
Layer-wise Asynchronous Training of Neural Network with Synthetic Gradient

Xupeng Tong, Hao Wang, Ning Dong, Griffin Adams

CONTENTS



BACKGROUND



INTRODUCTION



SYNCHRONOUS TRAINING



ASYNCHRONOUS TRAINING



INSIGHTS

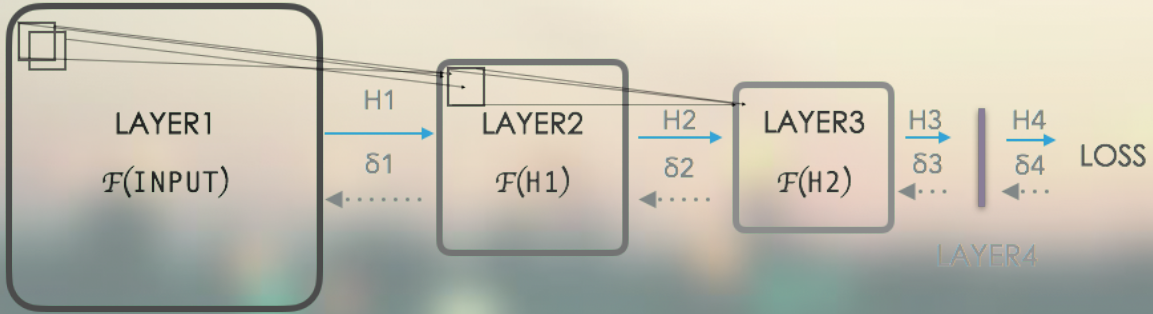
01

Part One

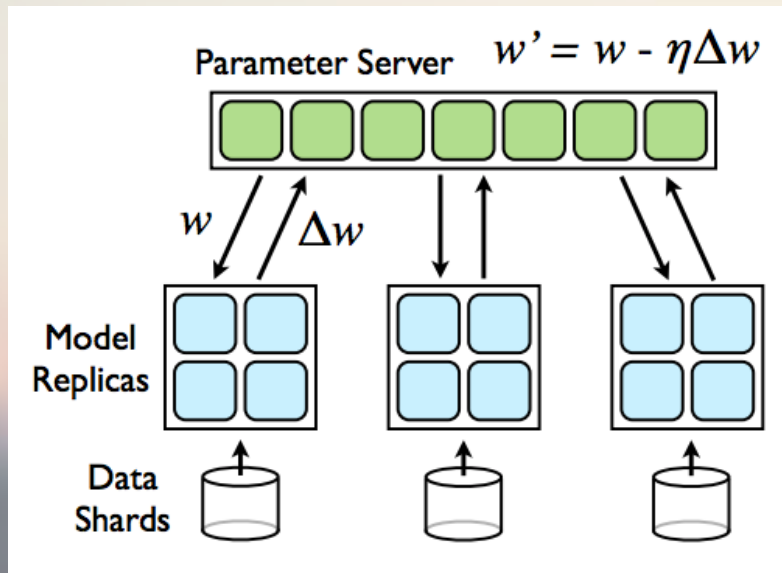
BACKGROUND

1 BACKGROUND

Back Propagation of Training CNN



Asynchronous SGD



