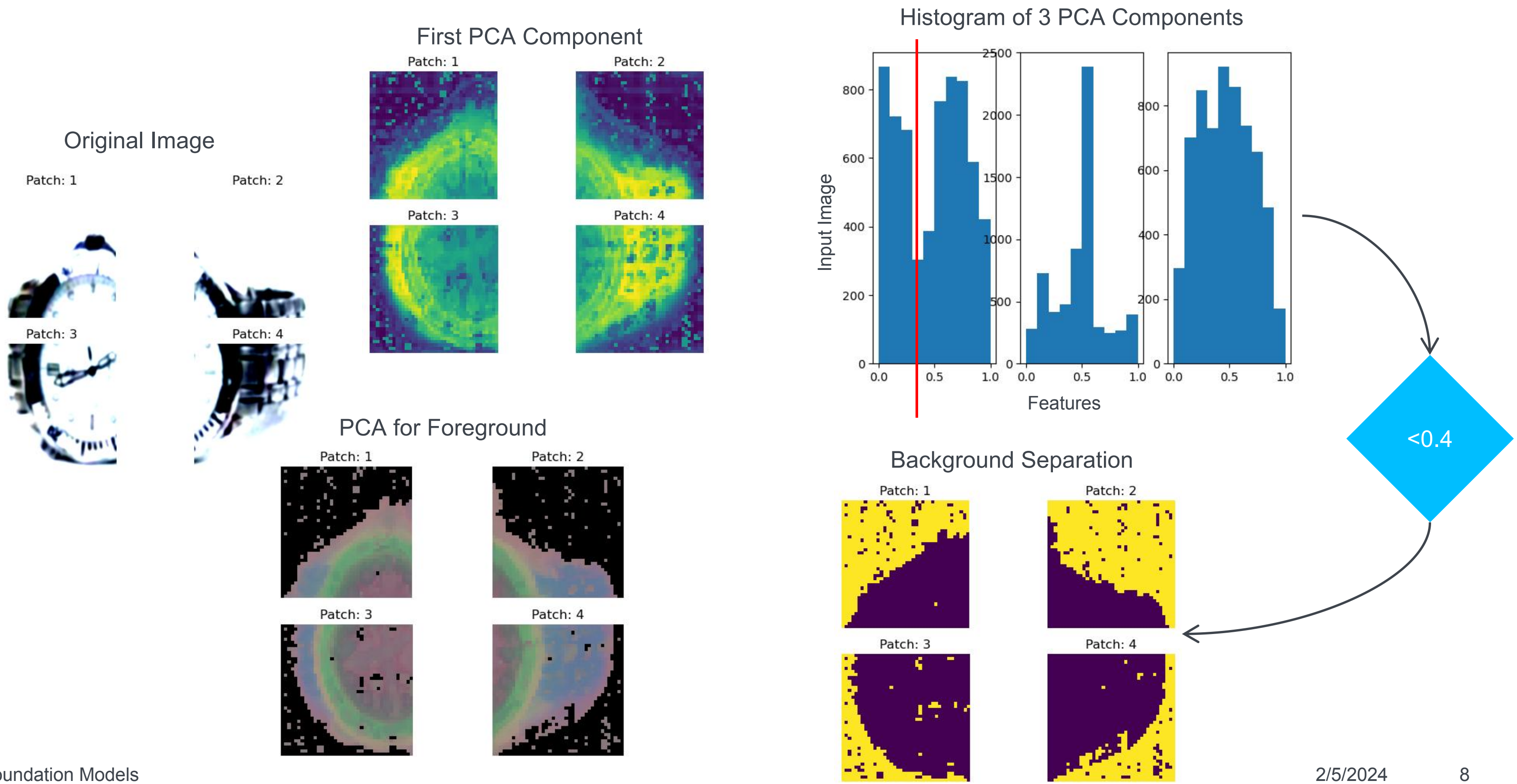
CIFAR-10	Caltech-101
10 Classes	101 Object Categories
32*32 Image Resolution	Higher resolution with varying sizes
Common objects and animals like airplanes, cars, birds, cats and dogs	Wide range of categories such as faces, animals, vehicles and household objects
Simpler dataset	Challenging dataset

- KITTI, a benchmark dataset for Autonomous Driving was also used for a specific analysis using SAM.

