Harvard John A. Paulson School of Engineering and Applied Sciences

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

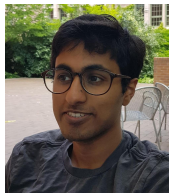
Google Research

Matryoshka Representation Learning

NeurIPS 2022

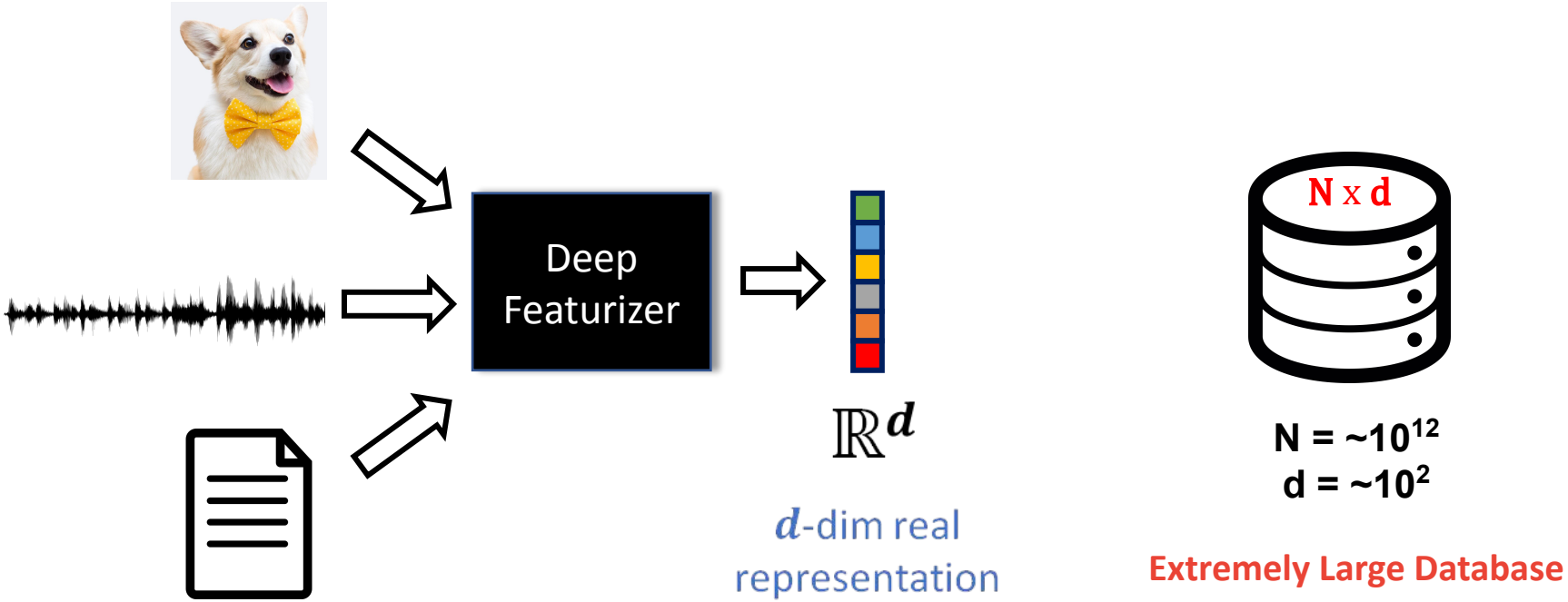

Aditya Kusupati



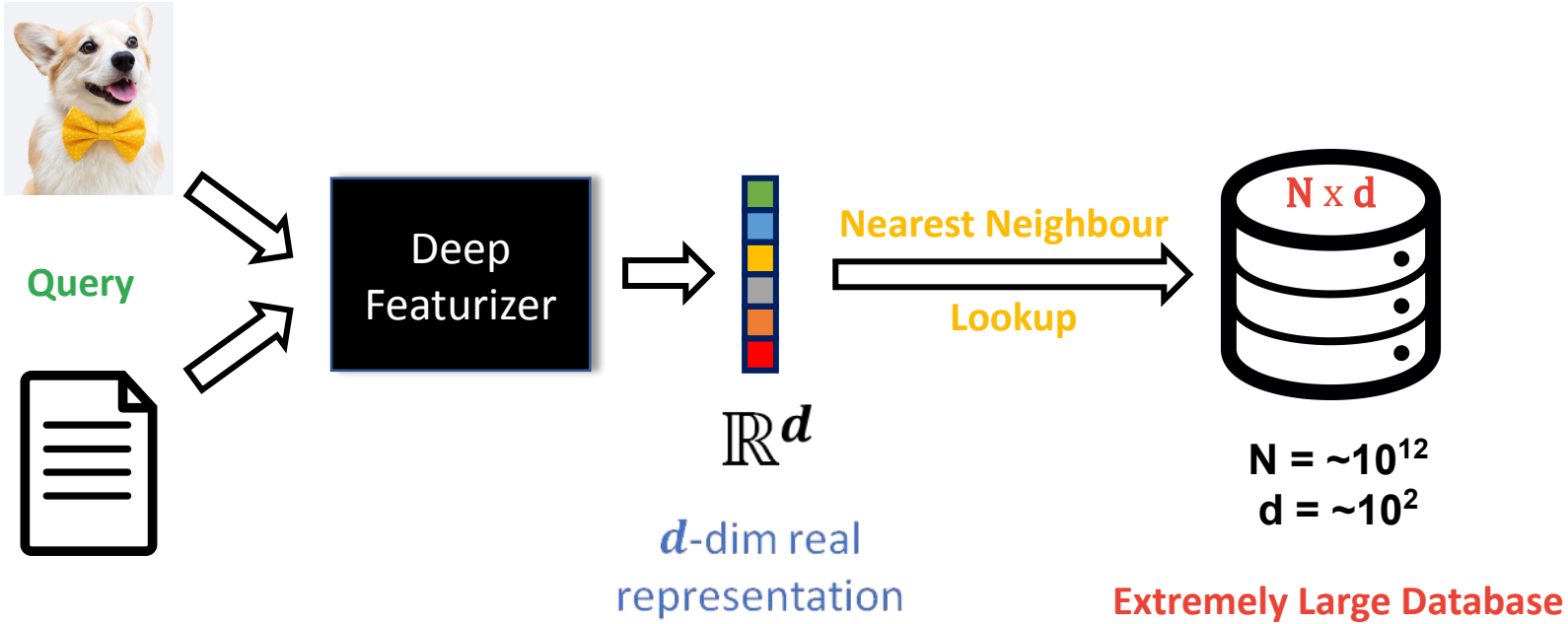




Learned Representations



Motivation: Query-based Retrieval



- Applicable for large-scale classification with millions of labels

Web-scale Challenges

- Extremely large databases – **100s of TB**
  - Linear dependence on representation size (*d*)