Downpour SGD

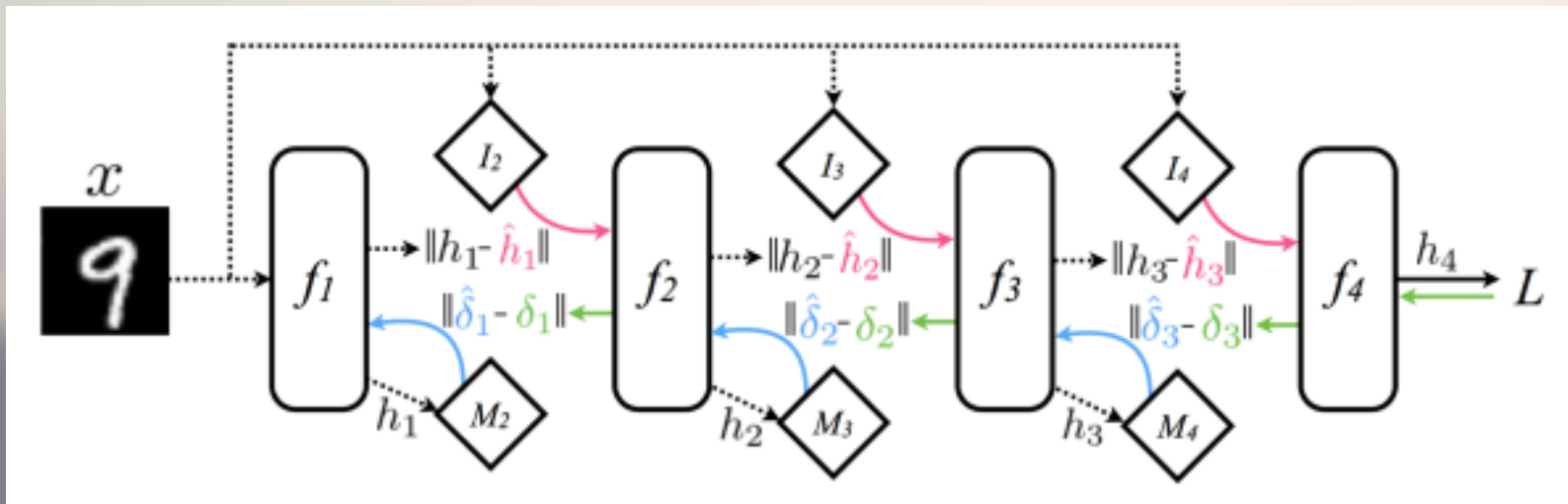
Dean, Jeffrey, et al. "Large scale distributed deep networks." *Advances in neural information processing systems*. 2012.

02

Part Two

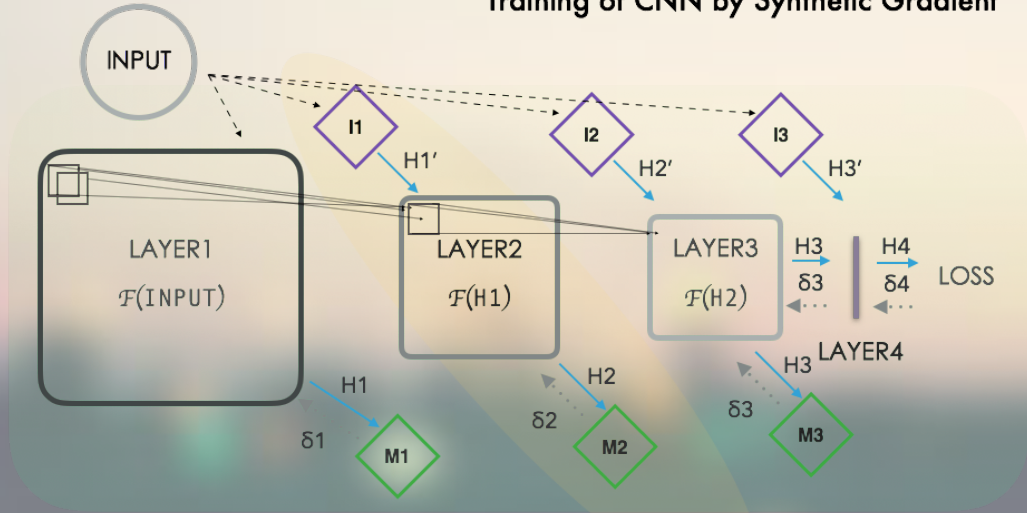
INTRODUCTION

2-1 OVERVIEW



Jaderberg, Max, et al. "Decoupled neural interfaces using synthetic gradients." *arXiv preprint arXiv:1608.05343* (2016).