Patch Analysis using DINOv2

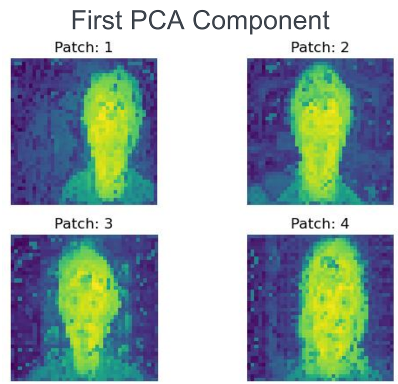
Results

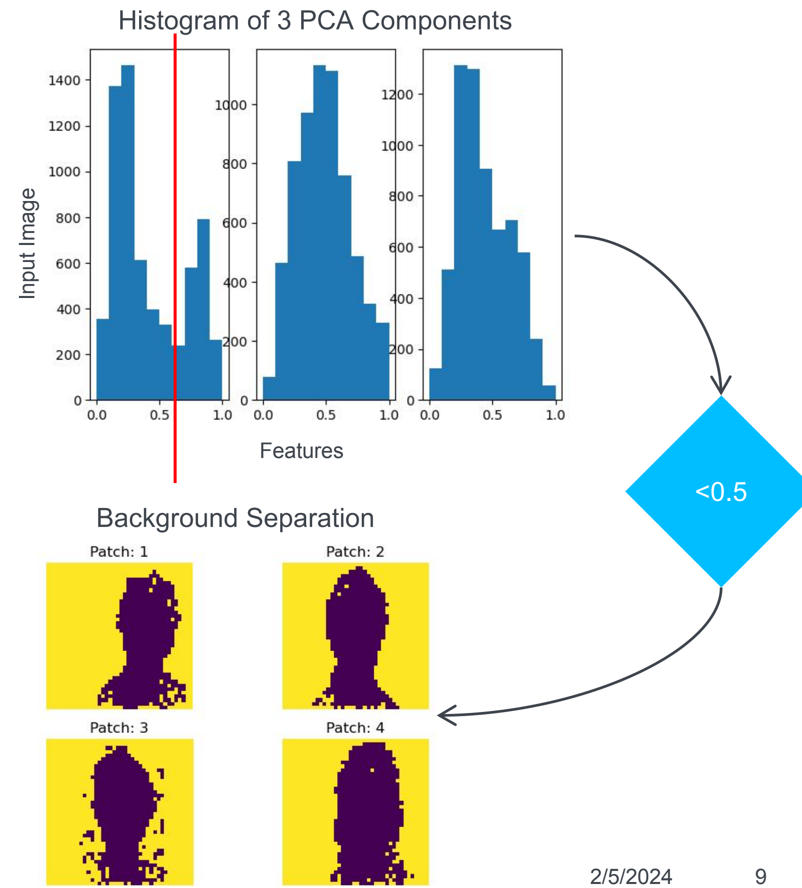
Intra-patch Visualization after PCA using DINOv2 Embeddings

