Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] Sequence to Sequence Learning with Neural Networks

1. Paper Summary

It has been theoretically and empirically proved that the deep neural networks are flexible enough to approximate any functional forms. However, there was a big limitation that inputs and targets should be the vectors of fixed dimensionality. Since the length of sequences are usually not known a-priori, many real-world problems, especially in language model, posed a challenge for deep neural networks.

At that time, recurrent neural network and its variants were emerged to handle this issue. Among them, LSTM becomes a natural choice for language model applications due to its ability to learn long range temporal dependencies. Therefore, to resolve the sequence to sequence problems, the author also utilizes 4-layered-LSTM to obtain large fixed-dimensional vector representation of input sequence and to extract the output sequence from that vector. Here, the goal of LSTM is to estimate the conditional probability $\mathrm{p}(y_1, \dots, y_{T'}|x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, \dots, y_{t-1})$ of the output sequence given input sequence. Unlike the RNN Encoder-Decoder model proposed by Cho, the initial hidden state of the decoder LSTM is set to the last hidden state of the encoder LSTM that is denoted as v.

The validity of the model architecture was verified through the WMT'14 English to French translation task. While training, the log probability of a correct translation T given the source sentence S as maximized so that the overall loss function can be defined as $1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$ where $\mathcal{S}$ is a training set. Then, the most likely translation was produced based on a beam search which restrict the number of partial hypotheses for efficiency; $\hat{T} = arg \max_{T} p(T|S)$. The optimal beam of their model turns out to be 2 in the experiment.

Moreover, as a final remark, the order of the words of the input sentence was reversed to decrease the minimal time lag so that it makes the backpropagation easy to establish communication between input and output. This remarkably improved the overall performance particularly on the long sentences, which implies better memory utilization.

2. Discussion

Here, I want to offer two discussion points. To begin with, how can the sequence to sequence learning framework be adopted to image generation process? Here, the sequence can be defined in two ways; First case is the series of image data along the time that may have sampled from a single video. Second case is the patches from a single image that may have obtained through sliding windows. In both cases, if each data is naively regarded to be independent, the correlation among them is ignored, which needs the sequence to sequence modeling framework. Next, while producing the output sentence from the model, how can we impose the variation in style of the sentence by auxiliary data? Every person has its own distinct style of speaking and writing. This characteristic particularly differs a lot across ages and gender. To devise a service to be used in a real world, the style of the generated sentence needs to be deliberately changed based on its use. I suggest hierarchical modeling so that as the hyperparameter on the style changes the parameter of the model can be adaptively changed.