Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] K-Sparse Autoencoders

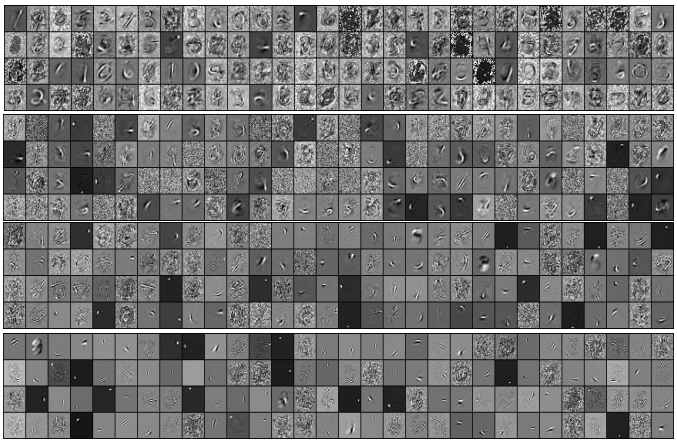
1. Experiment details

- The main theme of the paper is investigating the effect of sparsity. The sparse encoding process finds k largest activations among the values of hidden units and set the rest to 0. By varying the level of sparsity, the filter would learn different patterns of the inputs. To qualitatively understand the effect of this sparse encoding, I decided to plot the filters of K-sparse Autoencoder that is trained on MNIST dataset. The number of hidden units was chosen to be 1000 as in paper so that the weight matrix becomes 1000x784. Here, 784-dimensional row vector indicates the weight on each input pixel when computing the corresponding hidden unit. Therefore, when it is reshaped to 28x28, the filter that extracts meaningful information from the inputs can be visualized. Following the paper, I chose to plot only the first 120 filters.

- Comparing to vanilla autoencoder with the same architecture, the biggest difference may occur in the hidden unit activities. For the case of vanilla autoencoder, every node in the hidden layer is active. Therefore, the activations of all 1000 hidden units would not vary that much so that it restricts the capacity of model so that the richness of the learnt features are not guaranteed. In contrast, in K-sparse autoencoder, only 10 to 70 hidden units are active during the feed-forward process so that the degree of parameter update in certain edges would get bigger. This allows the values of hidden units to vary in wider range so that more global feature can be learned. As a result, the model can avoid overfitting and achieve better generalization performance. So, as an experiment, I decided to draw the histogram of hidden unit activities in Figure 4 to observe the range of activations.

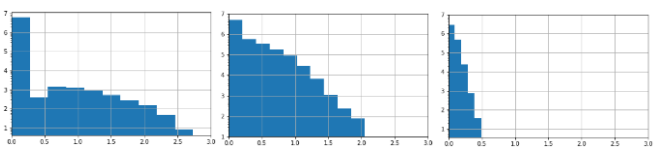
2. Reproduction results

- As the above filter plot implies, with large values of k (=70), the filters capture too local features of the input, which is not appropriate to use along with naïve linear classifier. In contrast, with small values of k (=15), it results in too global features which cannot be a meaningful representation. In case when k=25 or k=40, the learnt features were neither too local nor too global, which would contribute to improve the performance of classification task by fine-tuning. Therefore, I found the fact that the moderate level of sparsity must be forced and also verified the validity of sparse encoding in a qualitative manner. As a future analysis, I would like to suggest the quantitative analysis to measure the disentanglement in terms of information theory such as Maximum Mean Discrepancy.



Filter plots: k=10, 25, 40, 70 in order

- Moreover, the log-histogram of hidden unit activities were plotted in three cases when k=15, k=70 and k=1000. Note that K-sparse Autoencoder with k=1000 is identical to the vanilla autoencoder. Before analysis, I want to mention that comparing to the experiment in paper, the total number of samples were less due to the time constraint. However, it turns out to be sufficient enough to capture the tendency. When k=1000, the values of hidden unit activations were highly concentrated at low values. Whereas, when k=15 or k=70, the range of the activations was at least 4 times wider, which further verify the validity of sparse encoding just as the first experiment. As an additional side remark, I found that the shape of the histogram is smoother when k=70 so that the frequency of activations of any interval is similar. In contrast, in case when k=15, there is a big spike near the value 0. At first glance, k=70 seems to be a better choice, but the maximum value of activations was larger when k=15. This, again, shows that moderate and elaborate choice of k is essential.



Log-histogram of hidden units activities: K = 15, 70, 1000 in order