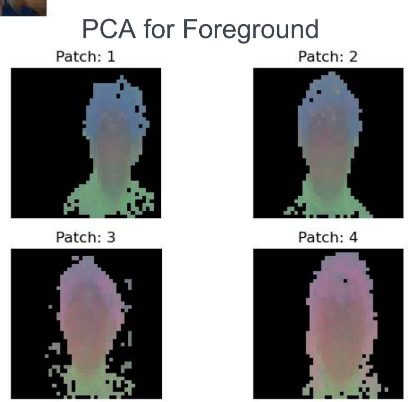


Results

Inter-patch Visualization after PCA using DINOv2 Embeddings



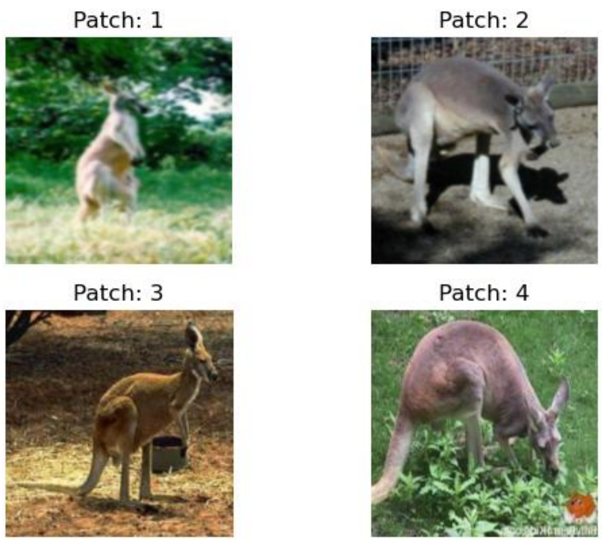




Results

Patch Matching using DINOv2 Embeddings

Kangaroo



~=

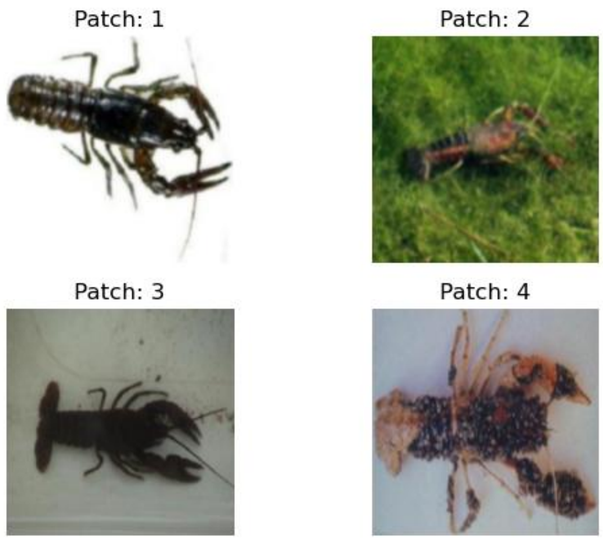
Pizza

Cosine Similarity : 0.18
Euclidean Distance : 72.67

Results

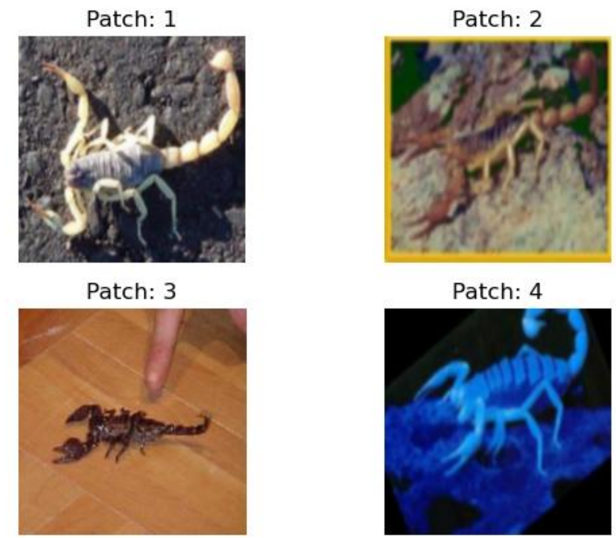
Patch Matching using DINOv2 Embeddings