  - Embedding look-up much more expensive than featurization
- Require Approximate Nearest Neighbour Search (ANNS)
  - **Post-hoc** compressed index
- **Incapable** of Multi-Granularity
  - Use same *high-d* embedding for all tasks
  - Retrain a model for *low-d* based on deployment constraints
  - Eg: *768-d* ViT image representation for all tasks

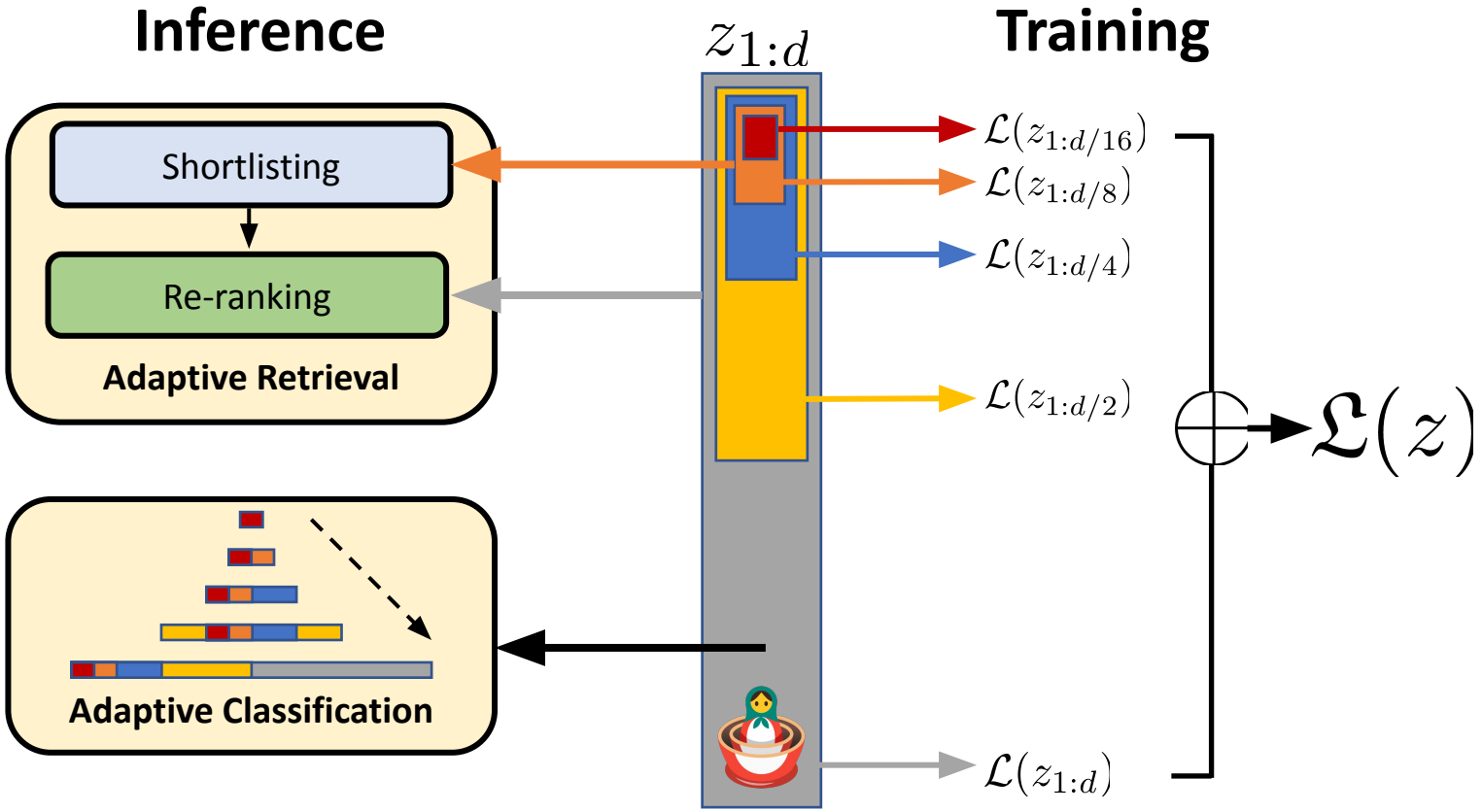
Adaptive Deployment – Goals

- **One representation** vector for all downstream tasks
  - No post-hoc compression or expensive feature selection
  - No retraining for specific resource constraints

