ELEG 5491: Homework #2, 2019 Edition

Due on March 28, 2019, 11:59pm, HK Local Time

Xiaogang Wang

Problem 1

[30 points]

Figure 1 is the architecture of CNN customized by Krizhevsky et al. to better leverage model parallelism in multi-GPU training. The architecture consists of two columns each allocated on one GPU. The forward propagation goes through multiple stages. At which stages, do the two GPUs need to be synchronized?

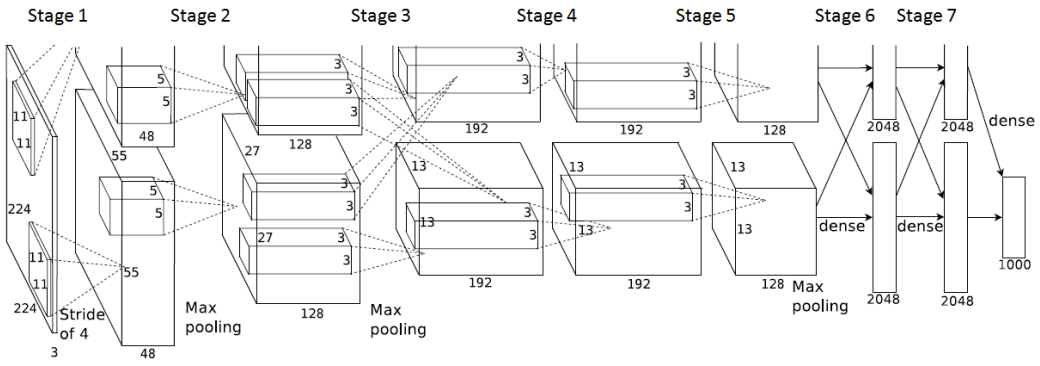

Figure 1: Problem 1

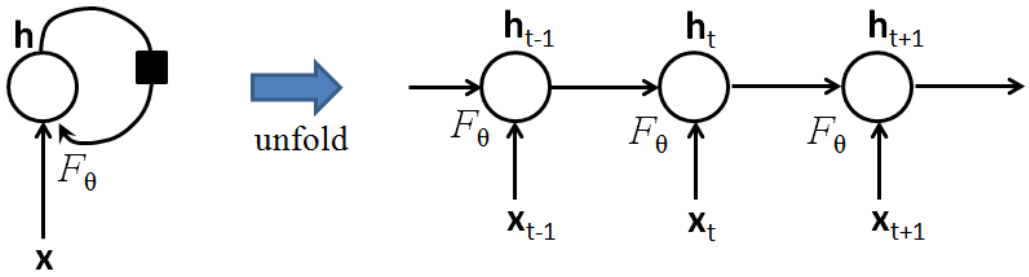

Figure 2: Problem 2

Problem 2

[**70 points**]

A RNN model is shown in Figure 2. $\mathbf{x}_1, \ldots, \mathbf{x}_t$ are input variables. A hidden variable $\mathbf{h}_t = F_\theta(\mathbf{h}_{t-1}, \mathbf{x}_t)$ contains information about the whole past sequence. It defines a function which maps the whole past sequence $(\mathbf{x}_t, \ldots, \mathbf{x}_1)$ to the current state $\mathbf{h}_t = G_t(\mathbf{x}_t, \ldots, \mathbf{x}_1)$.

- Give the explicit expression of G_t in terms of F_θ, $\mathbf{x}_1, \ldots, \mathbf{x}_t$. [**30 points**]
- There are two major difficulties when directly modeling $\mathbf{h}_t = G_t(\mathbf{x}_t, \ldots, \mathbf{x}_1)$: (1) G_t is different for different t (i.e. sequence length). It limits the generalization power of the model and there may not be enough training samples for each sequence length. (2) The model complexity of G_t may increase exponentially with t. Provide two reasons to explain how RNN solves the two challenges. [**40 points**]