Crayfish



~=

Scorpion

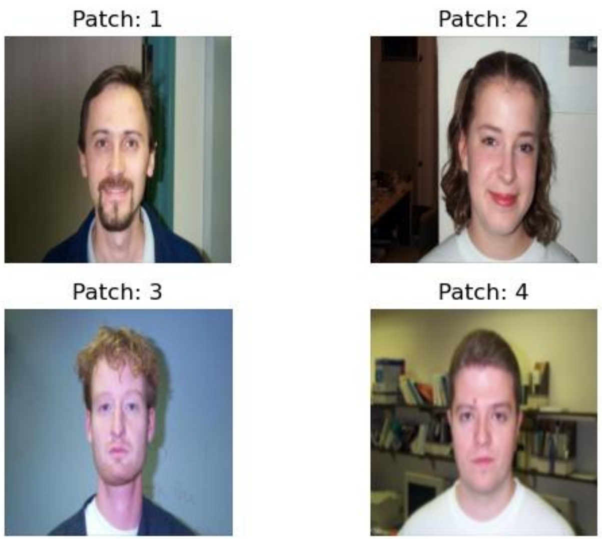


Cosine Similarity : 0.27
Euclidean Distance : 68.30

Results

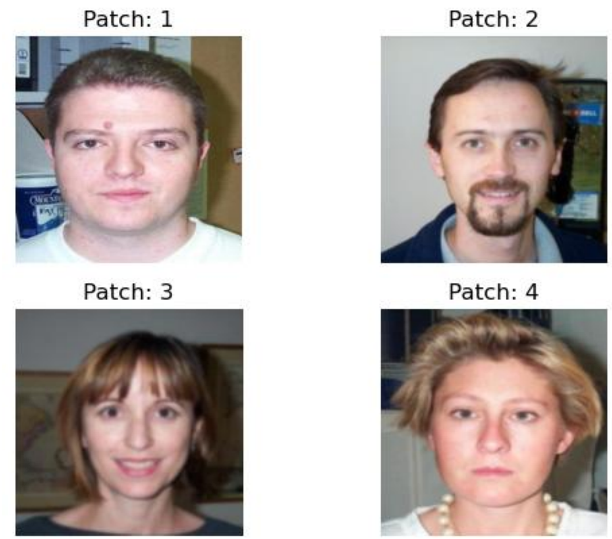
Patch Matching using DINOv2 Embeddings

Human Faces



~=

Human Faces

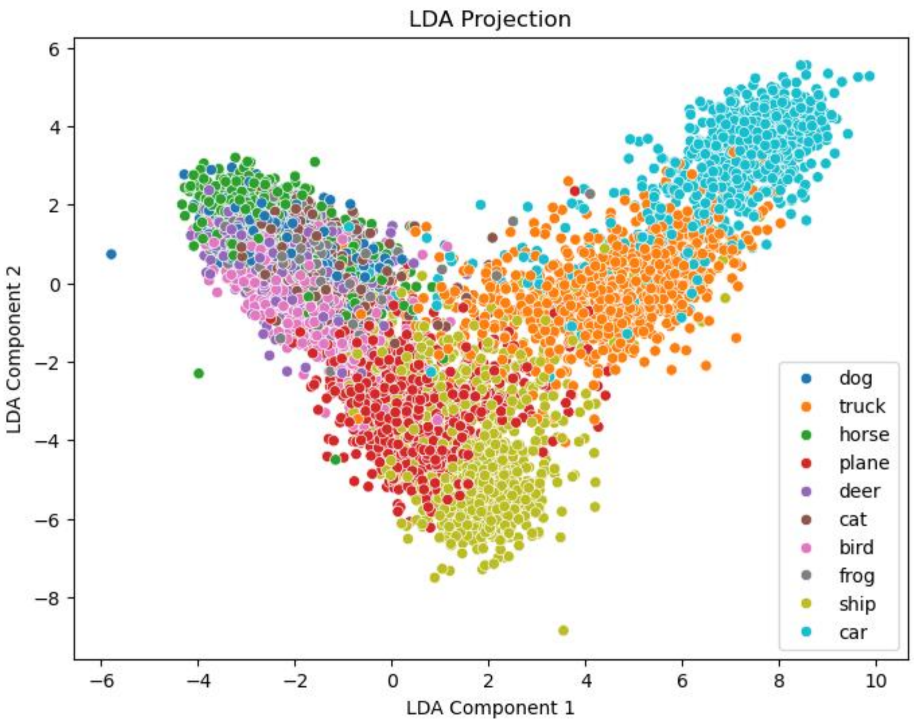


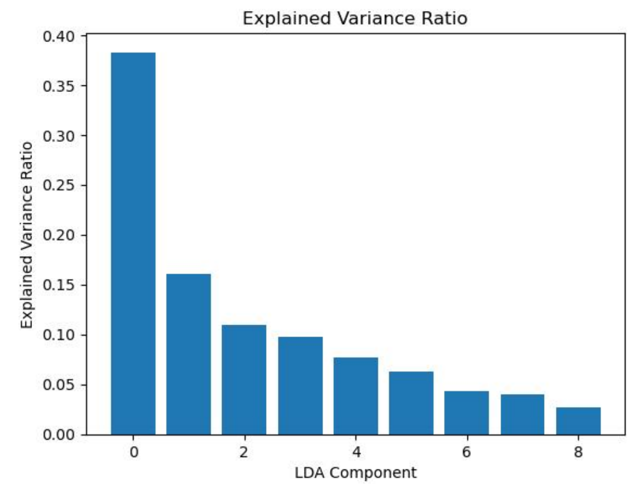
Cosine Similarity : 0.35
Euclidean Distance : 65.19

