Improving Image Recognition by Retrieving from Web-Scale Image-Text Data

Ahmet Iscen, Alireza Fathi, Cordelia Schmid

Google Research

THU-AM-268

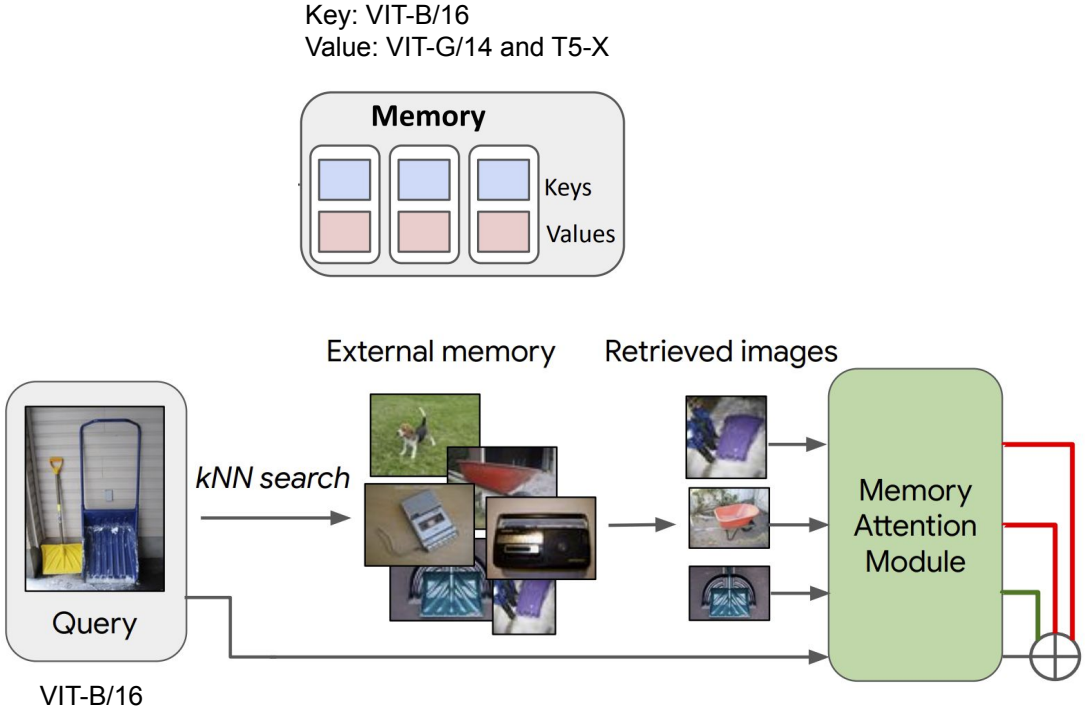
Google Research

Motivation

- Long-tailed recognition
  - Some classes are frequently seen others are rarely seen
- External memory usage
  - **No need to memorize rare cases in the model parameters**
- Learn to filter out irrelevant memory items with a light-weight model

Google Research

Idea



Google Research

Comparison

			ImageNet-LT				Places-LT			
Method	Retrieval	Backbone	Many-shot	Mid-shot	Low-shot	All	Many-shot	Mid-shot	Low-shot	All
BASELINES										
Linear Classifier		ViT-B16 🔒	76.5	72.6	66.5	73.5	44.5	44.4	44.0	44.3
MLP Classifier		ViT-B16 🔒	80.1	74.1	66.9	75.2	48.6	46.1	41.3	46.0
Mean k-NN	✓	ViT-B16 🔒	75.9	75.8	**75.7**	75.8	44.3	45.2	45.5	44.9
EXISTING METHODS										
PaCo [10]		ResNext-101	68.2	58.7	41.0	60.0	36.1	47.9	35.3	41.2
VL-LTR [43]		ViT-B16	84.5	74.6	59.3	77.2	**54.2**	48.5	42.0	50.1
RAC [29]	✓	ViT-B16	-	-	-	-	48.7	48.3	41.8	47.2
RAC† [29]	✓	ViT-B16 🔒	80.9	76.0	67.5	76.7	50.3	48.3	42.5	47.9
RAC† [29]	✓	ViT-B16	**85.9**	79.3	69.3	80.5	51.9	49.8	46.8	50.0
Ours	✓	ViT-B16 🔒	80.6	77.5	74.5	78.3	50.9	49.9	47.5	49.9
Ours + FT	✓	ViT-B16	85.4	**81.5**	**76.4**	**82.3**	52.4	**52.0**	**48.5**	**51.4**

