– Supplementary Material – QuEST: Quantized Embedding Space for Transferring Knowledge

Himalaya Jain[1], Spyros Gidaris[1], Nikos Komodakis[2,3], Patrick Pérez[1], and Matthieu Cord[1,4]

[1]Valeo.ai [2]University of Crete [3]LIGM [4]Sorbonne University

A On using cosine similarity in assignment predictor

In the assignment predictor we use cosine similarity to predict the K-dimensional soft-assignment vector $\boldsymbol{p}_S^{(h,w)}$ from the feature vector $\boldsymbol{f}_S^{(h,w)}$ (Eq. 5). The reason for choosing this similarity measure over the Euclidean distance is that the former L2-normalizes the features and the visual word weights, which we observed to lead to better behavior. We believe that this is due to the fact that the L2-normalization acts as a regularizer for the weights of the visual words in the assignment predictor in case of unbalanced k-means clusters: without the L2-normalization, more frequent teacher-words would lead to bigger weight magnitudes for the corresponding student-words. Also, with cosine similarity, it is easier to control the peakiness of the predicted word distribution since the range of its output values is fixed and a priori known (i.e., between -1 and 1).

B Implementation details for vector quantization

In our quantization-based distillation method we use k-means to learn the visual teacher-words vocabulary V. Here we provide implementation details regarding how we apply the k-means clustering algorithm.

k-means implementation. For k-means, we use the implementation provided by the publicly available FAISS [3] library.

Applying k-means on ImageNet. The training set of ImageNet is quite large (i.e., it has around 1.28M images). Therefore, when evaluating our distillation method on it, to learn efficiently the visual teacher-words vocabulary V, we apply k-means only to a randomly sampled subset of 0.2M images from this set. Given the spatial size of feature maps, this subset provides a sufficiently large corpus of vectors to learn a vocabulary of size $K = 4096$, as we use in our experiments.

Applying k-means on CIFAR-100 and CIFAR-10. For CIFAR-100 and CIFAR-10 experiments, we apply k-means to the entire training sets.

C Training details

C.1 Model compression

For the ImageNet experiments in Section 4.1, following [4] and [2], we train for 100 epochs with the initial learning rate of 0.1 which is reduced every 30 epochs with a decay rate of 0.1. For ResNet34 to ResNet18, we use a batch size of 256, while for ResNet50 to MobileNet we use 210 as batch size due to GPU memory constraints. For the hyper-parameters of our distillation method, we use $\alpha = 1$, $\beta = 1$, $\tau = 0.2$ and $K = 4096$.

For all the experiments on CIFAR-100, we follow the protocol of [4] for training the student networks. Specifically, in all cases we train the student for 240 epochs with batch size of 64 and an initial learning rate of 0.05, which we drop by a factor of 0.1 after 150, 180, and 210 epochs. The only exception is MobileNetV2 and ShuffleNetV1/V2, as in [4], where the learning rate is initialized to 0.01. The hyper-parameters of our method are $\alpha = 1$, $\beta = 1$, $\tau = 0.2$ and $K = 4096$.

For the CIFAR-10 experiments we follow the protocol of [5] and train the student for 200 epochs with a batch size of 128. The initial learning rate is set to 0.1 which decays by a factor of 0.2 at 60^{th}, 120^{th}, 160^{th} epoch. The hyper-parameters of our distillation method are set to $\alpha = 1$, $\beta = 1$, $\tau = 0.005$ and $K = 256$.

C.2 Transfer learning to small-sized datasets

For the transfer learning experiments in Section 4.2 of the main paper, we used the hyper-parameters $\alpha = 1$, $K = 4096$, and τ equal to 0.2 and 0.002 for layer4 and layer3 of ResNet34 respectively (for both layers the τ value was chosen so that, as mentioned in the main paper, the softmax probability for the closest visual teacher-word is on average around 0.996). We found that in the transfer learning experiments it is important to tune properly the β hyper-parameter so as to prevent overfitting on the classification task of the training images. To that end, as it is recommended in the evaluation protocol of [1], we used 20% of the training images as validation images and we tuned the β hyper-parameter on them. Specifically, for the ResNet18 student experiments we used $\beta = 10.0$ and for the VGG9 student experiments we used $\beta = 20.0$. To train the students we follow the training protocol of [1], i.e., 200 training epochs with an initial learning rate of 0.05 which is dropped by a factor of 10 after 150 epochs. The batch size is 128 for the ResNet18 student and 32 for the VGG9 student.