- Accurate and efficient *low-d* embeddings
  - Baked within the *high-d* embedding – **Free**
  - **Reduced costs** for expensive & high-recall shortlisting
  - As **accurate** as independently trained counterparts

- *High-d* embedding for **cheap** & precise re-ranking

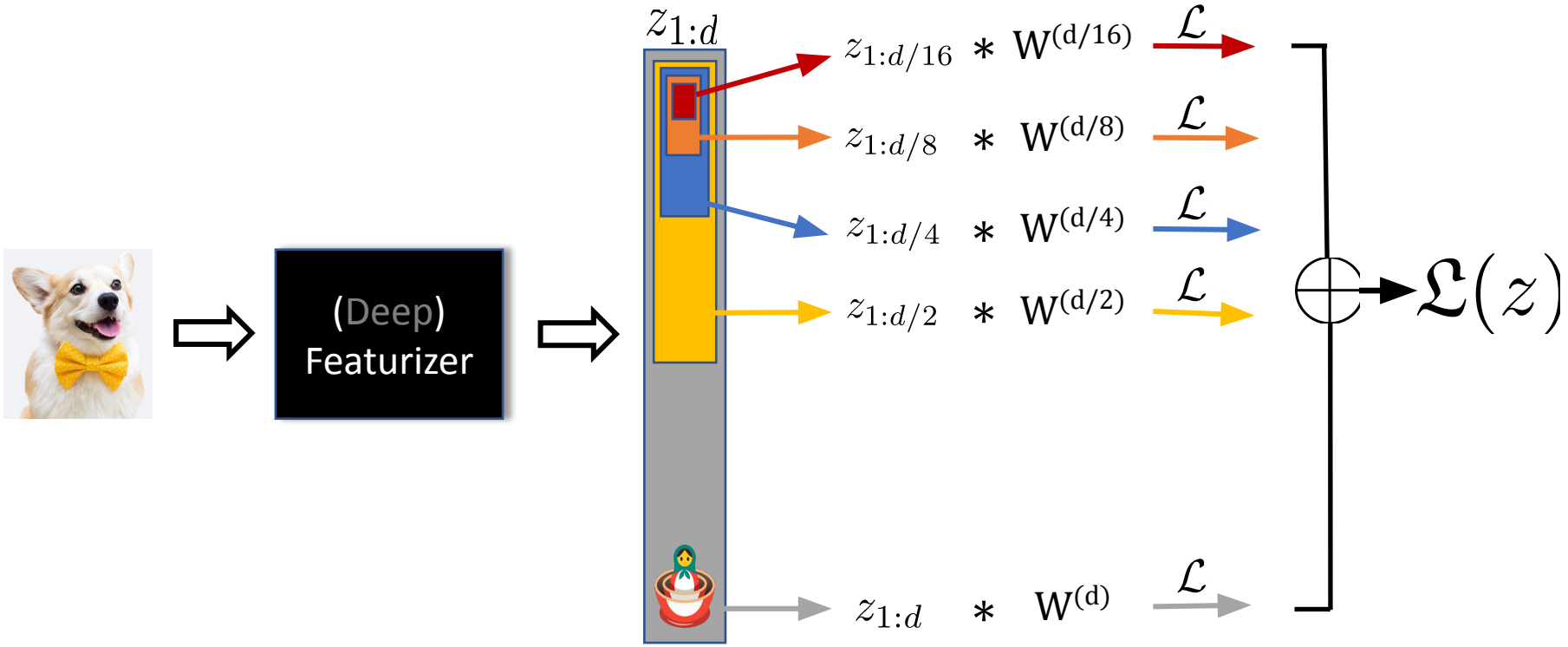
Matryoshka Representation Learning - MRL



🪆 Matryoshka Representation Learning - MRL

- Solve the same learning task at **multiple granularities** ($log(d)$)
- Easily adaptable to any representation learning setup
  - Scale, modality and task agnostic **– 1B images with ease**
- **First *k* dims** form the required *low-d* embeddings
  - As accurate as retrained *low-d* counterparts
- Enable **adaptive** deployment
  - Accurate large-scale classification and retrieval
  - Based on task and resource constraints

MRL: Supervised Learning



Applications & Experiments

- **Classification**
  - Representation quality – data, scale and modalities
  - Adaptive classification

- **Retrieval**
  - Adaptive Retrieval

- **Analysis & Ablations**
  - Robustness, few-shot and long-tail learning
  - Analysis across representation sizes and ablations

Classification

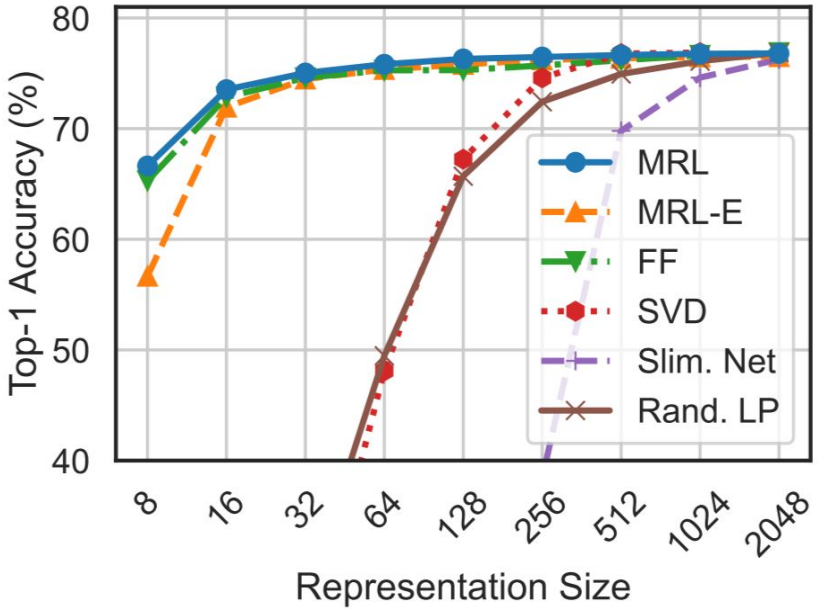
- **Models + Data**
  - ResNet50 + ImageNet-1K
  - ViT-B/16 + JFT-300M
  - ALIGN: VIT-B/16 (V) + BERT-Base