Google Research

Improving Image Recognition by Retrieving from Web-Scale Image-Text Data

Ahmet Iscen, Alireza Fathi, Cordelia Schmid

Google Research

THU-AM-268

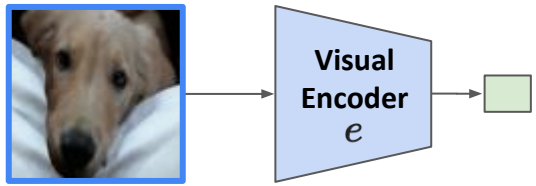
Google Research

Motivation

- Long-tailed recognition
  - Some classes are frequently seen others are rarely seen
- External memory usage
  - **No need to memorize rare cases in the model parameters**
- Learn to filter out irrelevant memory items with a light-weight model

Google Research

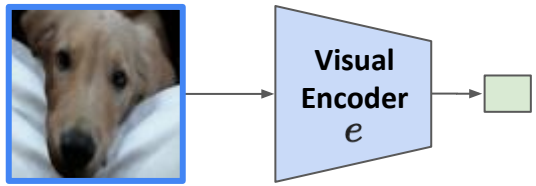
Approach



Query embedding $\mathbf{z}_i \in \mathbb{R}^d$

Google Research

Approach

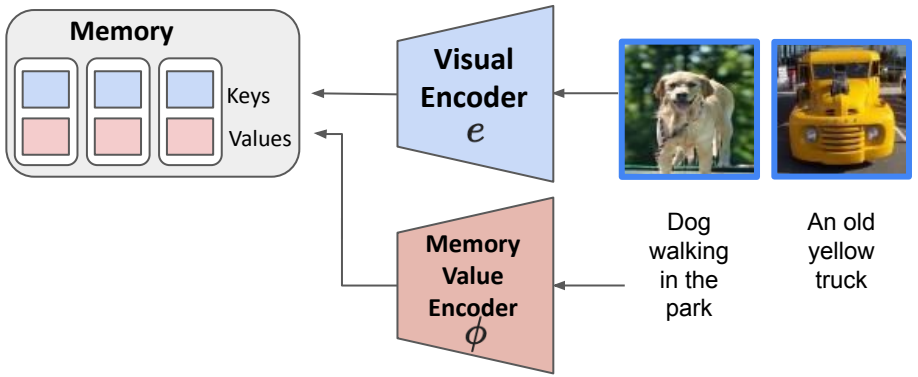


Query embedding $\mathbf{z}_i \in \mathbb{R}^d$

Memory key $\mathbf{m}_j \in \mathbb{R}^d$

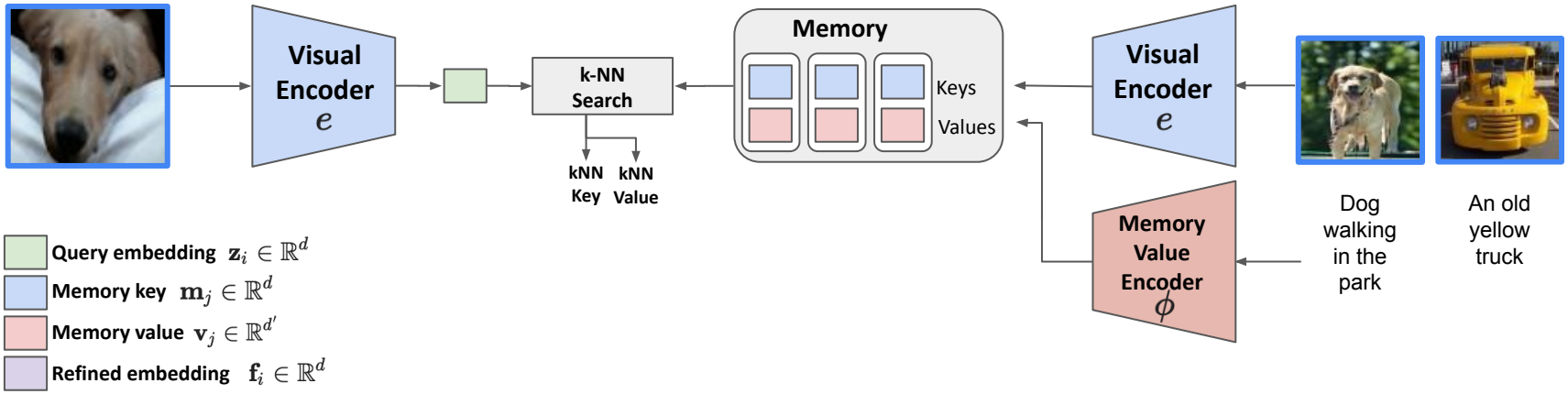
Memory value $\mathbf{v}_j \in \mathbb{R}^{d'}$

Refined embedding $\mathbf{f}_i \in \mathbb{R}^d$



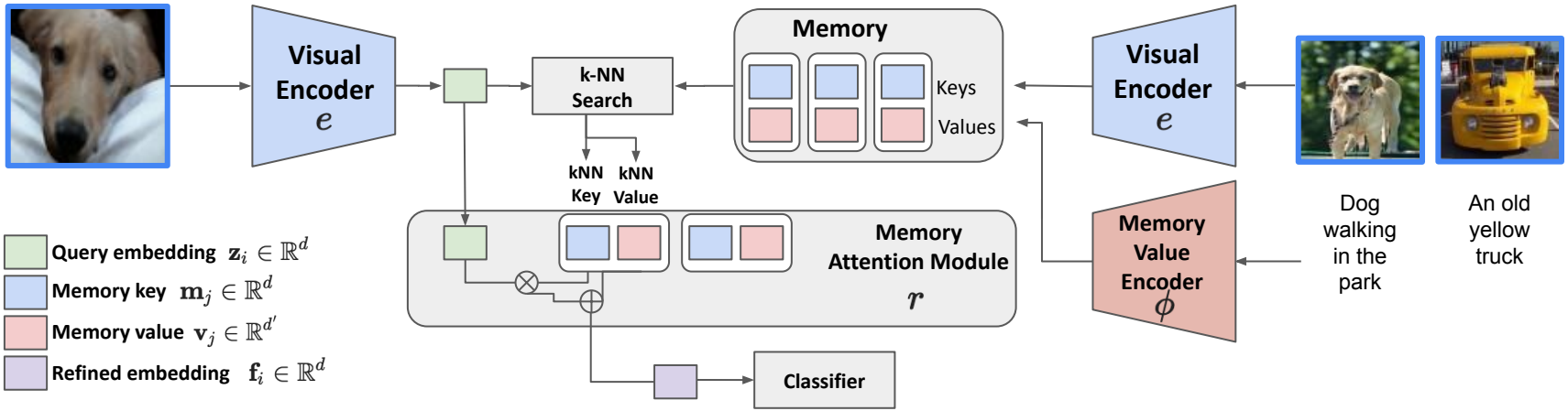
Google Research

Approach



Google Research

Approach



Google Research

Experimental Datasets

ImageNet-LT Dataset
Long-Tailed Image Recognition. 1000 classes, number of training images per class varies from 5 to 1280.