LDA Projection of CLIP and SAM Image Embeddings

Results

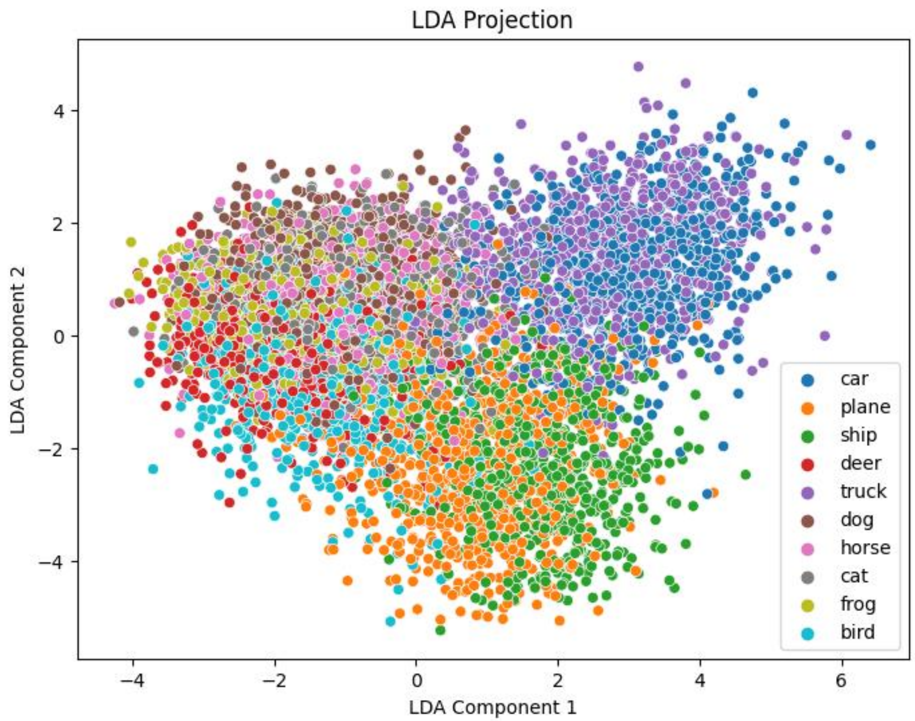
LDA Projection of CLIP Embeddings of CIFAR-10 Dataset

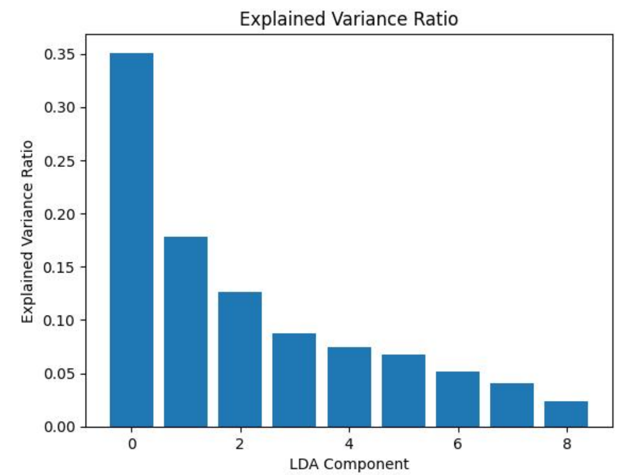




Results

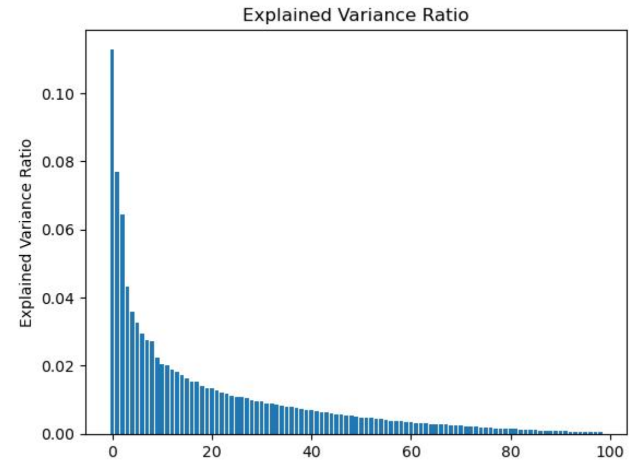
LDA Projection of SAM Embeddings of CIFAR-10 Dataset