- **Evaluation** on ImageNet-1K validation set
  - One-vs-All (*OVA*) top-1 & 1-Nearest Neighbour (*1-NN*) accuracy (%)
  - Interpolation between granularities

- **Adaptive Classification**
  - MRL classification with cascades across dimensions

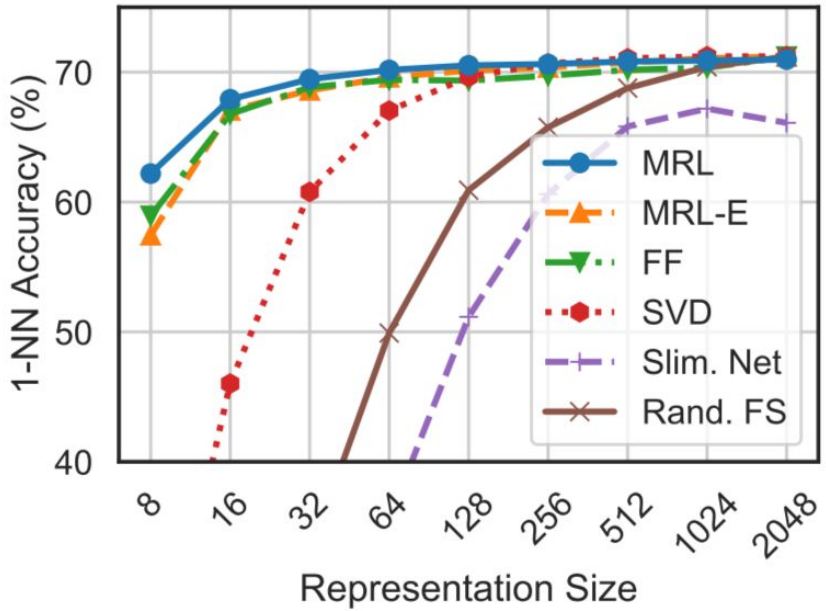
Classification Accuracy - *ImageNet* OVA

- ResNet50 models trained on ImageNet-1K
- Same accuracy as independently trained *low-d* models (FF)



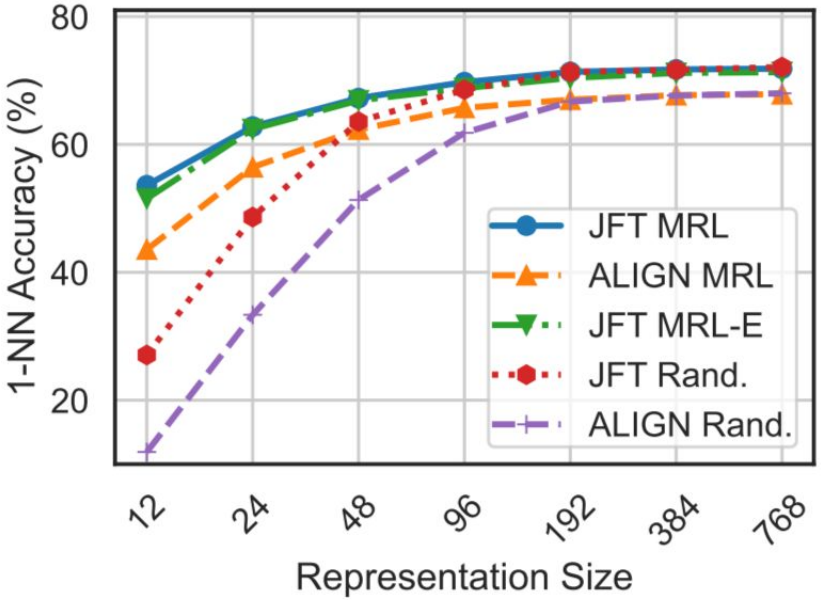
Representation Quality - *ImageNet* k-NN

- ResNet50 models trained on ImageNet-1K
- Other baselines fall off drastically at *low-dimensions*



