**Motivation: Query-based Retrieval**

- Solve the same learning task at multiple granularities (log(d))
  - Scale, modality and task agnostic \( \L(\z) \)
- First \( d \) dims form the required low-d embeddings
  - As accurate as retrained low-d counterparts
- Enable adaptive deployment
  - Accurate large-scale classification & retrieval based on constraints

**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

**Classification Accuracy**

- **ImageNet OVA**
  - ResNet50: Same accuracy as independently trained low-d models (FF)
- **ImageNet-1K**
  - ResNet50 models trained on ImageNet-1K
  - Other baselines fall off drastically at low-dimensions
- **ImageNet-4K (Try it!)**
  - 6x real-world speed-up for the best mAP@10
  - Funnel retrieval alleviates the need for optimal \( D_s \) & \( D_r \)

**Representation Quality**

- **ImageNet k-NN**
  - \( D_s = 8 \) achieved high precision
- **ViT+JFT & ALIGN k-NN**
  - Scales to \( 1B \) images w/o accuracy drop

**Training**

- \( \L(\z) \)
  - \( \z \) = cascades
  - \( \L(\z) \)
  - \( \z \) = adaptive learning

**Inference**

- **Shortlisting**
  - \( \L(\z) \)
  - \( \z \) = cascades
- **Re-ranking**
  - \( \L(\z) \)
  - \( \z \) = cascades