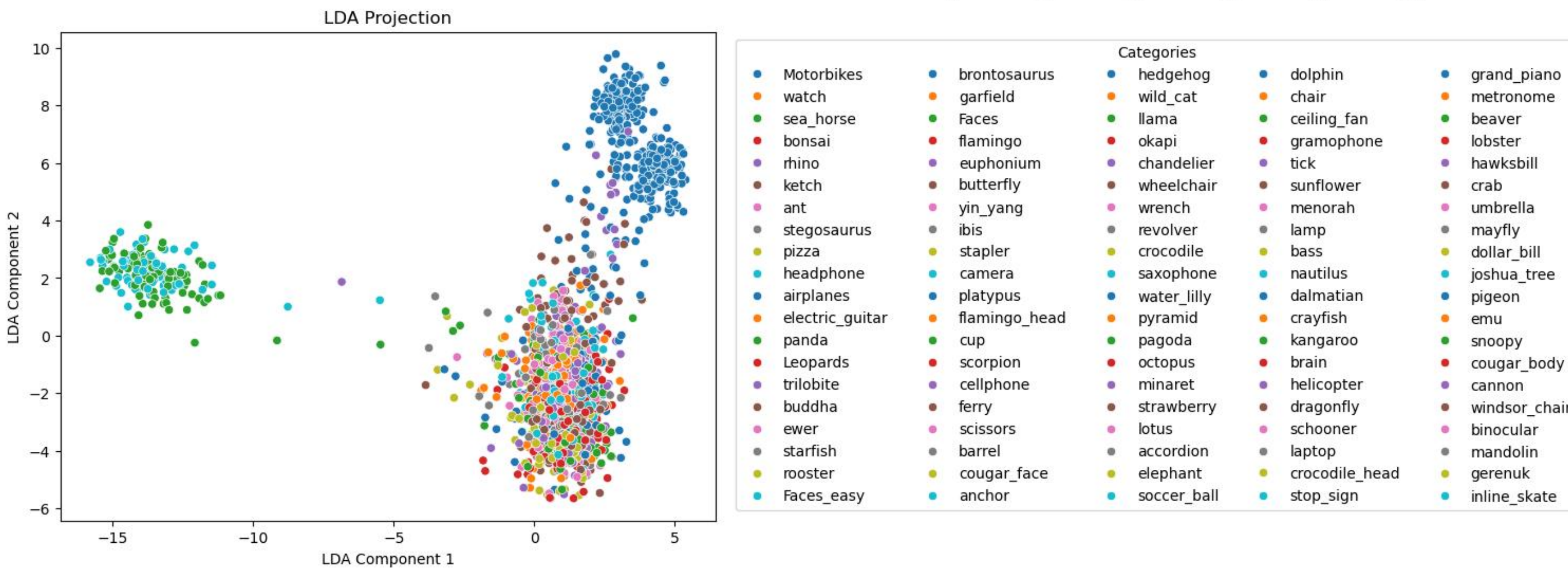




Results

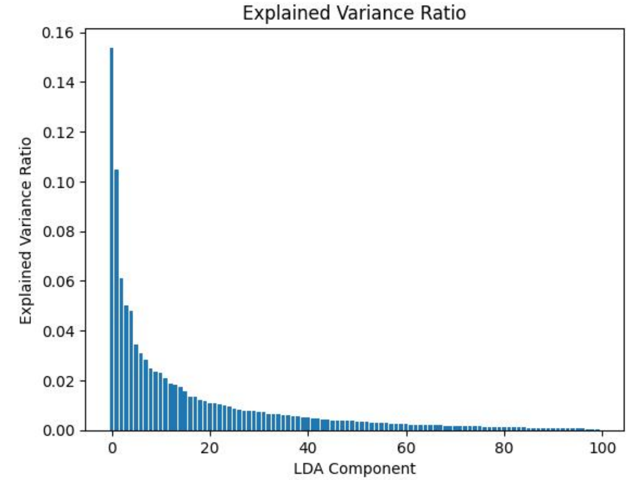
LDA Projection of CLIP Embeddings of Caltech-101 Dataset

