SUPERVISED ENCODING FOR DISCRETE REPRESENTATION LEARNING

Speaker: Cat P. Le

Cat P. Le*, Yi Zhou†, Jie Ding‡, Vahid Tarokh*

* Department of Electrical and Computer Engineering, Duke University

† Department of Electrical and Computer Engineering, University of Utah

‡ School of Statistics, University of Minnesota

Outline

- Introduction
- Supervised-Encoding Quantizer (SEQ)
  1. Encoder
  2. Quantizer
  3. Decoder
- Experimental Results
- Conclusion

Duke
UNIVERSITY

Introduction

- Most of the deep learning models do not provide sufficient **interpretability** for the learned features.

- This framework typically produces a black-box model for the classification task.
  - For instance, the network model only learn what it had been asked to classify.
  - The model don't provide any *visualization of features* or interpretable decision rules.

Introduction (cont')

- The objectives of proposed model:
  - Understanding the data not only by their labels but also their features
  - Providing interpretable graph of features
  - Generating new data (generative model)

Supervised-Encoding Quantizer

- The supervised-encoding quantizer (SEQ) model consists of an encoder, a quantizer, and a decoder.
  - This model is inspired by the autoencoder structure.
- The quantizer serves as a clustering mechanism in the feature space of the autoencoder.
- The encoder is a traditional convolutional neural network.
- The decoder is also a convolutional neural network, whose structure reflects encoder's.

Supervised-Encoding Quantizer

- **Encoding**: we pre-train the encoder by attaching a Softmax layer to its output and train the encoder via standard supervised training;
- **Quantization**: the encoded features produced by the pre-trained encoder are passed to the quantizer for clustering
- **Decoding**: we further train a decoder that can reconstruct the original data samples from the encoded features.

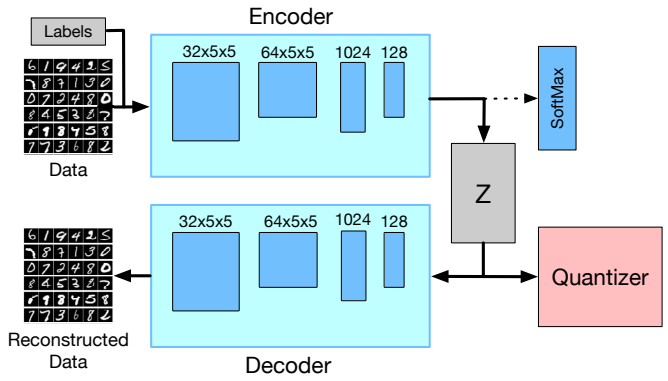


Figure 1. The architecture of supervised-encoding quantizer

Encoder

- To obtain meaningful embedding features, we pre-train the encoder with labeled data.
- (Encoder + Softmax) is similar to a traditional convolutional neural network and can be trained with cross-entropy loss and stochastic gradient descend.
- Remove the Softmax layer after training, and the output of encoder is guaranteed to be linearly separable

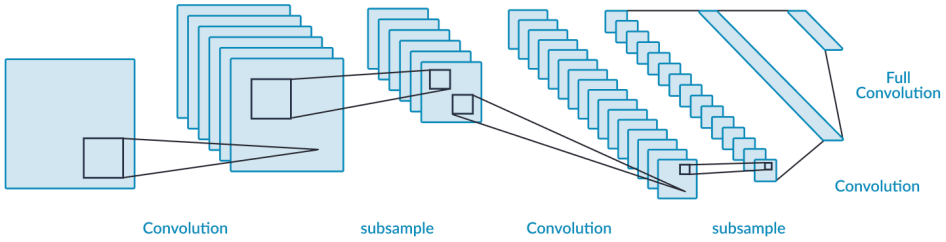


Figure 2. Convolutional Neural Network Architecture from: https://missinglink.ai/guides/convolutional-neural-networks/convolutional-neural-network-architecture-forging-pathways-future/

Quantizer

- Apply quantizer to the feature space (output of the encoder).
  - Quantization techniques: k-means, vector quantization, self-organizing map, grow-when-required network.
- By choosing the quantized clusters greater than the total number of class label, the quantizer can identify the sub-classes within each class of data.
  - The accuracy of the quantizer is upper bounded by that of the encoder.

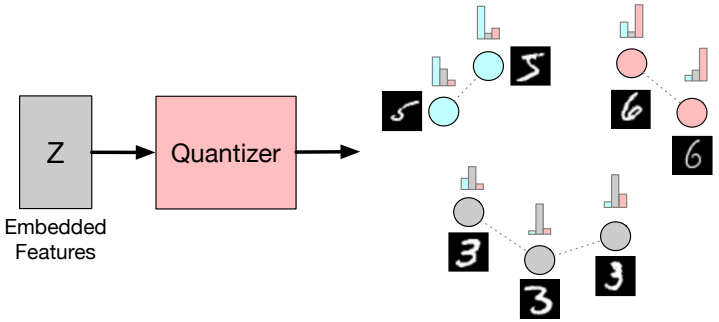


Figure 3. The mechanism of quantizer in SEQ

Decoder

- Apply decoder to the output of the encoder in order to reconstruct the data from the embedded features.
- To train the decoder, we fix all the parameters of the encoder and apply the MSE loss to measure the reconstruction error on the training data samples:

$$\mathcal{L}(\theta) := \|X - \mathcal{D}_\theta(\mathbf{sg}[\mathcal{E}(X)])\|^2$$

- Training decoder here is like training the autoencoder with the encoder part is fixed.

