Student ID : 20194293

Name : Go, Kyeong Ryeol

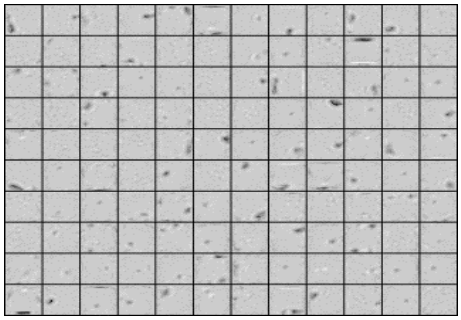
[AI 502] Extracting and Composing Robust Features with Denoising Autoencoders

1. Experiment details

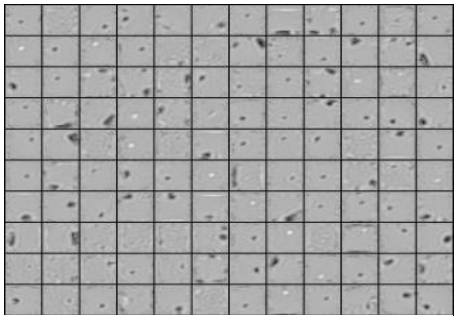
- The main theme of the paper is extracting robust features through corruption in inputs. The corruption process replaces certain portion of input unit to 0. By varying the level of corruption, the filter would learn different patterns of the inputs. To qualitatively understand the effect of this corruption process, I decided to plot the filters of Denoising Autoencoder that is trained on MNIST dataset. The number of hidden units was chosen to be 120 as in paper so that the weight matrix becomes 120x784. 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 and turned to grayscale, the filter that extracts meaningful information from the inputs can be visualized.

- Second experiment is on classification task with the stacked denoising autoencoders with 3 hidden layers which is denoted as SdA-3 in the paper. Autoencoder plays an important role in the parameter initialization of the network as a means of 'pretraining'. The idea was initially introduced in 'Greedy Layer-Wise Training of Deep Networks' and this can be done by learning only one weight matrix at a time in unsupervised way while other parameters are kept unchanged. Then, at the end, all the parameters are learned jointly in supervised way which usually referred to as 'fine-tuning'. The pretraining step mentioned above usually improve the performance on classification, as the optimization can be conducted in a nice starting point. As a reproduction, I construct SdA-3 network and work on MNIST with variations to verify its validity.

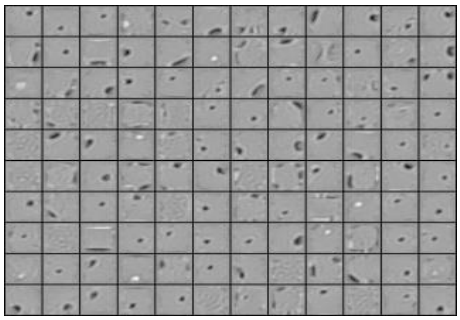
2. Reproduction results



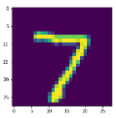
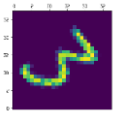
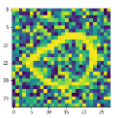
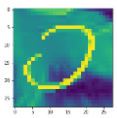
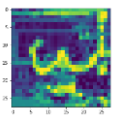
No destroyed inputs



25% destruction



50% destruction

Variation type	basic	rot	bg-rand	bg-img	rot-bg-img
Sample plot					
Accuracy (v)	97.51% (10%)	90.69% (10%)	88.53% (40%)	83.24% (25%)	53.42% (25%)

Classification accuracy (1-error rate) with SdA-3 on several MNIST variations

- As the above filter plot implies, without any noise, a number of filters has almost purely uniform values which indicates that it is not able to give distinctive effect across different pixels. In contrast, when higher noise level is accompanied, global features of the inputs were captured. Therefore, the validity of destruction was qualitatively verified. Here, the quantitative analysis would be also possible by measuring the disentanglement in terms of information theory such as Maximum Mean Discrepancy.

- Moreover, the usefulness of Denoising Autoencoder as a fine-tuner was verified on 5 variations of MNIST. Here, I plotted the samples and recorded the model accuracy on each variation type. Comparing to the Table 1 in the paper, most of the accuracy (=1-error rate) were in the 95% confidence interval. (FYI: I attached the accuracy curve along the epoch.) As the task gets difficult, the accuracy gets lower and this may be improved though recent proposed ideas such as Dropout or Batch Normalization.