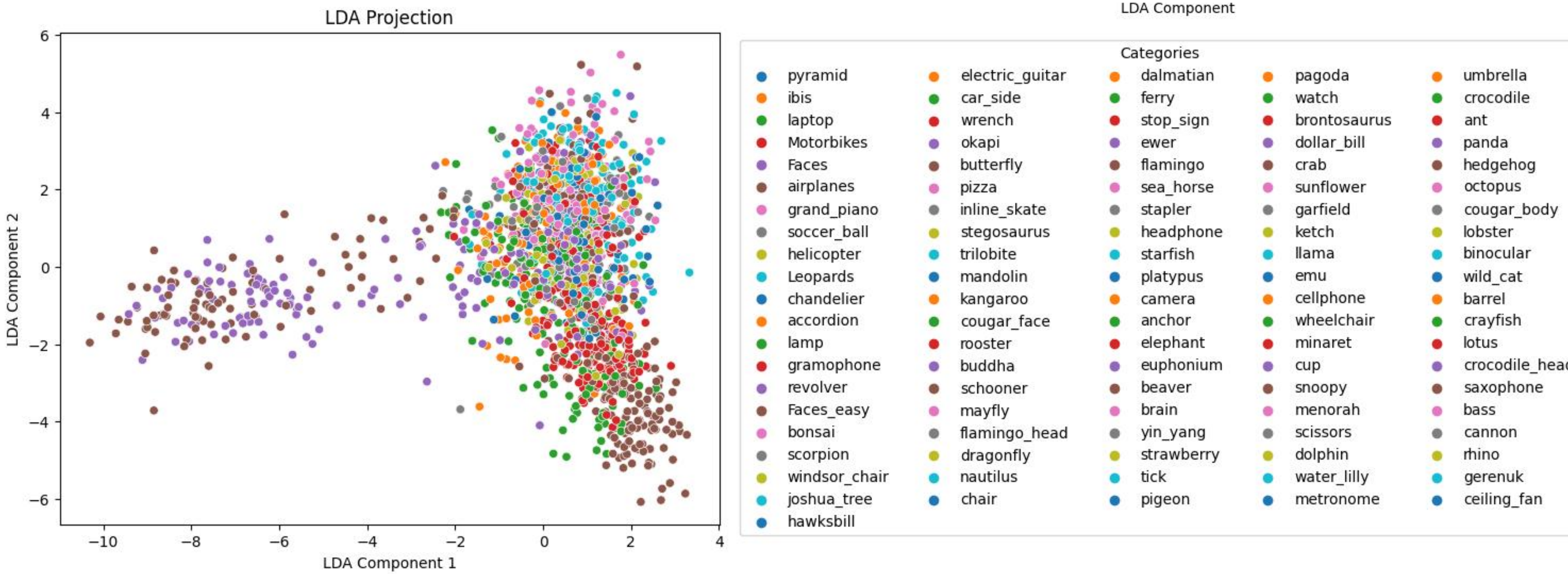




Results

LDA Projection of SAM Embeddings of Caltech101 Dataset





Discussion

Classifier Accuracy of LDA Projected Embeddings

Model	CIFAR-10	Caltech-101
CLIP	0.85	0.83
SAM	0.71	0.62

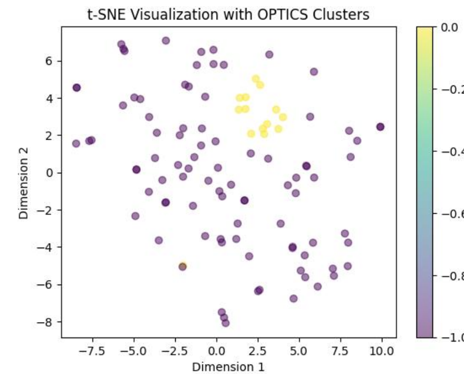
- Logistic regression classifier trained on reduced dimensionality CLIP embeddings performs better as compared to that trained on reduced dimensionality SAM embeddings.
- The larger size of CLIP’s embeddings implies that CLIP’s image encoder captures a wider variety of image features, which resulted in a more discriminative feature set after LDA.
- The complexity and size of SAM’s image encoder implies that it captures more nuanced features, which did not translate into a linearly separable space as effectively after LDA.
- Features relevant for semantic segmentation may not necessarily be optimal for classification.