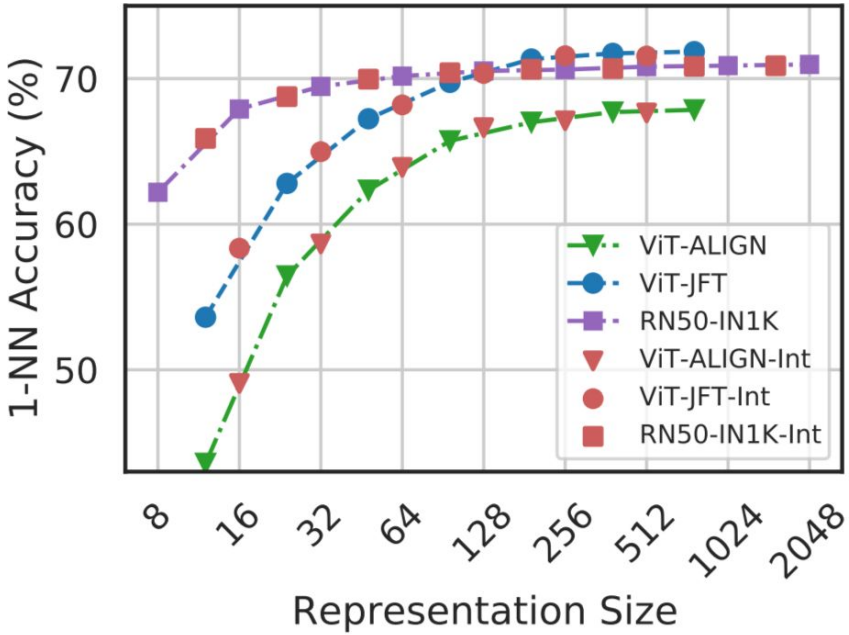
Representation Quality - *ImageNet* k-NN

- ViT-B/16 models trained on JFT-300M and ALIGN (V+L)
- Scales to **1B images** w/o accuracy drop



Granularity Interpolation - *ImageNet* k-NN

- MRL models interpolate for intermediate dimensions
- Allows for extremely fine-grained deployment

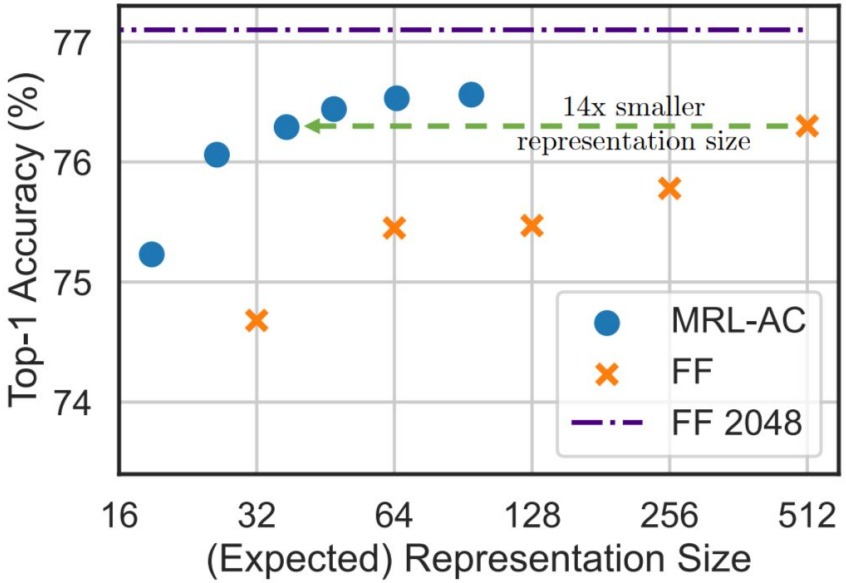


Adaptive Classification

- Cascade the classification of an instance
  - Increasing granularity within the Matryoshka Representation
- *Adaptively* choose the **confidence thresholds**
- Extremely useful for extremely-large output spaces
- Same accuracy as an highest capacity trained model

Adaptive Classification - *ImageNet-1K*

- ResNet50-MRL model trained on ImageNet-1K with cascades
- 14x smaller embedding size for same accuracy



Retrieval

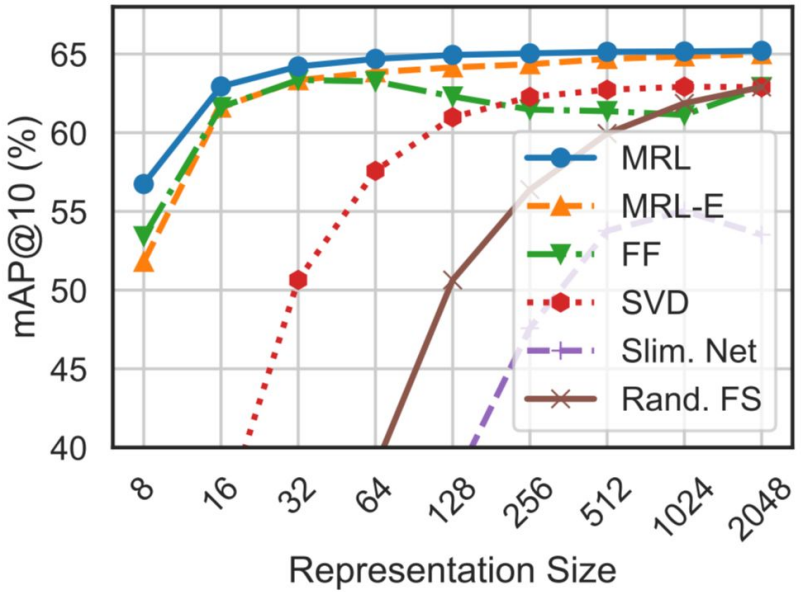
- Query image ⟶ Most relevant images from database
  - Same class / label
  - Evaluated using mean Average Precision @ k (k = 10, *mAP@10*)