D Model compression in CIFAR-100 experiments

Table 1 gives the number of parameters of all the networks used in Tables 2 and 3 for CIFAR-100 experiments. The table shows that we can get high reduction in parameters with significantly less drop in performance with our proposed method of knowledge distillation. In case of WRN-40-2 to WRN-16-2 and to ShuffleNetV1, we even get an improvement of 0.49% and 1.14% in accuracy over the teacher with compression of 68.81% and 57.91% respectively.

Table 1: Number of parameters of the teacher and student networks used in CIFAR-100 experiments and compression obtained by replacing the teacher with the student network. The compression is computed as percentage of reduction in number of parameters with respect to the teacher network

Model			Accuracy		
Teacher network	Student network	compression (%)	teacher	student	Ours
WRN-40-2 (2.26M)	WRN-16-2 (0.70M)	68.81	75.61	73.26	76.10
WRN-40-2 (2.26M)	WRN-40-1 (0.57M)	74.73	75.61	71.98	74.58
ResNet56 (0.86M)	ResNet20 (0.28M)	67.70	72.34	69.06	71.84
ResNet110 (1.73M)	ResNet20 (0.28M)	83.97	74.31	69.06	71.89
ResNet110 (1.73M)	ResNet32 (0.47M)	72.78	74.31	71.14	74.08
ResNet32x4 (7.43M)	ResNet8x4 (1.23M)	83.41	79.42	72.50	75.88
VGG13 (9.46M)	VGG8 (3.96M)	58.10	74.64	70.36	73.81
VGG13 (9.46M)	MobileNetV2 (0.81M)	91.41	74.64	64.6	68.79
ResNet50 (23.7M)	MobileNetV2 (0.81M)	96.57	79.34	64.6	69.81
ResNet50 (23.7M)	VGG8 (3.96M)	83.27	79.34	70.36	75.17
ResNet32x4 (7.43M)	ShuffleNetV1 (0.95M)	87.23	79.42	70.50	76.28
ResNet32x4 (7.43M)	ShuffleNetV2 (1.36M)	81.77	79.42	71.82	77.09
WRN-40-2 (2.26M)	ShuffleNetV1 (0.95M)	57.91	75.61	70.50	76.75

E Effect of vocabulary size on CIFAR-10 experiments

In Section 4.3 of the main paper, we discussed the impact of teacher vocabulary size K, based on the plot of accuracy versus K for CIFAR-100 in Figure 3(b). Here we provide the analogous plot for CIFAR-10 in Figure 1. As we mentioned in the paper, K between 128 to 256 leads to better performance.

F Qualitative results: Alignment of teacher and student quantized feature maps

As we claim in Section 3.3 of the main paper, our distillation loss enforces a quantization of the student feature maps into visual words that is in accordance to the quantization at the teacher side. As defined in paper's Section 3.2, we compute the soft-assignment maps $\boldsymbol{p}_T$ and $\boldsymbol{p}_S$ from $\boldsymbol{f}_T$ and $\boldsymbol{f}_S$ respectively (see equations (4) and (5) of main paper). In Figure 2, we illustrate here the alignment of the soft-assignment maps by providing image retrieval results where the query is represented by the teacher soft-assignment map $\boldsymbol{p}_T$ while each database image is represented by the student soft-assignment map $\boldsymbol{p}_S$. To compute the similarity, we flatten $\boldsymbol{p}_T$ and $\boldsymbol{p}_S$ from $K \times H \times W$-sized tensors (where $H \times W$ are the common spatial dimensions of the teacher and student networks and K is the vocabulary size) to KHW-dimensional vectors and then compute the dot product of the two vectors. We see that we manage to retrieve semantically and structurally similar images, which means that the two representations $\boldsymbol{p}_T$ and $\boldsymbol{p}_S$ match well.