Places-LT Dataset
Long-Tailed Image Recognition. 365 classes, number of training images per class varies from 5 to 4980.

WebVision Dataset
Learning with Noisy Labels. 2.4M images and 1000 classes with noisy labels.

iNaturalist21-Mini Dataset
Fine-grained classification. 10,000 classes and 50 images per class.

Google Research

Memory Sources

WebLI Dataset
1B image + text pairs

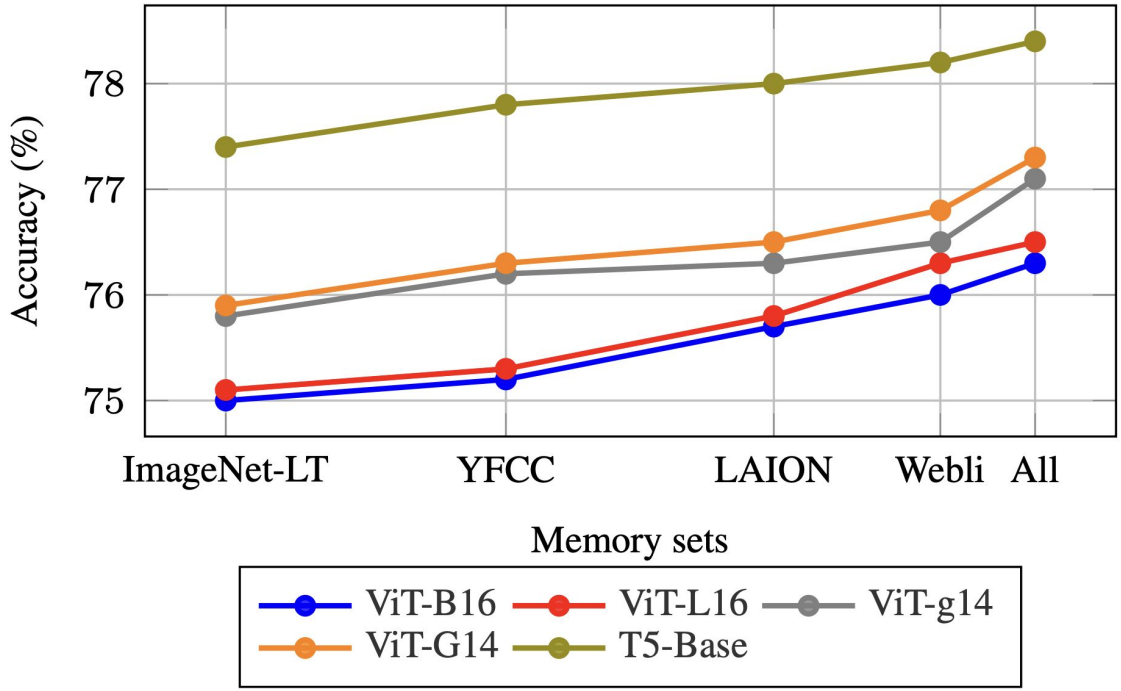
LAION Dataset
400M image + text pairs

YFCC100m Dataset
We use a 15m subset containing image + text

Training Dataset
Few hundred thousands image + label.