- ResNet50 models trained on ImageNet-1K

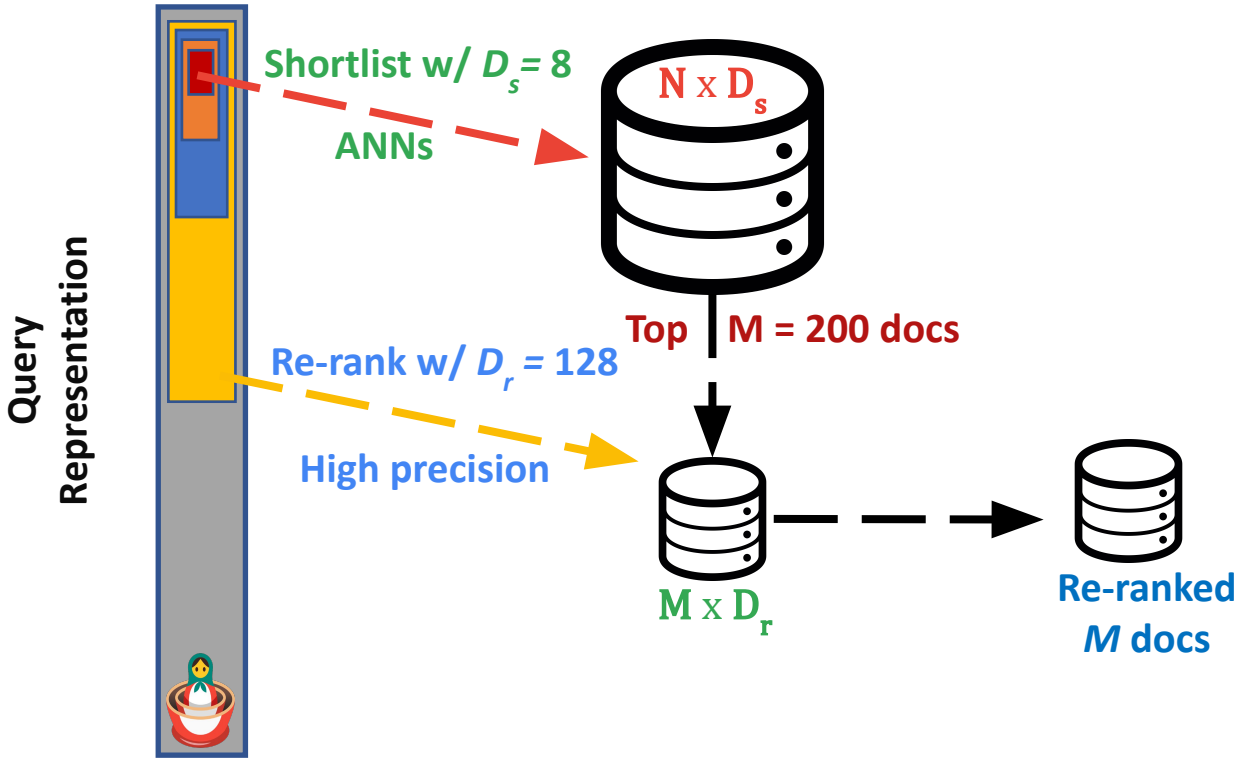
- Datasets
  - ImageNet-1K: ~1.3M database & 50K query set – 2.6 GFLOPs
  - ImageNet-4K (New!!): ~4.2M database & 200K query set – 8.6 GFLOPs
    - Publicly available & subset of ImageNet-21K w/o ImageNet-1K overlap

Retrieval mAP@10 – *ImageNet-1K*

- Better mAP@10 as independently trained *low-d* models (FF)
- Similar recall for *low-d* representation as *high-d*



Adaptive Retrieval

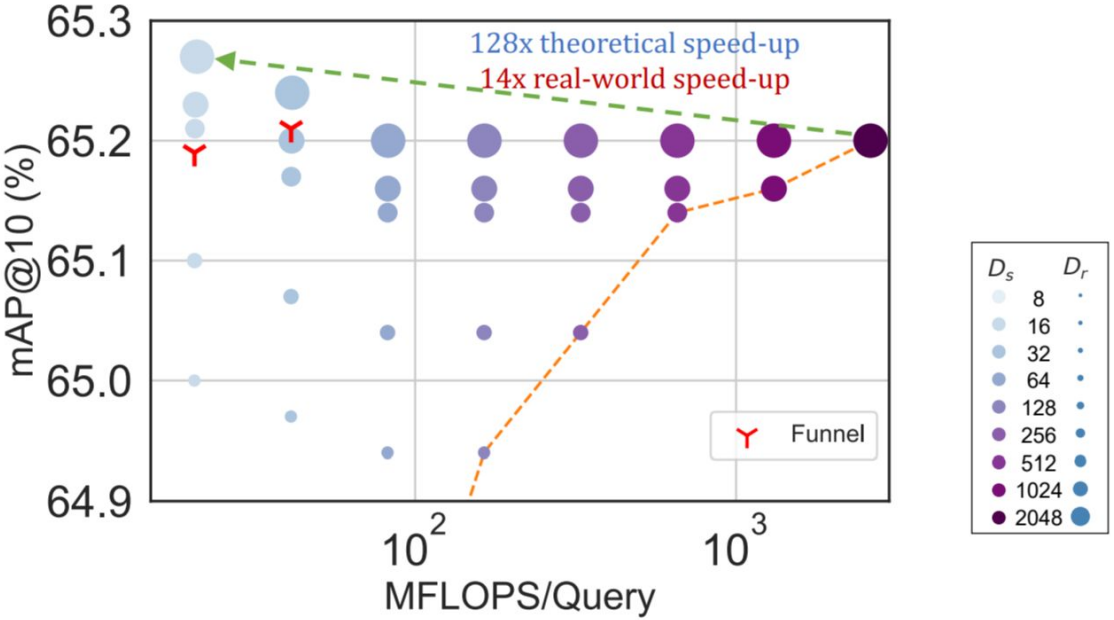


Adaptive Retrieval

- Replace single-shot retrieval with highest dimension
- Shortlist with *M = 200* data points with *low-d* ($\mathbf{D}_s$)
  - Re-rank with *high-d* ($\mathbf{D}_r$)
- Match the maximum mAP@10 from original embedding
- Significantly lower FLOPs and inference time

