Table 3. Number of parameters used in practice.

MODEL	NUMBER OF PARAMETERS
MNIST-500-500-100	1,441K
MNIST-530-530-100	1,559K
MNIST-500-500-100-MEM	1,550K
OCR-letters-200-200-50	164K
OCR-letters-200-200-50-MEM	208K

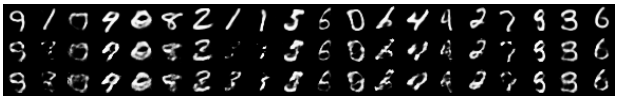


Figure 6. Generation from MEM-VAE when memory is disabled.

Table 4. Average preference values of selected slots.

"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
0.27	0.82	0.33	0.11	0.34	0.15	0.49	0.27	0.09	0.28
0.24	0.09	0.06	0.11	0.30	0.13	0.12	0.27	0.09	0.21
0.18	0.05	0.06	0.11	0.07	0.07	0.05	0.11	0.09	0.18

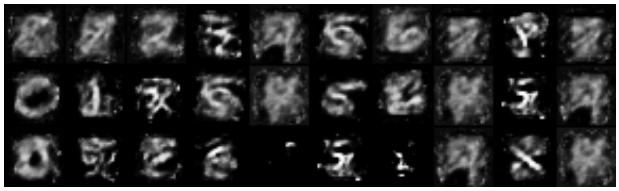


Figure 7. Visualization of selected slots.

A. Recognition Model

We feed the output of the last deterministic layer in the recognition models into a linear SVM to classify the MNIST digits to examine the invariance in features. We achieve slightly better classification accuracy (97.90% for VAE and 98.03% MEM-VAE), which means that additional memory mechanisms do not hurt or even improve the invariance of the features extracted by the recognition model.

B. Number of Parameters Used

As we employ external memory and attention mechanisms, the number of parameters in our building block is larger than that in a standard layer. However, the total number of parameters in the whole model is controlled given a limited number of slots in the memory. See Table 3 for a comparison, and we do not observe that our method suffers from overfitting.

C. Disabling Memory

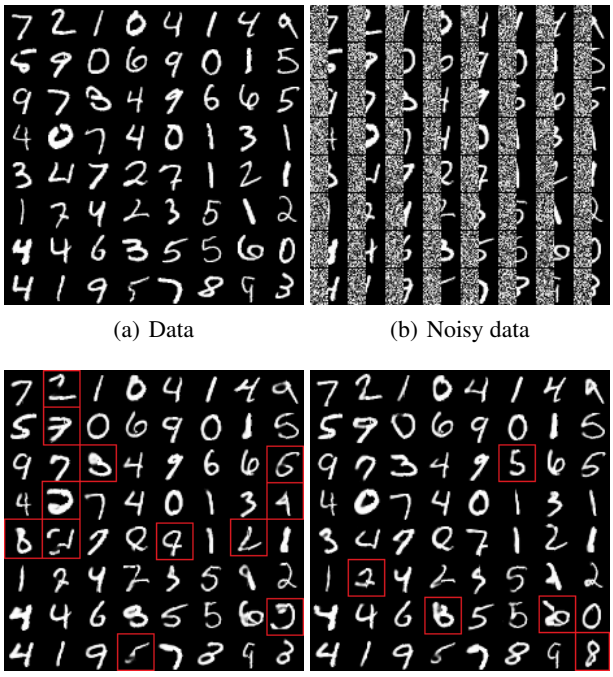
We investigate the performance of MEM-VAE when the memory is disabled (setting $\mathbf{h}_m$ as a vector filled with ones) as in Figure 6. The top row shows original samples; the middle row shows samples with memory of the first layer disabled; and the bottom row shows samples with memory of both layers disabled. It can be seen that, without information from memory, the main pattern of the generation does not change much but the local details are lost in some sense, which supports our assumption.

D. Visualizing the Memory Slots Directly

As mentioned in Section 5, MEM-VAE employs the sigmoid function and element-wise MLP as the attention and composition functions respectively. MEM-VAE is complicated and flexible, which has better quantitative results on log-likelihood estimation but is hard to be visualized directly as it has much nonlinearity. We introduce VIS-MEM-VAE, which uses the softmax function and element-wise summation as the attention and composition functions respectively. VIS-MEM-VAE achieves a slightly worse density estimation result (-84.68 nats on MNIST, still better than -85.67 nats of VAE) but the memory slots can be mapped to data level for visualization due to the simple composition function and sparse attention mechanism.

We average the preference values $\mathbf{h}_a$ of the test data and select top-3 memory slots that has the highest activations for each class. Note that the activations are normalized, i.e., the summation of the preference values on all memory slots equals to one for each class. We set the generative information to be a vector filled with zeros and the memory information to be one of the selected slots and then generate a corresponding image. The activations and images are shown in Table 4 and Figure 7, where each column represents a class (0-9 in left-right order). For example, the image at the first column and the second row in Figure 7 corresponds to the memory slot that has the second-highest averaged activation of class "0" and the value is 0.24.

It can be seen that most of the selected slots respond to one class or similar classes (some slots are shared by similar classes such as"4", "7" and "9") and the corresponding image contains a blurry sketch of the digit (or mixture of some digits) with different local styles, which indicates that the external memories can encode local variants of objects and can be retrieved based on generative information $\mathbf{h}_g$. A few images are less meaningful but the corresponding activations are relatively small (smaller than 0.08 or so).



(a) Data (b) Noisy data

(c) Results of VAE (d) Results of MEM-VAE

Figure 8. (a-b): Original test data on MNIST and perturbed data with left half missing respectively; (c-d): Imputation results of VAE and MEM-VAE respectively. Red rectangles mean that the corresponding model fails to infer digits correctly but the competitor succeeds.

E. Missing Value Imputation

We visualize the images recovered by VAE and MEM-VAE in Figure 8 given incomplete test data. It can be seen that MEM-VAE has better visualization than VAE with fewer meaningless images, clearer digits and more accurate inference.