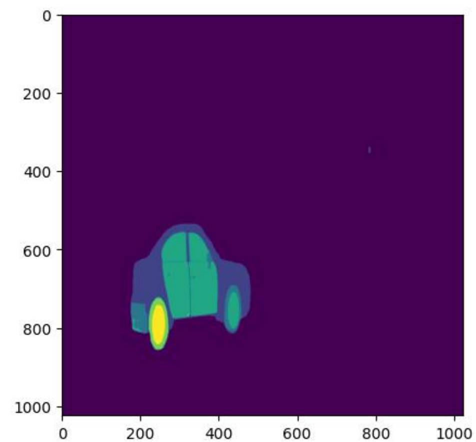
Cluster Analysis of SAM Image Embeddings

Results

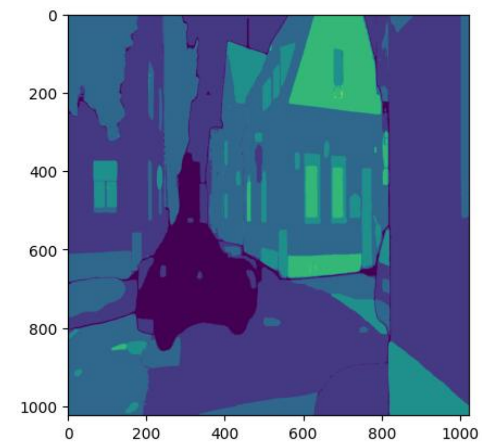
Clustering of SAM Embeddings

Masks generated by SAM for a sample KITTI image





Cluster 1: Car (Foreground)

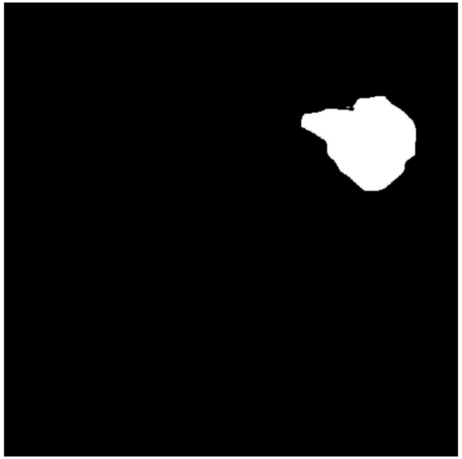


Cluster 2: Buildings, Road (Background)

Results

SAM’s poor zero-shot performance on medical images

Breast cancer - malignant tumor



Expected semantic segmentation

Semantic segmentation masks generated by SAM

Conclusion

Summary

- DINOv2, CLIP, and SAM, has provided valuable insights into the realms of dimensionality reduction, patch matching, and image clustering.
- The embeddings generated by CLIP effectively distinguish between different image categories, whereas the embeddings produced by SAM group similar categories together.
- The utilization of DINOv2 facilitated detailed intra-patch and inter-patch analyses, showcasing its effectiveness in identifying significant features within images.
- SAM's cluster analysis shed light on the quality of image embeddings, demonstrating their capability to form cohesive clusters corresponding to similar objects.
- Visual foundation models play a crucial role in advancing the capabilities of computer vision systems.

Acknowledgement

A BIG Thanks

- We would like to thank the Professors for their informative lectures and notes.
- We would also like to thank all our tutors for their valuable support and guidance.

References

1. Alec Radford, et al. Learning transferable visual models from natural language supervision, 2021.
2. Maxime Oquab, et al. Dinov2: Learning robust visual features without supervision, 2023.
3. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Doll´ar, and Ross Girshick. Segment anything, 2023.
4. Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).
5. Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, Apr 2022.
6. Andreas Geiger, et al. Vision meets Robotics: The KITTI Dataset, 2013.
7. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Nov 2019.

Universität Stuttgart

Contact

Vielen Dank!

Ashwin Murali @ ashwin.cse18@gmail.com

M.Sc Computer Science (Autonomous Systems)

Namrata Jangid @ namratarjangid@gmail.com

M.Sc Computer Science (Autonomous Systems)