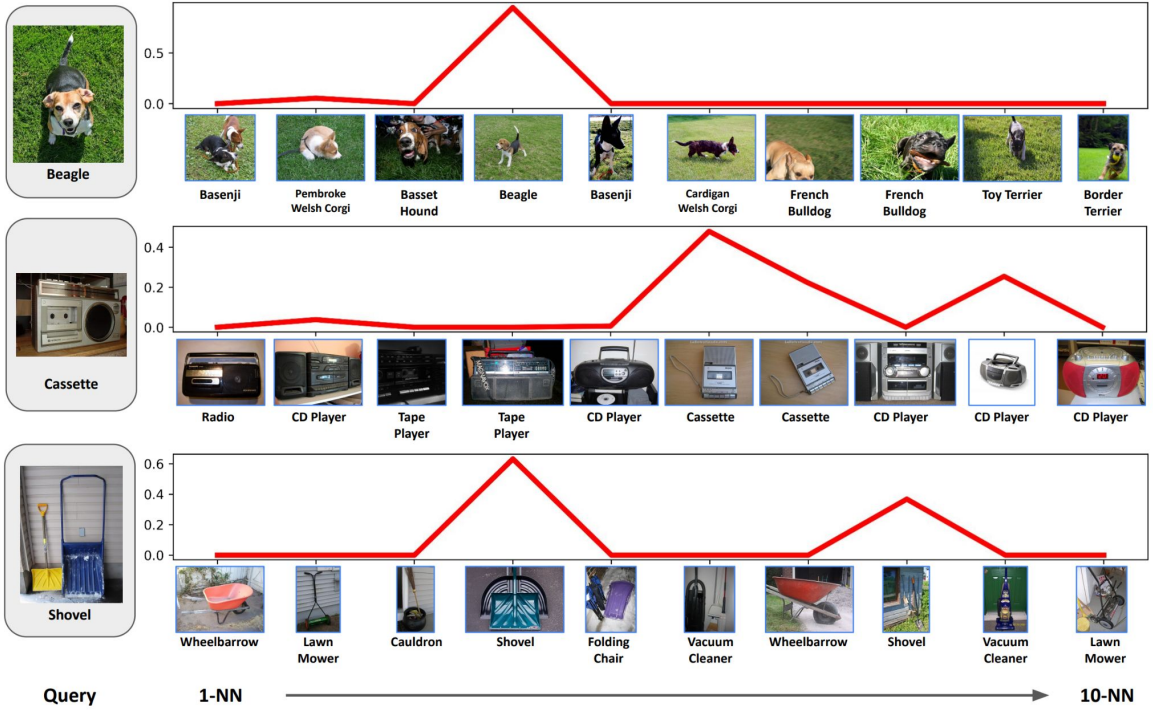
Google Research

Long-Tailed Image Recognition (ImageNet LT)