Training of CNN by Synthetic Gradient



Each Layer can be trained independently

03

Part Three

SYNCHRONOUS TRAINING

Minimizing the synthetic input/gradient simultaneously with the general loss

$$L_{\mathcal{M}} = \sum_i \|\delta_i - \hat{\delta}_i\|_2^2$$

$$L_{\mathcal{J}} = \sum_i \|h_i - \hat{h}_i\|_2^2$$

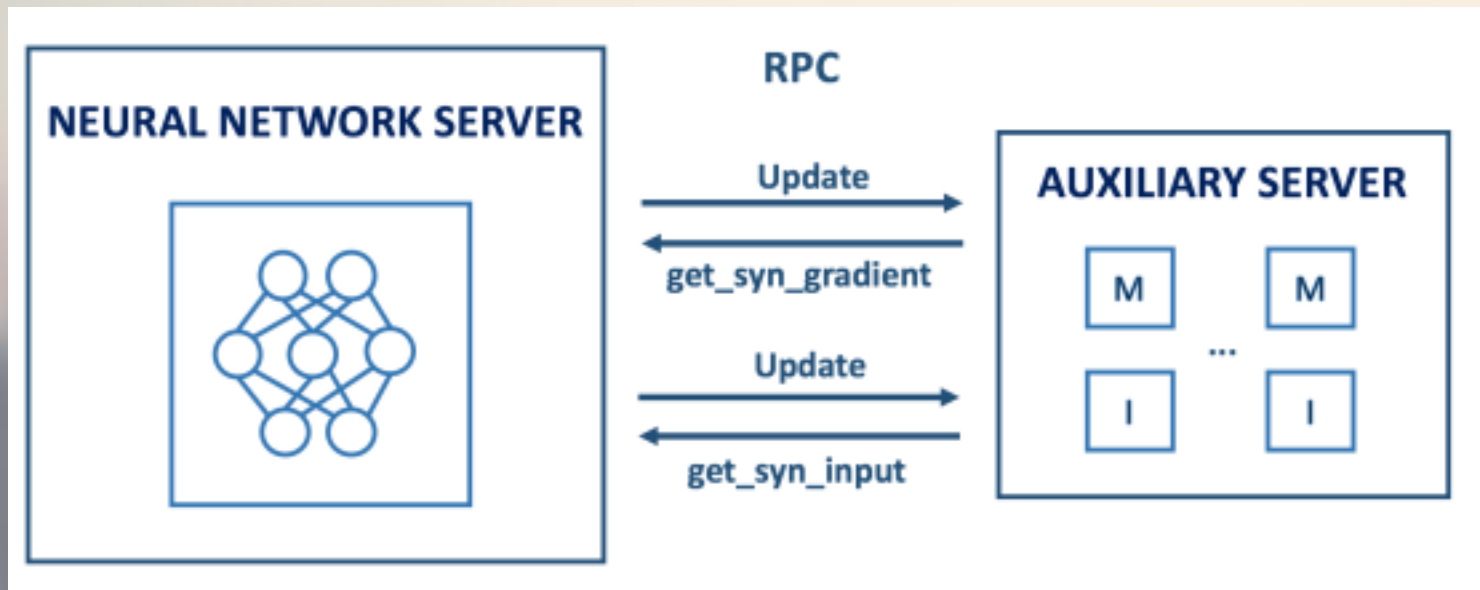
$$L(w_1, b_1, \dots, w_i, b_i, \dots, w_n, b_n) = L(W, B)$$