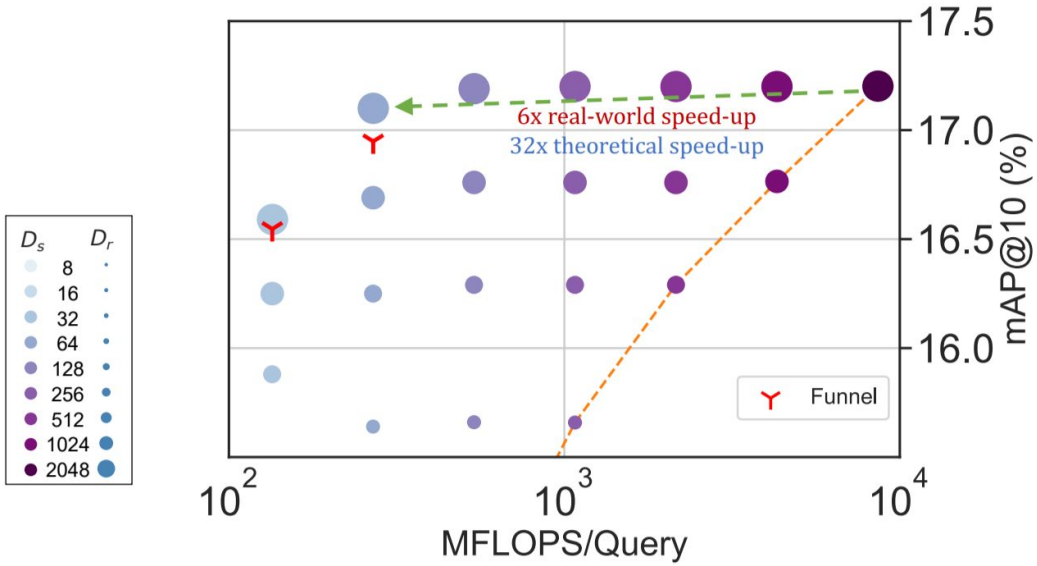
Adaptive Retrieval - *ImageNet-1K*

- 14x real-world speed-up for the best mAP@10
- All real-world implementations use HNSW for shortlisting



Adaptive Retrieval - *ImageNet-4K* (Try it!!)

- 6x real-world speed-up for the best mAP@10
- Funnel retrieval alleviates the need for optimal D_s & D_r



Robustness

- **As robust** as independently trained *low-d* (FF) models
  - ImageNetV2/R/A/Sketch classification accuracy
  - For all representation sizes of MRL across models

- MRL models have **more robust retrieval** vs FF models
  - Up to 3% better mAP@10 for ImageNetV2 query set (10K samples)

- MRL **improves** cosine similarity span for ALIGN
  - Increases span between positive and random image-text pairs

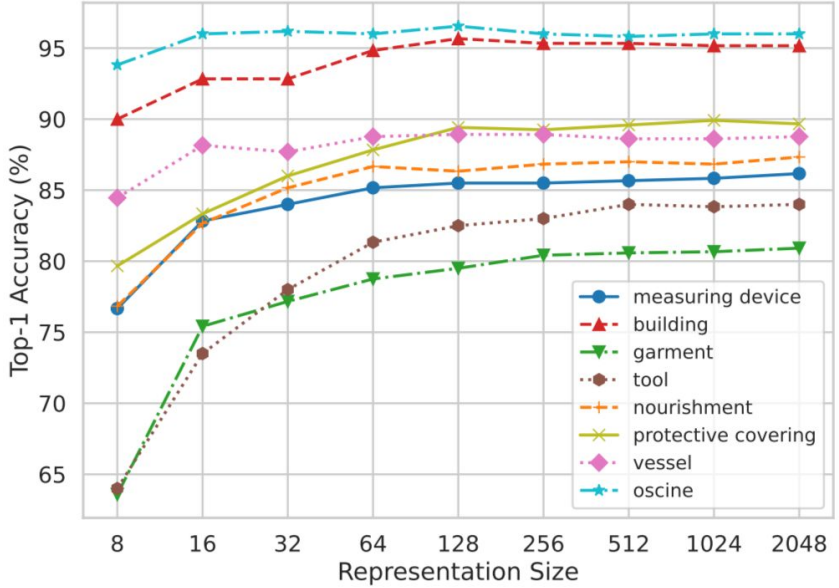
Few-shot and Long-tail Learning

- MRL representations perform **comparably** to FF
  - On ImageNetV2 using Nearest Class Mean (NCM)
  - Across varying shots and number of classes

- Long-tail & sequential evaluation on FLUID
  - **Up to 2%** higher accuracy on novel classes in the tail
  - *Low-d* as **accurate** as *high-d* for pretrain classes in head & torso
  - *high-d* required to **differentiate** the low-shot classes