Experimental Results

- SEQ performs well compared with other clustering methods due to the semi-supervised pre-trained encoder.

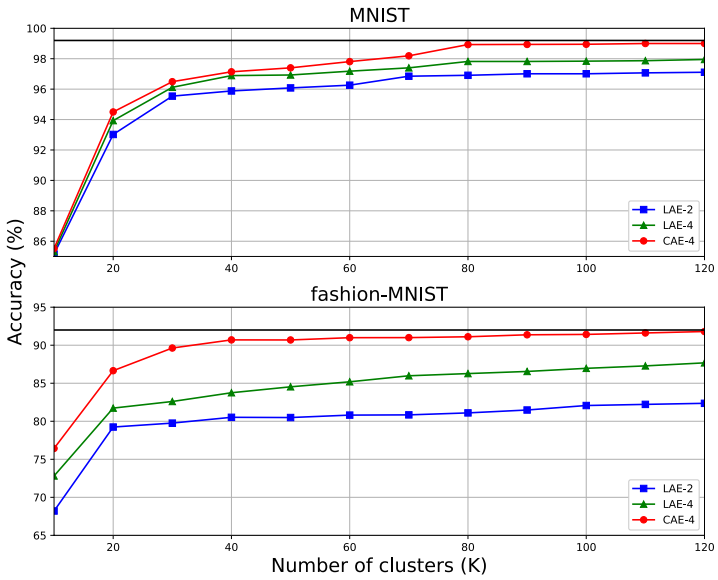


Table 1. The clustering performance on MNIST

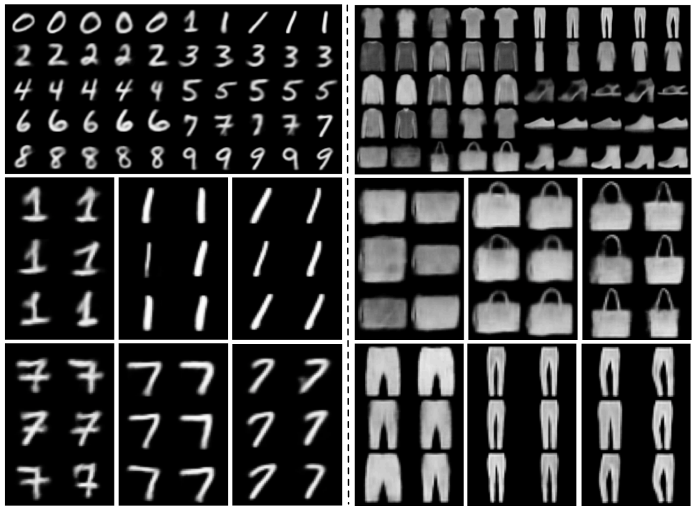
DEC	IDEC	DCEC	CAE-l_2 k-means	SEQ k-means
86.55	88.06	88.97	95.11	**99.74 (0.046)**

Figure 4. Performance of several SEQ model on MNIST (upper) and fashion-MNIST (lower)

Duke UNIVERSITY

Experimental Results (cont')

- It preserves the interpretability of features.
- The interpretable representation of images from a same clusterhave a similar style.



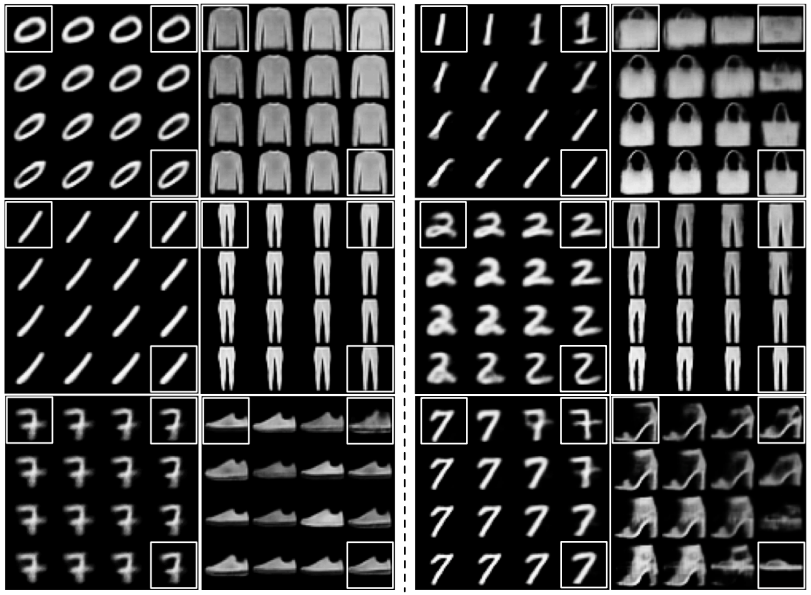
(a) MNIST (b) fashion-MNIST

Figure 5. Reconstructed images from small sub-classes in MNIST (left) and fashion-MNIST(right)

Duke UNIVERSITY

Experimental Results (cont')

- It can be used to generate specific type of data within a specific class.



(a) Similar style (b) Different styles

Figure 6. Generative images from similar type (left) and different type (right) of data in MNIST and fashion-MNIST

Duke UNIVERSITY

Conclusion

- SEQ shows that the feature space contains useful information
- SEQ can learn more about the data while maintain the meaningful of its features via the class label.
- The interpretability of feature can be used to further classify data into subclasses of different styles or generative purposes.

Duke
UNIVERSITY

Thank you