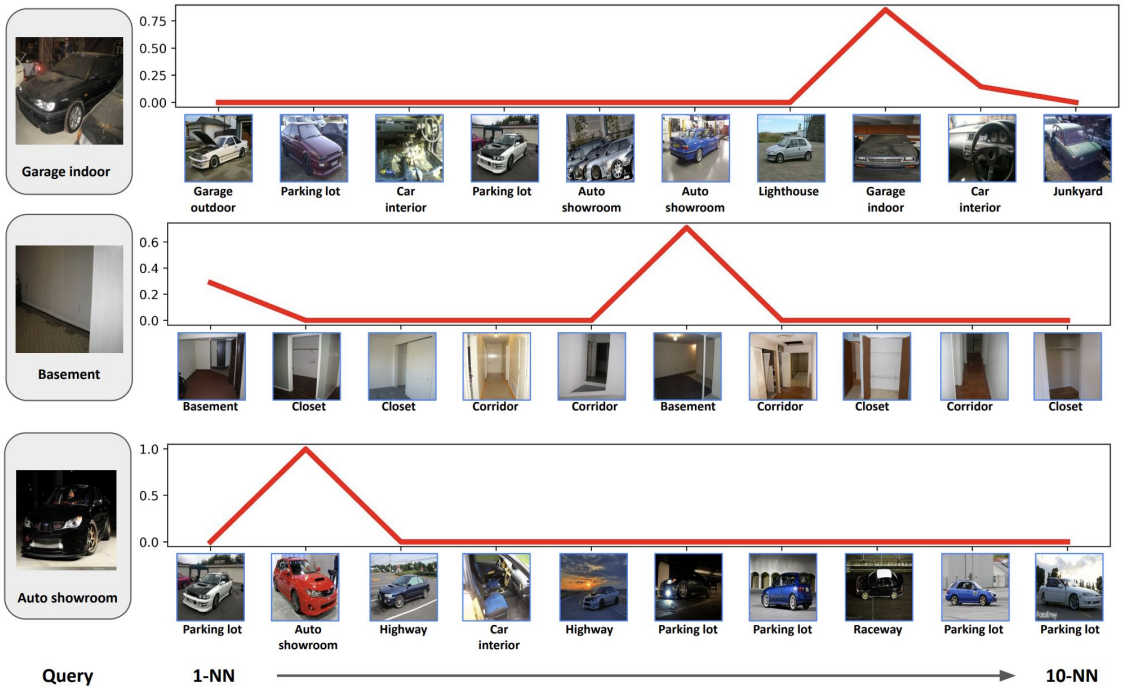
Google Research

Visualizing Attention Scores



Google Research

Visualizing Attention Scores



Google Research

Comparison

			ImageNet-LT				Places-LT			
Method	Retrieval	Backbone	Many-shot	Mid-shot	Low-shot	All	Many-shot	Mid-shot	Low-shot	All
BASELINES										
Linear Classifier		ViT-B16 🔒	76.5	72.6	66.5	73.5	44.5	44.4	44.0	44.3
MLP Classifier		ViT-B16 🔒	80.1	74.1	66.9	75.2	48.6	46.1	41.3	46.0
Mean k-NN	✓	ViT-B16 🔒	75.9	75.8	**75.7**	75.8	44.3	45.2	45.5	44.9
EXISTING METHODS										
PaCo [10]		ResNext-101	68.2	58.7	41.0	60.0	36.1	47.9	35.3	41.2
VL-LTR [43]		ViT-B16	84.5	74.6	59.3	77.2	**54.2**	48.5	42.0	50.1
RAC [29]	✓	ViT-B16	-	-	-	-	48.7	48.3	41.8	47.2
RAC† [29]	✓	ViT-B16 🔒	80.9	76.0	67.5	76.7	50.3	48.3	42.5	47.9
RAC† [29]	✓	ViT-B16	**85.9**	79.3	69.3	80.5	51.9	49.8	46.8	50.0
Ours	✓	ViT-B16 🔒	80.6	77.5	74.5	78.3	50.9	49.9	47.5	49.9
Ours + FT	✓	ViT-B16	85.4	**81.5**	**76.4**	**82.3**	52.4	**52.0**	**48.5**	**51.4**

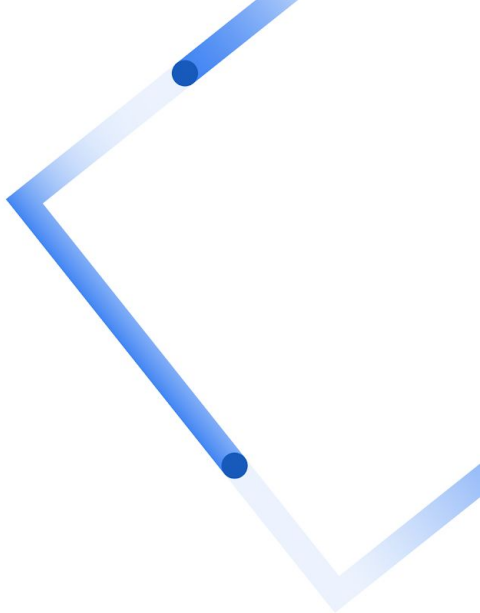
Google Research

Generalization to other tasks

	iNat2021-Mini	WebVision
BASELINES		
Linear Classifier	58.8	78.1
MLP Classifier	59.6	81.0
Mean k-NN	58.9	78.2
EXISTING METHODS		
MILe [36]	–	75.2
Heteroscedastic [9]	–	76.6
NCR [23]	–	76.8
CurrNet [17]	–	79.3
Ours	**66.2**	**83.6**

Google Research

Thank you



Google Research