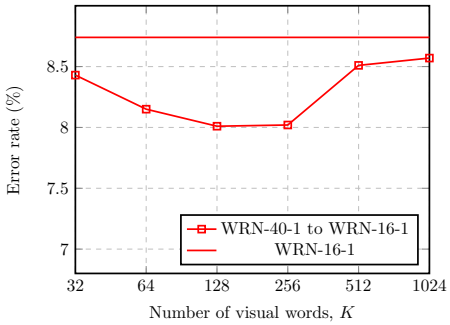

Fig. 1: **Effect of varying** K. Error rate vs. K on CIFAR-10 with WRN-16-1 as the student networks. The students are trained on the proposed distillation loss with WRN-40-1 as the teacher with varying number K of visual teacher-words. The solid straight line represents student trained without distillation loss

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 2
2. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE International Conference on Computer Vision (2019) 2
3. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734 (2017) 1
4. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020) 2
5. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision (2019) 2

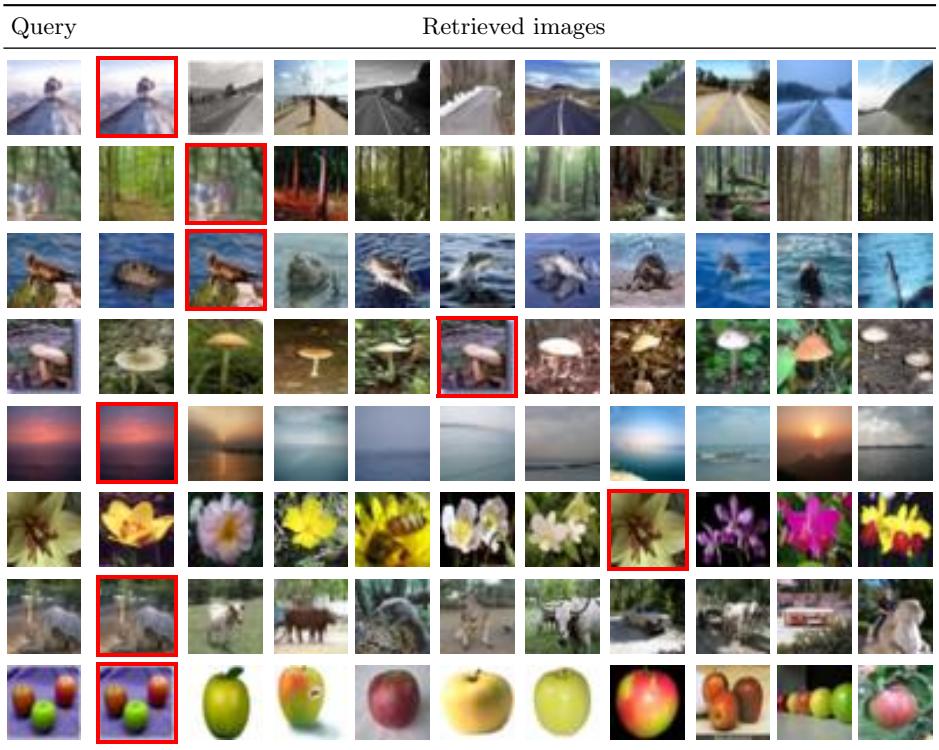

Fig. 2: **Image retrieval in the quantized embedding spaces**. For the query image we used the quantized features of a WRN-40-2 teacher network and for the database images we used the predicted quantized features of a WRN-16-2 student network trained with our distillation method. As database we used the $10K$ images of CIFAR-100 test set, and as queries we used randomly sampled images from this database. The figure shows the query on the left-most column and top-10 retrieved images (in that order) next to the query. We see that, as top results, we always retrieve the query itself (framed with red box) as well as other semantically and structurally similar images. This indicates that the two quantized embedding spaces are well aligned