04

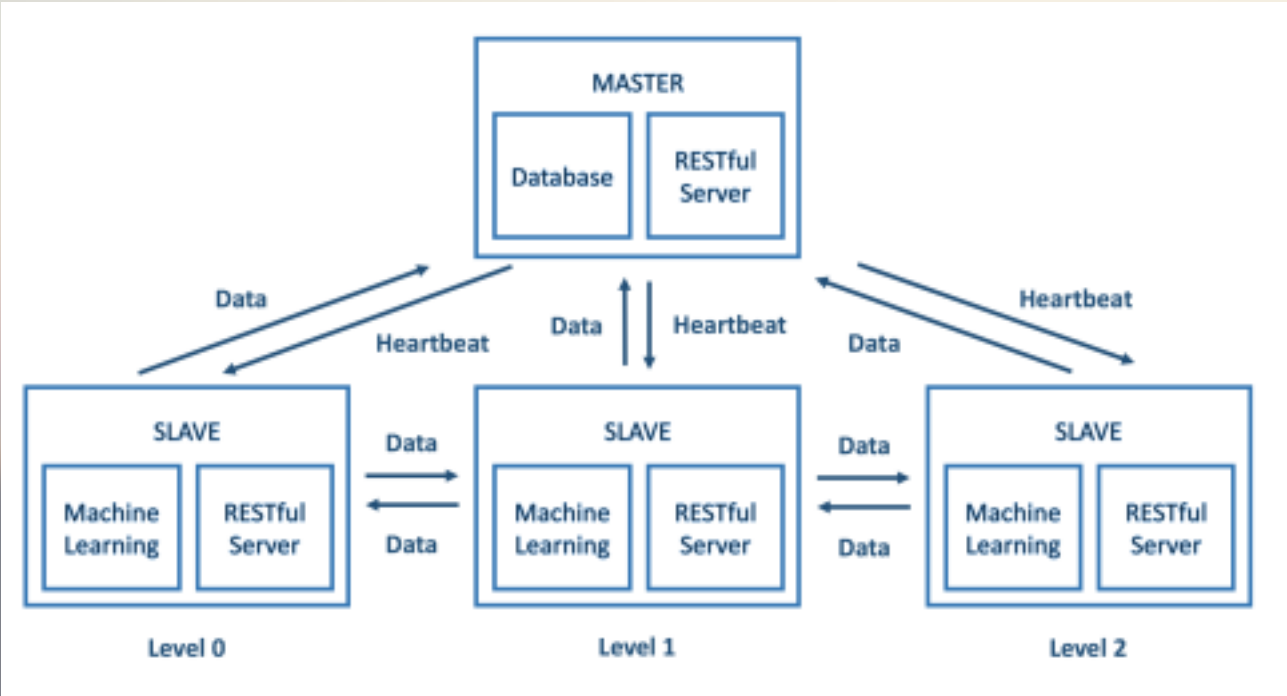
Part Four

ASYNCHRONOUS TRAINING

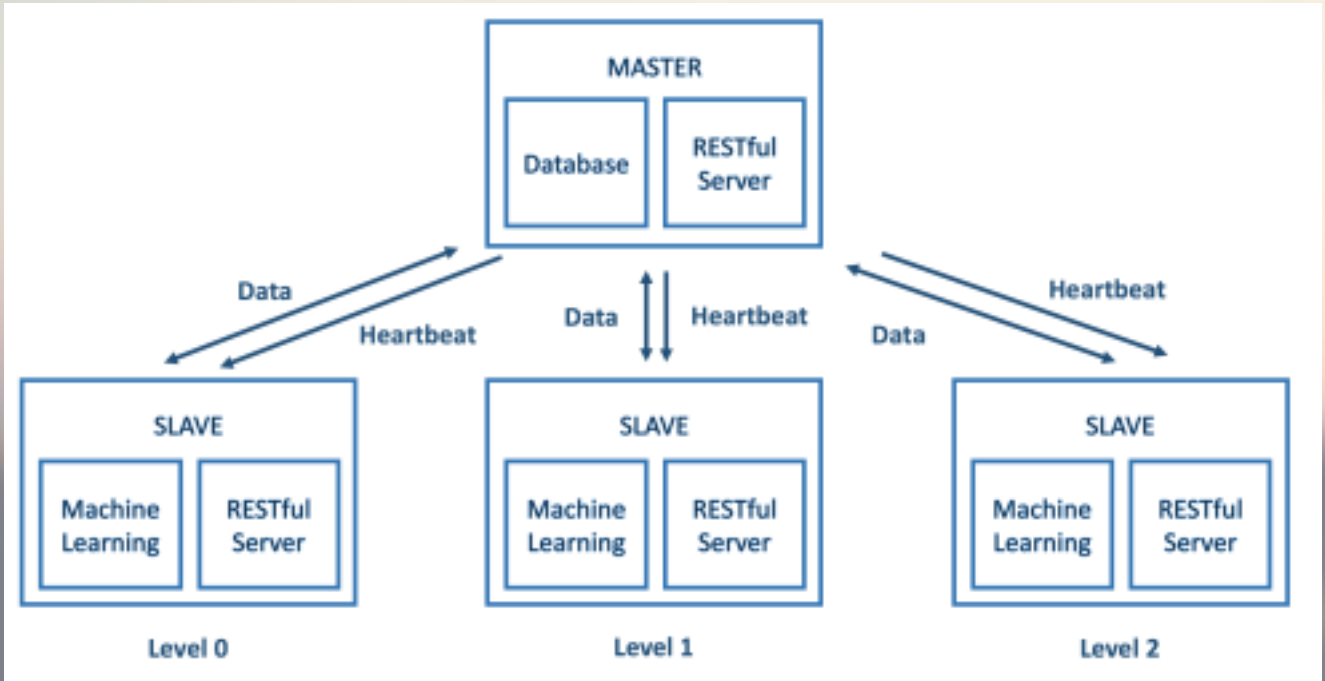
4-1 ONE POSSIBLE ARCHITECTURE



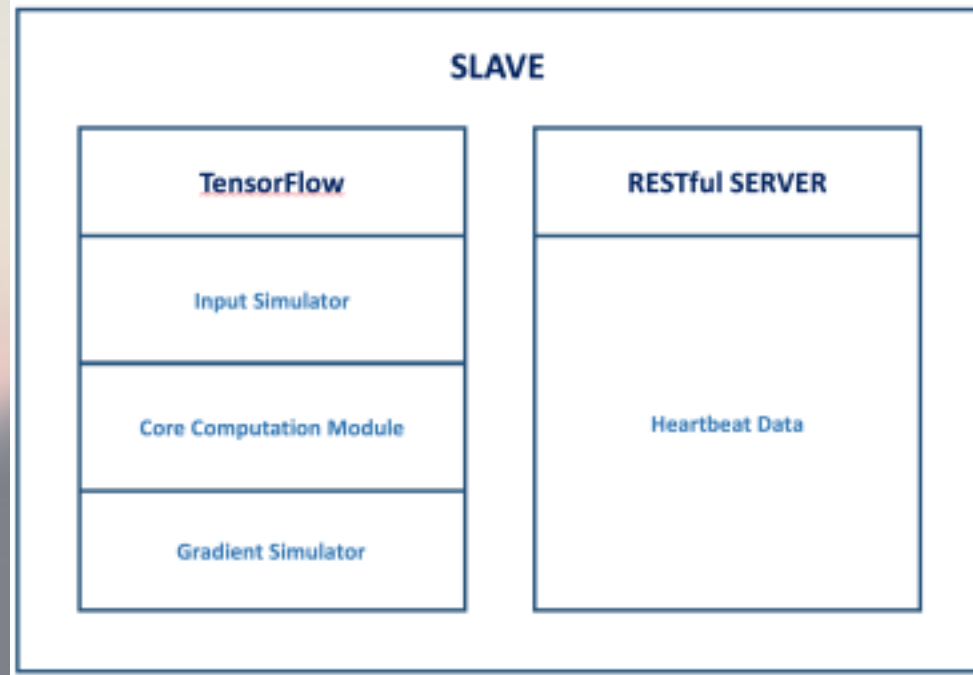
4-1 INFRASTRUCTURE - BASIC



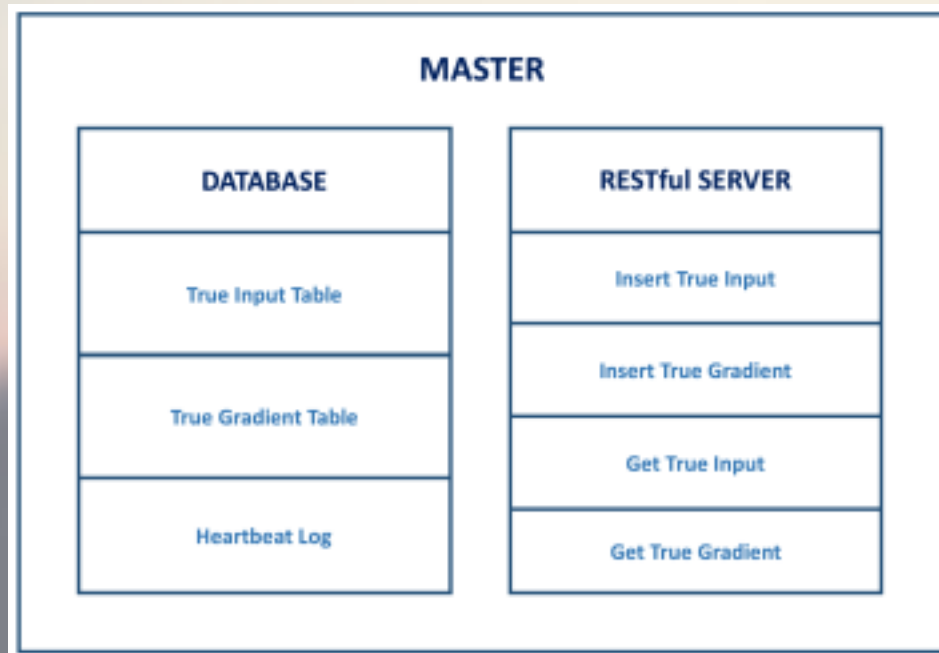
4-2 INFRASTRUCTURE - LIGHT



4-2 SLAVE



4-3 MASTER



05

Part Five

INSIGHTS

- Synthetic gradient and synthetic input as a new alternative of batch normalization / Dropout
- M and I auxiliary network introduces noises in the input/gradient of each layer
- No exact update is required!
- The model will learn how to reduce the variance within each batch
- while keeping the flavor of that specific batch.

Thank You
