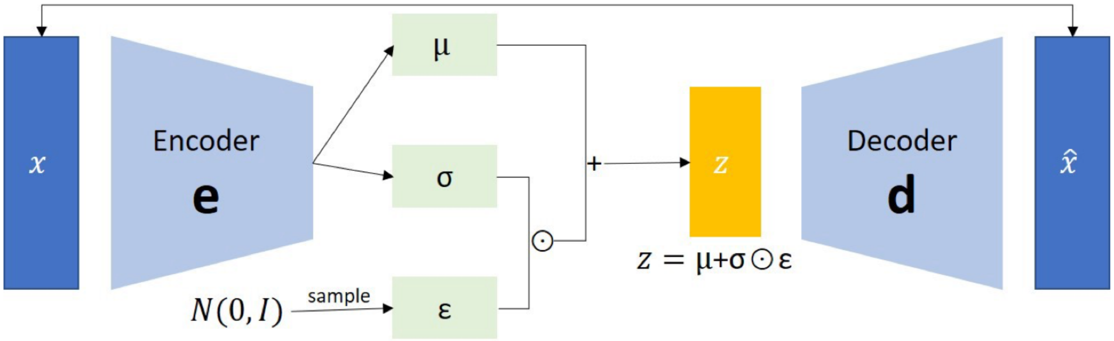
MSSE 277B: Machine Learning Algorithms
Homework assignment #9: VAE and GNN
Assigned Apr. 9 and Due Apr. 19

1. **Variational Autoencoder(VAE) applied to MNIST dataset.** (10 pt)
We will implement a VAE model for the MNIST dataset. The encoder and decoder of the VAE model are convolutional neural networks. The VAE model will be trained to reproduce (reconstruct) the images.
(1) (2 pt) Use the provided code to load MNIST dataset and normalize the data by dividing the maximum value.
(2) (5 pt) Implement a VAE model. The encoder will have 4 convolutional layers, each with 4, 8, 16, 32 channels, kernel size of 4x4, padding of 1 and stride of 2. The decoder is the reverse of that. In the bottleneck region, the encoder output is flattened and mapped to two latent vectors μ and σ each represented with 32 hidden neurons by two separate linear layers. Then the latent state z with 32 hidden neurons is formulated by applying reparameterization with addition of noise ε, which is then passed to decoder.



(3) (3 pt) Use binary cross entropy (BCE) plus KL divergence (KLD) as your loss function.

$$L(X, \hat{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \mathrm{BCE}(X, \hat{X}) + \frac{1}{2}\sum_{i=1}^{32}(\mu_i^2 + \sigma_i^2 - \ln \sigma^2 - 1)$$

Train this model with the MNIST dataset and use the provided reconstruction code to show that your model is able to reproduce the images.

2. **Predicting Molecular Enthalpy of Formation with GNNs. (10 pt)** QM9 is a dataset of over 130,000 molecules consisting of 9 heavy atoms drawn from the elements C, H, O, N, F. There are multiple output labels, but we'll be predicting the enthalpy of formation at 298.15 K.
(1) (2 pt) Use the provided code to download and load the QM9 dataset, and split 80% of the dataset as training set and the other 20% as test set. The molecular graph is constructed by treating atoms as nodes and building edges between every pair of atoms, no matter whether they are connected by a chemical bond or not. The edge feature is a scalar, which is the inverse of the atomic distance. The node features are from this paper (https://arxiv.org/pdf/1704.01212.pdf). Check the dataset and what is the dimension of the node features?
(2) (5 pt) Finish the code to define a GNN with one message passing layer as follows:

First, the input node and edge features are embedded to $\mathbb{R}^{N_v}$ and $\mathbb{R}^{N_e}$ respectively, with a linear layer and a ReLU activation function. In this model, please use $N_v = N_e = 64$.

$$v_i^{(1)} = \sigma(W^T v_i^{(0)} + b)$$
$$e_{ij}^{(1)} = \sigma(W^T e_{ij}^{(0)} + b)$$

where v_i is the feature of node (atom) i and e_{ij} is the feature of edge between node i and j. The superscript (0) refers to the input node/edge features. σ is the ReLU activation function.

Then, the edge features are updated by concatenating features of its two consisting nodes and features of the previous state, then passing through a linear layer with a ReLU activation function. The output dimension of this layer should be the same as the dimension of the embedded edge feature, i.e. $N_e = 64$.

$$e_{ij}^{(2)} = \sigma[W^T(v_i^{(1)} \oplus v_j^{(1)} \oplus e_{ij}^{(1)}) + b]$$

Similarly, the node features are updated by concatenating the features from the previous state and the summation of features of all the connected edges, then passing through a linear layer with a ReLU activation function. Here $N(i)$ means the set of nodes that is connected with node i with an edge. Again, the output dimension of this layer should be the same as the dimension of the embedded node feature, i.e. $N_v = 64$.

$$v_i^{(2)} = \sigma\left[W^T\left(v_i^{(1)} \oplus \sum\nolimits_{j\in N(i)} e_{ij}^{(2)}\right) + b\right]$$

Finally, the updated node features are passed through a readout layer that maps to a scalar v^{readout} and the predicted formation enthalpy of the molecule is given by summing over these scalars.

$$H = \sum\nolimits_i v_i^{\text{readout}} = \sum\nolimits_i (W^T v_i^{(2)} + b)$$

(3) (3 pt) Use the training set to train the model with Adam optimizer for 3 epochs (if time permits, you can try more epochs), batch size 512, learning rate $10^{-3}$and L2 regularization $\lambda = 10^{-5}$. Use the mean squared error (MSE) as the loss function. Then evaluate on the test set and comment on how good or bad the prediction is.