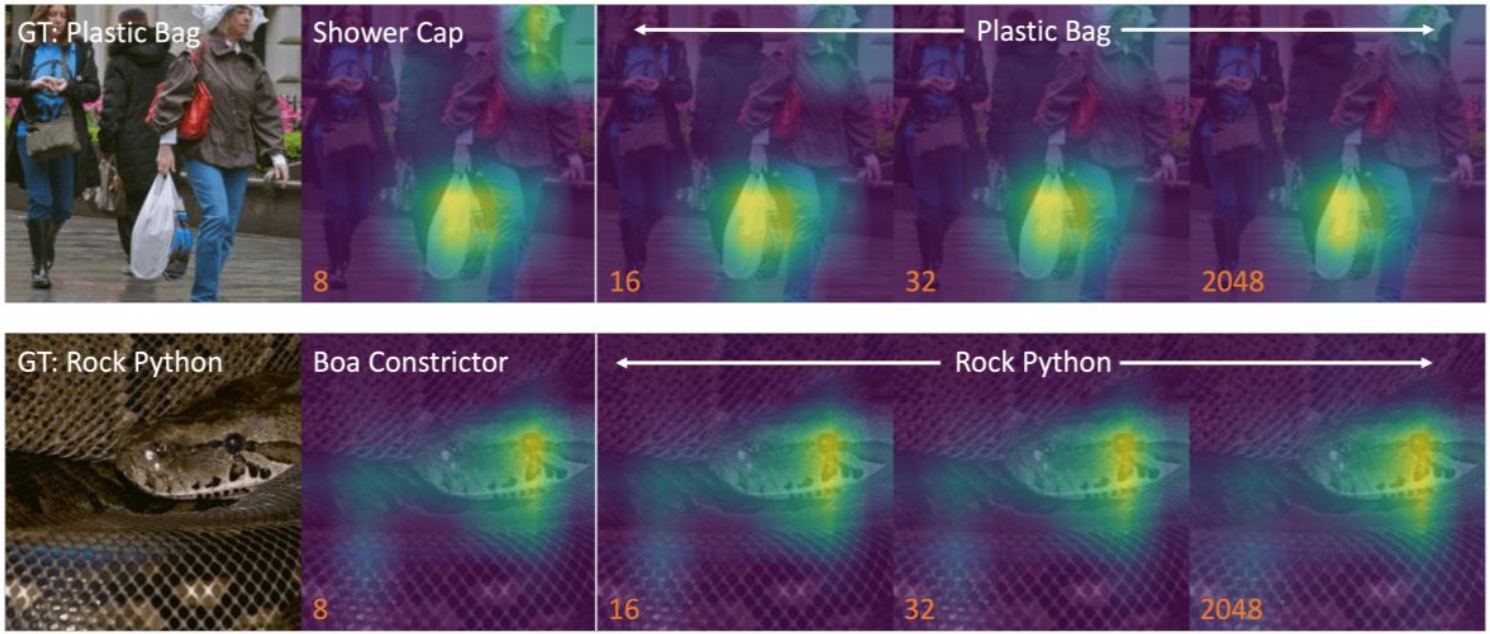
Analysis across Representation Sizes

- Per-class accuracy often increases – barring some exceptions
  - Super-class accuracy more consistent
- Routing leveraging disagreement ⟶ improves accuracy by 4.6%



Analysis across Representation Sizes

- Graceful failing of *low-d* representations – clutter & super-class



Ablations

- Case for tuning relative loss weights
- Boosting 8-d by **2x** & 16-d by **1.5x** improves low-d accuracies

Model	MRL		MRL-8boost		MRL-8+16boost	
Rep. Size	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
8	66.63	84.66	**69.53**	86.19	69.24	85.96
16	73.53	89.52	73.86	89.44	**73.91**	89.55
32	75.03	91.31	**75.28**	91.21	75.10	91.14
64	75.82	92.27	**75.84**	92.22	75.67	92.06
128	**76.30**	92.82	76.28	92.74	76.07	92.52
256	76.47	93.02	**76.48**	92.97	76.22	92.72
512	**76.65**	93.13	76.56	93.09	76.35	92.85
1024	**76.76**	93.22	76.71	93.21	76.39	92.98
2048	**76.80**	93.32	76.76	93.28	76.52	93.05

Ablations

- MRL for pretrained models – fine tuning ResNet50 on ImageNet-1K

Rep. Size	4.2 conv3, fc	4.2 conv3, conv2, fc	4.2 full, fc	fc	All (MRL)
8	36.11	54.78	60.02	5.15	66.63
16	58.42	67.26	70.10	13.79	73.53
32	67.81	71.62	72.84	32.52	75.03
64	72.42	73.61	74.29	52.66	75.82
128	74.41	74.67	75.03	64.60	76.30
256	75.30	75.23	75.38	69.29	76.47
512	75.96	75.47	75.64	70.51	76.65
1024	76.18	75.70	75.75	70.19	76.76
2048	76.44	75.96	75.97	69.72	76.80

Future Work

- Optimizing the relative loss weightings
  - **Pareto** optimal accuracy vs-efficiency trade-off
- Specific losses across fidelities for adaptive deployment
  - e.g. high recall for 8-dimension and robustness for 2048-dimension
- Learning a search data-structure, like differentiable k-d tree
  - w/ Matryoshka Representation to enable **task & cost aware** retrieval
- Joint optimization of multiobjective MRL + end-to-end learnable search data-structure – **data-driven adaptive web-scale retrieval**

Conclusions

github.com/RAIVNLab/MRL
adityakusupati.com

- **MRL**  : A general purpose method adaptable to any representation learning setup to obtain flexible representations
  - Scale, task and modality agnostic

- Multi-granularity databases for free at web-scale
  - *Low-d* database – learned index fitting in RAM
  - Complementary to **ANNS** and learned indices like **LLC**

- Up to 14× faster yet accurate web-scale classification & retrieval

- Framework for analyzing information bottlenecks