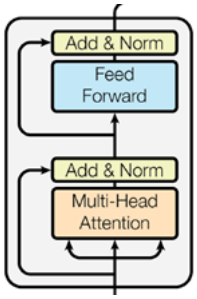
EE538 Neural Networks

Homework 8

Due: 12:59 on June 11th, 2021

1. For the multi-head Transformer model from Slide 3 to Slide 6 of Lecture Note 10, please answer the followings.



   (a) Estimate the number of adaptive elements to be trained for a single Transformer module in the right. Please use the N_{xxx} notation (e.x., N_{voc}) in Slide 4. If necessary, you may define additional parameters. (5 points)
   (b) Estimate the number of adaptive elements for the input and output embedding. (5 points)
   (c) Estimate the number of adaptive elements for the whole transformer model including input/output embedding, encoder, and decoder modules. (5 points)
   (d) Calculate the number of adaptive elements for the estimates in (a), (b), and (c), and the whole transformer model. You may use $N_{voc} = 50{,}000$, $N_{que} = N_{key} = N_{val} = 64$, $N_{head} = 8$, $N_{pos} = 384$, and the MLP at Step 5 on Slide 5 has one hidden-layer with $N_{hid} = 1{,}024$. Also, assume that both the encoder and decoder blocks include $N_{end} = N_{dec} = 6$ transformer modules. (10 points)

2. Let's try to extend the linear ICA in Slide 8 of Lecture Note 10 into nonlinear ICA.
   (a) Now the encoder (i.e., feature extractor) may be represented as $\boldsymbol{u}=\boldsymbol{f}(\boldsymbol{x})$. Derive a pdf $p(\boldsymbol{u})$ in terms of $p(\boldsymbol{x})$ and $\boldsymbol{f}(.)$. (10 points)
   (b) With a known $p(\boldsymbol{x})$, derive a learning rule to minimize Mutual Information between the joint pdf $p(\boldsymbol{u})$ and $\prod_i p(u_i)$. (10 points)
   (c) Assuming you use a one-hidden-layer Perceptron to learn $\boldsymbol{u}=\boldsymbol{f}(\boldsymbol{x})$, simplify the results of (b). For further simplification, you may assume both $\boldsymbol{x}$ and $\boldsymbol{u}$ are 2-elements vectors, and the hidden-layer has 2 neurons. (10 points)

3. For the Variational AE, please answer the followings.
   (a) For a Gaussian pdf with mean=0 and std=1 in 2-dimensional (x_1,x_2) space, and let a complex number defined as $z = x_1 + i\,x_2$. Please generate 300 samples with $y=z/10+z/|z|$. (5 points)
   (b) Train a 2-hidden-layer Perceptron to learn a mapping from (x_1,x_2) to $(\mathrm{Re}(y),\mathrm{Im}(y))$. (10 points)
   (c) Derive the KL-divergence between two Gaussian distributions as

$$\mathcal{D}[\mathcal{N}(\mu_0,\Sigma_0)\|\mathcal{N}(\mu_1,\Sigma_1)] = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^\top \Sigma_1^{-1}(\mu_1-\mu_0) - k + \log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\right).$$ (10 points)

   (d) Describe a learning procedure for the Conditional VAE. (10 points)
   (e) Derive a learning rule for the Conditional VAE. (10 points)